

On the Painterly Depiction of Materials

An Interdisciplinary Study on the Depiction and Perception of Materials within Paintings

van Zuijlen, M.J.P.

DOI

[10.4233/uuid:ef4bae0f-0263-472e-a868-f0c4a2fdb908](https://doi.org/10.4233/uuid:ef4bae0f-0263-472e-a868-f0c4a2fdb908)

Publication date

2021

Document Version

Final published version

Citation (APA)

van Zuijlen, M. J. P. (2021). *On the Painterly Depiction of Materials: An Interdisciplinary Study on the Depiction and Perception of Materials within Paintings*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:ef4bae0f-0263-472e-a868-f0c4a2fdb908>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

On the Painterly Depiction of Materials

An Interdisciplinary Study on the Depiction and Perception of
Materials within Paintings

On the Painterly Depiction of Materials

An Interdisciplinary Study on the Depiction and Perception of
Materials within Paintings

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
donderdag 23 september, 2021 om 10:00 uur

door

Mitchell Johannes Petrus VAN ZUIJLEN

Master of Science in Psychology,
Universiteit Utrecht, Nederland,
geboren te Nieuwerkerk aan den IJssel, Nederland.

Dit proefschrift is goedgekeurd door de

promotor: Prof.dr. S.C. Pont
copromotor: Dr. M.W.A. Wijntjes

Samenstelling promotiecommissie:

| | |
|----------------------|---|
| Rector Magnificus, | voorzitter |
| Prof.dr. S.C. Pont, | Technische Universiteit Delft, promotor |
| Dr. M.W.A. Wijntjes, | Technische Universiteit Delft, copromotor |

Onafhankelijke leden:

| | |
|------------------------|---|
| Prof.dr. J. Dik, | Technische Universiteit Delft |
| Prof.dr. A.S. Lehmann, | Rijksuniversiteit Groningen |
| Prof.dr. I. Motoyoshi, | The University of Tokyo, Japan |
| Dr. J.J.R. van Assen, | Technische Universiteit Delft |
| Prof.dr. E. Karana, | Technische Universiteit Delft, reservelid |

Overige leden:

| | |
|------------------|--|
| Prof.dr. K. Bala | Cornell University, United States of America |
|------------------|--|

Prof.dr. K. Bala has contributed to the scientific study contained within chapter 3 of this thesis.

Copyright © 2021 by M.J.P. van Zuijlen

ISBN 978-94-6419-317-6

An electronic version of this dissertation is available at

<http://repository.tudelft.nl/>.

This work is part of the research program “Visual Communication of Material properties” with project number 276-54-001, which is financed by the Netherlands Organization for Scientific Research (NWO).

-
- Keywords:** Material perception, material properties, art history, crowdsourcing
- Printed by:** To be determined
- Front & Back:** The cover contains nine sections of seven paintings, the same sections are reproduced on the back. All sections of paintings are reproduced under open-access licenses. The digital image for the first (starting from the left) painting is courtesy of The Pinacoteca di Brera, Milan, Italy. The digital image for the fifth painting is courtesy of the Museo Nacional del Prado, Madrid, Spain. The remaining digital images are courtesy of the Rijksmuseum, Amsterdam. Attribution for each image:
- 1) The Kiss by *Francesco Hayez*, 1859.
 - 2 and 4) The Cannon Shot by *Willem van de Velde (II)*, c. 1680.
 - 3 and 9) View in the Bentheim Forest by *George Andries Roth*, 1870.
 - 5) The Cardinal by *Raphael*, 1510.
 - 6) Maurits, Prince of Orange by *Michiel Jansz van Mierevelt*, c. 1615.
 - 7) Marten Soolmans by *Rembrandt van Rijn*, 1634.
 - 8) View of Houses in Delft by *Johannes Vermeer*, c. 1658.

*It is good to have an end to journey toward,
but it is the journey that matters in the end.*

Ursula K. Le Guin

CONTENTS

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 1.1 | Material perception | 1 |
| 1.2 | Paintings and perception | 4 |
| 1.3 | A dataset of artistic material depictions | 5 |
| 1.4 | This Thesis. | 6 |
| 2 | Painterly depiction of material properties | 15 |
| 2.1 | Introduction | 16 |
| 2.2 | Methods | 18 |
| 2.2.1 | Ethics | 18 |
| 2.2.2 | Stimulus collection | 18 |
| 2.2.3 | Perceptual Experiment | 20 |
| 2.3 | Results | 24 |
| 2.3.1 | Data quality; intra- and inter- correlations. | 24 |
| 2.3.2 | Material judgments | 24 |
| 2.3.3 | Material attribute correlations. | 26 |
| 2.3.4 | PCA | 28 |
| 2.3.5 | Image statistics | 33 |
| 2.4 | Discussion | 35 |
| 2.5 | Supplementary materials | 43 |
| 3 | Materials In Paintings | 53 |
| 3.1 | Introduction | 54 |
| 3.1.1 | A simple taxonomy of image datasets. | 54 |
| 3.2 | Methods | 59 |
| 3.2.1 | Collecting paintings | 59 |
| 3.2.2 | Image-level coarse-grained material labels | 60 |
| 3.2.3 | Extreme click bounding boxes | 61 |
| 3.2.4 | Fine-grained labels | 63 |
| 3.3 | Results and demonstrations | 63 |
| 3.3.1 | Dataset statistics | 64 |
| 3.3.2 | Perceptual demonstrations | 65 |
| 3.3.3 | Cultural visual ecology demonstrations | 67 |
| 3.3.4 | Computer vision demonstrations | 74 |
| 3.4 | Conclusion | 79 |
| 4 | The Materials In Painting dataset | 81 |
| 4.1 | What is it? | 81 |
| 4.1.1 | The creation of the dataset | 83 |

| | | |
|----------|---|------------|
| 4.2 | The website | 86 |
| 5 | Soft like velvet and shiny like satin | 93 |
| 5.1 | Introduction | 94 |
| 5.2 | Experiment 1 | 97 |
| 5.2.1 | Methods | 97 |
| 5.2.2 | Results | 99 |
| 5.2.3 | Intermediate conclusions and discussion | 103 |
| 5.3 | Experiment 2 | 104 |
| 5.3.1 | Methods | 104 |
| 5.3.2 | Results | 108 |
| 5.4 | General discussion | 111 |
| 5.5 | Conclusions | 117 |
| 5.6 | Supplementary materials | 123 |
| 6 | On perceiving the luster of pearls. When less is more. | 145 |
| 6.1 | Introduction | 146 |
| 6.1.1 | Overview of the experiments | 151 |
| 6.2 | Experiment 1 | 151 |
| 6.2.1 | Participants and stimuli | 151 |
| 6.2.2 | Procedure | 153 |
| 6.2.3 | Results and discussion | 153 |
| 6.3 | Experiment 2 | 154 |
| 6.3.1 | Participants and stimuli | 154 |
| 6.3.2 | Procedure | 155 |
| 6.3.3 | Results and discussion | 155 |
| 6.4 | Experiment 3 | 156 |
| 6.4.1 | Participants and stimuli | 156 |
| 6.4.2 | Procedure | 157 |
| 6.4.3 | Results and discussion | 158 |
| 6.5 | Experiment 4 | 160 |
| 6.5.1 | Participants and stimuli | 160 |
| 6.5.2 | Procedure | 160 |
| 6.5.3 | Results and discussion | 161 |
| 6.6 | General discussion | 161 |
| 6.7 | Conclusion | 164 |
| 6.8 | Attribution | 164 |
| 6.9 | Supplementary materials | 170 |
| 7 | Conclusion | 187 |
| | Acknowledgements | 195 |
| | Summary | 197 |
| | Samenvatting | 199 |
| | List of Publications | 201 |

1

INTRODUCTION

MATERIAL PERCEPTION AND ART, OR THE ART OF PERCEPTION

Humans are capable of visually recognizing materials extremely quickly and accurately (Sharan et al., 2009, 2014) to the extent that this task appears trivial. However, the task is actually very complex. The visual appearance of materials is dependent on complex interactions between a variety of physical and optical factors (Anderson, 2011). At least, this is true for the real world: paintings are instead created by the artists directly manipulating image features on the canvas, which might therefore deviate from the “complex interactions” found in natural scenes. Cavanagh (2005) points out that the real world must follow the rules of physics, but that no such restrictions apply to the worlds depicted within paintings. Despite sometimes violating the rules of physics within their depictions, paintings can evoking strong perceptions of a rich 3D world filled with objects and materials. Through endless artistic experimentation (Gibson, 1978) painters have found successful ways to capture the appearance of materials within paintings. Studying these painterly depictions of materials can grant insights into human material perception that might not be achieved through the study of “normal” materials.

In the next sections, we will first introduce material perception to provide a general context for this thesis. Next, we will provide a discussion of how painterly depictions of materials can be a valuable contribution to this field and we will argue the need for a large-scale dataset of painterly depictions of materials and the multi-disciplinary benefits this could generate. Finally, we discuss the main aims of this thesis and shortly introduce each individual chapter.

1.1. MATERIAL PERCEPTION

Our world is filled with objects and materials (Adelson, 2001). The perception of objects has received much scientific attention, while the perception of materials has only recently started receiving more attention (Adelson, 2001; Fleming, 2014, 2017). Despite this dispar-

ity in attention, the perception of materials is a ubiquitous challenge that the human visual system is exposed to numerous times a day. Material perception allows us to visually distinguish between fresh and rotting food, judge if the pavement is slippery, or if a bench is dry enough to sit on. In other words, we need material perception in order to successfully interact with our world.

But what exactly is material perception? No agreed-upon definition exists, but material perception is generally understood as the multi-modal ability, or set of abilities, that allow us to categorize materials (Fleming, 2017; Hu et al., 2011; Sharan et al., 2009, 2014), estimate material attributes (Ferwerda et al., 2001; Fleming, 2014, 2017; Marlow et al., 2012; Zhang et al., 2015) and estimate an object's or material's physical state (Komatsu and Goda, 2018; Sawayama et al., 2017). Here categorization relates to differentiating between for example wood and fabrics. Estimating material attributes informs us if a material is glossy, soft, etc., and estimating physical state allows us to infer if an object is wet, dusty or dirty. See Fig. 1.1 for examples of each. Material perception is inherently multi-modal: by merely viewing a material, we can already form expectations of what the material might feel or sound like when interacted with. However, discussing all the modalities that are involved with material perception is outside the scope of this thesis, and for the remainder of this text, the focus shall be on visual material perception.

Based on our ability to perform material perception without conscious effort, we might naively assume that material perception is straightforward and uncomplicated. Indeed, it has been shown that humans are capable of categorizing materials accurately and quickly (Sharan et al., 2009, 2014), and that when judging and estimating material attributes, participants tend to correlate strongly with each other (Fleming, 2014), further implying the task of material perception is an easy one. However, if we actually consider the task itself, instead of our ability to perform it, we quickly uncover a hidden complexity. The input for visual material perception consists of the light-rays that are reflected from our environment to our retina. The actual light that reaches our eyes depends on multiple interacting factors such as the viewing angle, lighting condition, shape, optical properties of the material (Anderson, 2011), and the physical state of the material (e.g., dusty or wet) (Komatsu and Goda, 2018).

A classical explanation for material perception holds that the visual system is able to estimate, or model, the physical scene in order to discount effects thereof to recover intrinsic properties of the material (Marr, 2010; Pizlo, 2001; Poggio and Koch, 1985; Poggio et al., 1985). The origin of this idea is generally attributed to Helmholtz (1866/1962), who conjectured that we "eliminate" the effect of the illumination to "reveal the body". In essence, this *inverse optics* view claimed that the visual system is able to create an accurate mental model of the physical scene by running physics in reverse (Poggio et al., 1985). However, current views consider inverse optics to only be feasible for simple, straightforward scenes (Fleming et al., 2004; Nishida, 2019). For example, Fleming et al. (2004) argues that it would be extremely difficult to recover complex surfaces, such as mirrors, as "all visible features belong to the environment surrounding the object".

A more recent approach, known as *image statistics*, reasons that generating accurate model of the environment is too computationally complex and that the human visual system needs not know (i.e., model) the environment as it may instead rely on image features to infer physical material properties (Fleming, 2014; Fleming and Bülthoff, 2005; Motoyoshi



Figure 1.1: An example of the large visual variety found in materials. In the top row, stone, metal, and fabric are three examples of different material categories. In the middle row, three examples of wood show the perceptual variety that can be found within a single category. The bottom row shows that even the exact same material can differ across physical state, with the left image being dry, and the right wet.

et al., 2007; Nishida, 2019). That is, instead of requiring the generation of complex models of the natural scene the visual system can instead rely on much simpler learned statistical rules of natural scenes. For example, Nishida (2019) argues that when the visual system encounters a blurry reflection on a metal object it can infer that the metallic object is not polished - as the natural scene reflected is originally sharp.

More approaches exist that attempt to explain the connection between the light reaching the eye and the material perception that this evokes. To our knowledge however, there is currently no theory, approach or model that can robustly and accurately predict the evoked output, i.e., perception, for arbitrary input images. Interestingly however, there are visual experts that are capable of doing the inverse, i.e., generating input images that can robustly evoke perceptions: painters.

Painters have learned how to capture the world within paintings through endless artistic experimentation and observation (Gibson, 1978). When we observe a painting we perceive the world through the artist’s eye; a painting is not just a reflection of a natural scene, it is instead a reflection of a painters perception thereof. In other words, the endless artistic experimentation and observation of painters is itself shaped by the painters’ perception. As such, paintings can be considered as a double perception interface: an image made *by* and *for* perception (Cavanagh, 2005). Studying the depiction and perception of materials within paintings provides a valuable novel, multidisciplinary approach to understand visual material perception due to this double perception interface.

1.2. PAINTINGS AND PERCEPTION

Paintings have a long history of being used as stimuli to study the human visual system (Buswell, 1935; Yarbus A L, 1967) and the inverse is also true: art historians have a long history of using vision science to study art (Arnheim, 1965; Gombrich, 1960). Paintings typically display strong statistical correlations with natural scenes (Graham and Field, 2008; Redies et al., 2008), but they merely reflect the physical, 3D world. Painters do not intent to capture the world as truthfully as possible¹. Instead, paintings are explicitly created for human perceptions and in the process of painting, artists can directly manipulate the 2D image features in order to evoke a desired perceptual response. This results in a more perceptually motivated stimulus that is ideal for visual perception experiments, as the stimulus is explicitly created for and by perception (Cavanagh, 2005), i.e., a double perception interface. Furthermore painters are free to bend or ignore the rules of physics (i.e., by directly manipulating the 2d image features) when creating a painting in order to “further the painting’s intended effect” (Cavanagh, 2005; Graham and Meng, 2011) which can lead to scenes that are very unlikely, if not impossible, to be encountered in the real world.

One of the primary goals of this thesis is to enable a multidisciplinary audience to study material depiction and perception within paintings by creating a large-scale dataset of artistic depictions. Therefore, in this thesis, we introduce the Materials In Paintings (MIP) dataset, which is a large-scale dataset of paintings containing material information.

¹It is interesting to note that the human visual system does not try to “capture” the world as truthfully as possible either. A “perfect” perceptual system that accurately perceives “reality” would for example not be able to perceive paintings as it would merely perceive a planar surface marked with with splashes of color. Instead of pursuing such “perfection” evolution has shaped our human visual system and our perception towards maximizing biological fitness.

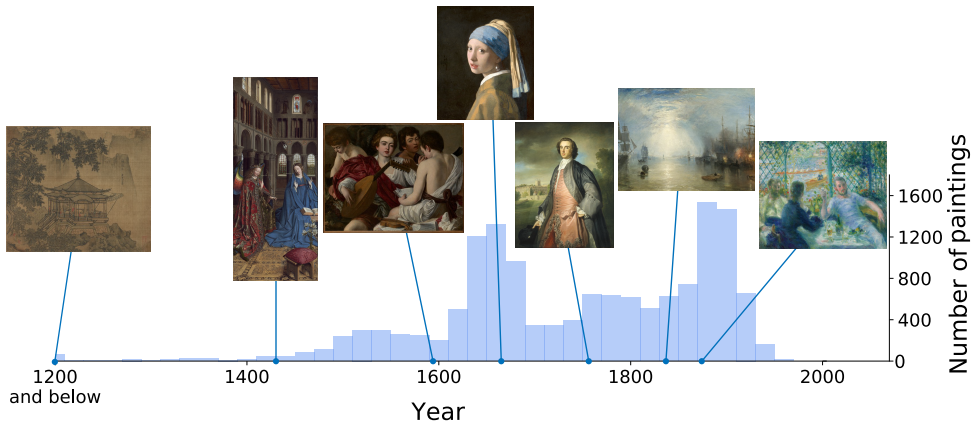


Figure 1.2: The distribution of paintings over time. The paintings shown exemplify the large variability of paintings contained within the dataset. The lack of paintings after the 1900s is due to the copyright restrictions for cultural heritage images. Paintings shown from left to right are: *A Pavilion* by unknown artist, 1153. The Cleveland Museum of Art, USA. *The Annunciation* by Jan van Eyck, 1435. National Gallery of Art, USA. *The Musicians* by Caravaggio, 1595. The Metropolitan Museum of Art, USA. *Girl with a Pearl Earring* by Johannes Vermeer, 1665. Mauritshuis, The Hague, Netherlands. *A Graduate of Merton College, Oxford* attributed to George Knapton, 1754. National Gallery of Art, USA. *Keelmen Heaving in Coals by Moonlight* by Turner, 1835. National Gallery of Art, USA. *The Rowers' Lunch* by Pierre-Auguste Renoir, 1875. The Art Institute of Chicago, USA.

We tried to gather a large set of paintings that encompasses a large selection of art history. As can be seen in Fig. 1.2, our collection contains paintings across a large range of time and styles.

1.3. A DATASET OF ARTISTIC MATERIAL DEPICTIONS

A decade ago only few paintings were digitally available, but with the push of digitizing cultural heritage over the last decade, many previously unavailable paintings have now become freely accessible online (Bernhard, 2016; Saleh and Elgammal, 2015). Possibly the most famous dataset of paintings is the WikiArt dataset, created by a non-profit organisation with the goal “to make world’s art accessible to anyone and anywhere” (“Visual Art Encyclopedia”, 2020). The WikiArt dataset has been widely used for a variety of scientific purposes (Bar et al., 2014; Elgammal et al., 2018; Saleh and Elgammal, 2015; Strezoski and Worring, 2017; Tan et al., 2017). However, some of these images are copyrighted, which in some cases can complicate the use of WikiArt images for scientific purposes. Luckily, large-scale datasets of paintings exist that were explicitly created for scientific use and exclusively use paintings from the public domain. These datasets typically contain low-quality images, as this is usually preferred within the computer vision field for deep learning purposes. For example, the *Art500k* dataset contains more than 500k low-resolution artworks which were used to train a model to identify style in paintings (Mao et al., 2017). In another example, Crowley and Zisserman (2014a) and Crowley and Zisserman (2014b) used a set of 10k medium-resolution paintings to study object recognition within paintings.

While the previously mentioned datasets contain paintings, they do not explicitly con-

tain material information. To our knowledge, with the exception of the Materials In Painting (MIP) dataset introduced in this thesis within chapter 3, no datasets exist that contains paintings with material information. There are however a small number of datasets that contain material information for photos or for digital renderings. One such database was highly influential for the creation of this thesis: OpenSurfaces (Bell et al., 2013), which contains around 70k crowd-sourced polygon segmentations of materials sourced from photos. In a later study, the authors upgraded OpenSurfaces by providing an additional 3 million point-samples across 23 material classes (Bell et al., 2015). The MIP dataset has used, updated, and adapted the software and annotational pipe-line described in Bell et al. (2013). Another example of influential datasets containing material information is the Flickr Material Database (FMD), which was originally created to study how quickly people can recognize materials.

Many of the datasets previously mentioned were originally created for a single discipline. For example, Opensurfaces was created for computer vision, but could easily be applied for material perception research. Other dataset have been presented as more multidisciplinary, such as for example, the FMD which has been used for both human vision and computer vision research by the original authors (Sharan et al., 2013; Sharan et al., 2009), which demonstrates the possible multi-disciplinary nature that databases can have. Similarly, the MIP dataset introduced in this thesis is intended to be multi-disciplinary. The benefits for human vision are described above, as well as in chapters 5 and 6. For computer vision the MIP dataset provides data previously unavailable, namely the painterly depiction of materials. In Lin, Zuijlen, et al. (2021) we explicitly discuss the insights the MIP dataset can grant computer vision. Furthermore, in Lin, van Zuijlen, et al. (2021) we use the MIP dataset to improve model robustness for computer vision applications. For (digital) art history we consider the depictions within paintings as fundamentally interesting and we consider this field a third major discipline that the MIP can contribute towards. It can, for example, enable the study of material depictions on a large scale, which is a topic we further explore in chapter 3.

1.4. THIS THESIS

The main rationale of this thesis is that the study of painterly depiction of materials can be beneficial for vision science, art history, and computer vision. The motive here is that painters are assumed to be masters of depiction and that their painterly skills and/or expertise can be leveraged into scientific insights. As such, the primary aim of this thesis is twofold. First, to study the perception of materials in painterly depictions and second to enable the study of painterly material depictions for a broad, interdisciplinary audience.

Each of the chapters work towards these aims. Below we will shortly introduce and describe the rationale for each chapter.

Chapter 2. One of the central points of this thesis is that studying materials depicted within paintings can lead to valuable insights for our understanding of material perception. However, this belief rests on the implicit assumption that insights into the perception of painterly materials can transfer to the perception of materials in other media, like photos or drawings. Is the perception of materials in paintings indeed comparable to the perception

of materials across other media? Do materials in paintings evoke comparable perceptions of material attributes relative to other media?

In Fleming et al. (2013), the authors measured perceived material attributes, such as softness, coldness, and hardness for various photographed materials and found that the distributions for these material attributes are "well defined, distinct, and systematically related to material class membership". In this chapter, which contains the first peer-reviewed paper of this thesis, we posed the question if similar, distinct relationships exist between materials and material attributes for painterly material depictions. To test this we studied the perception of painterly depictions by measuring the perceived material attributes for a number of materials and compared this to the result reported by Fleming et al. (2013).

Without discussing the results in depth here, it is important to note that we found that painterly depictions of materials also display distinct, well defined, and systematic distributions of material attributes. Moreover, these distributions for painterly materials were found to be comparable to those reported for natural stimuli by Fleming et al. (2013). We interpreted this to strongly imply that insights gained into the perception of painterly depiction might, at least to some extent, transfer to the perception of materials as a whole. This opens up possibilities for multidisciplinary research, which we further explored in the next chapter.

Chapter 3. In the previous chapter, in order to measure the perception of painterly materials, we started with the collection of painterly materials. Specifically, we collected a substantial set of paintings and collected material annotations using crowd-sourcing over the internet. Furthermore, we showed that the perception of materials within paintings appears to be comparable to the perception of materials in photographs. This implies that insights into the perception of painterly materials can transfer to the perception of materials in different media.

A primary aim of this thesis is to promote the multidisciplinary study of materials. We consider the study of the painterly depiction of materials to be inherently multidisciplinary. The depiction of materials within paintings is of fundamental interest to art history, and the perception thereof is of fundamental interest to perception. Furthermore, the depiction of materials is also of interest to human vision science and vice versa. For example, how did painters depict such smooth and shiny fabrics? And what visual cues present within paintings are used by the human visual system to trigger these perceptions?

In addition to vision science and art history, the depiction and perception of painterly materials can also be of interest to computer science, specifically computer vision and computer graphics. If a large enough set of painterly material depictions becomes available, this data could be used to train and improve existing neural networks. Can the painterly depiction of materials improve a network's robustness? Or can it improve the generalizability of networks across different domains (i.e., different medias of depiction). Can networks be trained to recognize materials within paintings? Can painterly depiction techniques be "learned" and employed by generative networks?

As a starting point, to answer the questions posed above, we created the Materials In Paintings (MIP) dataset. The primary purpose of the paper, contained within this chapter, was to describe, introduce, and release the dataset to the scientific community. To demonstrate the multidisciplinary utility of the dataset we furthermore conducted various

experiments and analyses across perception, art history and computer vision.

Chapter 4. In the previous chapters we have worked towards, and completed, a large scale dataset of material depictions within art. To maximize the reach and accessibility of this dataset we made the data available at an easily navigable location and provided a number of filters and search tools to enable detailed exploration of the data. As such, the dataset is hosted on a dedicated server and can be visited on materialsinpaintings.tudelft.nl, where the entirety of the dataset can be viewed, browsed, and downloaded.

In the previous chapter we have provided a scientific discussion of the dataset. In this chapter we provide a more practical and design-driven discussion. First a brief, high-level overview of the final 'result' is given by answering the following questions: What is the MIP database? What is in it? Where does the data come from? Next, we provided an informal discussion of the design decisions and reflected on the process that has led to the final version of the dataset; in chapter 2, for example, we used polygonal segmentation, while, in chapter 3, we primarily focused on bounding boxes. The explicit discussion of the challenges and/or problems that have caused such changes typically do not belong within of a scientific article. Nevertheless, the discussion of such reflections could give some insight into the final design of the dataset. Finally, we provided a short practical guide on how to browse and explore the dataset.

Chapter 5. In the previous chapters, we have taken a big picture approach to material perception. First, we studied numerous materials and material attributes. Secondly, we created and released an extensive set of annotated material depictions within art. Doing so we worked towards one of the main aims of this thesis: to make the study of materials multidisciplinary. In the following chapters we apply this multidisciplinary approach by studying materials depicted within paintings.

Studying the depiction and perception of materials in this multidisciplinary approach can lead to novel insights. For example, painters are free to directly manipulate the image features of their depictions. Artists can even ignore certain cues entirely while still rendering convincing material perceptions.

In this chapter we specifically study the depiction and perception of fabrics. Fabrics are interesting as this material displays an especially large variability in appearance across different exemplars. Nevertheless, from chapter 2 we know that fabrics in paintings, similar to photographed fabrics, on average display a distinct and well defined distribution of perceived material attributes. This raises the question if different types of fabrics display a comparable, well defined distribution of perceived material attributes, or do different fabrics have distinct distributions? We attempt to answer this question by using stimuli from the MIP dataset for two fabrics: velvet and satin. We consider these fabrics as especially interesting as the appearances of these fabrics are heavily influenced by their reflective properties, in almost entirely opposite ways.

Chapter 6. In this chapter we studied the depiction and perception of pearls within paintings. Some materials, such as fabrics from the previous chapter, display considerable visual variability in shape, texture, and color. In contrast with this, pearls might naively be considered a simple material in the sense that their visual appearance is quite robust. However,

optically, pearls are fascinating due to their physical structure and their luster, a term used to denote the various interactions with light that are characteristics of pearls. This physical structure gives rise to subtle but striking differences between different instances of pearls.

In this chapter we try to study the perception of pearls depicted in paintings. The rationale and multidisciplinary approach here is similar to the previous chapters: we consider painters to be experts at depicting materials. During this depiction they might emphasize relevant perceptual features while ignoring or reduce non relevant perceptual features. In this way, painters might function as a sort of perceptual filter - they depict *by* and *for* perception. As such, the study of materials depicted within paintings is not just a study of perception *in the eye of the beholder*. Instead, here we study perception through a double perception interface: *twice through the eye of the beholder*.

BIBLIOGRAPHY

- Adelson, E. H. (2001). On Seeing Stuff: The Perception of Materials by Humans and Machines. *Proceedings of the SPIE*, 4299, 1–12. <https://doi.org/10.1117/12.429489>
- Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, 21(24), R978–R983. <https://doi.org/10.1016/j.cub.2011.11.022>
- Arnheim, R. (1965). *Art and visual perception: A psychology of the creative eye*. Univ of California Press.
- Bar, Y., Levy, N., & Wolf, L. (2014). Classification of artistic styles using binarized features derived from a deep neural network. *European conference on computer vision*, 71–84.
- Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2013). OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG)*, 32.
- Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2015). Material recognition in the wild with the Materials in Context database. *Computer Vision and Pattern Recognition (CVPR)*.
- Bernhard, M. (2016). Gugelmann Galaxy . An Unexpected Journey through a collection of Schweizer Kleinmeister. *International Journal for Digital Art History*, 2.
- Buswell, G. T. (1935). *How People Look at Pictures*. <https://doi.org/10.2307/771371>
- Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, 434(7031), 301–307.
- Crowley, E. J., & Zisserman, A. (2014a). In search of art. *European Conference on Computer Vision*, 54–70.
- Crowley, E. J., & Zisserman, A. (2014b). The state of the art: Object retrieval in paintings using discriminative regions. *British Machine Vision Conference (BMVC)*.
- Elgammal, A., Mazzone, M., Liu, B., Kim, D., & Elhoseiny, M. (2018). The shape of art history in the eyes of the machine.
- Ferwerda, J. A., Pellacini, F., & Greenberg, D. P. (2001). Psychophysically based model of surface gloss perception. *Proc. SPIE*, 4299(June 2001). <https://doi.org/10.1117/12.429501>
- Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research*, 94, 62–75. <https://doi.org/10.1016/j.visres.2013.11.004>
- Fleming, R. W. (2017). Material Perception. *Annual Reviews*, 3, 365–88.
- Fleming, R. W., & Bülthoff, H. H. (2005). Low-Level Image Cues in the Perception of Translucent Materials. *ACM Transactions on Applied Perception*, 2(3), 346–382. <https://doi.org/10.1145/1077399.1077409>
- Fleming, R. W., Torralba, A., & Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of Vision*, 4(9), 10. <https://doi.org/10.1167/4.9.10>
- Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8), 9. <https://doi.org/10.1167/13.8.9>
- Gibson, J. J. (1978). The ecological approach to the visual perception of pictures. *Leonardo*, 11(3), 227–235.

- Gombrich, E. H. (1960). *A study in the psychology of pictorial representation*. Pantheon Books.
- Graham, D., & Field, D. (2008). Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities. *Spatial vision*, 21(1-2), 149–164.
- Graham, D. J., & Meng, M. (2011). Artistic representations: Clues to efficient coding in human vision. *Vis. Neurosci*, 28(4), 371–379.
- Helmholtz, H. (1866/1962). *Treatise on physiological optics*. (vol. 2, trans. j. p. l. southall) (Vol. 2). New York: Dover.
- Hu, D., Bo, L., & Ren, X. (2011). Toward robust material recognition for everyday objects. *BMVC*, 2, 6.
- Komatsu, H., & Goda, N. (2018). Neural mechanisms of material perception: Quest on shitsukan. *Neuroscience*, 392, 329–347.
- Lin, H., van Zuijlen, M., Pont, S. C., Wijntjes, M. W., & Bala, K. (2021). What can style transfer and paintings do for model robustness? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11028–11037.
- Lin, H., Zuijlen, M. V., Wijntjes, M. W., Pont, S. C., & Bala, K. (2021). Insights from a large-scale database of material depictions in paintings. *International Conference on Pattern Recognition*, 531–545.
- Mao, H., Cheung, M., & She, J. (2017). Deepart: Learning joint representations of visual arts. *Proceedings of the 25th ACM international conference on Multimedia*, 1183–1191.
- Marlow, Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, 22(20), 1909–1913. <https://doi.org/10.1016/j.cub.2012.08.009>
- Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.
- Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. 447(May), 206–209. <https://doi.org/10.1038/nature05724>
- Nishida, S. (2019). Image statistics for material perception. *Current Opinion in Behavioral Sciences*, 30, 94–99.
- Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision research*, 41(24), 3145–3161.
- Poggio, T., & Koch, C. (1985). Ill-posed problems early vision: From computational theory to analogue networks. *Proceedings of the Royal society of London. Series B. Biological sciences*, 226(1244), 303–323.
- Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, 638–643.
- Redies, C., Hasenstein, J., & Denzler, J. (2008). Fractal-like image statistics in visual art: Similarity to natural scenes. *Spatial vision*, 21(1-2), 137–148.
- Saleh, B., & Elgammal, A. (2015). Large-scale classification of fine-art paintings: Learning the right metric on the right feature.
- Sawayama, M., Adelson, E. H., & Nishida, S. (2017). Visual wetness perception based on image color statistics. *Journal of Vision*, 17(5), 7–7.

- Sharan, L., Liu, C., Rosenholtz, R., & Adelson, E. H. (2013). Recognizing materials using perceptually inspired features. *International Journal of Computer Vision*, 103(3), 348–371. <https://doi.org/10.1007/s11263-013-0609-0>
- Sharan, L., Rosenholtz, R., & Adelson, E. H. (2009). Material perception : What can you see in a brief glance? *Vision Sciences Society Annual Meeting Abstract*, 9(8), 2009.
- Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision*, 14(9), 1–24. <https://doi.org/10.1167/14.9.12>
- Strezoski, G., & Worring, M. (2017). Omniart: Multi-task deep learning for artistic data analysis.
- Tan, W. R., Chan, C. S., Aguirre, H. E., & Tanaka, K. (2017). Artgan: Artwork synthesis with conditional categorical gans. *IEEE International Conference on Image Processing (ICIP)*, 3760–3764.
- Visual art encyclopedia. (2020). <https://www.wikiart.org/en/about>
- Yarbus A L. (1967). *Eye movements and vision*. [https://doi.org/10.1016/0028-3932\(68\)90012-2](https://doi.org/10.1016/0028-3932(68)90012-2)
- Zhang, F., de Ridder, H., & Pont, S. C. (2015). The influence of lighting on visual perception of material qualities. 9394(0), 93940Q. <https://doi.org/10.1117/12.2085021>

2

PAINTERLY DEPICTION OF MATERIAL PROPERTIES

Painters are masters of depiction and have learned to evoke a clear perception of materials and material attributes in a natural, three-dimensional setting, with complex lighting conditions. Furthermore, painters are not constrained by reality, meaning that they could paint materials without exactly following the laws of nature, while still evoking the perception of materials. Paintings have to our knowledge not been studied on a big scale from a material perception perspective. In this paper, we studied the perception of painted materials and their attributes by using human annotations to find instances of 15 materials, such as wood, stone, fabric, etc. Participants made perceptual judgments about 30 unique segments of these materials for 10 material attributes, such as glossiness, roughness, hardness, etc. We found that participants were able to perform this task well while being highly consistent. Participants, however, did not consistently agree with each other and the measure of consistency depended on the material attribute being perceived. Additionally, we found that material perception appears to function independently of the medium of depiction - the results of our PCA analysis agreed well with findings in former studies for photographs and computer renderings.

2.1. INTRODUCTION

Materials represent the ‘stuff’ that things are made of (Adelson, 2001). We interact daily with these ‘things’, either physically (e.g., manual interaction) or visually (e.g., assessing ripeness, quality, or value). While the importance of material perception for humans seems evident, we lack a full understanding of the underlying mechanisms. In a previous study, Fleming et al. (2013) investigated the relationships between attribute ratings and material classes (e.g., wood, glass, foliage, etc.) for photographs. In this paper, we extended on this study by using a big data approach to measure the perception of material properties in paintings. Our investigation is motivated by the assumption that to depict materials convincingly, painters presumably hold insights into visual cues that lead to the perception of various attributes.

Painters are masters of depiction and are capable of evoking a clear perception of a three-dimensional world, with complex lighting and recognizable materials. Interestingly, while the appearances of real materials are limited by the rules of physics, materials as depicted in paintings have no such constraints. Incongruencies between paintings and reality often go unnoticed by the viewer (Cavanagh, 2005). Instead of strictly following physics, painters have extracted the essential visual cues needed to trigger the perception of materials. Di Cicco et al. (2019) studied visual cues for gloss, which were implicitly discussed in a painting manual by the 17Th-century painter Willem Beurs (Beurs, 1692). They found that predictors that explained a large portion of the variance in gloss perception had implicitly been described within this 17Th-century manual. This shows that painters held insights into perception and that studying art could lead to new insights for perception scientists.

While art reveals insights into perception, conversely perception can be used to understand art. For example, several important art historical publications (Arnheim, 1965; Baxandall, 1995; Gombrich, 1960) use knowledge about perception to analyze art. Anecdotally, this approach can also be seen in artistic attributions such as in the case of *Still Life with Grapes and a Bird*, which is attributed to Antonio Leonelli by the Metropolitan Museum curator. In his attribution, the curator comments on “The tendency to geometrize the forms with shading that rigorously enhances their rotundity [...] the emphasis on surface effects—the grained wood [...] the clearly delineated shadows”. It is interesting to see that many of the curator’s terms are conceptually very similar to those used in perception science. The overlap between the perceptual sciences and art means that a fuller understanding of perceptual concepts could be beneficial for both fields.

Yet, how to study and quantify the depiction of materials in paintings? There are several standard psychophysical methods that potentially apply to the study of depicted materials, such as matching tasks, similarity ratings, or attribute ratings. The first method requires a material probe, which is an interactive image that can be adjusted to match the material attributes of the target stimulus. The probe can be parameterized by an analytical physical model (e.g. Ward), weight parameters of data-driven Bidirectional Reflectance Distribution Functions (BRDFs, Matusik et al., 2003a) or additive mixing of basis images representing canonical modes (Griffin, 1999; Zhang et al., 2016). As such, material matching tasks require predefined models for each material or sets of basis (BRDF) samples or (canonical mode) images to represent a wide range of materials. These methods are suitable for testing a wide range of materials, but not for all materials. For example, varying 3D textures (based on Bidirectional Texture Functions, BTFs) systematically and fluently is technically

extremely hard. Moreover, there is no method yet to vary 3D textures in a tractable interface such that all materials are covered. Therefore, material matching is not suited to study the wide variety of material attributes found in paintings as we aim to do here.

A second method, similarity ratings, relies on systematic variations of the stimulus set. Pellacini et al. (2000) asked participants to rate the apparent difference in gloss between pairs of images, without defining gloss. They then used multi-dimensional scaling to infer the dimensionality of gloss. Often, the similarity is not specified to the observer and can comprise of any combination of subjective criteria. Radonjić et al. (2015) asked participants to judge which of two test patches rendered under varying illuminations was more similar to a third patch under a fixed illumination, to investigate the relative contribution of illumination on color-constancy. The fact that comparisons are made between pairs or triplets implies that a very large number of trials (i.e. quadratically increasing with sample size for pairs) is needed. For the large number of materials that we aim to study this method is thus not feasible.

Last, a popular method relies on attribute scaling. In this method, a participant either rates single images explicitly (e.g. how glossy is this material?) or makes implicit forced-choice pairwise comparisons (e.g. which of these two images is glossier?). However, making comparisons inflates the trial number, so we decided to choose attribute ratings for single images to study materials depicted within paintings. This raises a straightforward question: which attribute names should be chosen that most completely covers the perception of the wide variety of materials present in paintings?

While a large variety of attributes has been investigated previously in perception literature, the majority of these attributes are studied in isolation, such as glossiness (Chadwick and Kentridge, 2015; Ferwerda et al., 2001; Kim et al., 2012; Marlow and Anderson, 2013; Marlow et al., 2012; Wiebel et al., 2015; Wijntjes and Pont, 2010), translucency (Fleming and Bülthoff, 2005; Motoyoshi, 2010; Xiao et al., 2014) or transparency (Fleming et al., 2011; Motoyoshi, 2010; Nakayama et al., 1990). These studies often investigate how the perception of attributes are affected by various distal cues such as shape (Fleming et al., 2004; Marlow and Anderson, 2015) and light (Adams et al., 2004; Fleming et al., 2003) or proximal (image structure) cues (Marlow and Anderson, 2013; Motoyoshi et al., 2007; Sharan et al., 2008).

Perceptual attributes are also studied in computer science, albeit with different motivations. Since attributes – such as gloss, translucency, and roughness – seem to be intuitive, perceptual parameters (i.e., attributes) are often preferred over physical parameters for rendering interfaces. To develop intuitive interfaces Serrano et al. (2018) collected attribute ratings for fourteen material attributes, also including high-level class descriptors such as plastic-like, fabric-like, and metallic-like. They mapped these perceptual attributes to an underlying PCA-based representation of BRDFs (i.e. physical parameters) and showed that their functionals were good predictors of the perceived material attributes.

Aside from perception and graphics, industrial design also makes use of quantitative attribute descriptions of materials. Designers use attributes to investigate and design user experiences (Karana et al., 2009). Interestingly, these three disciplines use partially overlapping but also distinct vocabularies to describe attributes. In computer science (e.g., Matusik et al., 2003b; Serrano et al., 2018) a large portion of attributes refer to material classes, such as metallic-like, plastic-like, ceramic-like, while perception and design studies often focus

more on sensorial qualities like fragility, hardness, and elasticity. The material classes in our study were based on human annotations. The attributes we selected for our study are mostly based on the perception and design studies. We arrived at a set of 15 materials and 10 attributes, the details of which will be explained in the methods section.

In this paper, we studied perceived material classes and how they vary in perceived material attributes for a large set of paintings. We first present our methods relatively extensively, as this partly consisted of collecting the painting images. Furthermore, we detail how we collected a large set of annotated segmentations of paintings via online experiments. Then we present the results, first addressing the subjects' consistency, as validation of our method, followed by a detailed analysis of the collected material judgments.

2.2. METHODS

Our stimulus collection serves a broader goal than the study reported here. The collection and annotation of artworks is part of ongoing research that will be comprehensively published at a later stage. In the current study, we perform a perceptual experiment in which we use a subset of the artworks and annotations we collected. We nevertheless report all the details on the collection of data for sake completeness.

In the following paragraphs, we detail our artwork annotation pipeline. This is followed by the perception experiment, in which participants judged material attributes for various material classes.

2.2.1. ETHICS

The study conformed to the declaration of Helsinki and was approved by the ethical review committee of the Technical University of Delft. All data was collected anonymously.

2.2.2. STIMULUS COLLECTION

In the context of a project where we are creating a database of depicted materials and their properties, we collected material segments from paintings. As this process was not part of the current studies' scope, we report it in the supplementary material. Below we report a summary.

We created a list of materials, based on the previous research mentioned in the introduction, plus observations of the paintings and our desire to cover as many materials in those paintings as possible. The list contains 15 materials as can be seen in Table 2.1:

| | | |
|--------|---------|---------|
| animal | ceramic | fabric* |
| flora† | food | gem |
| glass* | ground | liquid† |
| metal* | paper* | skin |
| sky | stone* | wood |

Table 2.1: The 15 materials used in this study. Items marked with * are in also in the Flickr Material Database (FMD, Sharan et al., 2013; Sharan et al., 2009). The two items marked with † are also present in the FMD but with a more narrow category, namely foliage and water instead of flora and liquid respectively.

The material list has six items in common with the Flickr Material Database (FMD,

Sharan et al., 2013; Sharan et al., 2009). These common items have been indicated above with an asterisk (*). Additionally, two materials from the FMD, foliage and water, were incorporated within our list as part of our broader labels flora and liquids – indicated with a †. These eight materials were also used by Fleming et al. (2013) with the original names as defined in the FMD. The remaining seven materials were included for a variety of reasons. Animal and food are two instances of materials that we included as an overarching concept, encompassing many different materials such as fruits, vegetables, and bread for food, and materials such as fur, claws, feathers, and scales for animal. We included gem to contain items such as pearls and precious stones. Ground and Sky were included because they often cover large portions of the painting’s surface. Note that 1) we defined ground as things such as dirt and gravel, without grasses or shrubbery, since those should be identified as flora and 2) we counted clouds as belonging to the sky. Lastly, we included (human) skin, instead of an overarching human concept such as is done with animal and food. We made this decision because skin is a very interesting material in its own right (sometimes even referred to as the “holy grail” of rendering) both from a classic perceptual point of view (Mattis et al., 2007; Stephen et al., 2011) as well as from a computer science perspective (Igarashi et al., 2007; Jensen et al., 2001) and an art-historical point of view (Lehmann, 2008).

ANNOTATION PIPELINE

In Bell et al. (2013b) OpenSurfaces was published, a database with annotated and segmented materials. This database is a public resource and is available at <http://opensurfaces.cs.cornell.edu/>. Bell et al. (2013b) created this database to fill the need within computer graphics to accurately model materials within context. Besides the database, they made their annotation pipeline, i.e. their process of collecting data, open-source. We have adapted their annotation pipeline to fit our purposes for the collection of material segmentations and annotations.

COLLECTING STIMULI

The collection of stimuli was executed in multiple steps. Here we provide a summary of each step. Each step is discussed in-depth within the supplementary materials. **Step 1**, collecting paintings: we collected digital images and the associated meta-data for paintings from 7 online museum galleries. **Step 2**, collecting materials: we used Amazon Mechanical Turk (AMT) to measure inferences of what materials were depicted within the painting. Participants had to indicate which paintings contained a requested material. For each painting, we collected at least 5 responses for each material and required an agreement of 80% to consider a painting to contain the material. **Step 3**, segment collection: participants segmented materials from paintings. In each task, the participant would see a painting and be requested to segment one instance of a specific material that was indicated to be present within the painting in step 2. **Step 4**, quality check: the quality of the created segments was checked by a minimum of 5 participants. **Step 5**, material check: It is possible that a participant wrongfully segmented wood, when tasked with segmenting metal, therefore in this step we asked participants to indicate what they perceived the material of the segment to be. **Step 6**, manual selection: in the end, we manually selected the 90 best segments per material.

2.2.3. PERCEPTUAL EXPERIMENT

Using the selected segments discussed above, we had a total of 198 AMT participants rate 10 perceptual attributes for each of these segments. The perceptual attributes are listed below. All participants were located within the USA, according to AMT and each participant had previously completed at least 1000 tasks on the AMT platform of which at least 95% had been accepted by the creators of those tasks.

ATTRIBUTES AND IMAGE STATISTICS

We created a list of ten perceptual material attributes. Our attribute list has five items in common with Fleming et al. (2013), i.e., those indicated with an asterisk. When applicable we have copied the original attribute definitions and we have created our attribute definitions to be similar to the other attributes used in Fleming et al. (2013). Additionally, we split colorful into multicolored and vivid. We expected colorful might be difficult for naïve participants, since it could be interpreted as “many, low-intensity colors” or “a single, very intense color” for multicolored and vividness respectively. Additionally, we included translucent to the existing transparent attribute, because we found that some participants were aware of the optically defined difference between these two, while some were not. Altogether this resulted in the following list and definitions:

- Bendable: How bendable is the material? Low values indicate that the material is highly rigid and could not easily be bend; high values indicate that a small force would be required to bend the material.
- cold*: To what extent would you expect the surface to feel cold to the touch? Low values indicate that the material would typically feel warm or body temperature; high values indicate that the material would feel cold to the touch.
- fragile*: How fragile or easy to break is the material? Low values indicate that the material is highly resistant and could not easily be broken; high values indicate that a small amount of force would be required to break, tear, or crumble the material.
- glossy*: How glossy or shiny does the material appear to you? Low values indicate a matte, dull appearance; high values indicate a shiny, reflective appearance.
- Hairy: If you were to reach out and touch the material, how hairy would it feel? Low values indicate that the surface would feel hairless; high values indicate that it would feel hairy.
- hard*: If you were to reach out and touch the material, how hard or soft would it feel? How much force would be required to change the shape of the material? Low values indicate that the surface would feel soft; high values indicate that it would feel hard.
- Multicolored: How multi-colored does the material appear to you? Low values indicate a monochrome (single-colored) appearance; high values indicate many colors.
- rough*: If you were to reach out and touch the material, how rough would it feel? Low values indicate that the surface would feel smooth; high values indicate that it would feel rough.
- transparent/translucent: To what extent does the material appear to transmit light? Low values indicate an opaque appearance; high values indicate the material allows a lot of light to pass through it.

- **vivid:** How vivid does the material appear to you? Low values indicate a dull, grayish appearance; high values indicate a strong vivid color.

Next, we also defined and calculated 4 simple, image histogram statistics for each of the material segments.

- **The contrast:** Defined as the Michelson contrast: $contrast = \frac{l_{max} - l_{min}}{l_{max} + l_{min}}$, where l_{max} and l_{min} are taken as the 95th and 5th percentile of the luminance distribution of the material segment (Michelson, 1891).
- **Skewness:** The skewness of the luminance distribution of the material segment.
- **Colorful:** The colorfulness, measured as the ratio of voxels filled in three-dimensional RGB color space to the total number of voxels, where the RGB color space was rescaled to 0 to 15, as opposed to the conventional 0 to 255, for each of the material segments.
- **Mean luminance:** the mean luminance of the image segment.

STIMULI

From the 90 segments per material – as discussed above – we randomly selected 30 segments, for each of the 15 materials, making a total of 450 stimuli. We chose to include this randomization to reduce the chance of experimenter bias, considering we originally selected the 90 segments per material. We subdivided these 450 segments into 5 sets of 90, where each set contained 6 segments per material. In other words, each set had 6 segments of wood, 6 segments of metal, etc. These sets were used in experimental blocks. We chose to partition the data into these 5 sets, to reduce the number of trials per participant. Without partitioning the data into these 5 sets, every participant would have needed to complete (450 stimuli X 3 repetitions=) 1350 trials, which we consider too many for web-based experiments. With these 5 sets, participants only need to complete 270 trials. The specific choice of 5 sets, over for example 9 sets, is arbitrary. Splitting the experiments into these sets implies that we calculated inter-rater reliability within each set.

We presented the segments in a section of the original painting. We created a square context box around the segment, which is, in essence, a bounding box around the segment with margin. The context box size was calculated as the maximum of the width or height of the segment, multiplied by 1.25. We took the maximum to ensure the context box is a perfect square. In some cases, this meant that the context box boundaries exceeded the dimensions of the original painting. To keep the aspect ratio consistent, we included this overflow as part of the segment and colored the overflow with the average of the color of the painting part within the bounding box. A few examples can be seen in Fig. 2.1.

PROCEDURE

Each of the 5 sets of images was rated on each of the 10 attributes, making a total of 50 set/attribute combinations. Each of 50 combinations would be rated by 10 different participants. Each participant would only see one set of images and rate this set on one attribute per task. Participants could choose how many of these combinations they would rate. This means that a single participant could, in theory, do each of the 50 experimental blocks once and that the total number of participants should be between 10 (i.e. each participant did all 50 set/attribute combinations) and 500 (i.e. each participant did only 1

task). In practice, 198 participants performed the task on average 2.5 times each, with 110 participants only performing one task. The full distribution is presented in Fig. 2.2.

Each task contained three repetitions of the 90 images, making a total of 270 trials. We used a Fisher-Yates Shuffle to create three shuffled permutations of the set and concatenated these three permutations in each task. This allows us to measure the intra-observer correlation (with 3 repetitions) next to the inter-observer correlations (with 10 repetitions)

2

TASK

Participants on the AMT platform were capable of choosing and selecting what tasks they wanted to work on. Once participants had selected our tasks, they would first be shown a text-based tutorial. After a 10 second interval, participants were able to start the task. First, the tasks displayed the main question in bold: “How [attribute] is this material?”, followed by our definition of that attribute. To give participants an impression of the range of stimuli we showed them a random selection of one-third of the stimuli, on which they were told to base their ratings. They could click the start button to start the first trial.

In each trial, the participants were shown the same question and definition as mentioned above, as well as one segment at a time, such as shown in Fig. 2.3. Upon the start of a trial, or when a participant clicked the show outline button, the outline would be indicated with a flashing red line around the edges of the segment for 1 second. On the right of the image was a vertical slider, ranging from 0 at the bottom with the label “not [attribute]” to 100 at the top with the label “[attribute]”. On the right of the slider was a small box indicating the current value selected. Participants could move the slider using the mouse. On a left-click the participants could progress to the next trial. A button allowed the participant to go to the previous trial.

EXCLUSION CRITERIA

As discussed above in the AMT section, we are capable of adding AMT qualifications, in an effort to improve data quality. First, we added three default qualifications, namely 1) that each participant needs to have completed at 1000 tasks and 2) that each participant needed to have at least 95% of those tasks approved and 3) that the participants were located within the United States of America.

Furthermore, we noticed in pilot experiments that some observers seemed to respond both fast and random. As their actual response cannot be an exclusion criterion (we cannot know what the perceive), we deemed it wise to use response time as selection criterion: if observers on average responded below 1 second, their data were excluded for further analysis.

ANALYSIS

For the analysis of the data, we used several statistical methods and techniques. We will look into the intra- and interobserver correlations. Furthermore, we use Principal Component Analysis (PCA) on the perceptual data. This technique applies an orthogonal transformation to data to produce a new set of uncorrelated variables, i.e. components. These components are ordered on the explained variance within the original data, where the first component explains the largest portion of the variance within the original data. Last, we also make use of a Procrustes analysis, which tries to find the best fit for a set unto a target set by minimizing the linear distance between points in the original set and the target set.

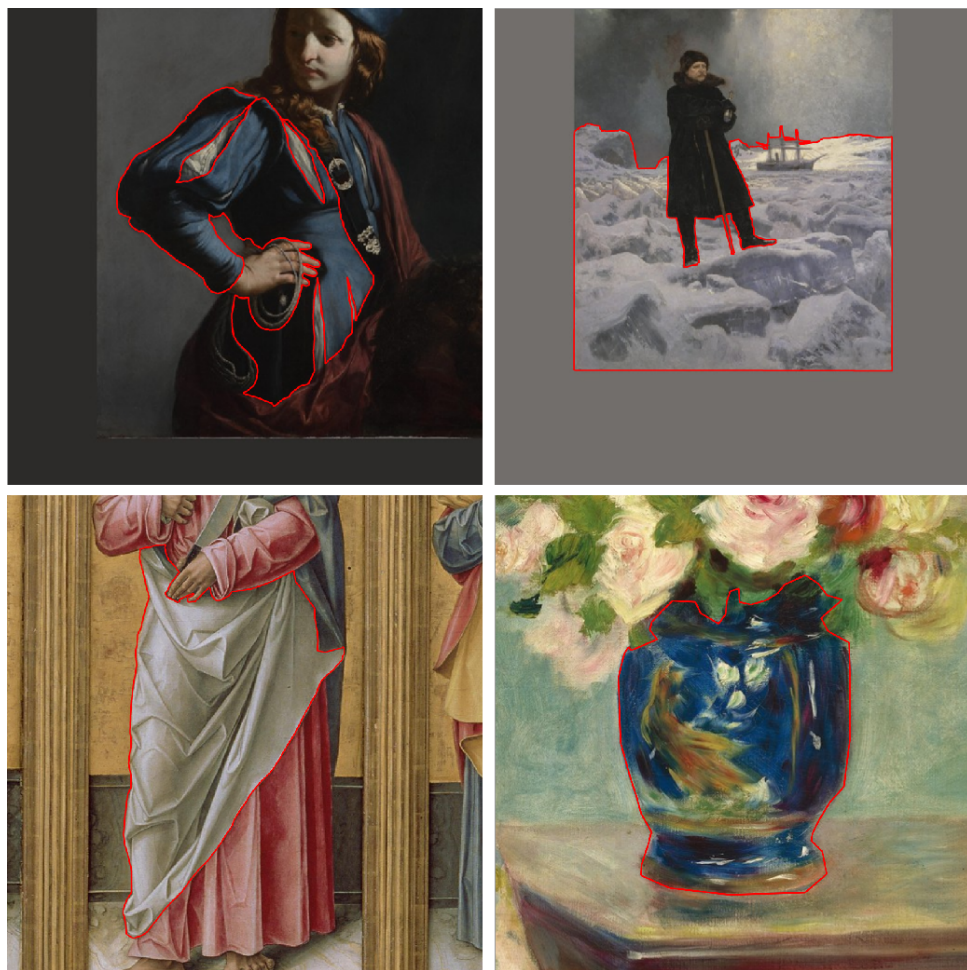


Figure 2.1: Examples of four stimuli. For the top two, the context size exceeds the dimensions of the original painting, and the overflow has been colored with the average RGB color value of the painting contained within the bounding box. For the bottom two, the context size does not exceed the original painting dimensions and is thus only a section of the painting without any overflow. The red outlines indicate the segments. From top-left to bottom-right: detail of David with the Head of Goliath (c. 1645) by Guido Cagnacci; The Explorer A.E. Nordenskiöld (1886) by Georg von Rosen; detail of Polyptych with Saint James Major, Madonna and Child, and Saints (1490) by Bartolomeo Vivarinil; and detail of Mlle Charlotte Berthier (1883) by Auguste Renoir.

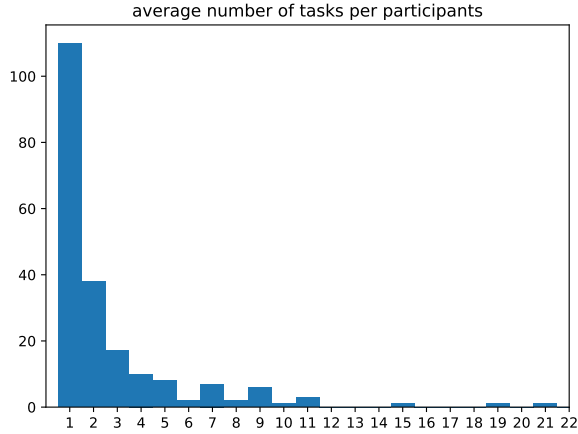


Figure 2.2: Distribution of completed rating tasks per participants with number of tasks on the x-axis and number of participants on the y-axis.

2.3. RESULTS

2.3.1. DATA QUALITY; INTRA- AND INTER- CORRELATIONS

First, we analyzed the internal consistency by calculating the intra- and interobserver correlations. Each task contained 3 repetitions of each stimulus and was judged by 10 different participants for each material attribute. The average intra-observer correlation is 0.76 (std = .08), which is higher than the average of 0.48 (std = .16) for the inter-observer correlation.

We plotted the correlations in Fig. 2.4, where each point corresponds to one of the 50 set/attribute combinations, with the intra-observer correlation as a function of the inter-observer correlation. Note that each of these 50 combinations was rated by a different group of 10 participants. We fitted an ellipse around the 5 points which belong to the same attribute. The distribution along the intra-observer axis shows that participants are, in general, consistent and that there is very little difference between the material attributes. The distribution of the inter-observer correlations shows a larger spread, implying participants do not always agree amongst each other. The inverse, the small spread on the averaged intra-observer correlations indicate the high agreement rate within participants. Additionally, the material attributes cluster together, but the clusters are spread out over the inter-observer correlation dimension implying that the magnitude of (dis-)agreement between participants is material attribute dependent.

2.3.2. MATERIAL JUDGMENTS

We collected a total of 135,000 human judgments about how much a specific stimulus depicted a specific attribute. We have plotted the distributions of these ratings per attribute in Fig. 2.5. At a glance, it becomes clear that the distributions are generally broad and flat, except for some attributes at 0. The stimuli cover the whole range for each attribute, and


How vivid is this material? Instructions

How vivid is this material?

Definition: How vividly colored does the material appear to you? Low values indicate a dull, grayish appearance; high values indicate a strong vivid color.

Trial 2 out of 270 Show outlines Back

Base your judgement only on the material in the red outline.



Vivid

Value

100

Not vivid

Figure 2.3: Example of the perceptual judgment task. At the top, the question and definition are repeated, which participants would have seen in the instructions. The task shows one segment at a time, as part of the original painting. In the live version, the red outline appears flashing at around 10hz at the onset of each trial (or when the participant pressed the corresponding button) to indicate the segment boundaries and disappears after a second. The slider can be moved by moving the mouse up and down, while a left mouse-click progresses the experiment to the next trial. The painting is a section of *The Annunciation* (c.1660) by Godfried Schalcken

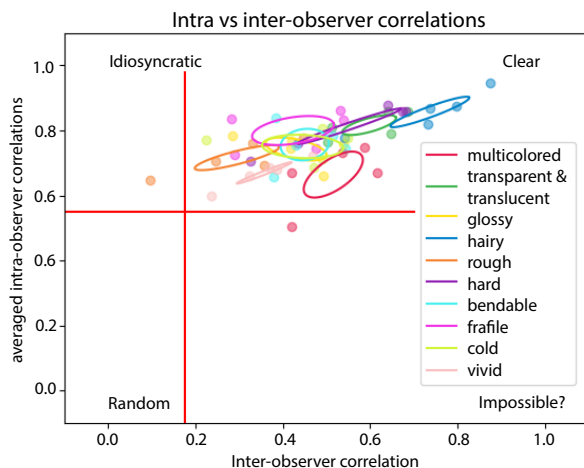


Figure 2.4: Each of the 50 set/attribute combinations expressed in a 2D intra/inter-observer correlational space. The data are color-coded to indicate the material attribute that was judged. Ellipses (1 SD) are fitted for each material attribute based on the 5 experimental blocks relating to that attribute. The red lines represent the one-sided 5% alpha significance level, with 88 and 8 degrees of freedom for intra and inter-observer correlations respectively.

when an attribute is present it is more or less equally likely to be present in any quantity.

We visualized the averaged distributions of material attributes for each material in Fig. 2.6. Here, we found some remarkable similarities and those reported by Fleming et al. (2013) and therefore we reproduced these in Fig. 2.7. To make an accurate comparison, it should be noted again that our study did not use the same set of attributes as did Fleming et al. (2013). Our signature included hairy and bendable, whilst excluding naturalness and prettiness. Furthermore, we split colorfulness into vividness and multicoloredness and included translucency into transparency. What we observe is that the distributions seem to follow the same pattern for the materials that are in common between our study and Fleming et al. (2013). To quantify this relationship, we performed a non-parametric Wilcoxon signed-rank test in which we paired the mean values for the materials and attributes that the current study has in common with the study of Fleming et al. (2013). Note that we equate Fleming's transparency with our 'transparent/translucent' and Fleming's colorfulness with our vivid and multicolored. The test showed that there was no significant difference between the attribute ratings for photographs and paintings $Z(56)=790$, $p = .94$

2.3.3. MATERIAL ATTRIBUTE CORRELATIONS

Correlations likely exist between the material attributes: a change in one attribute could lead to a predictable change in other attributes. We quantify these relations by calculating the correlations using Bonferroni adjusted alpha levels of .001, .0001, and .00001 (.05/45, .005/45 and .0005/45 respectively). These correlations have been visualized in Fig. 2.8. The highest correlations are found between roughness and hardness ($r = .54$, $p < .0001$), and

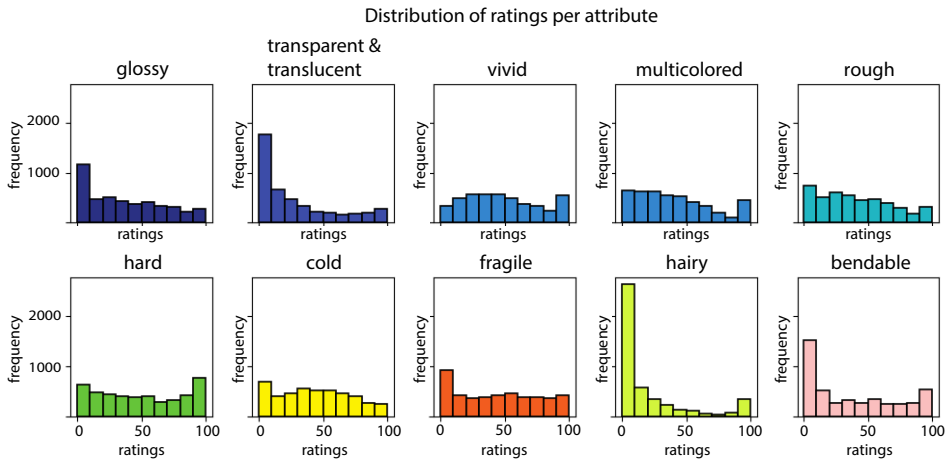


Figure 2.5: Distribution of all the judgments per attribute for all materials. The colors are in reference to the colors used by Fleming et al. (2013).

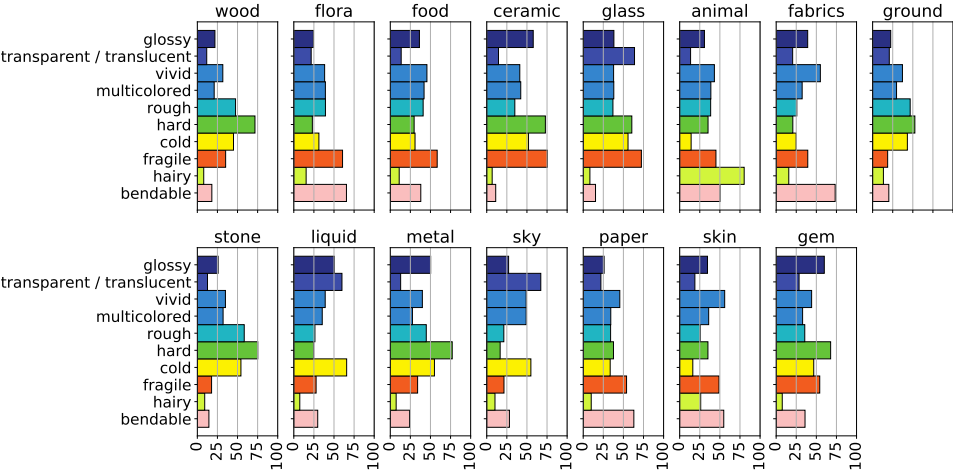


Figure 2.6: The averaged ratings for each attribute per material

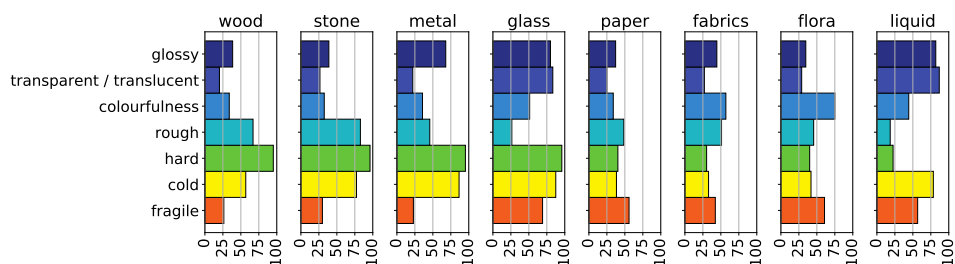


Figure 2.7: A recreation of the average rating for the attributes and materials from Fleming et al. (2013), for the materials and attributes that are shared between our study and Flemings study. Note that in our study, we split up colourfulness into vivid and multicoloredness. Figure adapted with permission; original copyright belongs to ARVO.

between vividness and multicoloredness ($r = .5$, $p < .0001$). The lowest correlation is found between hardness and bendableness ($r = -.6$, $p < .0001$). The majority – 33 out of 45 – of the attributes pairs only displayed a small (i.e. $r < .3$) correlation. This implies that while there is overlap, most attributes cover a distinct area of a high-level material-feature space.

2.3.4. PCA

To analyze the relationship between material attributes and to see if material attributes can predict material class identity, we applied a Principal Component Analysis (PCA) to uncover the underlying multi-dimensional attribute feature space. This technique applies an orthogonal transformation to remap the original data set in such a way that the new dimensions (components) are linearly uncorrelated, and ordered by the quantity of variance, where the first dimension explains the most variability within the original dataset. We have visualized the first two components in Fig. 2.9, which explain 52% of the variability within the data. Adding a third, fourth, or fifth component captures 68%, 76% and 83% of the variability respectively. These numbers are roughly comparable to the two numbers Fleming et al. (2013) reports: 62% for the first two PCs and 93% for the first five PCs. However, it should be noted that our measured dimensions are not identical (see Attributes and image statistics in the method section). We have plotted a full scree plot in Fig. 2.10 and added the factor loadings for the first 4 components in Table 2.2.

We also ran a PCA for each material, i.e. with only the 30 datapoints belonging to that specific material, as opposed to all 450 datapoints for all materials. We visualized these for paper, skin, flora, and fabric in Fig. 2.11. The remaining material plots are included in the supplementary materials.

We have included all the factor loadings for all the PCAs (1x global and 15x material specific) within the supplementary materials. Next, we applied a Procrustes analysis to map each material-specific PCA onto the associated data-points within the global PCA space, i.e. the 30 segments for one material-specific PCA were mapped unto the 30 corresponding segments within the global PCA. Here, the residual error quantifies how much a material-specific PCA deviates from the global PCA. Or, inversely, how similar the variance within one material is in comparison to the global variance found between all materials. We

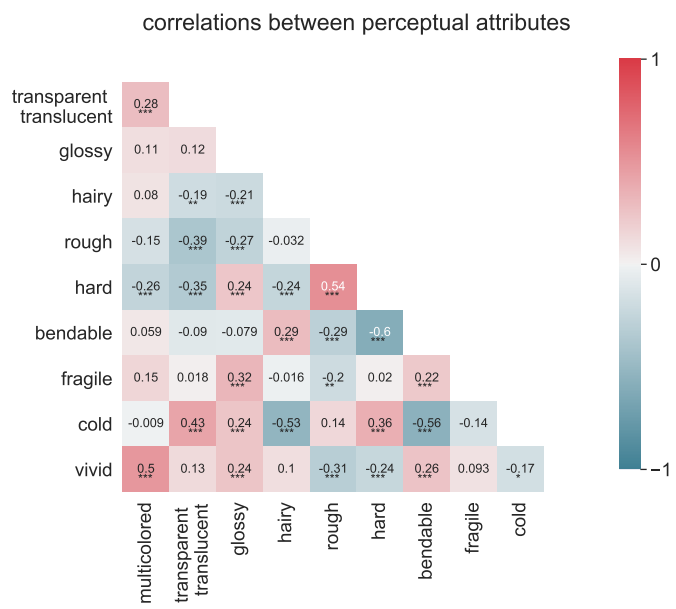


Figure 2.8: Correlation matrix heatmap, we have masked the values along the diagonal, which would always simply be 1 and the symmetrically identical values. * indicates $p < .001$, ** indicates $p < .0001$ and *** indicates $p < .00001$.

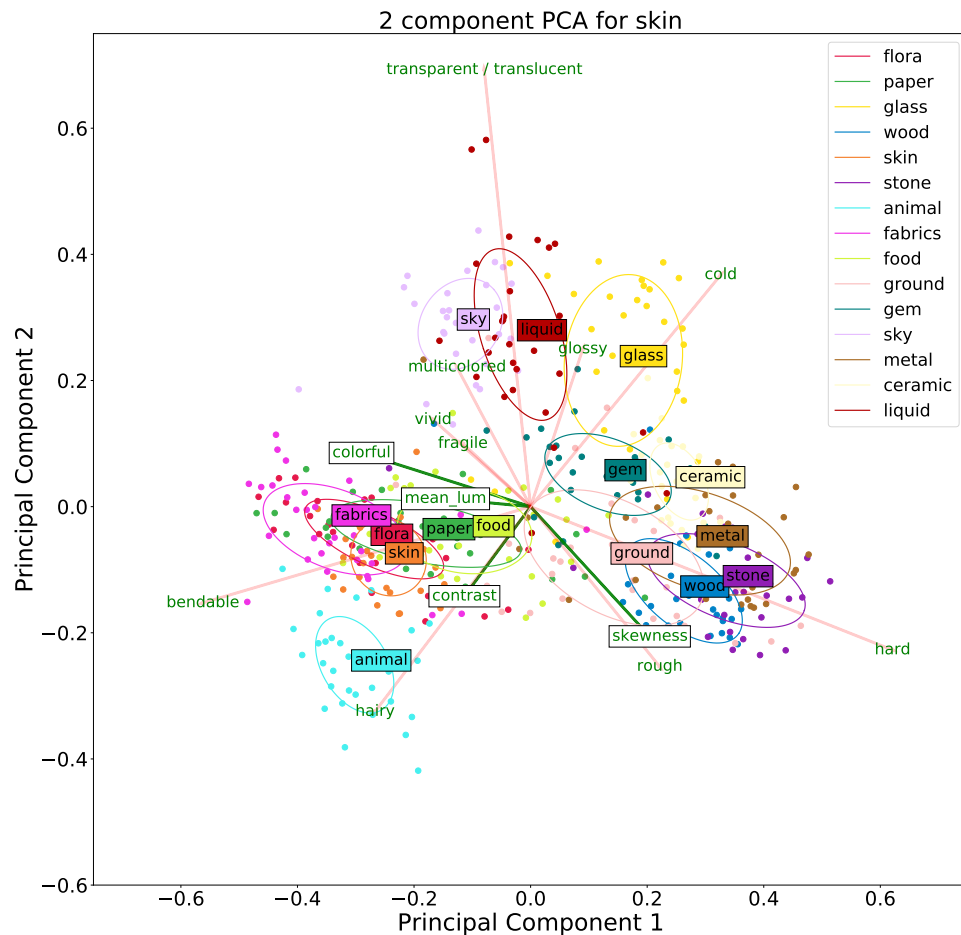


Figure 2.9: A visualization of the first two PCA dimensions. The color of the points relates to material class identity. The factor loadings of the original dimensions are plotted as red vectors. Lastly, we fitted ellipsoids ($sd = 1$) for each material class. Note that the PCA is not fed any class data, the clustering of material classes observed is thus purely based on the perceptual data.

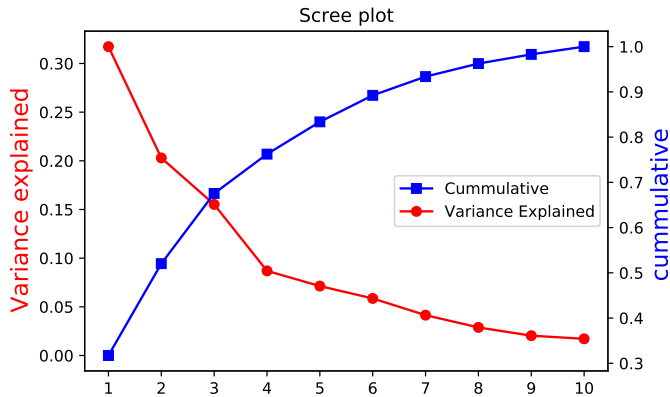


Figure 2.10: Scree plot for the PCA visualized in Fig. 2.9

| Attribute | PC1 | PC2 | PC3 | PC4 |
|---------------------------|--------|--------|--------|--------|
| multicolored | -0.126 | 0.221 | 0.070 | 0.511 |
| transparent / translucent | -0.080 | 0.693 | -0.200 | 0.019 |
| glossy | 0.090 | 0.250 | 0.471 | 0.157 |
| hairy | -0.267 | -0.325 | -0.063 | 0.583 |
| rough | 0.221 | -0.253 | -0.078 | 0.023 |
| hard | 0.621 | -0.227 | 0.305 | 0.152 |
| bendable | -0.562 | -0.153 | 0.171 | -0.341 |
| fragile | -0.116 | 0.100 | 0.757 | -0.167 |
| cold | 0.326 | 0.368 | -0.093 | -0.088 |
| vivid | -0.167 | 0.138 | 0.141 | 0.445 |

Table 2.2: Factor loadings for the first 4 principal components.

applied the Procrustes analysis on the first two components, as opposed to all ten. The reason for this is simple: a PCA works by applying an optimized transformation on a data set, while the Procrustes analysis tries to find an optimized transformation to map one dataset onto another. Consider that the material-specific PCA dataset is a subset of the global PCA dataset. This means the raw data for the PCA's are the same but have undergone different transformations within a 10-dimensional space.

Thus, applying a 10-D Procrustes analysis would perfectly map the material-specific subset onto the global PCA leaving a residual of exactly 0. Instead, we take the two primary components which explain the major part of the variability. Note that the loadings of the first two PCA dimensions can change from the global to the material-specific models and that materials with a larger variability can have a larger influence on the global variability relative to materials with less variability. The residuals of the Procrustes analysis are listed in Table 2.3. We also used randomly generated data points drawn from a uniform dis-

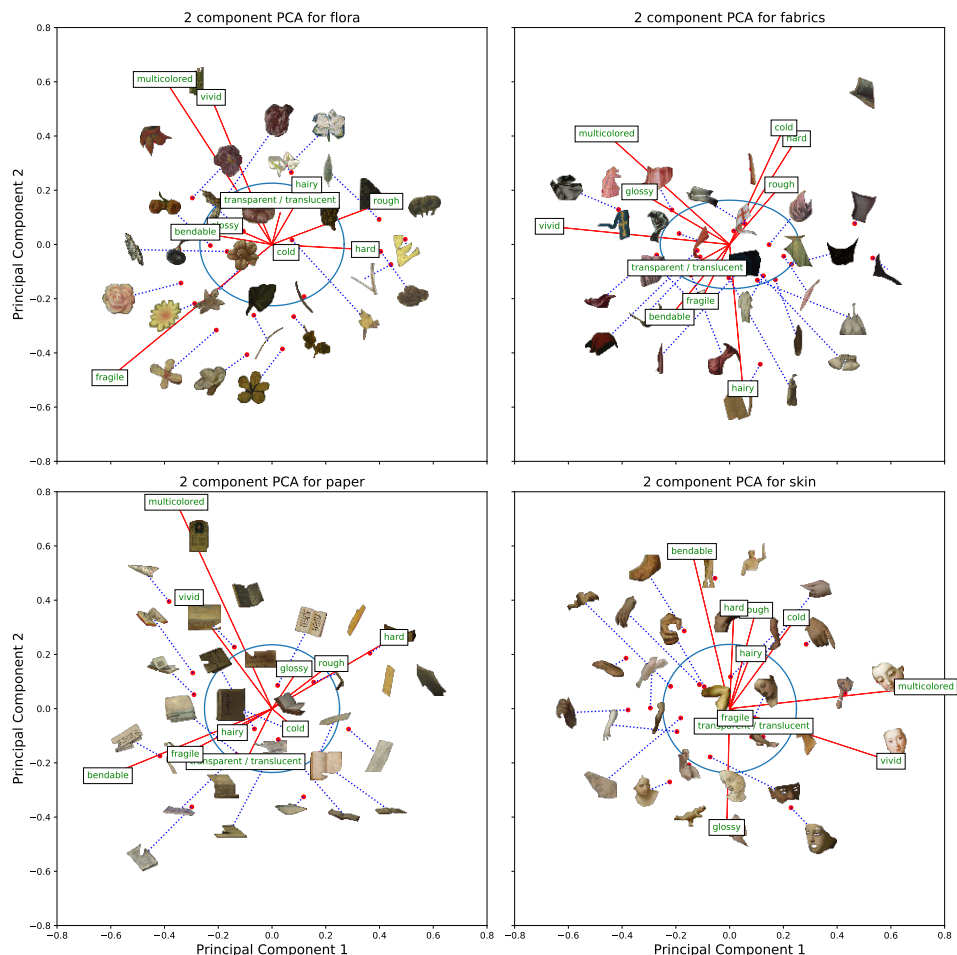


Figure 2.11: Four visualizations of the first two primary components for the material-specific PCA for flora, fabric, paper, and skin. Each PCA was run with only the 30 stimuli per material. The red vectors indicate the factor loadings of each attribute. We plotted the actual stimuli within the PCA space. The blue lines connect the stimuli to their actual position within the space when the stimuli would otherwise overlap. The ellipse is fitted around the points (1 sd).

tribution and mapped these to each of the material subsets within the global PCA using the Procrustes analysis. We repeated these 10,000 times, for each material, to find an averaged residual error of 0.9508 which functions as a comparison. The results are visualized and ordered in table 3 2 and show that the residuals range from 0.14 to 0.74 and are all smaller than for the random set. This shows that intra-material variations are described relatively well by the variation in the global PCA space, but for some materials better than others.

| Material | Residual |
|----------|----------|
| fabric | 0.14 |
| metal | 0.19 |
| stone | 0.22 |
| ground | 0.23 |
| glass | 0.25 |
| food | 0.28 |
| paper | 0.36 |
| wood | 0.36 |
| liquid | 0.37 |
| ceramic | 0.37 |
| flora | 0.5 |
| animal | 0.46 |
| sky | 0.53 |
| gem | 0.72 |
| skin | 0.74 |
| Random | 0.95 |

Table 2.3: Table of residuals of the Procrustes analysis. Lower residuals indicate more generic materials.

2.3.5. IMAGE STATISTICS

As detailed in the Methods section we calculated simple histogram-based image statistics for each image stimuli. We correlated the material attributes with these image statistics, both averaged over materials and per material. We adjusted the alpha levels using Bonferroni correction to .0013, .00013 and .000013 (.05, .005, and .0005 divided by 40 respectively). Over all materials generalized, we found some correlations. Colorful correlated with multicolored ($r = .44$, $p < .00013$) and with vivid ($r = .42$, $p < .00013$), suggesting our color metric indeed captures multicoloredness to a certain degree. Furthermore, mean luminance correlated with transparent/translucent ($r = .36$, $p < .00013$) and with hardness ($r = -.31$, $p < .00013$). These correlations and the remaining, smaller correlations have been visualized in Fig. 2.12. The significant correlations per material have been listed in Table 2.4. Here the colorful-multicolored and colorful-vivid relationships are often found to be significant. In addition, the skewness of the luminance distribution and mean luminance are found to be related to specific attributes in a material-dependent manner.

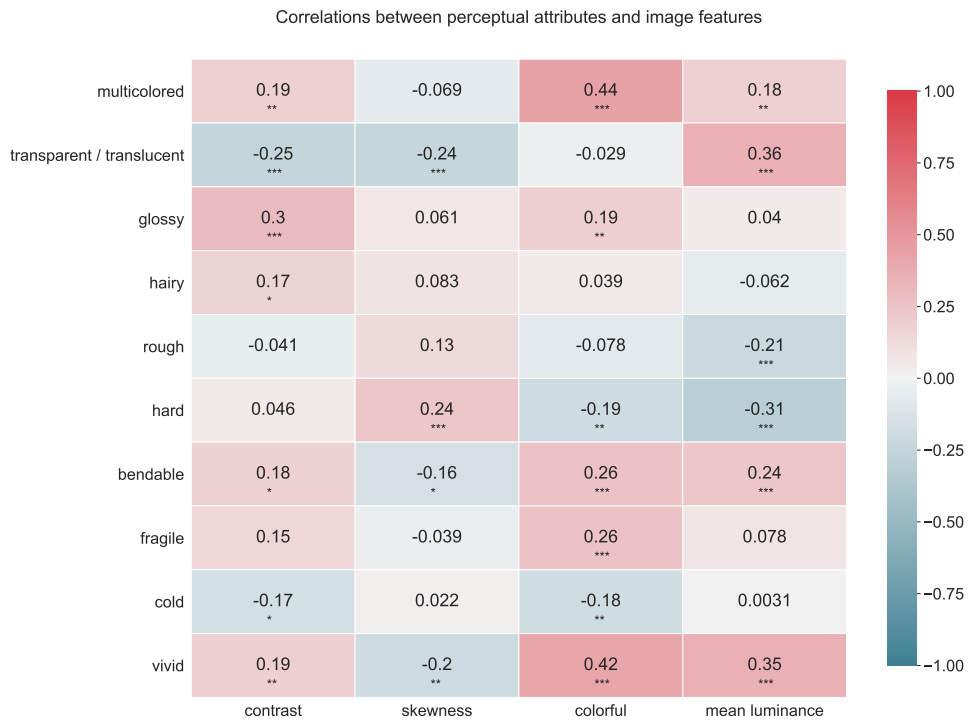


Figure 2.12: Correlation between the perceived perceptual attributes and image statistics (i.e. image statistics). * indicates $p < .0013$, ** indicates $p < .00013$ and *** indicates $p < .000013$

| Material | Perceptual Attribute | image Statistic | <i>r</i> | sig |
|----------|---------------------------|--|----------|-----|
| ceramic | multicolored | colorful | 0.75 | *** |
| glass | vivid | mean luminance | 0.6 | * |
| gem | multicolored | colorful | 0.75 | *** |
| gem | bendable | Skewness of the luminance distribution | −0.81 | *** |
| gem | bendable | mean luminance | 0.83 | *** |
| gem | fragile | Skewness of the luminance distribution | −0.73 | *** |
| gem | fragile | mean luminance | 0.72 | *** |
| sky | vivid | colorful | 0.69 | ** |
| fabric | transparent / translucent | Skewness of the luminance distribution | −0.58 | * |
| wood | multicolored | colorful | 0.73 | *** |
| wood | vivid | colorful | 0.63 | * |
| wood | multicolored | mean luminance | 0.68 | * |
| metal | multicolored | colorful | 0.57 | * |
| ground | multicolored | colorful | 0.68 | ** |
| ground | vivid | colorful | 0.58 | * |

Table 2.4: Significant correlations between perceived material attributes and image statistics. * indicates $p < .0013$, ** indicates $p < .00013$ and *** indicates $p < .000013$

2.4. DISCUSSION

In this study, we collected human perceptual judgments for 10 material attributes for paintings of 15 material classes. The consistency within participants indicates that participants understand and are capable of performing the task, which shows that our experimental setup allows for measuring the perception of material attributes in paintings via AMT and supports the validity of the data, while the inconsistency between participants shows that individual differences exist in how participants interpret the depictions. Additionally, we found that the material signatures and the material feature PCA spaces show many similarities to those of Fleming et al. (2013) based on material photographs, as well as of Zhang et al. (2019) based on mixtures of canonical reflectance modes, which implies that the perception of material properties functions independently of the medium of depiction and the structure found represents generic key components underlying material perception. Lastly, we looked at the residuals that result from mapping material-specific PCA data onto the global material PCA and found that the variation within materials is partially explained by the variation between materials, but that this varies depending on the material. Below we will discuss these findings in more detail.

In the task, participants were shown a square bounding box and asked to use only the segment, outlined in red (see Fig. 2.3) when they make their judgments. Participants may make their judgments based on the object category inferred from this red outline. However, it should be noted that for the vast majority of segments, the materials are partially occluded by other materials or objects and tend not to be informative of the object identity, see for example those in Fig. 2.1 and Fig. 2.3. Besides being a measure of the internal validity, the high consistency displayed by participants shows that the perception of material attributes is distinct and that participants have a clear perception of these attributes. Despite this clear perception, disagreement between participants does exist. The magnitude of this disagree-

ment – which ranged from $r=.01$ to $r=.87$ – appears to depend on the perceptual attributes. Roughness induced the highest level of idiosyncrasy, while hairiness is the most consistent between – and also within – participants. The overall pattern of (in)consistencies between participants for the perceptual attributes in our results appears to be very similar to those reported by Fleming et al. (2013), however, it is interesting to see that roughness is one of the most consistent in their results, while in our study it is the least consistent between participants. It is unclear why these results differ. Possibly, roughness is too multidimensional to be measured in one scalar measure; even for a single type of surface structure it was found that its roughness perception was multidimensional (Padilla et al., 2008).

In the experiment conducted by Fleming et al. (2013) they found that materials tend to display statistical regularities, such as glass tending to look glossy, transparent, smooth, hard and so on, while water also tends to look glossy and transparent, but not at all hard. They postulated that these distinctive features can be interpreted as a signature of a material class. In this study, we found that the material signatures for painted materials are also distinct and that some materials appear to be more similar to each other than others. For example, wood and stone have a very similar material signature, and glass and liquid are almost identical except that glass is – obviously – harder and more fragile. Also, many of the between-attribute correlations seem intuitive, such as the negative correlation between hardness and bendableness, and the negative correlation between hairy and cold. Furthermore, we find some remarkable similarities with the material signatures reported by Fleming et al. (2013), which shows that the perception of photographed and painted materials results in similar associations, which suggests a generic underlying mechanism.

When looking at the first two components of the PCA we find that materials tend to cluster together, but that clusters for different materials can overlap. This implies that the perceptual judgments are material specific in terms of perceptual attributes, but that extracting a specific material identity based solely on the perceptual attributes measured in this study would likely be prone to errors. Possibly by adding a more extensive list of perceptual attributes, a predictor model could predict the material class identity. Furthermore, when looking at the PCA visualization, it is again interesting to note the similarity between the data presented here and the PCA dimensions reported by Fleming et al. (2013). This implies that material perception functions independently of the medium of depiction. One could argue that this can be explained by semantic knowledge: material classification is extremely fast, and after classification we gain access to semantic information, which in turn could have a top-down influence on the perception of perceptual attributes (Sharan et al., 2014; Wiebel et al., 2013; Xiao et al., 2011). Then, the estimation and perception of material properties could be argued to be driven by a top-down influence from material recognition. This top-down influence would then also be independent of the medium. To test this idea, Fleming et al. (2013) conducted a second experiment, where participants rated the material attributes of semantic stimuli (i.e., only material class names). They found that material property ratings for the semantic-only represented material classes were very close to the cluster centers for photographic representations. It would be naïve to claim that semantic top-down influences can fully explain material perception since we are capable of making judgments based on material properties within a material class (which fruit looks fresher? which sweater looks softer?). It does, however, imply that our perception might be influenced by semantic top-down influences when viewing materials. Furthermore, Zhang et

al. (2019) had participants perform a material probing task on a canonical set of computer rendered base images, where material perception could only rely on material reflectance since there was no semantic information. They found a PCA space that is similar to our PCA space and the one reported by Fleming et al. (2013). Thus, while semantic information might explain a portion of material perception, it does not explain the perception of intra-class variations. Thus, while semantic information might explain a portion of material perception, it does not explain the perception of intra-class variations. Furthermore, since the global PCA structure cannot entirely be explained by semantic information, it is implied that a portion of the global PCA structure (i.e. the portion not explained by semantic information) is independent of the medium of depiction

The PCA space visualizes the majority of the perceptual variability of the materials and in doing so, shows how materials are – and importantly, how they are not – related. We were interested in seeing how similar the variability within one material is in comparison to the variability found between all materials. To do so, we took the variability within one material and analyzed how well this mapped onto the variability between all materials. To quantify, we performed the Procrustes analysis. Here, the lower the residual error is for a material, the closer the variability within the material resembles the variability between materials. The first, intuitive result is that different materials vary differently across the perceptual attributes we measured. This effect could be highly dependent on the stimulus set. However, if we consider the previous results, namely that different stimulus sets have remarkably similar PCA spaces, even with different methods of depictions (e.g. paintings in our study, photographs in Fleming et al. (2013) and reflectance modes mixtures in Zhang et al. (2019)) and that the material signatures are very similar for photographic and painted images (see Fig. 2.9 and Fig. 2.10). Furthermore, the finding that different materials varied differently across perceptual attributes, further suggests that the similarities we find are not just a semantic effect. If it was merely a semantic effect, it would be more likely that the Procrustes residuals would show little variability between materials. Looking at specific materials, the residuals showed that fabric, metal, and stone are relatively generic materials: the variability within these closely resembled the variability between all materials. Gem and skin were found to be much more distinctive materials, as the variability did not resemble the global variability. In summary, the residuals of the Procrustes showed that different materials varied differently across perceptual attributes. Some intra-material variations are quite generic, i.e. they closely resembled the global material PCA space. However, other materials are more unique and resembled the global variability.

It has previously been proposed that variations in the perception of specific material attributes could be explained by image statistics (Baumgartner and Gegenfurtner, 2016; Motoyoshi et al., 2007, but this has also been debated (e.g. Kim and Anderson, 2009). Considering the large amounts of data collected in our study, we decided to calculate several simple, histogram-based image statistics, to see whether those could explain variations of the perceptual attributes. It appeared that the small set of image statistics we used did not correlate strongly with the perceptual attributes across all materials. This could perhaps be expected, after finding that the Procrustes residuals varied across materials. We did find some weak and moderate correlations, however, and the correlation we found between transparency/translucency and the mean luminance of the stimuli seems an interesting finding. It has previously been argued that the average luminance is a poor predictor

for transparency in natural images and that this is due to the luminance of an object being strongly influenced by scene illumination and the objects spatial and directional properties, shape, and context (Fleming and Bühlhoff, 2005; Fleming et al., 2011; Fleming et al., 2004; Koenderink and van Doorn, 2001). When looking at the correlations per material, it is interesting to note that the majority of the correlations, as well as the strongest, were all found for gem. Perhaps, it is possible that this material simply shows stronger optical effects than other materials.

As previously noted, the image database we used is different from existing image sets, such as the FMD, because each image comes from a certain artist and a certain period. Although the sample size for the perception experiment is relatively small with respect to all paintings to our disposal, it does give us some idea of interesting future directions for the study of art and perception. For example, we conducted a small pilot, not reported here, where we found that gems are perceived as glossier for recent paintings relative to older paintings. A typical art historical hypothesis would include the invention of oil paint that supposedly increased the convincingness of materials. Yet, most of our paintings are after this invention and the material rendering revolution that van Eyck caused in the 15th century. But there can be many other reasons and possibly even patterns that have not yet been identified in art history. With our continued work on creating the painting database of material depictions we hope to further investigate these questions

ACKNOWLEDGMENTS

We appreciate the work and feedback of AMT participants that participated in our user studies. This work is part of the research program “Visual Communication of Material properties” with project number 276-54-001, which is financed by the Netherlands Organization for Scientific Research (NWO).

BIBLIOGRAPHY

- Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature neuroscience*, 7(10), 1057–1058.
- Adelson, E. H. (2001). On Seeing Stuff: The Perception of Materials by Humans and Machines. *Proceedings of the SPIE*, 4299, 1–12. <https://doi.org/10.1117/12.429489>
- Arnheim, R. (1965). *Art and visual perception: A psychology of the creative eye*. Univ of California Press.
- Baumgartner, E., & Gegenfurtner, K. R. (2016). Image statistics and the representation of material properties in the visual cortex. *Frontiers in Psychology*, 7(AUG), 1–13. <https://doi.org/10.3389/fpsyg.2016.01185>
- Baxandall, M. (1995). *Shadows and enlightenment*. Yale University Press.
- Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2013a). OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG)*, 32.
- Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2013b). OpenSurfaces: A Richly Annotated Catalog of Surface Appearance. *ACM Transactions on Graphics*, 32(4), 1. <https://doi.org/10.1145/2461912.2462002>
- Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2015). Material recognition in the wild with the materials in context database. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3479–3487.
- Beurs, W. (1692). *De groote waereld in't klein geschildert; of schilderagtig tafereel van s'weerlds schlderyen, kortelyk vervat in ses boeken; verklarend de hoofdverwen, haare versheide megelingen in oly, en derzelve gebruik*. Van Waesberge.
- Bohannon, B. J. (2016). Mechanical Turk upends social sciences. 352(6291).
- Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, 434(7031), 301–307.
- Chadwick, A. C., & Kientridge, R. W. (2015). The perception of gloss: A review. *Vision Research*, 109(PB), 221–235. <https://doi.org/10.1016/j.visres.2014.10.026>
- Di Cicco, F., Wijntjes, M. W., & Pont, S. C. (2019). Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *Journal of Vision*, 19(3). <https://doi.org/10.1167/19.3.7>
- Ferwerda, J. A., Pellacini, F., & Greenberg, D. P. (2001). Psychophysically based model of surface gloss perception. *Proc. SPIE*, 4299(June 2001). <https://doi.org/10.1117/12.429501>
- Fleming, R. W., & Bülthoff, H. H. (2005). Low-Level Image Cues in the Perception of Translucent Materials. *ACM Transactions on Applied Perception*, 2(3), 346–382. <https://doi.org/10.1145/1077399.1077409>
- Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 3, 347–368. <https://doi.org/10.1037/t25352-000>
- Fleming, R. W., Jäkel, F., & Maloney, L. T. (2011). Visual perception of thick transparent materials. *Psychological science*, 22(6), 812–820.

- Fleming, R. W., Torralba, A., & Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of Vision*, 4(9), 10. <https://doi.org/10.1167/4.9.10>
- Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8), 9. <https://doi.org/10.1167/13.8.9>
- Gombrich, E. H. (1960). *A study in the psychology of pictorial representation*. Pantheon Books.
- Griffin, L. D. (1999). Partitive mixing of images : a tool for investigating pictorial perception. *Journal of the Optical Society of America*, 16(12), 2825–2835.
- Igarashi, T., Nishino, K., & Nayar, S. K. (2007). *The appearance of human skin: A survey*. Now Publishers Inc.
- Jensen, H. W., Marschner, S. R., Levoy, M., & Hanrahan, P. (2001). A practical model for subsurface light transport. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 511–518.
- Karana, E., Hekkert, P., & Kandachar, P. (2009). Meanings of materials through sensorial properties and manufacturing processes. *Materials & Design*, 30(7), 2778–2784.
- Kim, J., & Anderson, B. L. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of vision*, 9(11), 1–17. <https://doi.org/10.1167/10.9.3>
- Kim, J., Marlow, P. J., & Anderson, B. L. (2012). The dark side of gloss. *Nature Neuroscience*, 15(11), 1590–1595. <https://doi.org/10.1038/nn.3221>
- Koenderink, J. J., & van Doorn, A. J. (2001). Shading in the case of translucent objects. *Human vision and electronic imaging vi*, 4299, 312–320.
- Lehmann, A.-S. (2008). Fleshing out the body: The 'colours of the naked' in workshop practice and art theory, 1400–1600. *Nederlands Kunsthistorisch Jaarboek*, 59, 86.
- Marlow, & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision*, 13(14), 1–23. <https://doi.org/10.1167/13.14.2>
- Marlow, & Anderson, B. L. (2015). Material properties derived from three-dimensional shape representations. *Vision Research*, 115, 199–208. <https://doi.org/10.1016/j.visres.2015.05.003>
- Marlow, Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, 22(20), 1909–1913. <https://doi.org/10.1016/j.cub.2012.08.009>
- Matts, P. J., Fink, B., Grammer, K., & Burquest, M. (2007). Color homogeneity and visual perception of age, health, and attractiveness of female facial skin. *Journal of the American Academy of Dermatology*, 57(6), 977–984.
- Matusik, W., Pfister, H., Brand, M., & McMillan, L. (2003a). Efficient Isotropic BRDF Measurement. *Proceedings of Eurographics/SIGGRAPH workshop rendering*, 241–248.
- Matusik, W., Pfister, H., Brand, M., & McMillan, L. (2003b). Efficient isotropic brdf measurement.
- Michelson, A. A. (1891). On the application of interference-methods to spectroscopic measurements.—i. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 31(191), 338–346.
- Motoyoshi, I. (2010). Highlight-shading relationship as a cue for the perception of translucent and transparent materials. *Journal of vision*, 10(9), 6. <https://doi.org/10.1167/10.9.6>

- Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *447*(May), 206–209. <https://doi.org/10.1038/nature05724>
- Nakayama, K., Shimojo, S., & Ramachandran, V. S. (1990). Transparency: Relation to depth, subjective contours, luminance, and neon color spreading. *Perception*, *19*(4), 497–513.
- Padilla, S., Drbohlav, O., Green, P. R., Spence, A., & Chantler, M. J. (2008). Perceived roughness of $1/f\beta$ noise surfaces. *Vision Research*, *48*(17), 1791–1797.
- Papadopoulos, D. P., Uijlings, J. R. R., Keller, F., & Ferrari, V. (2017). Extreme clicking for efficient object annotation. *International Journal of Computer Vision*. <https://doi.org/10.1109/ICCV.2017.528>
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods*, *46*(4), 1023–1031. <https://doi.org/10.3758/s13428-013-0434-y>
- Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 55–64.
- Radonjić, A., Cottaris, N. P., & Brainard, D. H. (2015). Color constancy supports cross-illumination color selection. *Journal of Vision*, *15*(6), 13–13.
- Serrano, A., Gutierrez, D., Myszkowski, K., Seidel, H.-p., & Masia, B. (2018). An intuitive control space for material appearance. *arXiv preprint arXiv, 1806.04950*. <https://doi.org/10.1145/2980179.2980242>
- Sharan, L., Li, Y., Motoyoshi, I., Nishida, S., & Adelson, E. H. (2008). Image statistics for surface reflectance perception. *Journal of optical society of america*, *25*(4), 846–865.
- Sharan, L., Liu, C., Rosenholtz, R., & Adelson, E. H. (2013). Recognizing materials using perceptually inspired features. *International Journal of Computer Vision*, *103*(3), 348–371. <https://doi.org/10.1007/s11263-013-0609-0>
- Sharan, L., Rosenholtz, R., & Adelson, E. H. (2009). Material perception : What can you see in a brief glance? *Vision Sciences Society Annual Meeting Abstract*, *9*(8), 2009.
- Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision*, *14*(9), 1–24. <https://doi.org/10.1167/14.9.12>
- Stephen, I. D., Coetzee, V., & Perrett, D. I. (2011). Carotenoid and melanin pigment coloration affect perceived human health. *Evolution and Human Behavior*, *32*(3), 216–227.
- Su, Z., Deng, D., Yang, X., & Luo, X. (2012). Color transfer based on multiscale gradient-aware decomposition and color distribution mapping. *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, (OCTOBER 2012), 753. <https://doi.org/10.1145/2393347.2396304>
- Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). The multidimensional wisdom of crowds. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, 1–9.

- Wiebel, C. B., Toscani, M., & Gegenfurtner, K. R. (2015). Statistical correlates of perceived gloss in natural images. *Vision Research*, 115, 175–187. <https://doi.org/10.1016/j.visres.2015.04.010>
- Wiebel, C. B., Valsecchi, M., & Gegenfurtner, K. R. (2013). The speed and accuracy of material recognition in natural images. *Attention, perception & psychophysics*, 75(5), 954–66. <https://doi.org/10.3758/s13414-013-0436-y>
- Wijntjes, M. W. A., & Pont, S. C. (2010). Illusory gloss on Lambertian surfaces. *Journal of Vision*, 10(9), 13–13. <https://doi.org/10.1167/10.9.13>
- Xiao, B., Walter, B., Gkioulekas, I., Zickler, T., Adelson, E., & Bala, K. (2014). Looking against the light: How perception of translucency depends on lighting direction. *Journal of vision*, 14(3), 1–22. <https://doi.org/10.1167/14.3.17>
- Xiao, B., Sharan, L., Rosenholtz, R., & Adelson, E. (2011). Speed of material vs. object recognition depends upon viewing condition. *Journal of Vision*, 11(15), 19–19.
- Zhang, F., De ridder, H., Barla, P., & Pont, S. (2019). A systematic approach to testing and predicting light-material interactions. *Journal of vision*, 19(4), 1–22.
- Zhang, F., de Ridder, H., Fleming, R. W., & Pont, S. (2016). MatMix 1.0 – using optical mixing to probe visual material perception. *Journal of vision*, 16, 1–37. <https://doi.org/10.1167/16.6.11>

2.5. SUPPLEMENTARY MATERIALS

Here we include a short description of the Amazon Mechanical Turk platform. We follow this by methodological details related to the gathering of the data that were not included in the main text. Furthermore, we have added here the material-specific PCA visualizations that were not included in the main paper. Next, we list the factor loadings of the two first principal components of all the PCAs that we ran, for all material specific PCAs and the global PCA in the bottom, which contained all materials.

AMAZON MECHANICAL TURK

For all data collection, we use the online labor marketplace Amazon Mechanical Turk (AMT), colloquially referred to as MTurk. On this platform, people can sign up as workers, or turkers, to perform micro-tasks for requesters. The platform has been popular for companies that require simple human-intelligence tasks such as transcriptions. In recent years, the MTurk platform has also seen increased use from the social sciences (Bohannon, 2016; Peer et al., 2014) and computer sciences (Bell et al., 2013a, 2015; Papadopoulos et al., 2017; Su et al., 2012).

Workers are free to choose what tasks to perform and have the option to stop at any moment if they so desire. They are paid per task, where each task can last a few minutes (the usual) up to multiple hours. The requester is given a chance to review the submitted work and has the option to either accept or reject the work based on the quality. Workers are not paid for any rejected work. For the experiments reported below we accepted all work as it is clearly impossible to define quality in our case: there is no right or wrong perception.

Workers can be selected before the task. This can be achieved by using qualifications that are based on demographic criteria such as geography, age, or reputational criteria such as requiring that a worker has completed a minimum number of tasks with a certain percentage of acceptance. For example, it is common – as these are the default values on AMT – to require that a worker has completed at least 1000 tasks, of which at least 95% percent has been accepted. With these qualifications, workers can be restricted from starting tasks.

STEP 1: COLLECTING PAINTINGS

We downloaded 17936 images from the digitized open-access painting collections from seven galleries. The galleries are listed in Table 2.5. Most of these images were collected via web scraping the websites of these seven galleries. Web scraping was done with Python, making use of the *Beautifulsoup*, *Selenium*, and *Urllib* modules. For the Rijksmuseum, we used their API to directly download each image individually. For the National Gallery of London, we used peer-to-peer downloading to download the collection. The URLs for each gallery have been supplied in the references.

For each image, we also obtained the original meta-data from the gallery, which contains for example information about the title of the artwork, the artist, and an estimated creation date.

Next, we filtered out 661 monochrome images using a linear regression on the 3D color data keeping a total of 17275 paintings. We have not applied any color management or color corrections techniques, nor do we have any information into what color management the respective galleries applied.

Table 2.5: Each gallery and the number of images we downloaded from their digitized open-access gallery.

| Gallery | Count | Country |
|----------------------------|-------|----------------|
| Rijksmuseum | 4657 | Netherlands |
| NationalMuseum | 2924 | Sweden |
| The Metropolitan | 2729 | USA |
| National Gallery of London | 2402 | United Kingdom |
| National Gallery of Art | 2132 | USA |
| Museo Nacional Del Prado | 2032 | Spain |
| Getty | 399 | USA |

STEP 2: COLLECTING MATERIALS

To identify what materials are perceived to be depicted in the collected paintings 1571 AMT participants completed a total of 36838 tasks. The majority ($n=511$) of participants only completed the task once, while some participants completed a much larger number of tasks: two participants completed more than 1000 tasks each and an additional 6 participants completed over 500 tasks each. Each task paid 0.02 USD and presented participants with 40 paintings and a target material. An example of this can be seen in Fig. 2.13, where the target material is fabric. Participants could scroll through the webpage to view all the paintings. The target material would be one of the materials listed in the Materials and Attributes section. Participants were instructed to use the mouse to click each painting where they perceived the target material to be present. If a painting was clicked the painting would gain an orange outline and a small textbox in the bottom right corner would change from “no [material]” to “[material]”.

Before starting the task, participants would be presented with a short tutorial. In the tutorial, the task was explained followed by a short description of the specific material was given. At the end of the tutorial, participants had to complete a small number of test trials. The number of test trials varied per material, between 3 to 5. In each test trial, the participants were shown one painting at a time and had to click the correct button, indicating whether the painting did or did not depict the requested material for that trial. Upon giving a wrong answer, a participant was immediately given feedback and asked to redo the same trial. Once all test trials were completed, the real task would commence. Once a tutorial was completed for a specific material, the participant would not be shown the tutorial for that material again.

For each painting/material combination, we collected responses from at least 5 participants. If at 80% or more of these participants responded that a painting depicted a material,

Instructions: Click on the images where you can see some fabrics

Instructions

Category: fabrics

If you cannot tell, don't select the photo. Please do not guess. Don't worry if you miss small things that are hard to see.

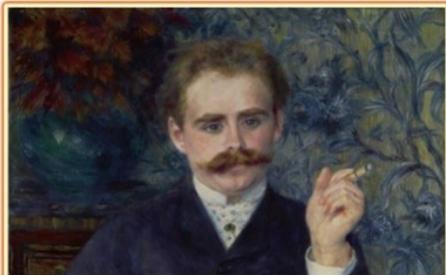


Figure 2.13: Example of the material-collection web task. After a tutorial, participants would see this screen. The yellow top bar shows the instructions in short, and the right-wards button opens a pop-up menu with additional instructions when clicked. The target material is indicated semantically. Participants would use the mouse to scroll through a total of 40 images. Once a painting was indicated (clicked) to contain the target material, the white outline (as can be seen in the top-left painting) changed to orange (as in the other 3 visible painting) and the label in the bottom right of the painting would change accordingly.

we considered that painting to depict that material for the continuation of this project.

STEP 3: COLLECTING SEGMENTATIONS

In the next step, AMT participants would be tasked with segmenting the target material, where the target material was one of the materials that is depicted in the painting, as indicated in step 2. Participants were told to “find the largest area displaying [material] and trace its outline as accurately as you can”, where [material] would be replaced by one of the materials listed in the Materials and Attributes section. A screenshot of the task can be seen in Fig. 2.14.

To make a segment, a participant could click anywhere on the presented painting, which would create a point on the painting. Consecutively placed clicks would be connected with a blue line, and a line would continuously be drawn between the last placed point and the mouse. Once the outline of the chosen material instance was traced, the last point could be connected to the first point by hitting the enter key, a right-mouse click, or by pressing the Close button. Once closed, a participant could enter the adjust mode, either pressing the Adjust button or the *a* key. In the adjust mode individual points could be moved using click-and-drag. Additionally, participants could zoom in or out by scrolling, could move the painting using the arrow keys, or undo and redo the last action using the buttons or by pressing *ctrl-z* or *ctrl-y* respectively. Once participants were satisfied with a segment, they could finish the task by pressing the submit button.

Initially, we placed no restrictions on which participants could perform the segmentation task. However, after we collected around 1700 segments from a total of 52 participants, we selected the five participants whom we judged to be the best, based on the observed quality of the created segments. From this point onwards, only these five selected participants were allowed to perform the task. This restriction increased the average number of points, i.e. vertices in the polygonal segmentation by 62% from 37 to 65, indicating that the average detail per segment increased. Additionally, the acceptance rate, i.e. the number of submitted segments that passed the quality control in the next step, improved by 20% from 0.58 to 0.7. The five selected participants completed 283, 455, 777, 1235, and 3082 tasks. In total, all participants together completed 7110 tasks, for 0.15 USD per task.

STEP 4: QUALITY CHECK

In the quality control check a total of 127 AMT participants performed 1349 tasks for 0.05 USD per task. In each task, a participant was shown 20 material segments, together with the paintings the segments were from. This can be seen in Fig. 2.15. Participants were asked to inspect each segment and click the painting if it fulfilled the requirements. The requirements were 1) that each segment only contains one material and 2) that the boundaries of the segment follow the outline of the depicted material. Examples were provided. Comparable to the task in step 2, once the segment was clicked, it would get an orange outline and a textbox in the bottom right would change from “multiple materials or bad boundary” to “one material and tight boundary”.

A minimum of 5 different AMT participants would judge each segment. Next, we did a quality check using the CUBAM algorithm, as described in (Bell et al., 2013a), which was created by (Welinder et al., 2010). CUBAM is intended to improve the quality of binary data by taking the (dis-)agreement between participants into account. The CUBAM

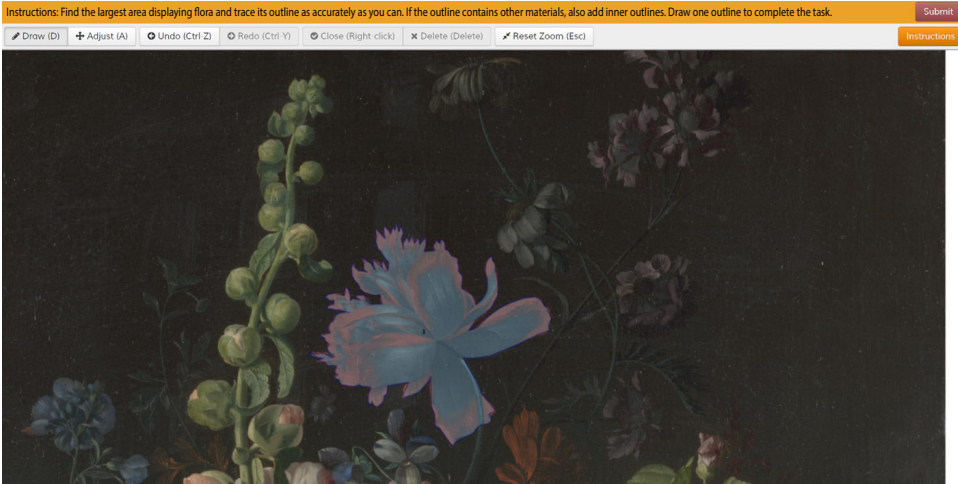


Figure 2.14: Example of the online segmentation tasks. At the top, a short summary of the instructions is given. Directly below is the menu, which also holds an option to view the instructions again. The remainder of the screen is used to display the painting, which is here displayed zoomed in. In the middle, a flower can be seen, which has been segmented, as indicated with a light-blue transparent layer. Once complete, a participant could press the submit button in the top right.

algorithm resulted in a single quality score value. Segments with an adequate quality score continued to the next step.

STEP 5: MATERIAL LABELING

In step 3 participants segmented specific materials. It is possible that instead of segmenting the required material a participant segmented a different material. To counter this possibility, in this step we re-labeled each segment with a material name. 178 AMT participants completed 749 of these labeling tasks, for 0.04 USD per task. In each task, 5 different participants were shown one segment at a time and asked to select to best fitting label from a list. The list contained the materials discussed in section Materials, as well as three additional options: “I can’t tell”, “More than one material” and “Not on list”. If there was an agreement on at least 3 out of 5 answers, we assigned that label to the segment.

STEP 6: MANUAL SELECTION

After completing the previous steps, we had collected around 4500 material segments. From these segments we selected 90 segmentations for each of the 15 material categories, aiming to create the most diverse set possible for each material. These sets were then used for the perceptual experiment.

FACTOR LOADINGS

The factor loadings for the first two principal components can be found in Table 2.6.

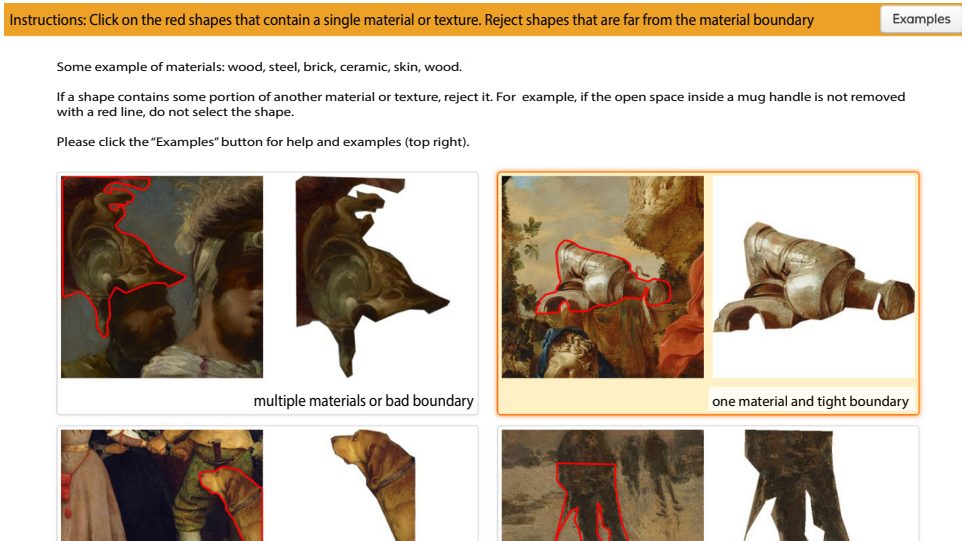


Figure 2.15: Example of a quality check task. At the top, a summary of the instructions is given, next to a button to see good and bad examples. Segments are displayed next to the section of painting where they are originally taken from, to show the material segment in context. AMT participants could click a pair to indicate the material segmentation was correct, i.e. contained one material and possessed a tight boundary. Participants could scroll down to see the rest of the segments. A submit button was below the last segments.

MATERIAL-SPECIFIC PCA VISUALIZATIONS

The material-specific PCA visualizations (i.e., Fig. 2.16, Fig. 2.17, and Fig. 2.18) for 11 out of 15 material categories. Note that the visualizations for fabrics, skin, flora, and paper can be found in the main text.

Table 2.6: A full list of the factor loadings for the first two principal components for each separate material, and one for all the materials together at the bottom. Some header names are abbreviated due to space constraints: MC is Multi-colored, TT is Transparent translucent, B is Bendable

| Material | | MC | TT | glossy | hairy | rough | hard | B | fragile | cold | vivid |
|----------|-----|------|------|--------|-------|-------|------|------|---------|------|-------|
| flora | PC1 | -.40 | .03 | -.15 | .13 | .42 | .35 | -.29 | -.6 | .05 | -.23 |
| flora | PC2 | .61 | .11 | .06 | .22 | .16 | -.02 | .05 | -.49 | -.03 | .54 |
| paper | PC1 | -.36 | -.10 | .08 | -.15 | .22 | .45 | -.61 | -.32 | .09 | -.31 |
| paper | PC2 | .76 | -.20 | .12 | -.09 | .17 | .26 | -.25 | -.17 | -.07 | .41 |
| glass | PC1 | -.34 | -.68 | -.26 | .12 | .01 | -.16 | .13 | -.31 | -.21 | -.39 |
| glass | PC2 | .72 | .05 | -.51 | .22 | .15 | -.28 | .20 | .02 | -.18 | -.03 |
| wood | PC1 | .48 | .35 | .20 | .08 | -.39 | -.51 | 0 | -.10 | -.11 | .41 |
| wood | PC2 | .25 | .21 | -.38 | .01 | .39 | -.33 | .52 | .40 | .16 | -.17 |
| skin | PC1 | .73 | .09 | -.01 | .08 | .10 | .02 | -.14 | .02 | .25 | .60 |
| skin | PC2 | .08 | -.06 | -.44 | .20 | .36 | .37 | .58 | -.04 | .34 | -.19 |
| stone | PC1 | .29 | .25 | -.04 | .06 | -.55 | -.61 | .07 | .14 | -.38 | .09 |
| stone | PC2 | .42 | .05 | .50 | -.12 | -.18 | .39 | -.16 | -.08 | .02 | .59 |
| animal | PC1 | .59 | .22 | .17 | -.58 | .12 | -.07 | .21 | .27 | .11 | .32 |
| animal | PC2 | -.27 | .14 | -.07 | -.4 | -.24 | -.54 | .12 | .24 | -.20 | -.53 |
| fabrics | PC1 | -.46 | -.16 | -.34 | .05 | .19 | .25 | -.22 | -.11 | .20 | -.67 |
| fabrics | PC2 | .41 | -.09 | .19 | -.53 | .22 | .39 | -.27 | -.21 | .43 | .07 |
| food | PC1 | .70 | .04 | .05 | .03 | -.05 | .17 | -.47 | -.12 | .13 | .48 |
| food | PC2 | -.24 | -.13 | -.06 | -.06 | .34 | .53 | -.61 | -.06 | -.01 | -.39 |
| ground | PC1 | .26 | .46 | -.06 | .20 | -.30 | -.66 | .34 | .12 | .09 | .08 |
| ground | PC2 | .60 | -.37 | -.15 | .27 | .05 | .03 | -.03 | .01 | -.44 | .46 |
| gem | PC1 | -.16 | .21 | .47 | .01 | -.11 | .26 | .61 | .44 | .060 | .25 |
| gem | PC2 | .57 | -.11 | .10 | -.02 | .27 | .30 | -.11 | -.18 | -.01 | .66 |
| sky | PC1 | -.49 | .03 | -.33 | -.06 | .02 | -.13 | -.03 | -.23 | .16 | -.74 |
| sky | PC2 | -.36 | -.45 | .20 | .37 | -.21 | -.04 | .31 | .27 | -.52 | -.1 |
| metal | PC1 | .40 | .29 | .18 | .06 | .03 | -.35 | .53 | .49 | -.07 | .26 |
| metal | PC2 | -.33 | -.03 | -.78 | .10 | -.03 | -.29 | .19 | .28 | -.09 | -.27 |
| ceramic | PC1 | .13 | .13 | -.83 | .08 | .28 | -.16 | -.07 | -.29 | -.17 | .21 |
| ceramic | PC2 | .59 | .19 | .19 | .04 | -.21 | -.06 | -.04 | .10 | .08 | .72 |
| liquid | PC1 | .19 | .32 | .67 | -.07 | -.30 | -.02 | .04 | .44 | .16 | .3 |
| liquid | PC2 | -.35 | -.20 | .08 | -.08 | .13 | .44 | -.02 | .62 | -.38 | -.29 |
| global | PC1 | -.13 | -.08 | .09 | -.27 | .22 | .62 | -.56 | -.12 | .33 | -.17 |
| Global | PC2 | .22 | .69 | .25 | -.32 | -.25 | -.23 | -.15 | .10 | .37 | .14 |

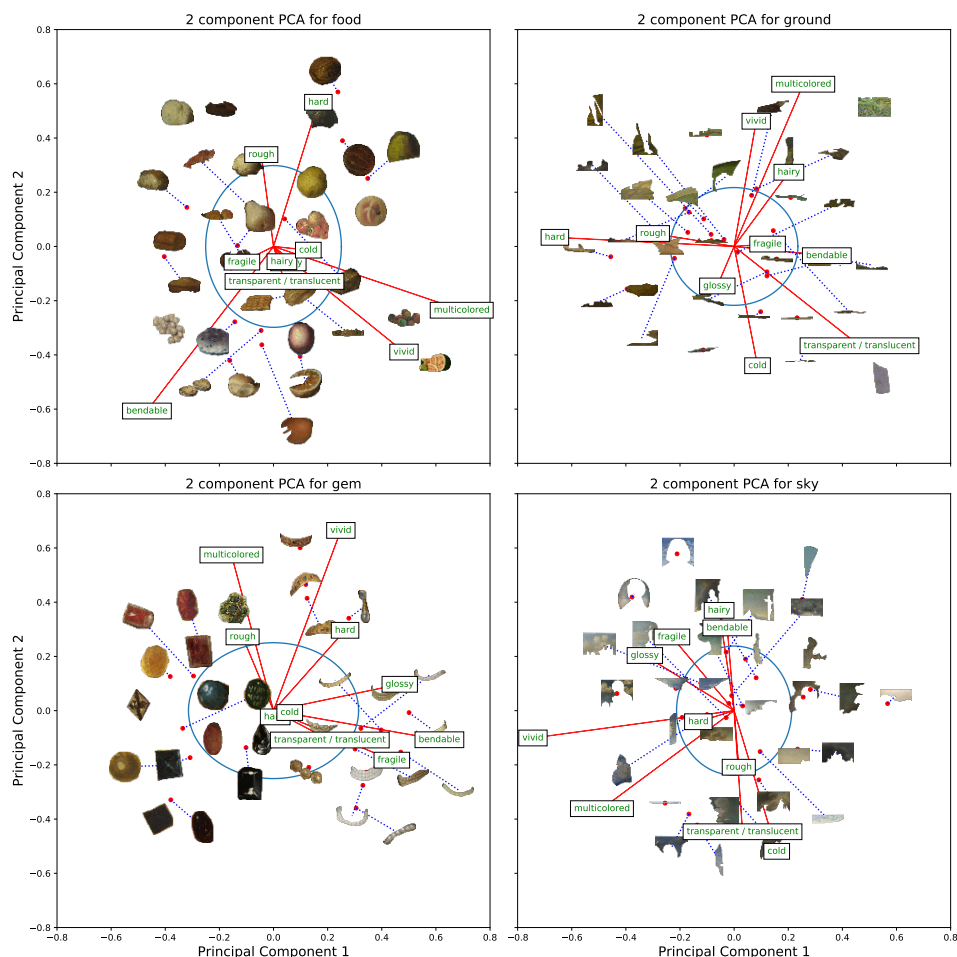


Figure 2.16: Four visualizations of the first two primary components for the material-specific PCA for food, ground, gem, and sky. Each PCA was run with only the 30 stimuli per material. The red vectors indicate the factor loadings of each attribute. We plotted the actual stimuli within the PCA space. The blue lines connect the stimuli to its actual position within the space when the stimuli would otherwise overlap. The ellipse was fitted around the points (1 sd).

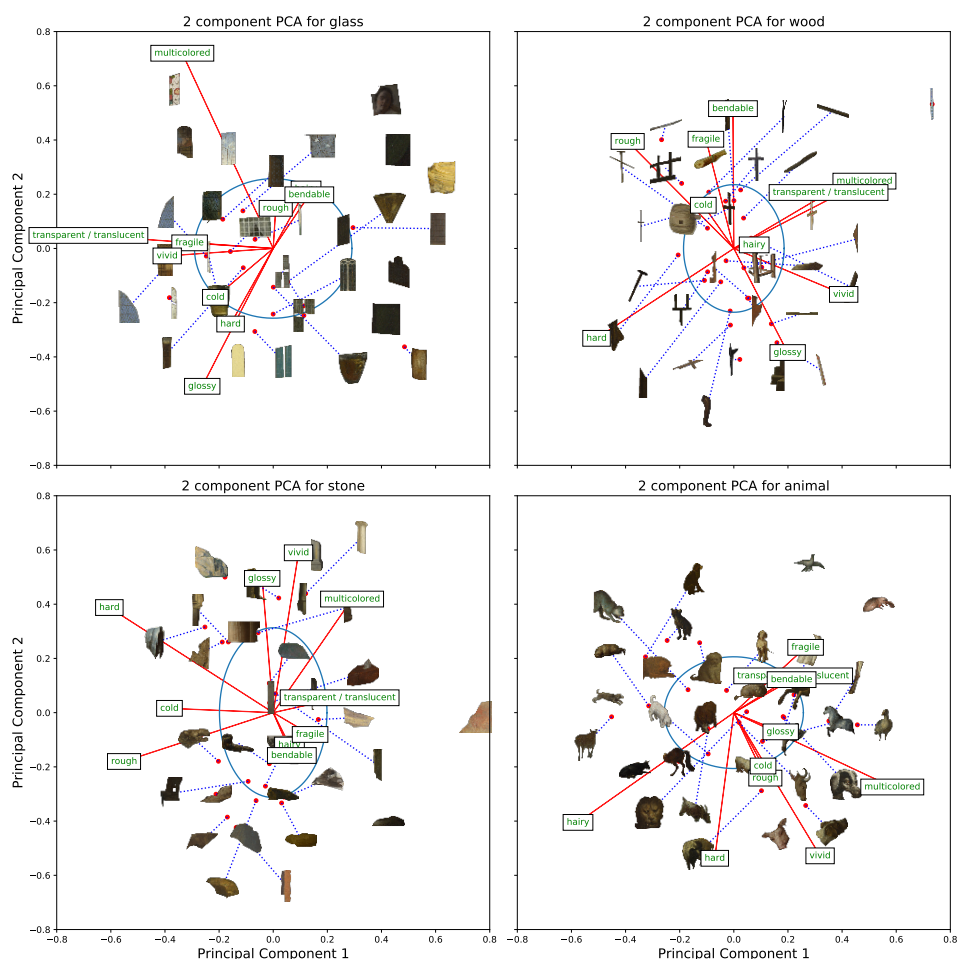


Figure 2.17: Four visualizations of the first two primary components for the material-specific PCA for glass, wood, stone, and animal. Each PCA was run with only the 30 stimuli per material. The red vectors indicate the factor loadings of each attribute. We plotted the actual stimuli within the PCA space. The blue lines connect the stimuli to its actual position within the space when the stimuli would otherwise overlap. The ellipse was fitted around the points (1 sd).

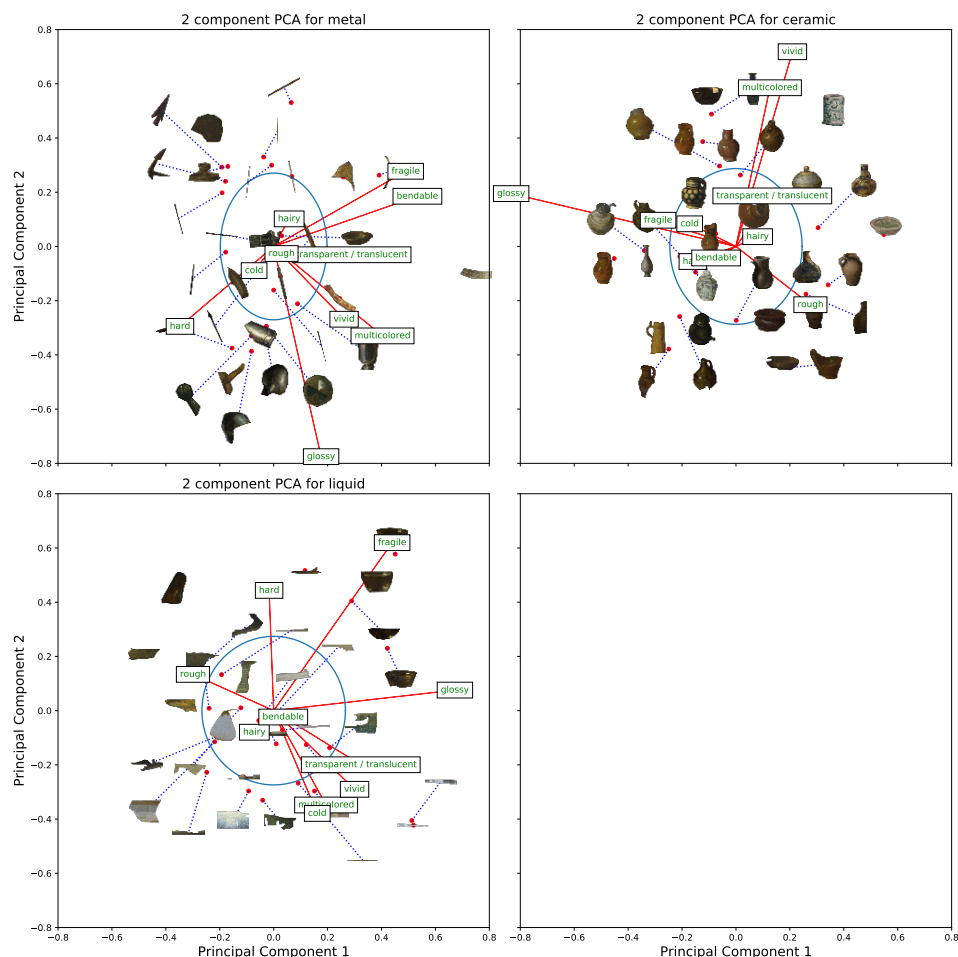


Figure 2.18: Three visualizations of the first two primary components for the material-specific PCA for metal, ceramic, and liquid. Each PCA was run with only the 30 stimuli per material. The red vectors indicate the factor loadings of each attribute. We plotted the actual stimuli within the PCA space. The blue lines connect the stimuli to its actual position within the space when the stimuli would otherwise overlap. The ellipse was fitted around the points (1 sd).

3

MATERIALS IN PAINTINGS

In this paper, we capture and explore the painterly depictions of materials to enable the study of depiction and perception of materials through the artists' eye. We annotated a dataset of 19k paintings with 200k+ bounding boxes from which polygon segments were automatically extracted. Each bounding box was assigned a coarse material label (e.g., fabric) and half was also assigned a fine-grained label (e.g., velvety, silky). The dataset in its entirety is available for browsing and downloading at materialsinpaintings.tudelft.nl. We demonstrate the cross-disciplinary utility of our dataset by presenting novel findings across human perception, art history, and computer vision. Our experiments include a demonstration of how painters create convincing depictions using a stylized approach. We further provide an analysis of the spatial and probabilistic distributions of materials depicted in paintings, in which we for example show that strong patterns exist for material presence and location. Furthermore, we demonstrate how paintings could be used to build more robust computer vision classifiers by learning a more perceptually relevant feature representation. Additionally, we demonstrate that training classifiers on paintings could be used to uncover hidden perceptual cues by visualizing the features used by the classifiers. We conclude that our dataset of painterly material depictions is a rich source for gaining insights into the depiction and perception of materials across multiple disciplines and hope that the release of this dataset will drive multidisciplinary research.

3.1. INTRODUCTION

Throughout art history, painters have invented numerous ways to depict the three-dimensional world onto flat surfaces (Kemp, 1992; Panofsky, 2020; Pirenne, 1970; White, 1957). Unlike photographers, painters are not limited to optical projection (Cavanagh, 2005; Willats, 1997) and therefore paintings have more freedom. This means that a painter can directly modify and manipulate the 2D image features of the depiction. When doing so, a painter's primary concern is not whether a depiction is optically or physically correct. Instead, a painting is explicitly designed for human viewing (Graham and Field, 2008a; Graham and Field, 2008c). The artist does not copy a retinal image (Perdreau and Cavanagh, 2011) (which would make the painter effectively a biological camera) but may apply techniques such as iteratively adapting templates until they 'fit' perceptual awareness (Gombrich, 1960).

As a result of this, the depiction contained can deviate from reality (Cavanagh, 2005). On one hand, this makes paintings unsuited as ecological stimulus (Gibson, 1978). On the other hand, as Gibson acknowledges, paintings are the result of endless visual experimentation, and therefore, indispensable for the study of visual perception.

The depiction and perception of pictorial space in paintings (Kemp, 1992; Panofsky, 2020; Pirenne, 1970; White, 1957; Willats, 1997) has historically received more attention than the depiction and perception of materials. It has previously been found that human observers are able to visually categorize and identify materials accurately and quickly for both photos (Fleming et al., 2013; Sharan et al., 2009, 2014) and paintings (van Zuijlen et al., 2020). Furthermore, for these painted materials, we can perceive distinct material properties such as glossiness, softness, transparency, etc (Cavanagh et al., 2008; Sayim and Cavanagh, 2011; van Zuijlen et al., 2020). A single material category (e.g., fabric) can already display a large variety of these material properties, which demonstrates the enormous variation in visual appearance of materials. This variation in materials and material properties has received relatively little attention. In fact, the perceptual knowledge that is captured in the innumerable artworks throughout history can be thought of as the largest perceptual experiment in human history and it merits detailed exploration.

3.1.1. A SIMPLE TAXONOMY OF IMAGE DATASETS

To explore material depictions within art there is a need for a dataset that relates artworks to material perception. Therefore, in this study, we create and introduce an accessible collection of material depictions in paintings, which we call the Materials in Painting (MIP) dataset. However, the use and creation of art-perception datasets is of broader interest.

We propose a simple taxonomy of three image dataset usages: 1) perceptual, 2) ecological, and 3) computer vision usage. In the remainder of the introduction below, we will contextualize our dataset within this taxonomy by first discussing existing image and painting datasets as well as the benefits our MIP dataset can provide for each of these three dataset usages. This shall be followed by a detailed description of the creation of the MIP dataset in the method section. Finally, we perform and discuss several small experiments that exemplify the utility of the MIP datasets for each of three dataset usages discussed.

Perceptual datasets. To understand the human visual system, stimuli from perceptual datasets can be used in an attempt to relate the evoked perception to the visual input. We

can roughly categorize three types of stimuli used for visual perception: natural, synthetic, and manipulated.

The first represent ‘normal’ photos of objects, materials, and scenes as they can be found in reality. Experimental design with such stimuli often attempts to relate the evoked perceptions to natural image statistics within the images or physical characteristics of the contents captured in the images. Some examples of uses of natural stimuli datasets include, but are not limited to, the memorability of pictures in general (Isola et al., 2011) or more specifically the memorability of faces (Bainbridge et al., 2013). In another example, images of natural, but novel objects were used to understand what underlies the visual classification of objects (Horst and Hout, 2016).

The second type, synthetic stimuli, are created artificially, such as digital renderings, drawings, and paintings. Synthetic stimuli might represent the real world, but often contain image statistics that deviate from natural image statistics. Paintings have for example often been used to study affect and aesthetics (Joshi et al., 2011; Machajdik and Hanbury, 2010; Sartori et al., 2015). In another example, Borkin et al. (2013) used a set of synthetic stimuli to test for memorability of data visualizations.

Both natural and synthetic images can be manipulated, which leads to the third type of stimuli. Manipulated stimuli are often used to investigate the effect of image manipulations by comparing them to the original (natural or synthetic) image. Here the manipulations function as the independent variables. For example Öhlschläger and Vö (2017) created a database of images that contain scene inconsistencies that can be used to study the compositional rules of our visual environment. In another example, a stimulus set consisting of original and texture (i.e., manipulated) versions of animals found that perceived animal size is mediated by mid-level image statistics (Long et al., 2018).

The advantage of using manipulated or synthetic images is that perceptual judgments can be compared to some independent variable, which is typically not available for natural images. Paintings are a special case here. They are a synthetic image of a 3D scene that is rendered using oils, pigments, and artistic instruments. However, the painting is also a mostly flat, physical object. Retrieving the veridical data is usually impossible for paintings. In other words, objects or materials depicted in photos can often be measured or interacted with in the real world but this is rarely possible for paintings. However, the advantage of using paintings is that it can often be seen, or (historically) inferred, how the painter created the illusory realism. Even if it cannot be seen with the naked eye, chemical, and physical analysis can be performed. In Di Cicco et al. (2020) a perceptually convincing depiction of grapes was recreated using a 17th century recipe. In this reconstruction, the depiction was recreated by a professional painter one layer at a time, where each layer represents a separate and perceptually diagnostic image feature that together lead to the perception of grapes. The physical limitations of painterly depictions relative to the physical 3D world, such as for example due to luminance compression in paintings (Graham et al., 2016; Graham, 2011; Graham and Field, 2008b; Graham et al., 2009) may lead to systematically different strategies for material depiction. Despite this, van Zuijlen et al. (2020) has shown that the perceptions of materials and material properties depicted in paintings are similar to those previously reported for photographic materials (Fleming et al., 2013).

Therefore, studying paintings in addition to more traditional stimuli like photos or renderings, can enrich our understanding of human material perception. It should be noted that

in this paper we focus on the image structure of the painting instead of the physical object. In other words, we focus on what is depicted within paintings and our data and analysis is limited to pictorial perception. In the remainder of this paper, when we mention *paintings*, we mean *images of paintings*.

Throughout history, painters have studied how to trigger the perceptual system and create convincing depictions of complex properties of the world. This resulted in *perceptual shortcuts*, i.e., stylized depictions of complex properties of the world that trigger a robust perception. The steps and painterly techniques applied by a painter to create a perceptual shortcut can be thought of as a perception-based recipe. Following such a recipe results in a perceptual shortcut, which is a depiction that gives the visual system the required inputs to trigger a perception. Many of the successful depictions are now available in museum collections. As such, the creation of art throughout history can be seen as one massive perceptual experiment. Studying perceptual shortcuts in art, and understanding the cues, i.e., features required to trigger perceptions, can give insights into the visual system. We will demonstrate this idea by analyzing highlights in paintings and photos.

Ecological datasets. To understand how the human visual system works it is important to understand what type of visual input is given by the environment. Visual ecology encompasses all the visual input and can be subdivided into natural and cultural ecology. Natural ecology reflects all which is found in the physical world. For example, to understand color-vision and cone cell sensitivities it is relevant to know the typical spectra of the environment. For this purpose, hyperspectral images (Foster et al., 2006; Nascimento et al., 2016) can be used, in this case to investigate color metamers (perceptually identical colors that originate from different spectra) and illumination variation. In another example, a dataset of calibrated color images were used to understand color constancy (Ciurea and Funt, 2003) (the ability to discount for chromatic changes in illumination when inferring object color). The Southampton-York Natural Scenes database, which contains a diverse set of images of natural scenes coupled with dense LiDAR range data, was used to relate image statistics to physical statistics (Adams et al., 2016). Another dataset contains photos taken in Botswana (Tkačik et al., 2011) in an area that supposedly reflects the environment of the proto-human and was used to investigate the evolution of the human visual system. Spatial statistics of today's human visual ecology are clearly different from Botswana's bushes as most people live in urban areas that are shaped by humans. For example, a dataset from Olmos and Kingdom (2004) was used to compute the distribution of spatial orientations of natural scenes (Girshick et al., 2011).

The content depicted within paintings only loosely reflects the *natural* visual ecology, but they do strongly represent *cultural* visual ecology. They have influenced how people see and depict the world and have influenced visual conventions up to contemporary cinematography and photography. Both perceptual scientists and art historians have looked for and studied compositional rules and conventions within art. A good example is the painterly convention that light tends to originate from the top-left (Carbon and Pastukhov, 2018; Sun and Perona, 1998), which is likely related to the human light-from-above prior (Berbaum et al., 1983; Gibson, 1950; Mamassian and Goutcher, 2001; Ramachandran, 1988).

New developments in cultural heritage institutions have made the measurement and study of paintings much more accessible. In recent years the digitization of cultural heritage

has led to a surge in publicly available digitized art works. Many individual galleries and art institutions have undertaken the admirable task to digitize their entire collection, and have often make a portion, if not the whole collection digitally available with no or minor copyright restrictions. The availability of digitized art works, combined with advancements in image analysis algorithms, has lead to Digital Art History, which concerns itself with the digitized analysis of artworks by for example analyzing artistic style (Saleh and Elgammal, 2016) and beauty (De La Rosa and Suárez, 2015), or local pattern similarities between artworks (Shen et al., 2019). In Sari et al. (2019), the authors for example developed a system that automatically detects and extracts garment color in portraits, which can for example be used for the digital analysis of historical trends within clothes and fashion.

Crowley and Zisserman (2014) pointed out that art historians often have the unenviable task of finding paintings for study manually. With an extensive dataset of material depictions within art, this task might become slightly easier for art historians that study the artistic depiction of materials, such as for example stone (Augart et al., 2018; Dietrich, 1990). The ability to easily find fabrics in paintings and its fine-grained subclasses such as velvet, silk, and lace could be used for the study of fashion and clothes in paintings in general (Hollander, 1993, 2016) or for paintings from a specific cultural context, such as Italian (Birbari, 1975), English, and French (Ribeiro, 1995) or even for the clothes worn by specific artists (De Winkel, 2005). The human body and it's skin, which clothing covers, is often studied within paintings (Bol and Lehmann, 2012; Hollander, 1993; Lehmann, 2008). For example, the Metropolitan Museum, published an essay on anatomy in the Renaissance, for which artworks depicting the human nude were used (Bambach, 2002). In this work on anatomy, only items from the Metropolitan Museum were used but with an annotated database of material depictions this could be extended and compared to other museum collections. Furthermore, through material categories such as food and flora category, the MIP could give access to typical artistic scenes such as stillives (Grootenboer, 2006; Woodall, 2012) and floral scenes (Taylor, 1995) respectively. It should be noted that 'stuff' like skin and food might not appear like a stereotypical material, however in this paper we adhere to the view of Adelson (2001), where each object, or 'thing', is considered to consist of some material, i.e., 'stuff'. Within this view non-stereotypical 'stuff' such as skin and food can certainly be considered as a material.

Computer vision datasets. Today, the majority of image datasets originate from research in computer vision. One of the first relatively large datasets representing object categories (Fei-Fei et al., 2004) has been used to both train and evaluate various computational strategies to solve visual object recognition. The ImageNet and CIFAR datasets (Deng et al., 2009; Krizhevsky, 2009) are regarded to be standard image recognition datasets for the last decade of research on deep learning vision systems.

Traditionally much visual research has been concerned with object classification but recently material perception has received increasing attention (Adelson, 2001; Bell et al., 2013, 2015; Caesar et al., 2018). A notable dataset that contains material information is OpenSurfaces by Bell et al. (2013), which contains around 70k crowd-sourced polygon segmentations of materials in photos. The Material In Context database improved on OpenSurfaces by providing 3 million samples across 23 material classes (Bell et al., 2015). To our knowledge, no dataset exists that explicitly provides material information within paint-

ings.

The majority of image datasets contain photographs, but various datasets exist that contain artworks. The WikiArt dataset for instance, which is created and maintained by a non-profit organisation, with the admirable goal “to make world’s art accessible to anyone and anywhere” (“Visual Art Encyclopedia”, 2020). The WikiArt dataset has been widely used for a variety of scientific purposes (Bar et al., 2014; Elgammal et al., 2018; Saleh and Elgammal, 2016; Strezoski and Worring, 2017; Tan et al., 2017). The Painting-91 dataset from Khan et al. (2014) consists of around 4000 paintings from 91 artists and was introduced for the purpose of categorization on style or artist. More recently, *Art500k* was released, which contains more than 500k low resolution artworks which were used to automatically identify content and style (Mao et al., 2017) within paintings.

The visual difference introduced by painterly depiction does not pose any significant difficulties to the human visual system, however it can be challenging for computer vision systems as a result of the domain shift (Patel et al., 2015; Wang and Deng, 2018; Wilson and Cook, 2020). Differences between painting images and photographic datasets include for instance composition, textural properties, colors, and tone mapping, perspective, and style. As for composition, photos in image datasets are often ‘snapshots’, taken with not too much thought given to composition, and typically intended to quickly capture a scene or event. In contrast, paintings are artistically composed and are prone to historical style trends. Therefore, photos often contain much more composition variation relative to paintings. Within paintings, composition can vary greatly between different styles. The human visual system can *distinguish* styles – for example, Baroque vs. Impressionism – and also implicitly judge whether two paintings are stylistically similar. Research in style or artist classification, as well as neural networks that perform style transfer, attempt to model these stylistic variations in art (Jing et al., 2020; Saleh and Elgammal, 2016).

Humans can also *discount* stylistic differences, for example, identifying the same person or object depicted by different artists. Similarly, work in domain adaptation (Patel et al., 2015; Wang and Deng, 2018; Wilson and Cook, 2020) focuses on understanding objects or ‘stuff’ across different image styles. Models that learn to convert photographs into painting-like or sketch-like images have been studied extensively for their application as a tool for digital artists (Jing et al., 2020). Recent work has shown that such neural style transfer algorithms can also produce images that are useful for training robust neural networks (Geirhos et al., 2019). However, photos that have been converted into a painting-like image are not identical to paintings; paintings can contain spatial variations of style and statistics that are not present in photos converted into paintings. Furthermore, painterly convention and composition are not taken into account by style-transfer algorithms.

Depending on the end goal for a computer vision system, it can be important to learn from paintings directly. Of course, when the end goal is to detect pedestrians for a self-driving car, learning from real photos, videos, or renderings of simulations can suffice. However, if the goal is to simulate general visual intelligence, multi-domain training sets are essential. Furthermore, if the goal is to create computer vision systems with a perception that matches human vision, training on paintings could be very beneficial. Paintings are explicitly created by and for human perception and therefore contain all the required cues to trigger robust perceptions. Therefore, networks trained on paintings are implicitly trained on these perceptual cues.

The multifaceted nature of datasets. While we have distinguished the broad purposes of datasets and exemplified each with representative datasets, it is important to keep in mind that these datasets can serve multiple goals across the taxonomy. For example, the Flickr Material Database (Sharan et al., 2009) was initially created as a perceptual dataset to study how quickly human participants were capable of recognizing natural materials. However, since then it has also often been used as a computer vision dataset, including by the original authors themselves (Sharan et al., 2013). In this study paintings are considered as especially interesting as they can be used for perceptual experiments, for digital art history, i.e., cultural visual ecology and can furthermore be used to train and test computer vision networks. The dataset presented in this paper is explicitly designed with this multidisciplinary nature in mind.

3.2. METHODS

Here we will first provide a short description of the dataset and the various stages of data collection, followed by an in-depth description of each stage.

Our dataset consists of high-quality images of paintings sourced from international art institutions and galleries. Within these images, human annotators have created bounding box around 15 material categories (e.g., fabrics, stone, etc). We further sub-categorized these material categories into over 50 fine-grained categories (e.g., velvet, etc). Finally, we automatically extract polygon segments for each bounding box. The annotated dataset will be made publicly available. All paintings, bounding boxes, labels, and metadata are available online.

The data collection was executed in multiple stages. Here we give an itemized overview of each stage and subsequently we discuss each stage in depth. The first two stages were conducted as part of a previous study (van Zuijlen et al., 2020), but we provide details here for completeness. Participants were recruited via Amazon Mechanical Turk (AMT). A total of 4451 unique AMT users participated in this study and gave written consent prior to data collection. Data collection was approved by the Human Research (ethics) Committee of the Delft University of Technology and adheres to the Declaration of Helsinki.

- First, we collected a large set of paintings.
- Next, human observers on the AMT platform identified which coarse-grained materials they perceived to be present in each painting (e.g., “is there wood depicted in this painting?”).
- Then, for paintings identified to contain a specific material, AMT users were tasked with creating a bounding box of that material in that painting.
- Lastly, AMT users assigned a fine-grained material label to bounding boxes (e.g., processed wood, natural wood, etc.).

3.2.1. COLLECTING PAINTINGS

We collected 19,325 high-quality digital reproductions of paintings from 9 online, open-access art galleries. The details of these art galleries are presented in Table 3.1. Images were downloaded from the online galleries, either using web scraping or through an API.

For the majority of these paintings we also gathered the following metadata: title of the work, estimated year of creation and name of the artist.

For 92% of boxes, we also have an estimate of the year of production. These estimates were made by the galleries from which the paintings were downloaded. The distribution of the year of production for all paintings are plotted in Fig. 3.1.

Table 3.1: List of galleries. A list of all the galleries, the country in which the gallery is located, and the number of paintings downloaded from that gallery.

| Gallery Name | Country | Count |
|--------------------------------|-------------|-------|
| The Rijksmuseum | Netherlands | 4672 |
| The Metropolitan Museum of Art | USA | 3222 |
| Nationalmuseum | Sweden | 3077 |
| Cleveland Museum of Art | USA | 2217 |
| National Gallery of Art | USA | 2132 |
| Museo Nacional del Prado | Spain | 2032 |
| The Art Institute of Chicago | USA | 936 |
| Mauritshuis | Netherlands | 638 |
| J. Paul Getty Museum | USA | 399 |

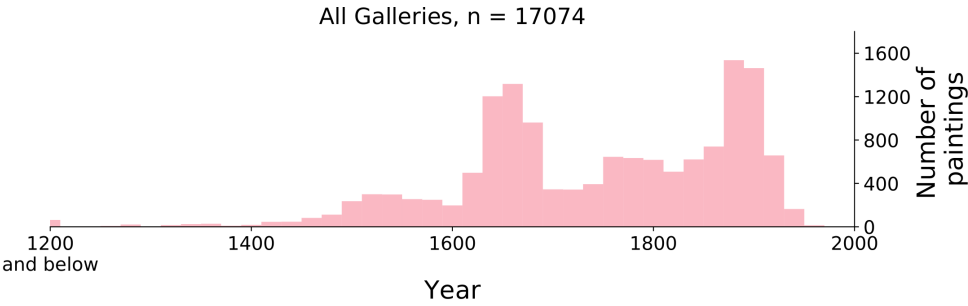


Figure 3.1: Histogram of the distribution of paintings over time. Each bin equals 20 years.

3.2.2. IMAGE-LEVEL COARSE-GRAINED MATERIAL LABELS

Next, we collected human annotations to identify material categories within paintings. We created a list of 15 material categories: animal, ceramic, fabric, sky, stone, flora, food, gem, wood, skin, glass, ground, liquid, paper, and metal. Our intention was to create a succinct list, that would nevertheless allow the majority of ‘stuff’ within a painting to be annotated. Our list is partially based on material lists used in Sharan et al. (2009) and Fleming et al. (2013), with which our set has 8 materials categories in common, and partially based on Bell et al. (2013), with which our list has 11 material categories in common. Note however minor difference in the category labels; the lists used in (Bell et al., 2013; Fleming et al.,

2013; Sharan et al., 2009 has ‘water’, which we have named ‘liquid’ instead.

Our working definition of materials here is heavily influenced by Adelson (2001) and Fleming (2017), where material does not just refer to prototypical materials that are used as construction materials such as *wood* and *stone*, but also to the ‘stuff’ that makes up ‘things’. For example, few people would consider *banana* as a material, but nevertheless this object has been made up of some type of *banana-material*, which humans are capable of recognizing, distinguishing, and estimating physical properties off. In this rational, we have included some less typical ‘stuff’ categories, such as for *food* and *animal*. Note however that we made an exception for *skin*, instead of a more overarching ‘human’ category as one might expect considering based on the previous. We made this choice because of the scientific interest in the artistic depiction (Lehmann, 2008), perception (Matts et al., 2007; Stephen et al., 2011), and rendering of skin directly (Igarashi et al., 2007; Jensen et al., 2001). Last, we realized that for many paintings a large portion was dedicated to the depiction of the sky or ground ‘stuff’, neither of which are considered a prototypical material, but on average both take up a large portion of paintings. Therefore, in an attempt to more densely annotate the whole region of the painting, we included *sky* and *ground*.

In one AMT task, participants would be presented with 40 paintings at a time and one target material category. In the task, participants were asked if the painting depicted the target material (e.g., *does this painting contain wood?*). They could reply ‘Yes, the target material is depicted in this painting.’ by clicking the painting and inversely, by not clicking the painting, participants would reply with ‘No, the target material is not depicted in this painting.’. Each painting was presented to at least 5 participants for each of the 15 materials. If at least 80% of the responses per painting claimed that the material was depicted in the painting, we would register that material as present for that painting. In total, we collected 1,614,323 human responses in this stage from 3,233 unique AMT users participating.

3.2.3. EXTREME CLICK BOUNDING BOXES

In the previous stage, paintings were registered to depict or not to depict a material. However, that stage does not inform us (1) how often the material is depicted, nor (2) where the material(s) are within the painting.

We gathered this information on the basis of extreme click bounding boxes. For extreme click bounding boxes, a participant is asked to click on the 4 extreme positions of the material: the highest, lowest, most left-, and most right-wards point (Papadopoulos et al., 2017). See Fig. 3.2 for an example. In the task, participants were presented with paintings that depicted the target material and tasked to create up to 5 extreme click bounding boxes for the target material.

To make bounding boxes within the task, the participants would use our interface, which allows users to zoom in and out, and pan around the image. The interface furthermore allowed participants to finely adjust the exact location of the extreme points by dragging the points around. Initially, the tasks were open to all AMT workers, but after around 2000 bounding boxes were created by 114 AMT users, with manual inspection, we found that the quality of bounding boxes varied greatly between participants. Therefore, we restricted the work to a smaller number of manually selected participants who were observed to create good bounding boxes. After this restriction, new boxes were manually inspected by the authors, and in a few cases additional participants were restricted due to a deterioration of

bounding box quality. Simultaneously additional participants were granted access to our tasks after passing (paid) qualification tasks. As a result, the number of manually selected participants varied between 10 and 20 participants. In total, 227,810 bounding boxes were created by participants.

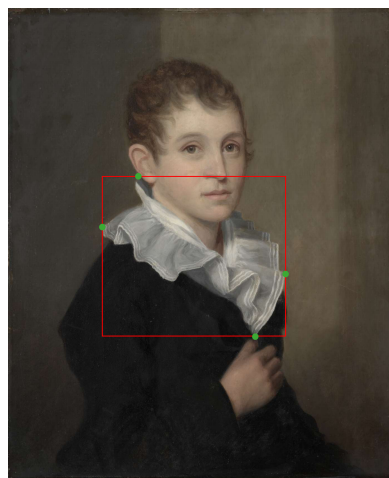


Figure 3.2: An example of four extreme clicks (marked in green) made by a user on a piece of fabric. These points correspond to the most left, most right, highest, and lowest points on the annotated item. The red-line displays the resulting bounding box. *Samuel Barber Clark*, by *James Frothingham*, 1810, Cleveland Museum Of Art, image reproduced under a CC0 license.

Automatic bounding boxes. While we consider our dataset to be quite larger, it only covers a small but representative portion of art history. It might be required to access materials in paintings that are not part of our dataset. To allow for this, we have trained a FasterRCNN (Ren et al., 2015) bounding box detector to localize and label material boxes in unlabelled paintings. We use the publicly available implementation from Wu et al. (2019) with a ResNet-50 backbone and feature pyramid network (R50-FPN). The model is fine-tuned from a COCO-pretrained model for 100 epochs using the default COCO hyperparameters from Wu et al. (2019). First we trained the detector on 90% of annotated paintings in the dataset. In section *Automatically detected bounding boxes* in the *Results and demonstrations* section below, we show our evaluation of the network, which was performed on the remaining 10% of annotated paintings. While we created this network to be able to detect paintings outside our dataset, we decided to apply the network on our dataset in order to more densely annotate our paintings. Therefore, after the evaluation, we ran the detection network on the entire set of paintings, i.e., training and testing data, in an attempt to more exhaustively annotate materials within paintings. From the automatic detected bounding boxes we first removed all boxes that scored $< 50\%$ confidence (as calculated by FasterRCNN). Next, we filtered out automatic boxes that were likely already identified by human annotators, by removing automatic bounding boxes that scored $\geq 50\%$ on intersection over union, i.e., automatic boxes that shared the majority of its content with human boxes. This resulted in an additional 96k bounding boxes.

3.2.4. FINE-GRAINED LABELS

In this step we supplemented the previously collected material labels with fine-grained material labels (see Table 3.2). For example, a bounding box labelled as *fabric* could now be labelled as *silk*, *velvet*, *fur*, etc.. We excluded bounding boxes that were too small (e.g., $\text{width in pixels} \times \text{height in pixels} \leq 5000$) and boxes that were labelled as *sky*, *ground* or *skin* for which fine-grained categorizations were not annotated. We collected fine-grained labels for the remaining 150,693 bounding boxes. Note that this only concerns the bounding boxes created by human annotations as no automatically detected boxes were assigned a fine-grained material label. For each of these 150,693 bounding boxes, we gathered responses from at least 5 different participants. If the responses reached an agreement of at least 70%, we would assign the agreed upon label to the bounding box. To guide the workers, we provide a textual description for each fine-grained category for them to reference during the task. We did not provide visual exemplars as we did not want to bias the workers into template matching instead of relying on their own perceptual understanding.

We found that it is non-trivial to define fine-grained labels in such a way that they are concise, uniform, and versatile (i.e., useful across different scientific domains) while still being recognizable and/or categorizable by naive observers. We applied the following reasoning to select fine-grained labels: first, we tried to divide the materials into an exhaustive list with as few fine-grained labels as possible. For example, for ‘wood’, each bounding box is either ‘processed wood’ or ‘natural wood’. If an exhaustive list would become too long to be useful, we would include an ‘other’ option. For example, for ‘glass’ we hypothesized that the vast majority of bounding boxes would be captured by either ‘glass windows’ or ‘glass containers’. However, to include all possible edge cases such as glass spectacles and glass eyeballs, we included the ‘other’ option.

A possible subset for ‘metal’ we considered was ‘iron’, ‘bronze’, ‘copper’, ‘silver’, ‘gold’, ‘other’. However, we feared that naive participants would not be able to consistently categorize these metals. An alternative would be to subcategorize on object-level, e.g. ‘swords’, ‘nails’, etc., but as we are interested in material categorization, we tried to avoid this as much as possible. Thus, for ‘metal’, and for the same reason ‘ceramic’, we required a different method. We chose to subcategorize on color, as often the color for these materials are tied to object identity.

Participants are shown one bounding box at a time and are instructed to choose which of the fine-grained labels they considered most applicable. Additionally, they are able to select a ‘not target material’ option.

We collected over one million responses from 1114 participants. This resulted in a total of 105,708 boxes assigned with a fine-grained label. See Table 3.2 for the numbers per category.

3.3. RESULTS AND DEMONSTRATIONS

We conducted a diverse set of experiments to demonstrate how our annotated art-perception dataset can drive research across perception, art history, and computer vision. First, we report simple dataset statistics. Next, we organized our findings under the proposed dataset usage taxonomy: perceptual demonstrations, cultural visual ecology demonstrations and computer vision demonstrations.

3.3.1. DATASET STATISTICS

The final dataset contains painterly depictions of materials, with a total of 19,325 paintings. Participants have created a total of 227,810 bounding boxes and we additionally detected 96k using a FasterRCNN. Each box has a coarse material label and 105,708 also have been assigned a fine-grained material label. The total number of instances per material categories (coarse- and fine-grained) can be found in Table 3.2. Further analysis of spatial distribution of categories, co-occurrences, and other related statistics will be discussed in a following section in the context of visual ecology.

Table 3.2: The number of annotated bounding boxes for each coarse- and fine-grained category. Note that not every bounding box is associated with a fine-grained label since participants were not always able to arrive at a consensus. See main text for details.

| Coarse-grained | Fine-grained | # Labels |
|----------------|-------------------------|----------|
| animal | | 11606 |
| | birds | 1822 |
| | reptiles and amphibians | 144 |
| | fish and aquatic life | 289 |
| | mammals | 7752 |
| | insects | 155 |
| | other animals | 10 |
| ceramic | | 3641 |
| | brown or red | 1088 |
| | white | 381 |
| | decorated | 289 |
| | other ceramic | 14 |
| fabric | | 31557 |
| | velvety | 261 |
| | lace | 491 |
| | silky/satiny | 1354 |
| | cotton/wool-like | 5712 |
| | brocade | 96 |
| | fur | 27 |
| | other fabric | 12 |
| flora | | 26693 |
| | trees | 12851 |
| | vegetables | 96 |
| | fruits | 1238 |
| | flowers | 2515 |
| | plants | 3699 |
| food | | 3690 |
| | cheese | 11 |

Continued on next page

Continued from previous page

| Coarse-grained | Fine-grained | # Labels |
|----------------|-------------------------------------|----------|
| | vegetables | 107 |
| | fruits | 1536 |
| | meat or poultry | 183 |
| | bread | 127 |
| | seafood | 183 |
| | nuts | 8 |
| | other | 14 |
| gem | | 10525 |
| | pearls | 719 |
| | gemstones | 715 |
| | other gems | 1 |
| glass | | 5546 |
| | glass window | 2243 |
| | glass container | 1003 |
| | other glass | 171 |
| ground | | 2552 |
| liquid | | 5737 |
| | body of water | 4583 |
| | liquid in container | 458 |
| | other liquid | 172 |
| metal | | 27708 |
| | colorless metal | 2933 |
| | yellowish metal | 4435 |
| | brownish or reddish metal | 510 |
| | multicolored or other colored metal | 215 |
| paper | | 3167 |
| | paper book | 1380 |
| | paper sheets | 585 |
| | paper scrolls | 114 |
| | other paper | 19 |
| skin | | 32323 |
| sky | | 12734 |
| stone | | 23157 |
| | processed stone | 9226 |
| | natural stone | 9429 |
| wood | | 26953 |
| | processed wood | 12810 |
| | natural wood | 10751 |

3.3.2. PERCEPTUAL DEMONSTRATIONS

We believe that one of the benefits of our MIP dataset is that selections of the dataset can be useful as stimuli for perceptual experiments. We demonstrate this by performing an

annotation experiment to study the painterly depiction of highlights on drinking glasses.

PERCEPTION-BASED RECIPES IN PAINTERLY DEPICTIONS

As previously argued, we believe that painterly techniques are a sort of perception-based recipe. Applying these recipes results in a stylized depiction that can trigger a robust perception of the world. Studying the image features in paintings can lead to an understanding of what cues the visual systems needs to trigger a robust perception.

Here we explore a perceptual shortcut for the perceptions of glass by annotating highlights in paintings and comparing these with highlights in photos. In paintings, it has previously been observed that highlights on drinking glasses are typically in the shape of windows, even in outdoor scenes (Miller, 1998). This highlight-shape can even often be found in contemporary cartoons (Anjyo and Hiramitsu, 2003; Pacanowski et al., 2008). This convention can be considered as a perception-based recipe, where the result is a window-shaped highlight that appears to be a robust cue that triggers the perception of gloss for drinking glasses.

We used bounding boxes from our dataset and photographs sourced from COCO (Lin et al., 2014). Participants for this study included 3 of the authors, and one lab-member naive to the purpose of this experiment.

Images. We used 110 images of drinking glasses, split equally across paintings and photos. First, we selected all bounding boxes in the *glass*, *liquid container* category in our dataset. From this set, we manually selected drinking glasses, since this category can also contain items such as glass flower vases. Next, we removed all glasses that were mostly occluded, were difficult to parse from the background - for example when multiple glasses were standing behind each other, and removed images smaller than 300x300 pixels. This resulted in a few hundred painted drinking glasses.

Next, we downloaded all images containing cups and wineglasses from the COCO (Lin et al., 2014) dataset, from which we removed all non-glass cups, occluded glasses, blurry glasses, and glasses that only occupy a small portion of the image, and small images. This left us with 55 photos of glass cups and wineglass. Next, we randomly selected 55 segmentations from our painted glass collection. Each image was presented in the task at 650 × 650 pixels, keeping aspect ratio intact.

During this selection phase, we did not base our decision on the shape of the glass. After the experiment, as part of the analysis, we divided the glasses into three shapes, namely spherical, cylindrical, and conical glasses. See Fig. 3.3 for an example of each shape.

Task. Participants annotated highlights on drinking glasses using an annotation interface. In the annotation interface, users would be presented with an image on which the annotated geometry was visible. This made it clear which glass should be annotated, in case multiple glasses were visible in the image. Users were instructed to annotate all visible highlights on that glass. Once the user started annotating highlights, the geometry would no longer be visible. Annotations could be made by simply holding down the left-mouse button and drawing on top of the image. Once a highlight was annotated a user could mark it as finished and continue with the next highlight, and eventually move to the next image.



Figure 3.3: Examples of the three glass shapes. From left to right: spherical, cylindrical, and conical. The red geometry annotations were manually created by the authors, and were used to standardize across glasses for the highlight analysis. The images have reduced contrast only to enhance the visibility of the red lines for this figure and were presented to participants with the original contrast. Paintings used, from left to right: *Portret van een jongen, zittend in een raamnis en gekleed in een blauw jasje*, by Jean Augustin Daiwaille. 1840. *Still Life with a gilded Beer Tankard*, by Willem Claesz. Hed, 1634. *The White Tablecloth*, by Jean Baptiste Siméon Chardin, 1731. The left and middle images courtesy of The Rijksmuseum. The right image courtesy of The Art Institute of Chicago. All images reproduced under a CC0 license.

Results. To compare the highlights between photos and paintings, we resized each glass to have the same maximum width and height, and then overlaid each glass on the center. Initially, we overlaid all images for both types of media (not visualized here) and found the resulting figure quite noisy. However, when we split the glasses on media and shape, clear patterns emerge for painted glasses Fig. 3.4.

As can be seen, painters are more likely to depict highlights on glasses adhering to a stylized pattern, at least for spherical and conical glasses. This pattern of highlights is perceptually convincing and very uniform in comparison with the variation found within reality.

Furthermore, we calculated the agreement between each pair of participants, as the ratio of pixels annotated by both participants (i.e., overlapping area) divided by the number of pixels that was annotated by either participant (i.e., total area). Averaged across participants, the agreement on paintings (0.33) was around 50% higher relative to the average agreement between participants on photos (0.21). This means that for our stimuli, highlights in paintings are less ambiguous when compared to photos.

3.3.3. CULTURAL VISUAL ECOLOGY DEMONSTRATIONS

The ecology displayed within paintings are representative of our visual culture. Our dataset consists of paintings spanning 500+ years of art history. This provides a unique opportunity to analyze a specific sub-domain of visual culture, i.e., that of paintings. Here we first analyze the presence of materials in paintings in the *Material presence* section and in the next section we analyse this over time. In *The spatial layout of materials*, we visualize the spatial distributions of materials in our dataset. In the last section, we analyze the automatically detected bounding boxes.

Highlight position between media for different shapes

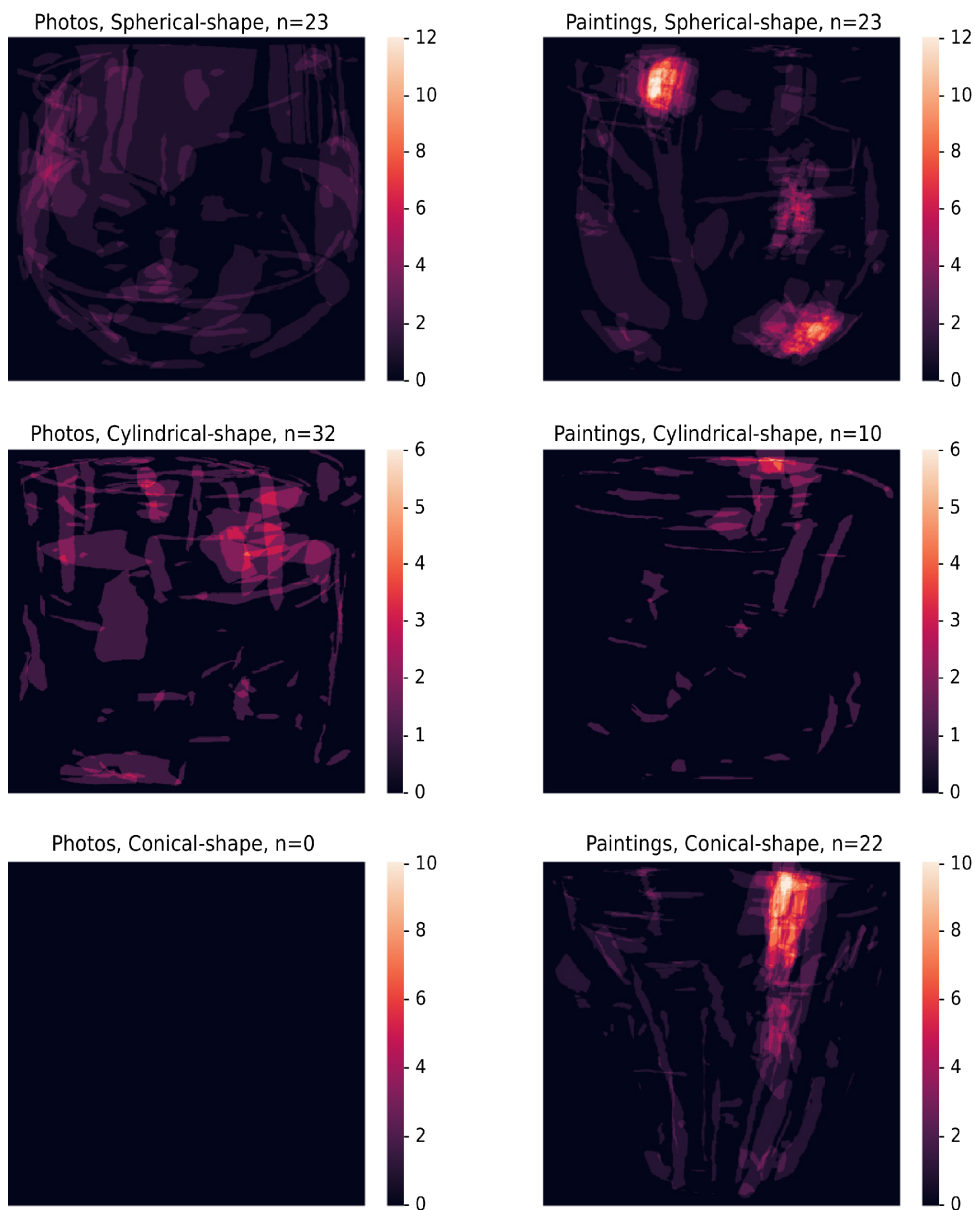


Figure 3.4: The overlaid highlights created by users, split on media and glass shape. In general, the photographic glass shapes display more variability and do not display a clear pattern. Note that for photos, no stimuli existed with a conical shape in our set which leads to a black image, since there were no highlight-annotations. On the right, for painted glasses, we see clear patterns in the placement of highlights for each glass shape.

Material presence. Within the 19,325 paintings, participants exhaustively identified the presence of 123,244 instances of 15 coarse materials. In other words, for each painting, participants indicated if each material is or is not present. The distribution of unique materials per painting is normally distributed with an average of 5.7 unique coarse materials present per painting (std = 2.8 materials). The most frequent materials are *skin* and *fabric*. The least frequent are ceramics and food. The relative frequency of each coarse material is presented in Fig. 3.5. We did not exhaustively identify fine-grained materials within paintings, so we will not report those statistics here.

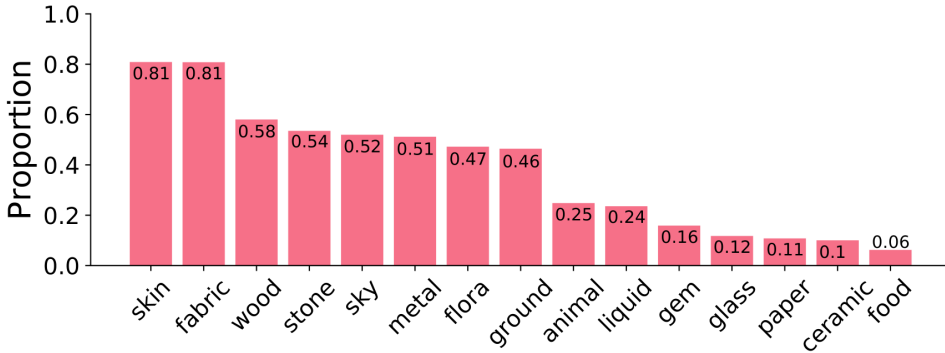


Figure 3.5: The proportion of paintings in our dataset that depict at least one instance of each material.

Based on prior knowledge of natural ecology, one might assume that some materials, such as *skin* and *fabric* might often be depicted together in paintings. To quantify the extent to which materials are depicted together, we create a co-occurrence matrix presented in Fig. 3.6, where each cell is the co-occurrence for each pair of materials as the number of paintings where both materials are present, divided by the number of paintings where either (but not both) materials are present. We can see for example, that if *skin* is depicted, there is a 94% change to also find *fabric* in the same painting.

Furthermore, one might expect that the presence of one material can have an influence on another material. For example, one might expect that *gem* might almost always be depicted with *skin*, but that *skin* is only sometimes depicted with *gem*. To quantify these relations, we calculated the occurrence of a material given that another material is present. We visualize this in Fig. 3.7. Here we see that if *gem* is present, then *skin* is found in 99% of the paintings, but that if *skin* is present, then *gem* is found in only 20% of the paintings. The same relationship is true for *gem* and *fabric*. This implies that gems are almost always depicted with human figures, however that human figures are not always shown with gems. Another example, when *liquid* is present, in 85% of the paintings, *wood* is also present. One might be reminded of typical naval scenes, or landscapes with forests and rivers. Inversely, when *wood* is present, only 34% of the paintings depict *liquid*. For *food* and *ceramics*, two materials which are present in less than 10% of paintings, we see that if *food* is present, *ceramics* has a 53% change to be present as well, but the inverse is only 33%. This implies that food is served in, or with, ceramic containers half of the time, but that this is only 1/3rd

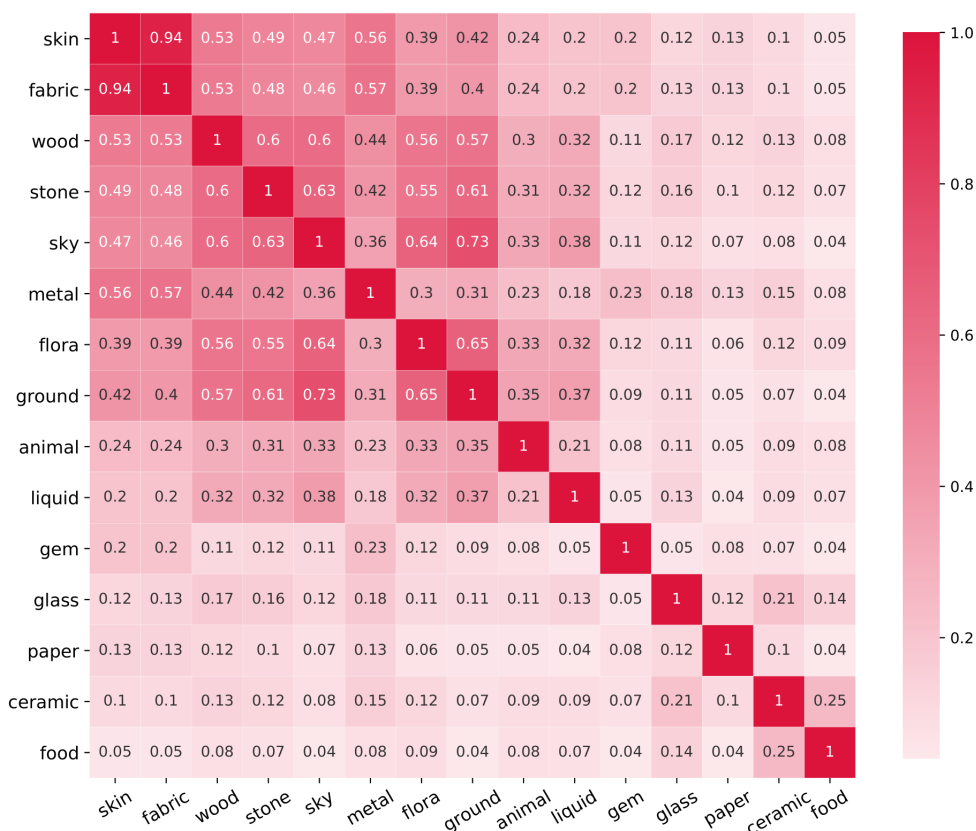


Figure 3.6: Co-occurrence matrix. Each cell equals the number of paintings where both materials are present divided by the number of paintings where one or the other material is present.

of what ceramics is used for.

Material presence over time. We have previously shown the distributions of materials in paintings in Fig. 3.5. When we created similar distributions (not visualised) for temporal cross-sections, for example for a single century, we found that these distributions were remarkably similar to the average distribution in Fig. 3.5. We used t-tests, to see if the distribution for any century was significantly different from the average distribution in Fig. 3.5 and found no significant effect. This means that despite the changes in stylistic and artistic techniques over time, the distribution of materials (such as in Fig. 3.5) remained remarkably stable over time for the period covered in our dataset.

The spatial layout of materials. Paintings are carefully constructed scenes and it follows that a painter would carefully choose the location at which to depict a material. With the knowledge that spatial conventions exists within paintings (e.g., lighting direction (Carbon



Figure 3.7: Likelihood matrix. This matrix visualizes the influence a material has on the likelihood of finding another material within the same painting, i.e., if one material on the y-axis is present, then how does this impact the presence of other materials on the x-axis? Calculated as the number of paintings where both materials are present, divided by the number of paintings that contain only one of the materials.

and Pastukhov, 2018; Sun and Perona, 1998), one can assume that these might extend to materials. The average spatial location and extent of materials is visualized by taking the (normalized) location of each bounding box for a specific material and subsequently plotting each box as a semi-transparent rectangle. The result is a material heatmap, where the brightness of any pixel indicates the likelihood to find a material at that pixel. In this section, we limit the material heatmaps to only include the bounding boxes created by human annotators. In the next section, we visualize the material heatmaps for automated boxes too.

Material heatmaps for the 15 coarse materials are shown in Fig. 3.8. The expected finding that *sky* and *ground* are spatially high and low within images serves as a simple validation or sanity-check of the data. It is interesting to see how *skin* and *gem* are both vertically centered within the canvases. It appears to suggest a face, with necklaces and jewelry adorning the figure. In general, each material heatmap appears to be roughly vertically

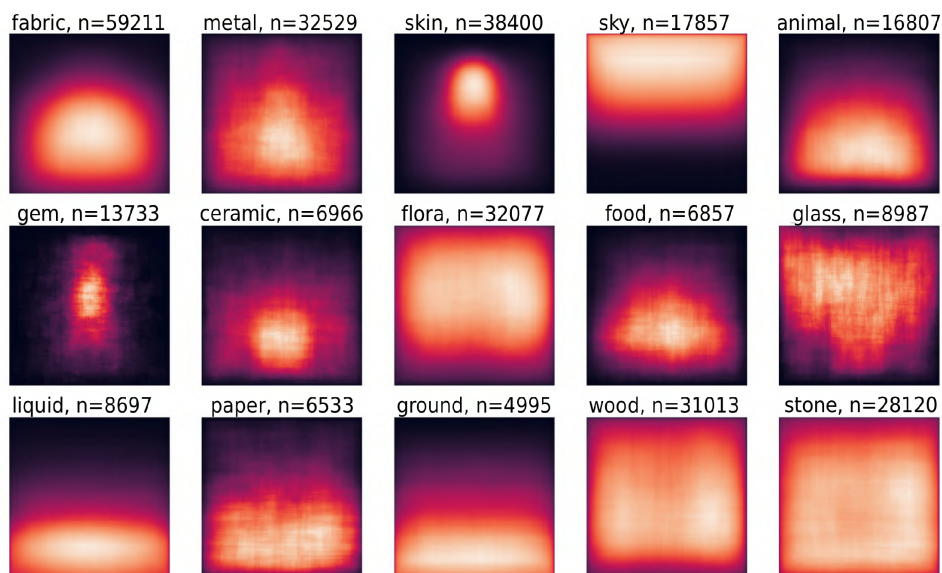


Figure 3.8: Material heatmaps, which illustrate the likelihood at any given pixel to find the target material at that pixel. Brighter colors indicate higher likelihoods.

symmetric. For *glass*, there does however appear to be a minor shift towards the top-left. This might be related to an artistic convention, namely that light in paintings usually comes from a top-left window (Carbon and Pastukhov, 2018). When we look at the heatmaps for the sub-categories for glass in Fig. 3.9, we see that it is indeed glass windows that show the strongest top-left bias.

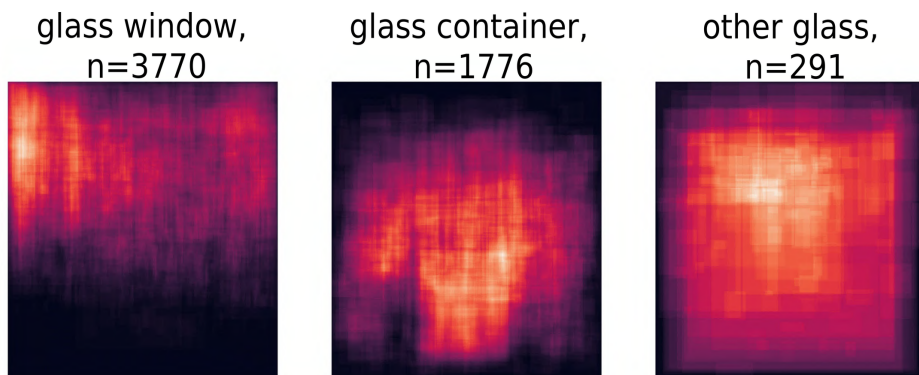


Figure 3.9: Material heatmaps for glass sub categories. For glass windows, it is interesting to see the clustering in the top-left corner, which is in agreement with the artistic convention of having light come from the top left.

Automatically detected bounding boxes. Besides the bounding boxes created by humans, we also trained a FasterRCNN network to automatically detect bounding boxes with 90% of the data as training data. On the remaining unseen 10% of paintings, the network detected 90,169 bounding boxes. We removed those with a confidence score below 50%, which resulted in 24,566 remaining bounding boxes. In the section below, all references to the automated bounding boxes refer to these 24,566 bounding boxes.

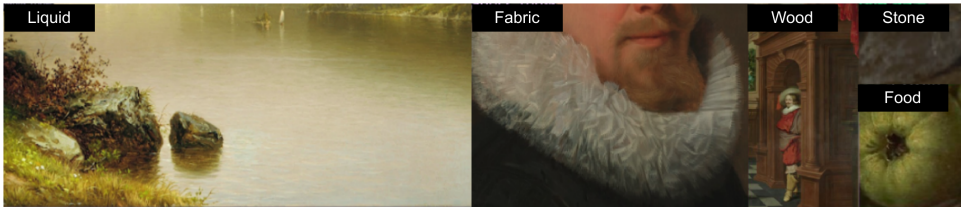


Figure 3.10: Examples of detected materials in unlabeled paintings. Automatically detecting materials can be useful for content retrieval for digital art history and for filtering online galleries by viewer interests.

A qualitative sample of detected bounding boxes is given in Fig. 3.10. Our human bounding boxes are non-spatially exhaustive in nature meaning that not every possible material has been annotated. As a result, the automatically created bounding boxes cannot always be matched against our human annotations and thus we cannot use this to evaluate their quality. In order to validate the automatic bounding box detection, we performed a simple user study to get an estimate of the accuracy per material class, which is visualized in Fig. 3.11. In the user study, a total of 50 AMT participants judged a random sample of 1500 bounding boxes. The 1500 bounding box stimuli were divided into 10 sets of 150 stimuli, where each set contain 10 boxes per course material class. Each individual participant only saw one set and each set was seen by 5 unique participants. Therefore, this can be thought of as 10 experiments, each with 5 participants and 150 stimuli, where participants performed the same task across each set/experiment. The participants were tasked to rate whether each stimuli is either a *good* or a *bad* bounding box, where a *bad* bounding box was defined as either 1) having the wrong material label (e.g., "I see wood, but the label says stone") or as having a bad boundary where the edges of the bounding box were not near the edges of the material. The order of stimuli was randomized between sets and participants. This leads to a total of 7500 votes, 500 per material classes. The ratio of good to bad votes per material classes can serve as a measure of accuracy, which has been visualized in Fig. 3.11.

The participant agreement averaged across bounding boxes was found to be 80%, i.e., on average 4 out of 5 participants agreed on their rating per bounding box. As a result of the user study, we found a mean average accuracy of 0.55 across participants. While not high, these results are somewhat interesting in that they show that a FasterRCNN model is capable of detecting materials in paintings, without any changes to the network architecture or training hyperparameters. It is certainly promising to see that an algorithm designed for object localization in natural images can be readily applied to material localization in paintings. Likely, the accuracy could be further improved by finetuning the network which we have not done in this paper.

It is interesting to note that the spatial distribution of automatically detected bound-

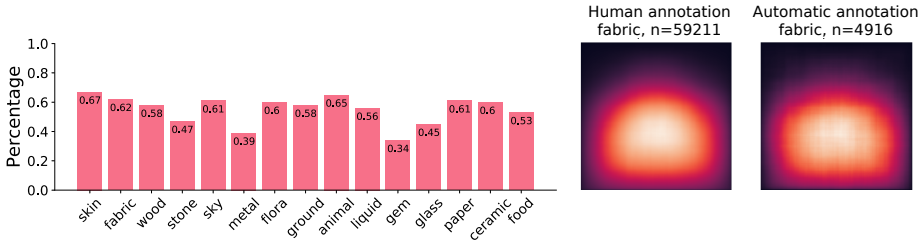


Figure 3.11: In the bar graph, the accuracy for automatically detected bounding boxes is displayed in the same order as in Fig. 3.5. The values were derived from human quality votes. On the right, we compare the material heatmaps for fabric between the automated and the human annotation bounding boxes. From left to right, top to bottom: *Lake George*, by John William Casilear, 1857. *Man with a Celestial Globe*, by Nicolaes Eliasz Pickenoy, 1624. *A Seven-Part Decorative Sequence: An Interior*, by Dirck van Delen, 1631. *Thomas Howard, 2nd Earl of Arundel*, by Anthony van Dyck, 1620. *The Poultry Seller*, by Cornelis Jacobsz. Delff, 1631. First and second digital image courtesy of The Metropolitan Museum of Art. Third and last image courtesy of The Rijksmuseum. Fourth image courtesy of the Getty's Open Content Program. All images reproduced under a CC0 license.

ing boxes looks very similar to the spatial distribution of the human annotated bounding boxes. We have visualized the material heatmap for one material, fabric, for the automated bounding boxes to show the similarity with the material heatmap for the same material created from human annotation bounding boxes. This has been visualized in the right side of Fig. 3.11

3.3.4. COMPUTER VISION DEMONSTRATIONS

In this section, we will first apply existing segmentation tools designed for natural photographs to extract polygon segmentations. Next, we perform an experiment to demonstrate the utility of paintings for automated material classification.

EXTRACTING POLYGON SEGMENTATIONS

A natural extension of material bounding boxes is material segments (Bell et al., 2013, 2015; Caesar et al., 2018). Polygon segmentations are useful for reasoning about boundary relationships between different semantic regions of an image, as well as the shape of the regions themselves. However, annotating segmentations is expensive and many modern datasets rely on expensive manual annotation methods (Bell et al., 2013; Caesar et al., 2018; Cordts et al., 2016; Lin et al., 2014; Zhou et al., 2017). Recent work has focused on more cost effective annotation methods (e.g., Benenson et al., 2019; Lin et al., 2019; Ling et al., 2019; Maninis et al., 2018). One broad family of methods to relax the difficulty of annotating polygon segmentations is through the use of interactive segmentation methods that transform sparse user inputs into a full polygon masks.

For this dataset, we apply interactive segmentation with the crowdsourced extreme clicks as input. To evaluate quality, we compared against 4.5k high-quality human annotated segmentations from van Zuijlen et al. (2020), which were sourced from the same set of paintings. We find that both image-based approaches like GrabCut (GC) (Rother et al., 2004) and modern deep learning approaches such as DEXTR (Maninis et al., 2018)

perform well. Surprisingly, DEXTR transfers quite well to paintings despite being trained only on natural photographs of objects. The performance is summarized in Table 3.3. The performance is summarized using the standard intersection over union (IOU) metric. IOU is computed as the intersection between a predicted segment and the ground truth segment divided by the union of both segments. IOU is computed for each class, and mIOU is the mean IOU over all of the classes. Samples are visualized in Fig. 3.12. Segments produced by these methods from our crowdsourced extreme points will be released with the dataset.

Table 3.3: Segmentations from extreme clicks. Grabcut (Rother et al., 2004) rectangles use bounding-box only initialization as a reference baseline. Grabcut Extr is based on the improved GC initialization from (Papadopoulos et al., 2017) with small modifications: (a) we compute the minimum cost boundary with the cost as the negative log probability of a pixel belonging to an edge; (b) in addition to clamping the morphological skeleton, we also clamp the extreme points centroid as well as the extreme points; (c) we compute the GC directly on the RGB image. DEXTR (Maninis et al., 2018) Pascal-SBD and COCO are pretrained DEXTR ResNet101 models on the respective datasets. Note that Pascal-SBD and COCO are natural image datasets of objects, but DEXTR transfers surprisingly well across both visual domains (paintings vs. photos) and annotation categories (materials vs. objects).

| mIOU (%) | | | | |
|-------------------|--------------|------------------|------------|----------------|
| Grabcut Rectangle | Grabcut Extr | DEXTR Pascal-SBD | DEXTR COCO | DEXTR Finetune |
| 44.1 | 72.4 | 74.3 | 76.4 | 78.4 |

| DEXTR Finetune IOU By Class (%) | | | | |
|---------------------------------|---------|--------|--------|-------|
| Animal | Ceramic | Fabric | Flora | Food |
| 76.9 | 86.8 | 79.1 | 77.0 | 87.5 |
| Gem | Glass | Ground | Liquid | Metal |
| 74.4 | 83.2 | 69.6 | 73.0 | 75.5 |
| Paper | Skin | Sky | Stone | Wood |
| 86.1 | 78.9 | 78.5 | 81.7 | 67.4 |

LEARNING ROBUST CUES FOR FINEGRAINED FABRIC CLASSIFICATION

The task of distinguishing between images of different semantic content is a standard recognition task for computer vision systems. Recently, increasing attention is being given to “fine-grained” classification, where a model is tasked with distinguishing images of the same coarse-grained category (e.g., distinguishing different species of birds or different types of flora (Van Horn et al., 2018; Wah et al., 2011; Wei et al., 2019). Classifiers for material categories can perform reasonably well on coarse-grained classification by relying on context alone. In comparison, fine-grained classification is more challenging for deep learning systems as contextual clues are often equal within fine-grained classes. For example, one might reason that the material of a shirt might be recognized as fabrics partly because of the context, i.e., being worn by a figure. However, in fine-grained classification context can be held consistent across classes (for example, both velvet shirts and satin shirts

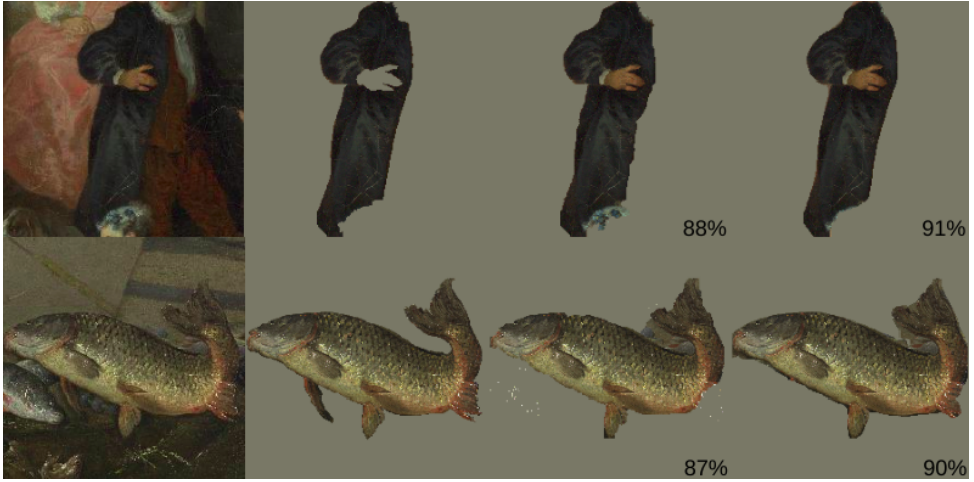


Figure 3.12: Segmentation visualizations. Left to right: Original Image, Ground Truth Segment, Grabcut Extr Segment, DEXTR COCO Segment. Both Grabcut and DEXTR use extreme points as input. For evaluation, the extreme points are generated synthetically from the ground truth segments. The IOU for each segmentation is shown in the bottom right corner. Top image: *Dance before a Fountain*, by Nicolas Lancret, 1724, Digital image courtesy of the Getty’s Open Content Program. Bottom image: *Still life with fish*, by Pieter van Noort, 1660, Het Rijksmuseum. Images reproduced under a CC0 license.

are worn). To successfully distinguish between these two fine-grained classes in a context-controlled setting, a classifier should use non-contextual features (at least more so relative to uncontrolled settings).

The rationale above leads to two interesting possibilities. First, we hypothesize that the painted depictions of materials can be beneficial for fine-grained classification tasks. Since artistic depictions focus on salient cues for perception, i.e., paintings are explicitly created for and by perception, it is possible that a network trained on paintings is able to learn a more robust feature representation by focusing on these cues.

Second, visualising the features used by a successful fine-grained classifier could potentially lead to the uncovering of latent perceptual cues. For example, in the *Perception-based recipes in painterly depictions* section above, we showed that window-shaped highlights are a robust cue for the perception of gloss on drinking glasses. However, it is assumed that the visual system used many such cues which are as of now unknown. Visualising what cues are used by classifiers might lead to the finding of cues used by the perceptual system.

Task. We experimented with the task of classifying cotton/wool versus silk/satin. The latter can be recognized through local cues such as highlights on the cloth; such cues are carefully placed by artists in paintings. To understand whether artistic depictions of fabric allow a neural network to learn better features for classification, we trained a model with either photographs or paintings. High resolution photographs of cotton/wool and silk/satin fabric and clothing (dresses, shirts) were downloaded and manually filtered from publicly available photos licensed under the Creative Commons from Flickr. In total, we downloaded roughly 1K photos. We sampled cotton/wool and silk/satin samples from our dataset to

form a corresponding dataset of 1K paintings. We analyzed the robustness of the classifier trained on paintings versus the classifier trained on photos in two experiments below. Taken together, our results provide evidence that a classifier trained on paintings can be more robust than a classifier trained on photographs, and that visualizing these features could lead to discovering perceptual cues utilized by the human visual system.

Generalizability of classifiers. Does training with paintings improve the generalizability of classifiers? To test cross-domain generalization, we test the classifier on types of images that it has not seen before. A classifier that has learned robust features will outperform a classifier that has learned features based on more spurious correlations. We tested the trained classifiers on both photographs and paintings across the two classes using 1000 samples per domain.

In Table 3.4, the performance of the two classifiers are summarized. We found that both classifiers perform similarly well on the domain they are trained on. However, when the classifiers are tested on cross-domain data, we found that the painting-trained classifier performs better than the photo-trained classifier. This suggests that the classifier trained on paintings has learned a more generalizable feature representation for this task.

We have reported the confusion matrices in Table 3.5 . The photo classifier applied to paintings is heavily biased towards satin predictions. We hypothesize that this is because the photo classifier is relying on spurious cues (such as image background or clothing shape) over more robust cues and thus that the shift from photos to paintings causes its mispredictions. Only 21% of cotton samples are correctly identified as cotton while 79% are identified as satin. This skew in precision/recall across the classes is also reflected by the F1 scores for each class. On the other hand, the painting classifier applied to photos is much more balanced in its predictions, with 57-59% of predictions being correct. The precision/recalls are also much better balanced as reflected by the F1 scores.

Table 3.4: Classifier performance across domains. Classifiers are trained to distinguish cotton/wool from silk/satin. The first column represents the classifier trained on photographs, and the second column represents the classifier trained on paintings. In the first row, the classifiers are tested on images of the same type they were trained on (i.e., trained and tested on photos, and trained and tested on paintings). In the second row, the classifiers are tested on the other medium, i.e., trained on photos and tested on paintings and vice versa.

| | Photo → Photo | Painting→ Painting |
|---------------|------------------|--------------------|
| MEAN F1 Score | 79.6% | 80.5% |
| | Photo → Painting | Painting→ Photo |
| MEAN F1 Score | 49.5% | 57.8% |

Human agreement with classifier cues. How indicative are the cues used by each classifier to humans? We hypothesized that training networks on paintings might lead to the use of more perceptually relevant image features. If this is true, then the features used by the classifier trained on paintings should be preferably by humans.

Table 3.5: Confusion matrix for the two classifiers. The top represents the classifier trained on photos, tested on paintings which is heavily biased towards satin. The bottom represents the classifier trained on paintings, tested on photos, which is more balanced in its predictions.

| Photo → Painting | | |
|------------------|--------|--------|
| | Cotton | Satin |
| Cotton | 20.83% | 79.17% |
| Satin | 9.72% | 90.28% |
| Per class F1 | 31.91 | 67.01 |

| Painting → Photo | | |
|------------------|--------|--------|
| | Cotton | Satin |
| Cotton | 58.82% | 41.18% |
| Satin | 42.86% | 57.14% |
| Per class F1 | 55.56 | 60.00 |

We produced evidence heatmaps with GradCAM (Selvaraju et al., 2017) from the feature maps in the network before the fully connected classification layer. We extracted high resolution feature maps from images of size 1024×1024 (for a feature map of size 32×32). The heatmaps produced by GradCAM show which regions of an image the classifier uses as evidence for a specific class. If the cues (i.e., *evidence heatmaps*, such as in Fig. 3.13) are clearly interpretable, this would imply the classifier has learned a good representation. For both models, we computed heatmaps for test images corresponding to their ground truth label. We conducted a user study on Amazon Mechanical Turk to find which heatmaps are judged as more informative by users. Users were shown images with regions corresponding to heatmap values that are above 1.5 standard deviations above the mean. Fig. 3.13 illustrates an example. Users were instructed to "select the image that contains the regions that look the most like <material>", where <material> was either cotton/wool or silk/satin. We collected responses from 85 participants, 57 of which were analyzed after quality control. For quality control, we only kept results from participants who spent over 1 second on average per trial.

Overall, we found that the classifier trained on paintings uses evidence that is better aligned with evidence preferred by humans (Fig. 3.14). This implies that training on paintings allows classifiers to learn more perceptually relevant cues, and it shows that this method might be useful to detect previously unknown perceptual cues.

Due to the domain shifts, training and testing a classifier on a single type of images will outperform a classifier trained and tested on different kind of images. Based on this, if paintings do not lead to a more robust feature representation we would expect the painting classifier to do best on paintings and the photo classifier to perform best on photos. Interestingly however, this does not hold when testing on photos of the satin/silk category (see last column of Fig. 3.14). We found that users actually have no preference for the cues from either classifier, i.e., the cues from the painting classifier appears to be equally informative as the cues from the photo classifier for categorizing silk/satin in photos. This suggests that

either (a) the painting classifier has learned human-interpretable perceptual cues for recognizing satin/silk, or (b) that the photo classifier has learned to classify satin/silk based on some spurious contextual signals that are difficult to interpret by humans. We asked users to elucidate their reasoning when choosing which set of cues they preferred. In general, users noted that they preferred the network which picks out regions containing the target class. Therefore, it seems that the network trained on paintings has learned better to distinguish fabric through the appearance of such fabrics in the image over other contextual signals (see Fig. 3.13).



Figure 3.13: Visualization of cues used by classifiers. Left to right: Original Image, Masked Image (Painting Classifier), Masked Image (Photo Classifier). The unmasked regions represent evidence used by the classifiers for predicting “silk/satin” in this particular image. See main text for details. Image from *Interior of the Laurenskerk at Rotterdam*, by Anthonie De Lorme, with figures attributed to Ludolf de Jongh, 1662. Digital image courtesy of the Getty’s Open Content Program, reproduced under a CC0 license.

3.4. CONCLUSION

In this paper, we presented the Materials in Paintings (MIP) dataset – a dataset of painterly depictions of different materials throughout time. The dataset can be visited, browsed, and downloaded at materialsinpaintings.tudelft.nl. The MIP dataset consists of 19,325 high resolution images of painting. Unlike existing datasets that contains paintings, such as for example (Khan et al., 2014; Mao et al., 2017), the MIP dataset contains exhaustive material labels across 15 categories for all paintings within the set. Additionally, human annotators have created 227,810 bounding boxes and we automatically identified an additional 96k bounding boxes. Each bounding box also contains a material label and half are additionally assigned with a fine-grained material label.

Although the findings reported in this study are valuable for their own sake, together they demonstrate the wide utility that a dataset of painterly depictions can serve. We hope that the MIP dataset can support research in multiple disciplines, as well as promote multidisciplinary research. We have shown that depictions in paintings are not just of interest for art history, but that they are also of fundamental interest for perception, as they can illustrate what cues the visual system may use to construct a perception. We have shown that computer vision algorithms trained on paintings appear to use cues more aligned with the human visual system, when compared to algorithms trained on photos. The benefits of

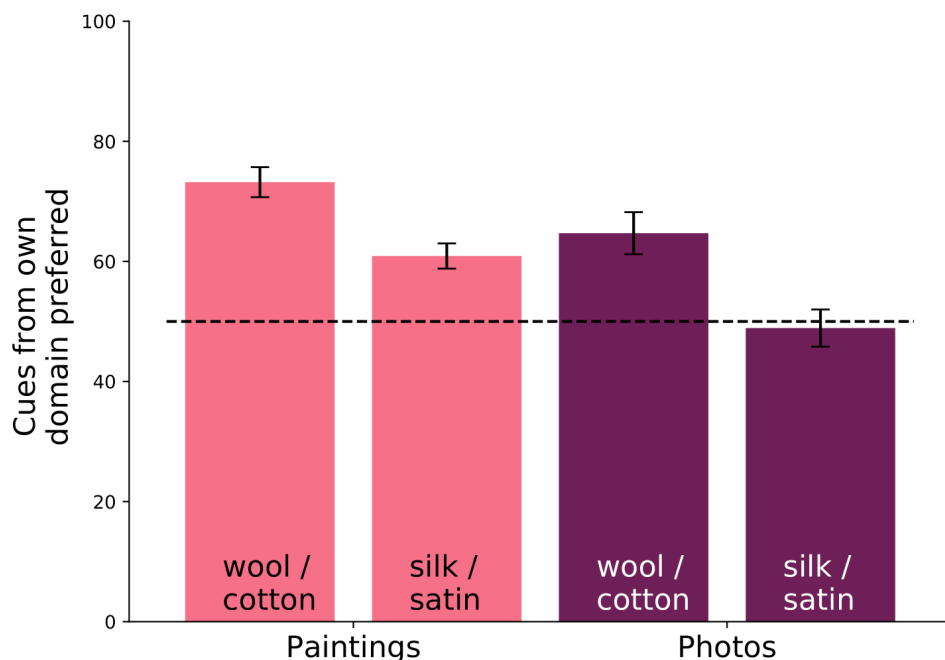


Figure 3.14: Human preference for classification cues used by each classifier. The y-axis represents how often humans prefer the cues from a classifier trained on the same domain as the test images. For example, the first bar indicates that in 73.2% of the cases, humans preferred cues from the classifier trained on paintings when classifying wool/cotton paintings (and thus, the inverse, that in 26.8% of the cases, humans preferred cues from the photo classifier.) Interestingly, note the last column – humans equally prefer cues used by both classifiers for classifying silk/satin photos despite the painting classifier never seeing a photo during training.

this might also extend to learning perceptually robust models for image synthesis.

Our findings support our hope that the MIP dataset will be a valuable addition to the scientific community to drive interdisciplinary research in art history, human perception, and computer vision.

ACKNOWLEDGMENTS

We appreciate the work and feedback of AMT participants that participated in our user studies. We further wish to thank Yuguang Zhao for his help with the design of the website.

4

THE MATERIALS IN PAINTING DATASET

4.1. WHAT IS IT?

The Materials In Painting (MIP) dataset consists of about 20.000 high-resolution digital images of paintings. The paintings are from 9 art galleries that have made (a portion of) their painting collection available to the public. A full list of all the galleries can be seen in Table 3.1 in the previous chapter. The paintings in MIP mostly depict western art but also include a small minority of Asian art. In addition to the paintings, we also have meta-data for most of the paintings provided by the galleries, including the title of the artwork, the attribution, and the year of production.

We have exhaustively labelled each painting with 15 material categories, meaning that, for each individual painting, we know if each of the 15 materials is present (i.e., "This painting has wood, metal, and stone, but no fabrics, glass..."). The 15 materials categories that we labelled are listed in Table 4.1.

Table 4.1: The 15 material categories used within the dataset. We also used sub-materials, which can be found in Table 3.2 in the previous chapter.

| | | | | |
|--------|---------|--------|--------|-------|
| animal | ceramic | fabric | flora | food |
| gem | glass | ground | liquid | metal |
| paper | skin | sky | stone | wood |

Next, we collected bounding boxes, the smallest possible box that completely encloses the region of interest. Here, the regions of interest were the depictions of the 15 material categories with the paintings. Bounding boxes were created using the Extreme Click method, in which a user clicks the 'extreme edges' of a material or object. The 'extreme edges' refer to the most left, highest, most right, and lowest point of the region of interest (see Fig. 4.1). About half of these boxes were also assigned a 'fine-grained' material label, meaning that boxes that were initially assigned a material label are now assigned a more

specific label; a box previously assigned as 'fabric', for example, now becomes 'fabric - cotton'.

4



Figure 4.1: An example of extreme clicks, i.e., the four green points. From the x,y coordinates of the extreme clicks we can draw a bounding box. In this example, the bounding box captures the metal pitcher. *Still Life with Ewer, Vessels, and Pomegranate*, by Willem Kalf. 1640. The J. Paul Getty Museum.

Last, we used our data to train a machine learning system and used this system to automatically label additional bounding boxes within the paintings.

Summarized in numbers, our dataset contains the following:

- 20k paintings
- 15 coarse-grained material categories created by human annotators

- 50+ fine-grained material categories created by human annotators
- 220k bounding boxes around those materials created by human annotators
- 100k bounding boxes around those materials created by an automated process
- 105k boxes assigned a fine-grained material category label created by human annotators

4.1.1. THE CREATION OF THE DATASET

In this section, we will provide a description of the design considerations and a discussion of the creation process of the dataset. In the previous chapters, we have not discussed the pilots nor earlier iterations of the data collection methods, as the discussion of pilots and the reflection thereof does not fit within scientific articles. Nevertheless, the lessons learned over these iterations have led to the final design of the data-collection and the dataset and could therefore be valuable.

The general idea of this dataset was to collect materials in paintings. Collecting paintings can be challenging from a technical perspective, as it requires in-depth knowledge of web-protocols and web-scraping tools. However, from a methodological perspective, it is quite simple, as the only relevant question is what paintings to include. The answer to this question came down to a related question: what paintings are available for download?

Only recently have art collections (museums, galleries, etc.) made their art digitally available. For reference, The Metropolitan Museum of Art (NY, USA) made their collection open-access in February 2017 while this PhD project started in January 2017. Since then, many more museums have digitized their collections and some of these institutions have also made their collection available. The push for open-access cultural heritage data continues to date; as of the time of writing (March 2021). Almost every month a new museum joins the open-access movement by releasing their data, with the most recent example being the Louvre.

At the start of the project however, only a few museums had made their collection available online. As such, at the start of the project we collected paintings primarily based on availability, with the only criteria being that the digital images of the paintings were of high-quality and contained the minimum meta-data, e.g., title, artist, and year. Most of the institutions did not have an option to download the entire dataset at once and therefore we needed to download each painting individually using web-scrapers. Because the online gallery (i.e., website) for each art institution is unique, we needed to write custom code for downloading the paintings from each art institution.

Once we had a large number of paintings available, we started with the collection of materials. In an early, pilot iteration of the material collection methodology, we simply had participants look at a painting and notate each material that they perceived. This process resulted in long lists, which were difficult to analyse and interpret. In addition to typos ('fabris', 'fbarics', and 'farbics'), for example, a problematic issue was the different levels of labels from participants. One participant might label a specific table simply as 'wood', while another might be more specific with 'oak'. It became clear we needed to provide participants with a limited set of options. Around the same time, we started experimenting with crowd-sourced data gathering on the Amazon Mechanical Turk platform, for the purpose

of performing simple and quick online experiments to attract hundreds or thousands of paid participants in a short time frame. We found that the best results for crowd-sourced data gathering are achieved when the task is as simple as possible. Combining this insight with the 'limited set of material options', we created the pilot for the first crowd-sourced material collection experiment. It simply showed paintings in succession and participants had to select from a limited menu which materials they perceived. In the next iteration, we further simplified this task to present participants with only a single "yes" or "no" question per painting, based on the annotational pipeline from OpenSurfaces (Bell et al., 2013). Some participant, for example, were repeatedly asked "Does this painting contain wood?", while other participants were asked the same question for fabrics, metal, etc. Initially, this process might sound like much more work for participants relative to selecting materials from a list; however, we found that having participants focus on a single material was actually much faster, as they don't need to mentally shift the task (i.e. "Is there wood? Is there wood? Is there wood?" vs "Is there wood? Is there stone? Is there fabric?"). Furthermore, with this approach, the task itself can be simplified. An example of the final task we used can be seen in Fig. 4.2. Furthermore, we had a minimal of 5 participants answer each question, e.g., 5 participants would be asked "Does this painting contain wood?" for the painting. This allowed us to average across participants responses and to minimize the effect of participant error. Only if the majority of participants replied that a painting depicted a material would that material be assigned with the corresponding material label.

Above we have reflected on the material labelling task and the design thereof. With the design of this task completed, we can ask the obvious follow-up question: what materials should be labelled using this task? A perusal of the literature across computer science, art history, and perception resulted in different lists of materials. We aimed to create a broad list that is beneficial across disciplines. However, as explained in the previous paragraph, we had at least 5 participants perform the material labelling task per painting, per material. As such, each additional material added to our list resulted in an additional number of tasks equal to the number of paintings multiplied by 5 participants. Therefore, we sought to create a list that was as succinct as possible, while containing all the 'stuff' depicted within paintings. The current list contains 15 materials. Most of these are prototypical materials including stone, wood, and fabric. However, the list also contains materials that are not intuitively considered as materials. For example, the food 'material' serves as an overarching collection of many food products, which we later specified with fine-grained labels, like food -> cheese. The same is true for the animal material. Furthermore, our material list includes sky and ground, two categories that are not prototypical material categories but which we chose to include as they typically cover a substantial portion of the surfaces of paintings. With the final list we aimed for a minimal set of material categories that could be used to annotate the majority of the surface for each painting.

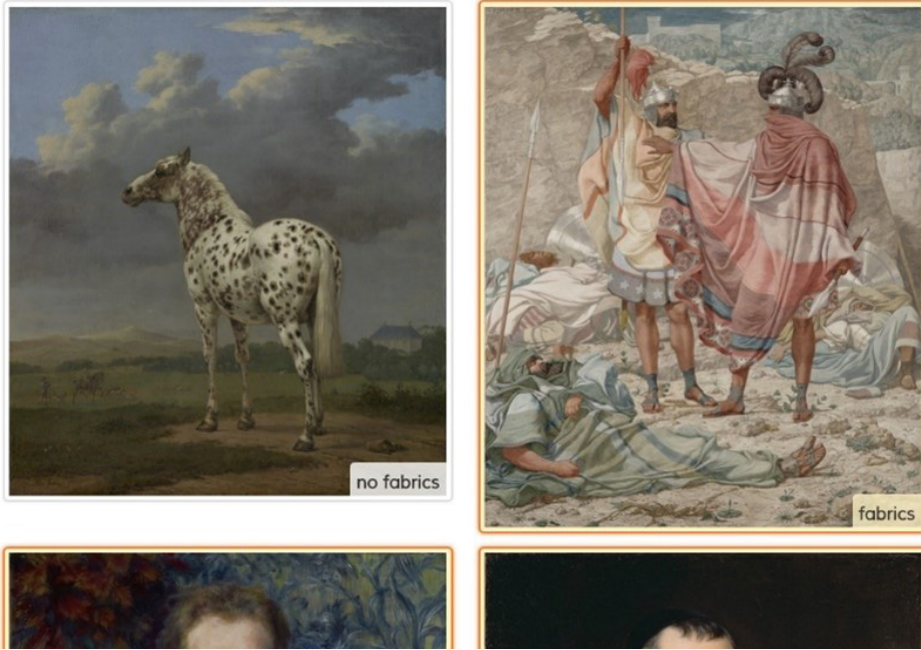
A subsequent major decision we had to make was related to the question: how to collect spatial information? At this point in the project, we had categorized materials within the paintings. However, there was not yet any information pertaining to the spatial location of these materials. Three main methods that we considered are presented in Fig. 4.3. The initial intention was to create polygonal segmentations as those are most informative. However, we found that in general, these segmentations are costly relative to the other two options presented in Fig. 4.3, in both time and money. Furthermore, materials in paintings

Instructions: Click on the images where you can see some fabrics

Instructions

Category: fabrics

If you cannot tell, don't select the photo. Please do not guess. Don't worry if you miss small things that are hard to see.



4

Figure 4.2: An example of the material labeling task. A participant was presented with around 30 paintings at a time. By clicking on the painting, the participant could indicate the material was present and inversely, not clicking meant the material was not present.

often have complex or fuzzy boundaries which can make tracing the boundaries very difficult. Therefore we chose to use the extreme click method in chapter 3 to create bounding boxes. In this method, participants always need to make exactly four clicks, which is much faster and easier compared to polygon segmentations which can be anywhere from three to hundreds of clicks. These four clicks are on the 'extreme' points of the material, i.e., the highest, lowest, most left, and most right pixels that belong to the material. This extreme click method makes it easier for participants to create bounding box relative to the 'typical' bounding box task (i.e., the left of Fig. 4.3) where participants need to place clicks on the "imaginary corners of a tight box around the" material (or object) of interest (Papadopoulos et al., 2017). In addition to being easier, faster, and less costly, extreme click bounding boxes also provide nearly identical information relative to polygon segmentations for many vision and art historical applications. However, they are regrettably much less informative

relative to polygon segmentations for most computer vision applications, but as we have shown in chapter 3, with computer algorithms it is possible to create a polygon segmentation from extreme clicks.

Initially, the extreme-click bounding box task was open to any participant on the AMT platform. Regretfully, while being much simpler relative to a polygon segmentation task, the extreme click bounding box tasks can still be quite difficult for participants. While the interface we designed makes creating the bounding box itself trivially simple, the difficulty lies in knowing what exactly should be contained within the bounding box. For example, segmenting 'skin', furthermore defined as 'human skin' might appear as relatively straightforward. However, considering the large collection of paintings and the variety of scenes depicted therein we can expect cases in which it is no longer straightforward. In distinct cases participants asked if they should consider Jesus, Buddha, or angels depicted within paintings as human and thus creating create bounding boxes around their 'skin'. In another case a participant indicated that a book depicted within a painting seems to bound in human skin and asked if this should be treated as such. Another example related to flayed skin laying at the foot of a martyr. For most other material similar examples exist. Furthermore, and more problematic than the task being difficult, many participants simply performed badly. For example, many participants created bounding boxes around materials we did not ask for in the task (e.g., creating bounding boxes for wood in a metal task). Other participants would simply create bounding boxes around figures or faces, independently of what material should be labelled within the task. We spent much effort re-writing and re-designing the instructions and tutorial, but eventually, we were forced to manually select workers that consistently created good bounding boxes. In hindsight, the best course of action likely would have been to set up an automatic system using sentinel items to select and restrict good and bad participants respectively.

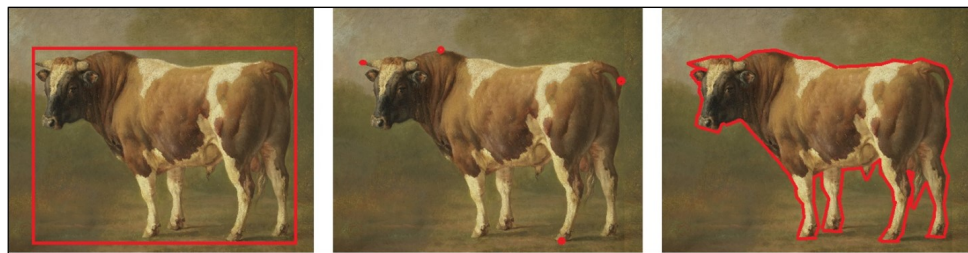


Figure 4.3: Three possible ways to create a polygonal segmentation around an object/material. From left to right: a bounding box, extreme clicks and a polygon segmentation. A polygonal segmentation per definition also contains the extreme clicks. Furthermore, from the extreme click coordinates the bounding box can easily be calculated. In chapter 2 we started with the collection of polygon segmentations, but moved on to extreme click segmentations in chapter 3.

4.2. THE WEBSITE

The MIP dataset can be viewed, browsed, and downloaded at materialsinpaintings.tudelft.nl. In this section we will provide a concise guide on how to navigate and use the website. As websites can be updated at any time, the visuals might change in future updates. As such,

screenshots provided here might not visually reflect the database website at a later point. However, the core functionality described here will likely remain the same.

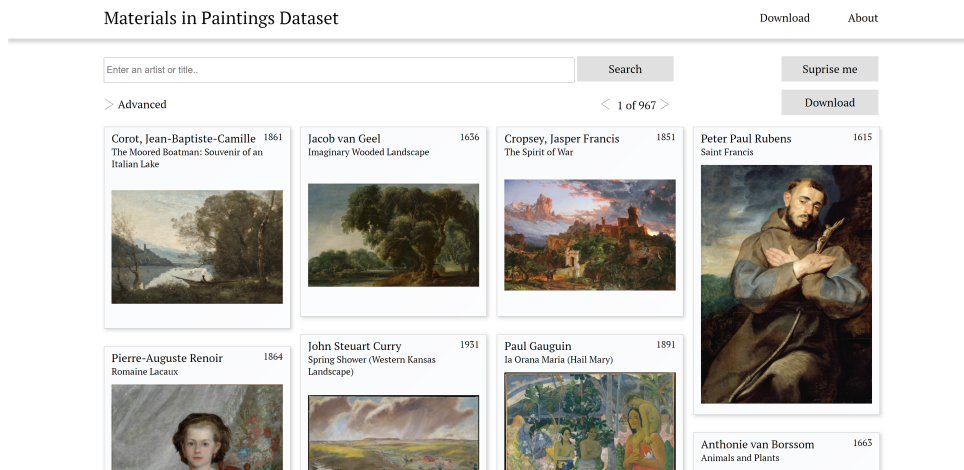


Figure 4.4: The Materials In Paintings homepage.

In Fig. 4.4, the homepage is shown. When users navigate to the website, they will be presented with this screen. The search bar at the top can be used for a basic search using string matching, for a certain artwork or artist. To execute a specific search the advanced search functionality can be used. Clicking on 'Advanced', the advanced search menu will become visible (see Fig. 4.5). With this menu open, a specific query can be created. At the top the option to query paintings or bounding boxes is presented. The available options in the menu will reflect the selected option. Note Fig. 4.6 for the painting and bounding box option on the left and right respectively. Specific materials can be selected for both data types. Each bounding box can have only one material which means that selecting multiple materials will query bounding boxes that show one of the selected materials (i.e., an OR filter). Paintings on the other hand can depict multiple materials and therefore there is an additional option when selecting materials for paintings. This option toggles searching for paintings that contain all the selected paintings (i.e., AND filter) or paintings that contain at least one of the selected materials (i.e., OR filter). Furthermore, for paintings there is also an option to exclude materials (i.e., a NOT filter). This can be used to query paintings that explicitly do not have a certain material (e.g., "show me all paintings that do not have wood"). For bounding boxes we can simply search for specific materials. In addition to this, both data types can also be filtered by specific artists, title or by the year of creation. Furthermore, for the bounding boxes, users can filter by creation type, i.e., created by human annotators or by an automated process. If the automatic boxes are selected, users can also filter by the score, which represents the automated network's confidence in the bounding box.

The advanced search functionality allows us to make a selection of paintings. By pressing the download button in the top-right corner (Fig. 4.5) we can download the metadata for this selection in a .csv file. Depending on the size of your selection, generating this file

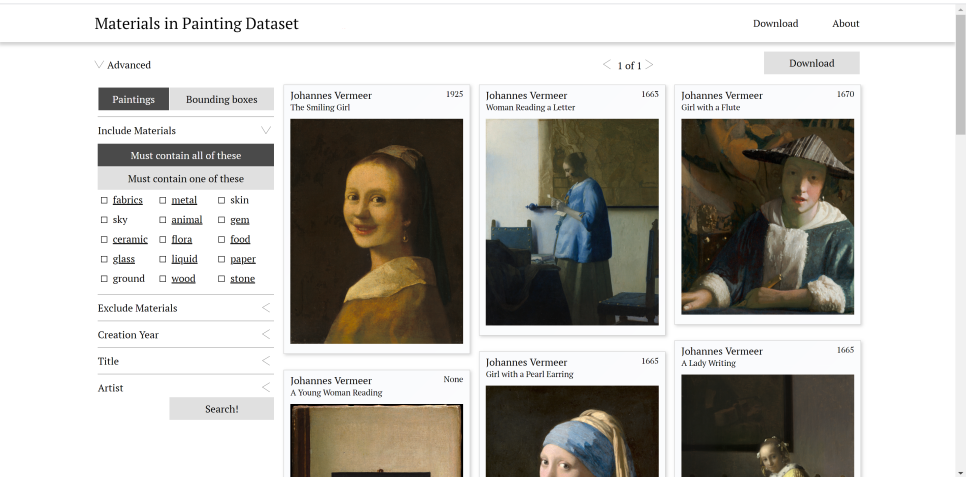


Figure 4.5: The advanced search menu. In this view we first used the search bar (no longer visible after opening the advanced menu, but visible in Fig. 4.4) to query for 'Vermeer'. The result of the previous query remains visible when opening the advanced search menu.

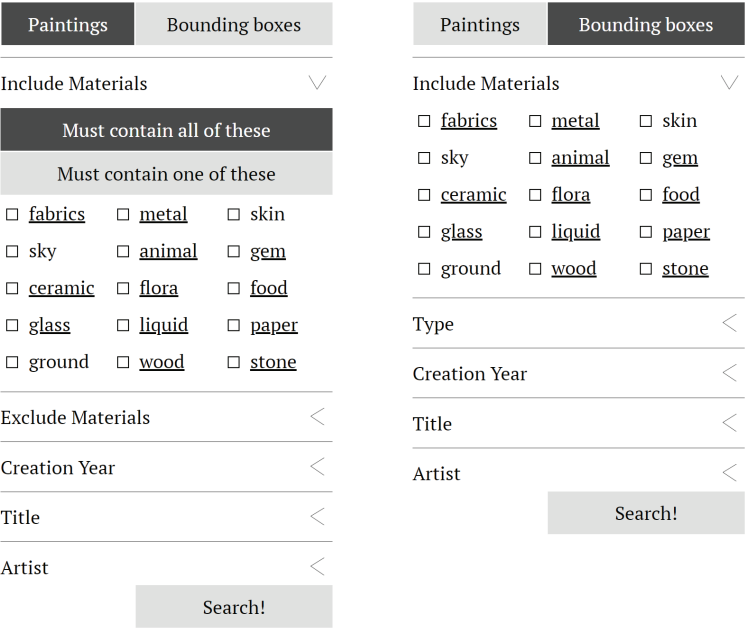


Figure 4.6: On the left is the advanced menu, with the options reflecting the choice of paintings. On the right the menu for bounding boxes is presented. The difference between these two are due to the different data types. For paintings the "Must contain all of these", "Must contain one of these" and the "Exclude materials" are essential AND, OR, and NOT filters respectively. For bounding boxes only an OR filter is available.

can take some time.

In the .csv files, each painting and each bounding box contains a unique ID, i.e., a specific painting will always have the same ID across .csv files and the same is true for each bounding box. This allows paintings and bounding boxes to be linked. In Table 4.2 a portion of a .csv file for a bounding box selection is reproduced. As can be seen, multiple bounding boxes are linked to the same painting, e.g., id "130220". Each .csv file contains much more information, including the title of the artwork, the artist, the year of production, the size of the image in pixels and the gallery which holds the original painting.

The next step is to download the actual images. There are two ways to download the images. Either, using the .csv files, each image can be downloaded individually. We have provided a python script to do so, which can also be found on materialsinpaintings.tudelft.nl. All boxes and paintings that can be downloaded in this way are at a maximum resolution of 1024 by 1024 pixels. This means that all boxes and paintings are resized to fit inside a box with a width and height of 1024 pixels, assuming the image did not already fit.

For large selections it is advisable to download all the images as a single .zip file. Also, if a resolution higher than 1024 by 1024 pixels is desired, a .zip files that contains all the images at the original (i.e., max) resolution is available. These .zip files are hosted on a dedicated server that is better equipped to handle large files. The availability of two data types (e.g., paintings vs bounding boxes), which are both available in two resolutions, compounded by bounding boxes which are available with and without padding, results in a possibly confusing number of files that can be downloaded. To help users find the desired .zip files we have provided a flow-chart on the 'Download' page on materialsinpaintings.tudelft.nl.

Table 4.2: A portion of the .csv file for a selection of bounding boxes by querying for 'Vermeer' in the artist field. The original field contains many more columns, such as the width and height of the box, the gallery from which it originates and the URL required to download the image.

| box id | painting id | main material |
|--------|-------------|---------------|
| 381278 | 129483 | fabrics |
| 351401 | 130220 | flora |
| 351400 | 130220 | flora |
| 351399 | 130220 | liquid |

BIBLIOGRAPHY

- Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2013). OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics (TOG)*, 32.
- Papadopoulos, D. P., Uijlings, J. R. R., Keller, F., & Ferrari, V. (2017). Extreme clicking for efficient object annotation. *International Journal of Computer Vision*. <https://doi.org/10.1109/ICCV.2017.528>

5

SOFT LIKE VELVET AND SHINY LIKE SATIN

Dutch 17th century painters were masters in depicting materials and their properties in a convincing way. Here, we studied the perception of the material signatures and key image features of different depicted fabrics, like satin and velvet. We also tested whether the perception of fabrics depicted in paintings related to local or global cues, by cropping the stimuli. In Experiment 1, roughness, warmth, softness, heaviness, hairiness, and shininess were rated for the stimuli shown either full figure or cropped. In the full figure, all attributes except shininess were rated higher for velvet, while shininess was rated higher for satin. This distinction was less clear in the cropped condition, and some properties were perceived significantly different between the two conditions. In Experiment 2 we tested whether this difference was due to the choice of the cropped area. Based on the results of Experiment 1, shininess and softness were rated for multiple crops from each fabric. Most crops from the same fabric differed significantly in shininess, but not in softness perception. Perceived shininess correlated positively with the mean luminance of the crops and the highlights' coverage. Experiment 1 showed that painted velvet and satin triggered distinct perceptions, indicative of robust material signatures of the two fabrics. The results of Experiment 2 suggest that the presence of local image cues affects the perception of optical properties like shininess, but not mechanical properties like softness.

Published as: Mitchell J. P. van Zuijlen*, Francesca Di Cicco*, Maarten W. A. Wijnjtes, Sylvia C. Pont; Soft like velvet and shiny like satin: Perceptual material signatures of fabrics depicted in 17th century paintings. *Journal of vision*, 21(5), 10-10 * Authors contributed equally



Figure 5.1: Satin (left) is visually more similar to aluminum foil (middle) than to velvet (right). However, satin certainly belongs to a different material class than aluminum. The first two images were downloaded from Morguefile.com and the third image from pxfuel.com, released under free license.

5

5.1. INTRODUCTION

Fabrics serve a wide array of functions in our daily life. We use fabrics to hold and carry things, to clean and dry surfaces, for decoration, and for clothing. With this wide array of functions, the material category of ‘fabric’ also comes with a wide variety of appearances. The visual appearance of fabrics depends on the type of fiber (e.g., natural or synthetic), the yarn (the continuous segment of fibers), and the weaving method (Koenderink and Pont, 2003; Pont and Koenderink, 2003; Zhao et al., 2011). Materials’ appearances are strongly dependent on light (Fleming et al., 2003; Pont and Te Pas, 2006) and shape (Ho et al., 2008; Marlow and Anderson, 2015; Schmidt et al., 2020). This is true also for the appearance of fabrics, which has been shown to depend on the illumination environment (Barati et al., 2015; Zhang et al., 2019) and on the folding shape (Xiao et al., 2016). Nonetheless, we can visually discriminate and identify different types of fabrics on the basis of their characteristic visual qualities, also known as “material signatures” (Fleming et al., 2013).

In this paper we focus on the appearance of velvet and satin. Velvet and satin both belong to the material category of fabric, but large differences exist within this same material class. Upon visual observation, one could find more similarities between the appearance of satin and aluminum foil, than between satin and velvet (Fig. 5.1). However, despite the visual similarity, nobody would classify aluminum as a fabric.

In this study, we studied the perception of painted fabrics in 17th century Dutch paintings, a class of paintings unanimously acknowledged for the convincing representation of materials and their properties. The economical yet effective rendering of material properties exploited by 17th century painters (Parraman, 2014) resonates with the mechanisms of the human visual system (Adelson, 2001; Casati and Cavanagh, 2019; Cavanagh, 2005; Di Cicco et al., 2019; Koenderink and van Doorn, 2001; Marlow et al., 2017; Sayim and Cavanagh, 2011; van Zuijlen et al., 2020; Wijntjes et al., 2012; Wijntjes et al., 2020). Painters carefully chose the image features to include and could choose to omit perceptually irrelevant or hindering features, as it was shown to be the case for the orientation of the highlights on grapes which do not need to be congruent with the object shape in order to communicate a glossy appearance (Di Cicco et al., 2019). Materials were often painted according to

standard, well-established instructions which assured the painter of getting the best possible rendering. Velvet, for example, could be convincingly depicted by simply inverting the typical patterns of light and shade (Lu et al., 1998; van Duijn and Roeders, 2012). Written records of such visual tricks can be found in “Het Schilder-boek”, a book describing the life and work of several painters, composed by Dutch painter and art historian Karel van Mander in 1604. He wrote: “In contrast to your other textile, where you render with light paint all the relief in the folds, this is completely different with velvet [drapery], as you make these entirely dark and paint flat highlights only on the reflecting side” (van Eikema Hommes et al., 2002; van Mander, 1604). Another relevant art historical source is “The big world painted small” by Willem Beurs (Beurs, 1692; Lehmann and Stumpel, *In press*). This book has already proven to be a useful tool to help understand pictorial procedures and the relevant image features for the rendering of materials (Di Cicco et al., 2020). In this collection of pictorial recipes, Beurs described how to paint satin and velvet emphasizing the different rendering of specular reflections, sharp and high contrast for satin and somewhat blurrier and with less contrast for velvet (Pottasch, 2020). These are examples of the value of investigating paintings and art historical writings for the sake of understanding the functioning of the human visual system.

Understanding the material attributes that form the signatures of the representation of different fabrics, like velvet and satin, is important for several applications. One example is online shopping, in which visual communication of the material qualities of fabrics is crucial to guide the consumers’ choice. The appearance in the image should match as closely as possible the appearance that would be perceived in a real shop. Failing to capture and convey the material attributes of the fabric is one of the major concerns of online retailing (Tuunainen and Rossi, 2002). On this topic, it has been shown that dynamic stimuli (videos) can better communicate the haptic properties of fabrics compared to static stimuli (images), because of the greater availability of information (Bouman et al., 2013; Wijntjes et al., 2019). Xiao et al. (2016) found that when observers can only rely on images to infer the material properties of fabrics, color and folding information interact to enhance the accuracy with which tactile properties are estimated. In the absence of folds, i.e., if the fabric is shown flat, chromatic information was found not to be discriminative enough. In perception-based computer graphics, it has been shown that the optical appearance of different fabrics contributes to the realism of the rendering more than their dynamics (Aliaga et al., 2015). The digital rendering of fabrics is gaining importance in the entertainment industry for movies and games (Zhao et al., 2016), and in online shopping with the option to virtually try-on clothes (Pons-Moll et al., 2017).

Velvet and satin have different mechanical and optical properties which give rise to their distinctive appearances. The appearance of velvet is due to asperity scattering, where light is scattered by the hairy layer on the surface, leading to a brightening of the contours (Koenderink and Pont, 2003; Pont and Koenderink, 2003). The reflectance properties of satin, which lead to its shiny appearance, depend on its constructional parameters (e.g., the yarn density and the weave pattern) (Akgun et al., 2014). In particular the weave pattern of satin is based on “floating” yarns, yarns that are weaved vertically over a horizontal weft. These floating yarns reflect the light from the fabric creating specular or split-specular reflections causing the shiny appearance (Barati et al., 2015). The specular peaks for satin are located at the regions of highest curvature, and under generic lighting conditions are

pointed towards the light source. For velvet however, the brightest regions are typically placed along its occluding contours, under generic lighting condition (Barati et al., 2015). The position of highlights, being related to the 3D shape of the object, also reveals the folding configuration of the fabric. This folding configuration is informative when estimating the optical and mechanical properties of a piece of fabric when presented with visual information only (Xiao et al., 2016).

Some physical properties of an object or material, such as softness or warmth, are not directly apparent by the optical cues present in the image. To infer these properties, the human visual system can either employ a bottom-up or a top-down approach. The first relies on the profile of image features that triggers material perception. The second approach would first require recognizing the object and the material class it belongs to, and then inferring the material attributes via prior knowledge and learned associations. However, it is not always necessary to identify the object in order to infer the material attributes. Schmidt (2019) and Schmidt et al. (2017) showed that identifying the material class already provides enough cues to derive material attributes via an “associative approach”. They conducted a rating experiment of several material attributes using unfamiliar shapes rendered with materials with different optical properties (e.g., marble, steel, velvet, etc.). They found that softness estimation relied on recognizing the different materials via the associative approach (e.g., it is velvet therefore it is soft). These two approaches, i.e., bottom-up and top-down typically, but not necessarily exclusively, use local and global visual information respectively. This then raises the question whether material perception relies on global or local visual information, or a combination of both. According to Schwartz and Nishino (2019), material attributes are inherently local, which is why a classifier trained on human similarity judgements could recognize these attributes from small image patches, like image crops. Marlow and Anderson (2013) proposed that human perception of glossiness depends on local image features of the highlights, such as coverage, contrast, and sharpness, but these features are in turn dependent on the global information of the shape and the illumination environment. Balas et al. (2020) proposed that the use of global or large scale visual information for material perception is developed with age, as they found that children’s performance in distinguishing between real and fake food was impaired when local information was disrupted but that this impairment was reduced or even absent when global information was disrupted. Schmidt et al. (2020) showed that local shape features affect the visual perception of softness and weight of unfamiliar, static objects. It is evident from the literature that the understanding of the visual systems’ use of local vs global visual information is still an open problem, therefore in this paper we tested and compared material perception providing either global or local information.

Another field in which it is relevant to distinguish and identify different fabrics is art history, as every element within paintings usually carries meaning. For example, in some drawings made around 1490 by a German artist known as the Master of the Coburg Roundels, the lively “fluttering loincloth” of the crucified Christ may signify his imminent resurrection (Lehmann, 2015). Another example is the dress of Eleanor of Toledo, painted by Bronzino in 1546, which symbolized the wealth and power of Florence and de Medici family in the 16th century (Thomas, 1994). According to Thomas (1994), “in order to understand the origin and purpose of the dress, we must first know the nature of the fabric” and he wondered whether the fabric was velvet or satin. The original hypothesis that the fabric was brocaded

satin was later confirmed when the tombs in de Medici's mausoleum were opened, as the dress was the burial gown of the Eleanor of Toledo (Thomas, 1994). However, it should be noted that art historical examples where the depictions of an object or material can actually be compared with the original object or material, are for obvious reasons extremely rare.

The first aim of this paper was to determine the perceptual material signatures of velvet and satin depicted in 17th century paintings. In Experiment 1, we further explored whether cropping the fabric out of its global form and providing only local information, caused a change in perception of its material properties. This indeed happened. In Experiment 2, we investigated whether the observed changes in material perception when judging a cropped image could be related to the choice of the cropped area, due to the presence or absence of triggering image cues. Finally, to explore which cues observers relied on to make their judgments, we correlated the perceived material properties in the different crops with image features of the highlights.

5.2. EXPERIMENT 1

5.2.1. METHODS

In Experiment 1 six material attributes (roughness, shininess, softness, weight, warmth, and hairiness) were rated for a set of paintings of fabrics, depicting either velvet or satin, to measure the extent of association of each attribute with the two types of fabric. The stimuli were presented in two viewing conditions, either with context where the full figure was presented or without context/object shape information, where crops of the fabric were presented. The different viewing conditions were aimed to test whether showing a fabric embedded in a recognizable object, such as a dress or a tablecloth, rather than in an anonymous form without context, would affect the perception of the material attributes.

STIMULI

We selected 19 fabrics from 17 high-resolution digital images of 17th century oil paintings. Two paintings depicted both velvet and satin and were therefore used twice. All paintings reproduced within this paper are available under open access at a CC0 or CC BY 4.0 license. The full list of all paintings used within this study, including those reproduced in this paper, can be found in in Fig. 5.16 through Fig. 5.34 in the supplementary materials.

The fabrics were categorized as either velvet ($n = 8$) or satin ($n = 11$) by the experimenters. The categorization was based on the expertise of all the authors in vision science and optics. We further supported this categorization with art historical sources identifying the fabrics of some of the paintings in our set of stimuli, as either satin or velvet (Gordenker, 1999; Liedtke, 2011; Pottasch, 13 Jun. 2020). In one viewing condition, the entire figure or object, including the background, was shown with a red arrow indicating the target fabric to rate (see the left image in Fig. 5.2). In the other viewing condition, each target fabric was cropped to a 600x600 pixels patch and presented on the screen at the same visual size as in the full figure condition, against a grey background (see the right image in Fig. 5.2). The cropped areas were chosen to be as informative as possible about the folding shapes. Throughout the rest of the paper, we will refer to the two viewing conditions as full figure condition and crop condition, respectively. See Figure S1 in the supplementary material for all the stimuli in both viewing conditions.



Figure 5.2: An example of each of the two conditions, within the interface. Left the full figure condition, in which the figure or object with the target fabrics is fully visible. On the right, the crop condition, where only a patch from the target fabric is visible, which is intended to deprive the visual system from context and shape information. Note that a participant would see only the left or right screen, never both.

OBSERVERS

Each participant rated all the stimuli in one viewing condition and for one material attribute. We collected data from 10 participants for each combination of the two viewing conditions and six attributes, for a total of 120 participants. Data were collected through the Amazon Mechanical Turk (AMT) platform. While AMT provides some benefits over conventional lab-settings, it is known to possibly result in noisy data as a result of a small, but considerable portion of participants that appear to perform badly in experiments. Based on previous experience with the AMT platform (van Zuijlen et al., 2020), we set an exclusion criterion to automatically remove data from participants whose median trial time was below 1 second (i.e., responding too fast). For each participant removed this way, we collected one more participant until we reached the targeted 10 participants per viewing condition/attribute combination. In total, 48 participants were removed this way, which in hindsight signals that this exclusion criterion might have been too strict. Participants were excluded in this way before any data analysis was performed. All participants were naïve to the purpose of the experiment. They agreed with the informed consent prior to the experiment. The experiments were conducted in agreement with the Declaration of Helsinki and approved by the Human Research Ethics Committee of the Delft University of Technology.

PROCEDURE

Experiment 1 consisted of a between-subjects design, with two viewing conditions and six perceptual attributes, namely roughness, shininess, softness, weight, warmth, and hairi-

ness. Before starting the experiment, participants received written instructions explaining the task. They were informed that they would be shown images of fabrics but not which type of fabric. Prior to the actual experiment, participants performed 15 practice trials, not only to become familiar with the interface but also to get an idea of the range of stimuli. Participants were randomly assigned to one of the viewing conditions and they were asked to rate one of the attributes. Each attribute was rated using a slider on a continuum ranging from 0 to 100: smooth vs. rough, matte vs. shiny, hard vs. soft, cold vs. warm, hairless vs. hairy, and light vs heavy. In both viewing conditions, each of the 19 stimuli was rated three times for a total of 57 trials. The trials were randomized across participants.

5.2.2. RESULTS

CONSISTENCY BETWEEN AND WITHIN OBSERVERS

In Experiment 1 each attribute was rated three times. The consistency within observers is visualized in Fig. 5.3 (left) and was calculated as the average pairwise (Pearson) correlation between the ratings over the three repetitions per observer, again averaged across observers. Next, we took the median across the three repetitions to smooth out the effects of potential outliers. Then, we normalized the data for each participant between 0 and 1 to rule out possible effects of unequal interval judgments. We used this median, normalized data for the remainder of the result section. For the consistency between participants, we calculated the intraclass correlation coefficient (ICC) using an average rating, consistency, two-way random effects model for each attribute and each condition (Koo and Li, 2016; McGraw and Wong, 1996). The ICC values and the 95% confidence intervals have been visualized in Fig. 5.3 (right). A full report of the ICC statistics can be found in Table S1. In Fig. 5.3 there is a clear trend of higher inter- and intra-rater agreement in the full figure condition compared to the crop condition, with the exception of roughness in the inter-rater agreement (Fig. 5.3, right). For the ratings of roughness, some participants in the crop condition may have attended to the visible roughness of the brushstrokes instead of judging the fabric. Furthermore, the ICC calculations show that the consistency between participants is significantly different from zero, thus above chance, for all attributes and in both viewing conditions, with the only exception of hairiness in the crop condition. However, the intra rater agreement on hairiness was high and significant in both viewing conditions.

MATERIAL SIGNATURES

We ran a two-way MANOVA to examine the effect of the viewing condition and the fabrics' material on the perception of the material attributes. We found a main effect for both viewing condition (i.e., full figure vs crop) at $F(6, 29) = 2.78$, $p < .05$, and material (i.e., velvet vs satin) at $F(6, 29) = 23$, $p < .001$. We also found an interaction effect between the two factors at $F(6,29) = 6.56$, $p < .001$. In Fig. 5.4, we visualized the average judgments of the material attributes, split by viewing condition (top) and material (bottom) and indicate significant differences (Bonferroni corrected) between the conditions. The perception of warmth and hairiness of satin, and hairiness and softness of velvet changed significantly between the two viewing conditions. For the full figure condition, velvet was judged to be significantly warmer, hairier, softer, heavier, and rougher, while satin was perceived to be shinier. For the crop condition, velvet was significantly warmer, hairier, and heavier, whereas satin was rated significantly shinier. There were no significant differences between

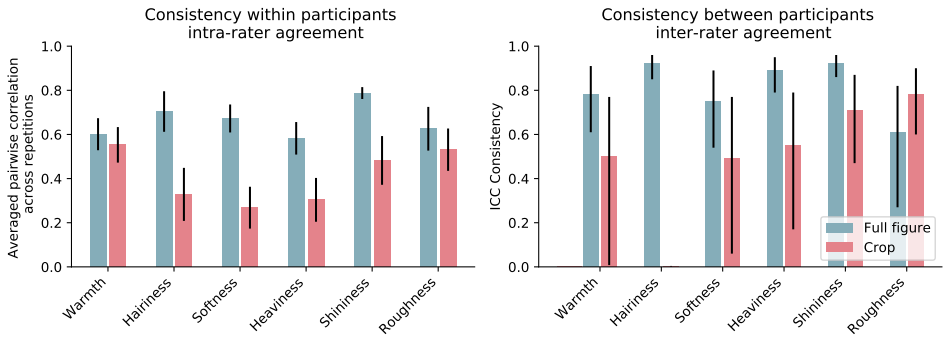


Figure 5.3: Consistency within and between participants. Left) The consistency within participants is calculated as the averaged pairwise correlation between each participants repetitions of the stimuli, and the error bars indicate the standard error. Right) The consistency between participants was calculated using intraclass correlations, and the error bars indicate the 95% confidence interval. The full report of the ICC analysis can be found in Table S1. Note that non-significant ICC are not visualized (i.e., hairiness in the crop condition).

5

satin and velvet, for the attributes of softness and roughness, in the crop condition.

To check if the material attributes were independent of each other or belonged to an underlying subset of dimensions, we computed a correlation matrix for both viewing conditions, visualized in Fig. 5.5. The correlation coefficients are reported in the cells of the matrices. Significant correlations at $p < .05$ are marked with an asterisk (*).

In the full figure condition, shininess was the only attribute that showed a negative significant correlation with each other attribute. All other attributes showed mutual positive, significant correlations except for roughness, which only correlated (negatively) with shininess.

In the crop condition, fewer correlations were found across all attributes. Roughness was again negatively and significantly correlated with shininess, as well as with softness and positively correlated with heaviness. Shininess was no longer correlated with hairiness, nor softness. Overall, this shows that the material attributes are not completely independent of each other, which implies they might be captured by a smaller set of dimensions.

PRINCIPAL COMPONENT ANALYSIS AND PROCRUSTES ANALYSIS

To visualize whether the two materials, velvet and satin, were perceived as having different material properties, we ran a PCA for both viewing conditions. Fig. 5.6 and Fig. 5.7 show the PCA biplots of the full figure and the crop conditions, respectively. These biplots indicate how the stimuli are related to the attributes. The stimuli were clustered using 95% confidence covariance ellipses, according to the depicted material, satin (light blue ellipse), or velvet (yellow ellipse).

To further compare the effect of cropping on the material properties perception, we performed Procrustes analysis. The PCA of the crop condition shown in Fig. 5.7 was matched to the PCA of the full figure condition (Fig. 5.6).

In the full figure condition, the first two principal components account for 84.2% of the variance. The factor loadings listed in Table 5.1, show that the first principal component

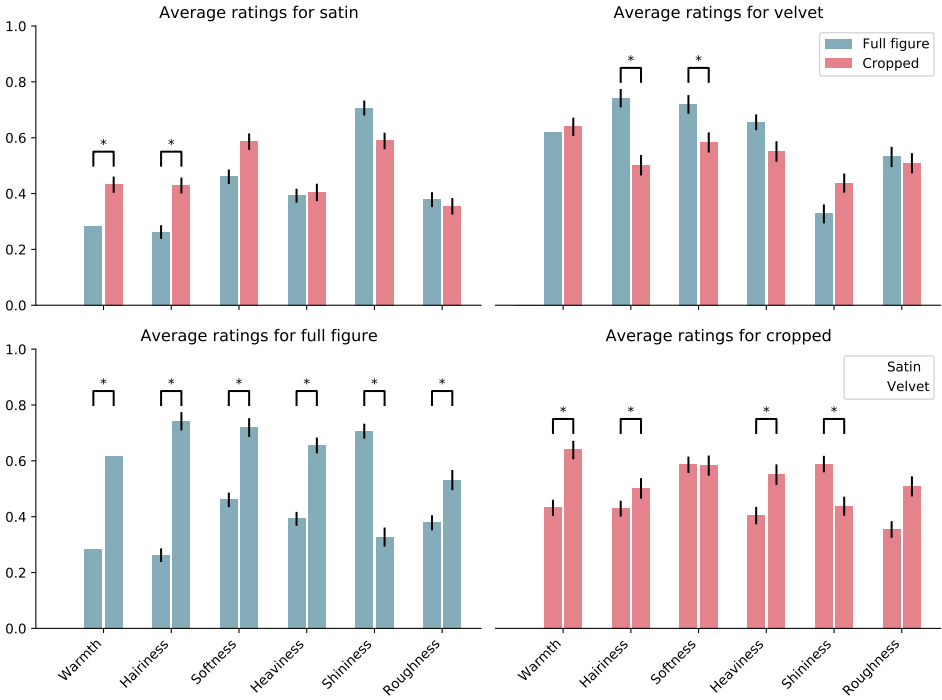


Figure 5.4: The perceptual judgments of satin and velvet, for both conditions. In the top plots, the data is split by viewing condition, while in the bottom plots data is divided by material. For each participants, we took the median rating across the stimuli repetitions, and then averaged across these values. Significance between condition (top) and material (bottom) is indicated at $p < .05$, Bonferroni corrected. Note that besides the significance, the top and bottom display the same data, only differently presented to make interpretations across conditions easier, and to avoid visual clutter of displaying all significant differences within a single plot.

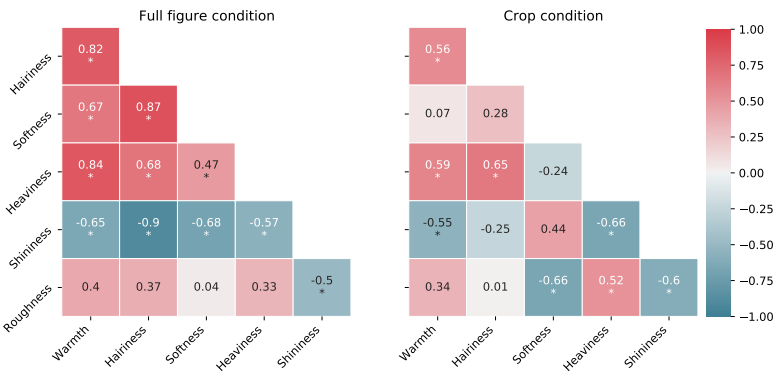


Figure 5.5: Correlation matrices of the attributes for both conditions. Color indicates the magnitude of the correlation coefficient. Asterisk (*) indicates a significant effect at $p < .05$.

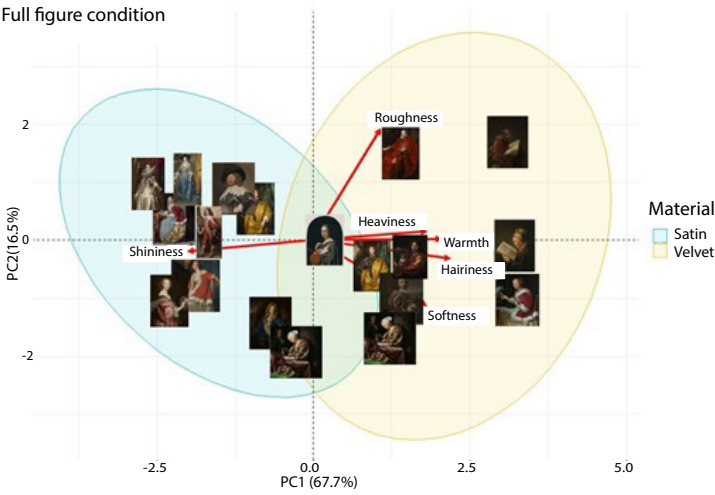


Figure 5.6: PCA biplot for the full figure condition. The materials are clustered within 95% confidence ellipses.

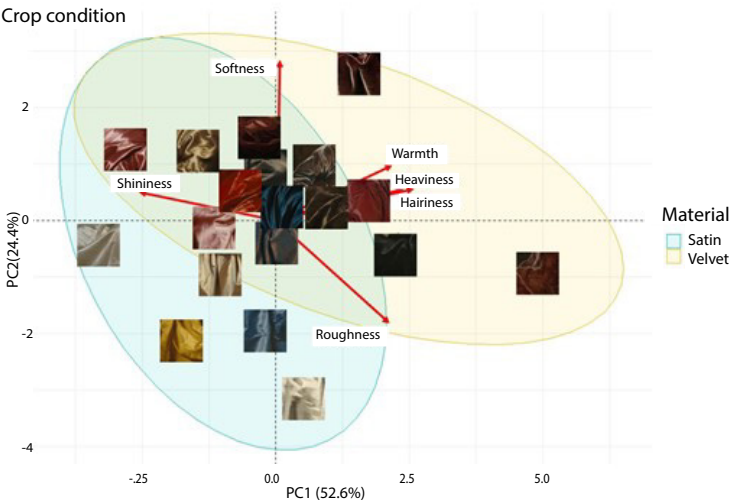


Figure 5.7: PCA biplot for the crop condition. The materials are clustered within 95% confidence ellipses.

is positively loaded by a cluster of attributes including hairiness, warmth, and heaviness. In the negative direction, shininess loads on the first component. The second principal component is mostly loaded by roughness.

Table 5.1: The factor loadings for the first two principle components of two PCAs, one for each condition.

| | PC1 | PC2 | PC1 | PC2 |
|------------------|--------------------|--------------------|-------------|-------------|
| | Full figure | Full figure | Crop | Crop |
| Warmth | 0.45 | 0.01 | 0.41 | 0.27 |
| Hairiness | 0.48 | -0.14 | 0.49 | 0.16 |
| Softness | 0.39 | -0.50 | 0.015 | 0.78 |
| Heaviness | 0.40 | 0.06 | 0.45 | 0.15 |
| Shininess | -0.44 | -0.1 | -0.48 | 0.14 |
| Roughness | 0.24 | 0.85 | 0.40 | -0.51 |

In the crop condition, the first two principal components explain 77% of the variance. The first component is mostly loaded in the positive direction by hairiness, heaviness, and warmth and by shininess in the negative direction. The second component is mostly loaded positively by softness and negatively by roughness.

A permutational test to check the significance of the Procrustes result ($r = 0.72$, $p < .001$), indicated that the overall distribution of the stimuli was similar between the PCA of the full figure condition and of the crop condition. However, the distribution of the stimuli in the PCA biplot (Fig. 5.7) shows much more overlap of the velvet and satin clusters, compared to the PCA of the full figure condition (Fig. 5.6). In addition, some stimuli clearly changed location between the two PCA spaces, indicating that their perception differed in the two viewing conditions. One example is shown in Fig. 5.8. The mean ratings of all the attributes for this fabric, averaged over the median rating of each participant, are shown in Fig. 5.8 for the two viewing conditions. The asterisk indicates that hairiness and shininess were perceived to be significantly different at $p < .05$ between the two viewing conditions.

5.2.3. INTERMEDIATE CONCLUSIONS AND DISCUSSION

We conclude that, within the attributes that we tested, the material signature of depicted velvet included warmth, heaviness, hairiness, and softness, and the signature of depicted satin included shininess. We further conclude that depriving the visual system of context and shape information significantly changed the perception of fabrics as depicted in 17th century paintings. Specifically, when depriving the visual system of shape and object information, the perception of material attributes can drastically change, as exemplified in Fig. 5.8. Moreover, the percepts became less consistent and more subjective as observed from the decrease in both inter- and intra-rater agreement. Furthermore, differences between materials expressed as the distributions of perceived material attributes became less distinct.

The cropped areas shown in the crop condition were chosen according to the amount of folding, in an attempt to maximize the amount of visual information. Considering that, we wondered to what extent the differences in perception found between the crop and full figure conditions were affected by this choice. One could argue that the depicted dress or robe from which crops are taken presents a certain shininess, roughness, etc., and thus different

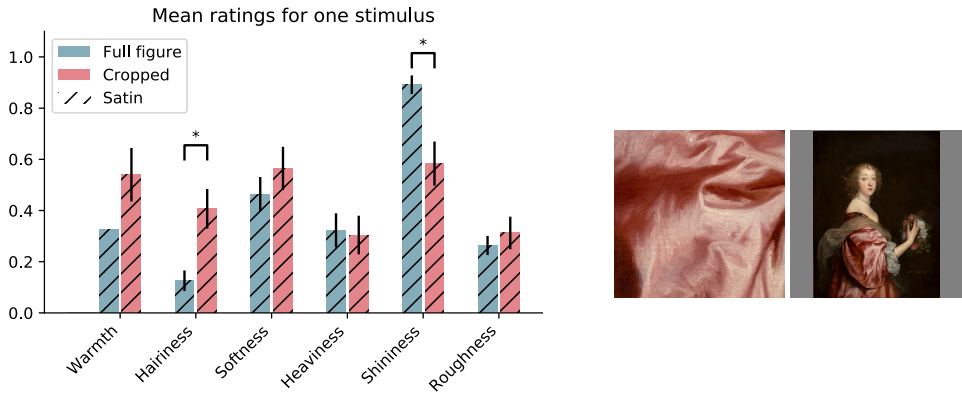


Figure 5.8: The mean ratings on the y-axis of the attributes for one specific stimulus. An asterisk (*) indicates a significant difference at $p < .05$. Right: the crop and full figure stimuli represented in the left bar chart. Error bars indicate the standard error. Painting: Catherine Howard, Lady d'Aubigny, by Anthony van Dyck. 1638, National Gallery of Art

5

crops from it would present these properties quite consistently without qualitative changes in perception between crops. However, on the other hand, local variations in shape (drapery) and effective lighting can cause major appearance variations, thereby causing differences in perception between the full figure condition, where participants could attend to all image features anywhere on the clothing, and the selected crop condition. For instance, a crop that coincidentally captures many highlights might be perceived to be shinier relative to a crop with few or no highlights, and, vice versa, it might also be possible that key image features were absent in our crops.

We follow-up on this question in Experiment 2 where we tested if the perception of crops changed depends on the choice of cropped area.

5.3. EXPERIMENT 2

5.3.1. METHODS

In Experiment 2, we investigated the extent to which perception of material attributes varies depending on the content of the crop and the presence or absence of local image features. We tested this with two material attributes that we also used in the previous experiment. The experiment consisted of a rating task of the two material attributes, followed by image analysis of the crops to extract highlights' features that could relate to the variations in perception between crops of the same fabric.

STIMULI

We used the 19 fabric stimuli from the full figure condition in Experiment 1 to make the stimuli in Experiment 2. From each image, we extracted a set of 9 to 21 equally sized crops, which covered the whole fabric (see Fig. 5.9 for an example). Thus, we made 19 sets of crops (velvet $n = 8$ and satin $n = 11$). To keep the visual size of the folds in the crops as

consistent as possible across different sets, the images were cropped with a constant ratio between the width of the whole fabric in the original image and the width of the crops. Images of all the crops can be found in the supplementary materials.

OBSERVERS

Identical to Experiment 1, data were collected on the AMT platform. Each of the 19 sets of crops was judged by a group of 5 participants for either shininess or softness. That is, participants would rate one set of crops for one material attribute. A total of 190 AMT users participated in the second experiment. All participants were naïve to the purpose of the experiment, and none had participated in the first experiment. Each participant agreed with the informed consent prior to performing the experiment. The experiments were conducted in agreement with the Declaration of Helsinki and approved by the Human Research Ethics Committee of the Delft University of Technology.

MATERIAL ATTRIBUTES

We used two material attributes in this experiment, both of which were also measured in Experiment 1. The first attribute was shininess, and the second was softness. Softness was found to be not correlated (see Fig. 5.5) with shininess and it can be seen to be nearly perpendicular to shininess in the crop condition PCA (Fig. 5.7). We interpreted this to mean that the majority of variability captured by softness is not explained by shininess, and vice versa, and that these two represented two main underlying dimensions of a perceptual material attribute space. Roughness was found to not be significantly different between velvet and satin and thus is unlikely to represent an underlying feature in this material space. The three remaining attributes used in Experiment 1 (warmth, hairiness, and heaviness) all inter-correlate and likely compose one underlying dimension. Therefore, with choosing shininess and softness we hope to capture the majority of the variation and underlying dimensions of the material feature space for fabrics with the least amount of attributes.

PROCEDURE RATING EXPERIMENT

In Experiment 2, participants were asked to rate one material attribute for each crop in one set of crops, taken from one of the 19 fabrics used in Experiment 1. After having read the instructions and having agreed to the informed consent, participants were asked to perform a size calibration, by adjusting a digital image of a credit card until it matches a physical payment card in the possession of the participants. Since all payment cards adhere to the standard size set forth by the International Organization for Standardization's 7810 ID-1 format (ISO/IEC 7810-ID-1), this allows us to rescale all images, so that each stimulus was presented at the same size, across different display settings for different participants. After the size calibration, participants performed a 10 second free-viewing task of the crops to get an idea of the range of the stimuli. Next, participants performed five practice trials followed by the actual experiment. For each trial, participant were tasked with rating shininess or softness with a slider on a continuum ranging from 0 to 100, corresponding to matte to shiny and hard to soft, as in Experiment 1. Each crop was rated three times, for a total number of trials ranging from 27 to 63 depending on the number of crops. The trials were randomized across participants.

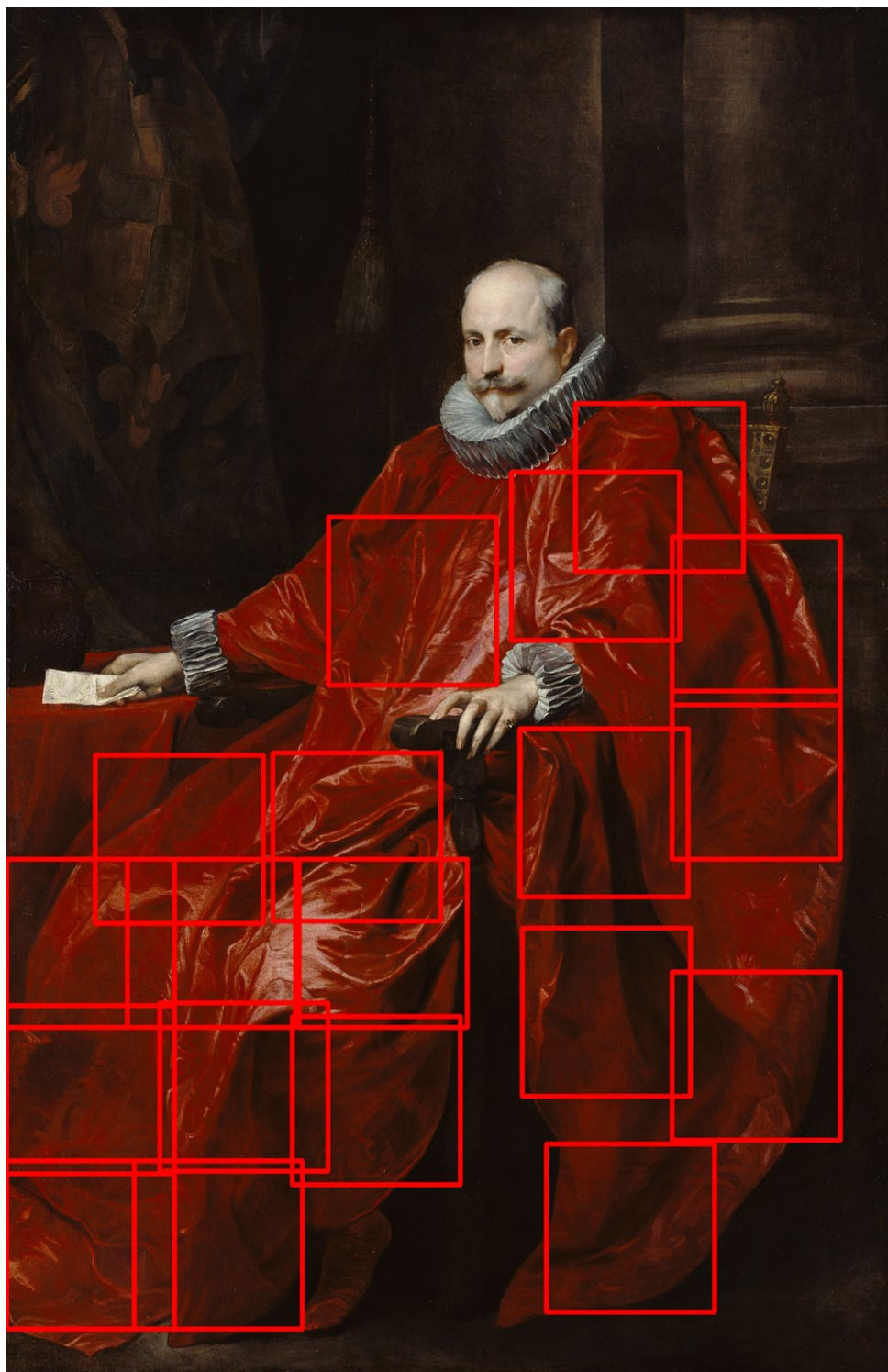


Figure 5.9: The original full figure stimuli, with red boxes that indicate the crops made for this stimulus. Each of the 19 stimuli from experiment 1 was subdivided into a set of crops as shown here. These sets of crops were used as stimuli in experiment two. Each crop within a set was the same size.

PROCEDURE IMAGE ANALYSIS OF HIGHLIGHTS

One way painters distinguished the depiction of velvet from satin, is through the rendering of the key image features of their reflectance properties (Gombrich, 1976). We hypothesized that, when judging the material properties of such depicted fabrics, humans attend to similar image features as perceptual cues.

Via photometric measurements of fabric samples, (Barati et al., 2015) assigned satin to a reflectance category combining specular and split-specular scattering, and velvet to the category of asperity scattering materials. From a perceptual rating experiment, (Barati et al., 2015) also found that the samples belonging to the asperity scattering category were perceived to be the softest, whereas the samples in the specular and split-specular scattering class were perceived to be the shiniest and the least soft. These findings support our hypothesis that softness and shininess are key attributes of velvet and satin, respectively.

The different scattering behaviors of velvet and satin result in distinctive optical cues. Previous studies have shown that image features of the highlights, such as coverage, contrast, and sharpness, can influence the perception of glossiness (Di Cicco et al., 2019; Marlow and Anderson, 2013; Marlow et al., 2017; Qi et al., 2014; Schmid et al., 2020). To test whether the perception of shininess and softness depended on the choice of the cropped area, and therefore on the image features of the highlights present in the crop, we computed the mean luminance of the crops, the relative coverage of the highlights and the mean contrast of the highlights. We did not measure sharpness because that was assumed to be relatively consistent between crops of the same painting.

The calculations of the highlight features, i.e., coverage and contrast, were done using binary images of the crops. The threshold values to binarize the images and isolate the highlights for the computations, were manually derived from the luminance histogram of each crop. Fig. 5.10 A shows the luminance histogram of the crop shown in Fig. 5.10 B. The highlight mode, one of the three general modes for a histogram-based measure of the surface structure proposed by Pont (2009), is indicated by a black bar (note that here the width and height of the bar have no other meaning beside providing a clear visual indication of the threshold value used to binarize the image, whereas in Pont (2009) these parameters were related to the width and the height of the mode). To binarize the images, we manually selected the threshold at the minimum value of the highlight mode (indicated by the red line in Fig. 5.10 A). The manual selection was done for every crop. Fig. 5.10 B shows the original crop and its binary image.

The contrast was calculated as Michelson contrast, using the 95th and the 5th percentiles of the luminance values instead of the absolute maximum and minimum for robustness; the percentage of coverage was calculated as the ratio of the areas covered by white and by black pixels in the binarized image.

All image analyses were done in Matlab 2018a (The MathWorks Inc., Natick, MA).

Note that the measures of the highlights' features reported here should be considered only rough approximations, due to the complexity of automatically and accurately segmenting the image regions that correspond to the highlights, especially in the case of paintings for which the ground truth is not known. Designing a robust algorithm to measure the image features of highlights that is generalizable to natural images such as photographs or paintings, is still an unsolved problem in the literature, due to the difficulty of defining and identifying what the visual system considers to be a highlight. In a previous study (Di Ci-

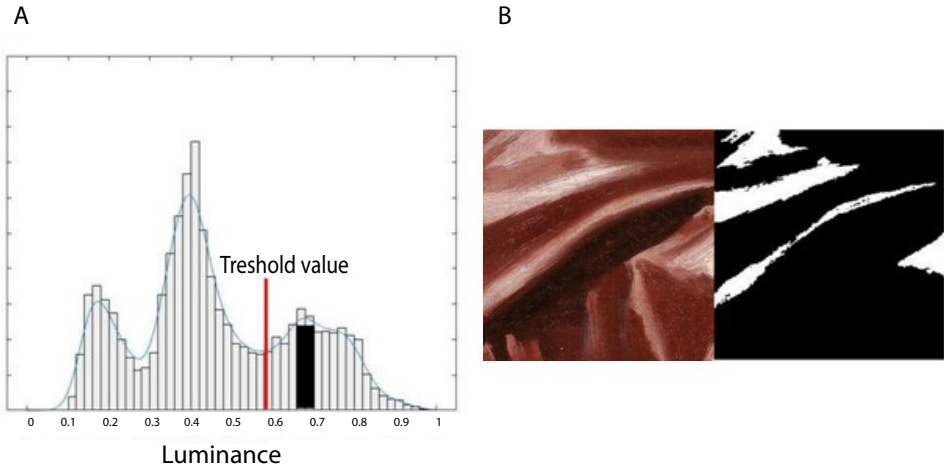


Figure 5.10: A) Luminance distribution and the highlight mode used as threshold value (black bar) to create the binarized image. B) The original stimulus, as presented to the participants in the rating experiment, and its binarized version..

5

cco et al., 2019), we addressed this issue by combining manual annotation of the highlights and self-developed algorithms for the semi-automatic computation of highlights' features directly from images of paintings. However, the paintings analyzed in that study were exclusively depicting grapes, meaning that each object showed a single, mostly round, specular reflection. This simplified the annotation and computation, and made the method more difficult to apply to paintings of fabrics with multiple reflections of various shapes. Marlow et al. (2012) approached the problem by using psychophysical measurements of contrast, coverage, and sharpness of highlights. They further compared the human judgements of the highlights' features with measures obtained via direct image computation, finding high correlations between the two types of measurements. Qi et al. (2014) employed a pixel-wise computation of the highlights' features based on luminance threshold for stimuli rendered with the same reflectance and illumination parameters. Recently, Schmid et al. (2020) developed a series of image-based calculations of the highlights' features that could be applied to stimuli with different shapes, but only with rendered images for which the diffuse and specular components can be defined.

5.3.2. RESULTS

CONSISTENCY BETWEEN AND WITHIN OBSERVERS

The intra- and inter-rater agreement (Fig. 5.11) were calculated for each of the 19 sets of crops for both material attributes. Because shininess and softness were rated three times per crop, prior to the data analysis we took the median over the three repetitions of the ratings to smooth out the effects of potential outliers. Then, the data were normalized to rule out possible effects of unequal interval judgments. Consistency within observers was calculated as the average correlation between the ratings over the three repetitions for each observer.

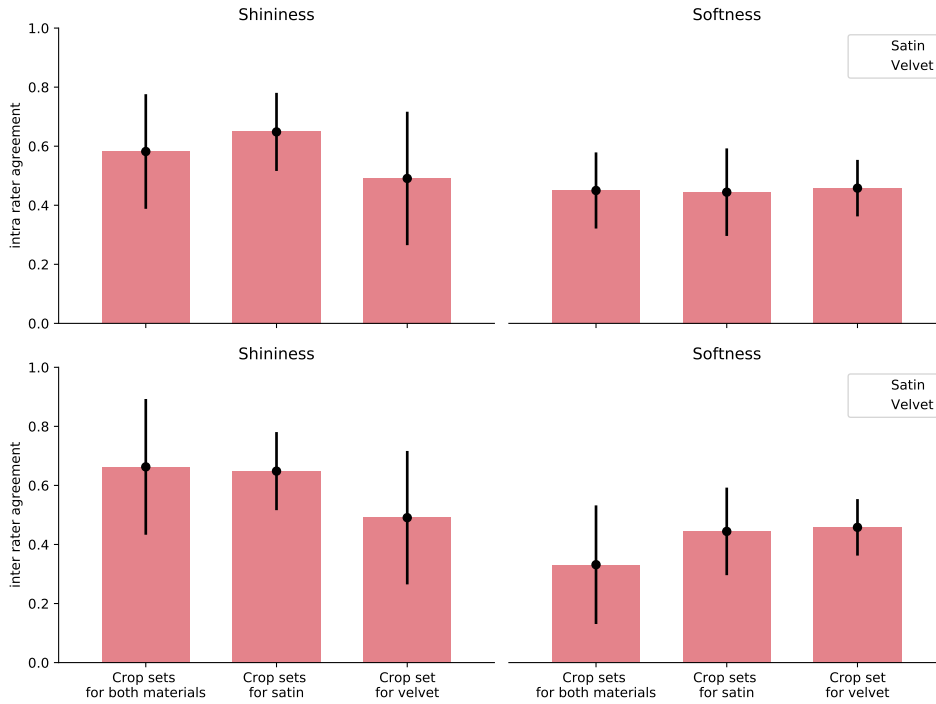


Figure 5.11: The intra- and inter rater agreement. The top contains the intra rater agreement(consistency within observers) with shininess on the left and softness on the right. The same ordering is applied at the bottom for the inter rater agreement (agreement between observers). The error bar indicates the standard deviation.

The consistency between participants was calculated as the mean correlation between all participants.

First, we report the consistency within and between participants split on material attribute. The mean consistency within participants for different sets of crops ranged between 0.16 and 0.81 ($M = 0.58$, $SD = 0.19$) for shininess, and it ranged between 0.18 and 0.77 ($M = 0.45$, $SD = 0.13$) for softness. The mean agreement for different sets of crops between participants ranged from 0.13 to 0.87 ($M = 0.66$, $SD = 0.23$) for shininess, and from 0.08 to 0.78 ($M = 0.33$, $SD = 0.2$) for softness. Next, we further split the data into material category. For satin, we found that the inter rater agreement on shininess ranged from 0.19 to 0.87 ($M = 0.71$, $SD = 0.18$), and on softness from 0.08 to 0.78 ($M = 0.36$, $SD = 0.22$). For velvet, shininess ranged from 0.13 to 0.85 ($M = 0.54$, $SD = 0.25$), and softness from 0.09 to 0.58 ($M = 0.29$, $SD = 0.15$).

The agreement both within and between participants varied greatly. This indicates that some sets of crops triggered a clear and consistent perception, whereas other sets were perceptually ambiguous.

ANOVA

The median ratings of shininess and softness were averaged over all participants for each set of crops, to calculate a one-way ANOVA to measure the effect of varying the cropped area. A significant effect for a set of crops indicates that the perceptual ratings differed between crops taken from a single fabric. Significant differences were evaluated at $p = .001$ after Bonferroni correction. The results of each individual ANOVA are reported Table 5.2. Overall, the crops of fifteen crop sets were significantly different for shininess, ten of which depicted satin, and five depicted velvet. Softness was significantly different for only three sets of the crops, two of which depicted satin and the remaining one velvet.

Table 5.2: Results of one-way ANOVAs of Experiment 2. The ordering of the stimuli corresponds to that of Fig. 5.35 and onward in the supplementary materials. The ANOVAs significant after Bonferroni correction are indicated by *. As can be seen, a significant effect (and thus a varying precept across the same fabric) was found more often for shininess than softness. Stimuli material identity is marked by a S for satin and a V for velvet.

| Stimuli # | Material | Shininess | | | Softness | | |
|-----------|----------|------------|-----------------|----------|------------|-----------------|----------|
| | | | <i>F</i> -value | <i>p</i> | | <i>F</i> -value | <i>p</i> |
| 1 | S | F(10, 55) | 1.8 | > .05 | F(10, 55) | 0.5 | > .05 |
| 2 | S | F(16, 85) | 3.8 | < .001 * | F(16, 136) | 1.1 | > .05 |
| 3 | S | F(12, 52) | 8.1 | < .001 * | F(12, 91) | 3.2 | < .01 |
| 4 | S | F(14, 60) | 10.1 | < .001 * | F(14, 45) | 0.6 | > .05 |
| 5 | S | F(16, 136) | 23.1 | < .001 * | F(16, 85) | 6.7 | < .001 * |
| 6 | S | F(19, 60) | 3.7 | < .001 * | F(19, 60) | 9.9 | < .001 * |
| 7 | S | F(11, 48) | 8.3 | < .001 * | F(11, 84) | 5.7 | < .01 |
| 8 | S | F(20, 84) | 15.8 | < .001 * | F(20, 126) | 1.3 | > .05 |
| 9 | S | F(12, 91) | 32.0 | < .001 * | F(12, 65) | 0.7 | > .05 |
| 10 | S | F(11, 96) | 6.2 | < .001 * | F(11, 84) | 1.2 | > .05 |
| 11 | S | F(16, 119) | 24.5 | < .001 * | F(16, 136) | 1.7 | > .05 |
| 12 | V | F(18, 76) | 13.1 | < .001 * | F(18, 133) | 0.9 | > .05 |
| 13 | V | F(10, 77) | 13.3 | < .001 * | F(10, 66) | 2.7 | < .01 |
| 14 | V | F(14, 90) | 10.7 | < .001 * | F(14, 75) | 0.5 | > .05 |
| 15 | V | F(8, 45) | 16.2 | < .001 * | F(8, 27) | 0.2 | > .05 |
| 16 | V | F(11, 84) | 3.4 | < .001 * | F(11, 72) | 1.8 | > .05 |
| 17 | V | F(12, 52) | 1.5 | > .05 | F(12, 52) | 5.8 | < .001 * |
| 18 | V | F(12, 52) | 1.2 | > .05 | F(12, 117) | 0.7 | > .05 |
| 19 | V | F(12, 52) | 0.9 | > .05 | F(12, 52) | 0.6 | > .05 |

The results from the ANOVAs showed that crops were perceived to vary significantly in shininess within most of the crop sets. We hypothesized that the observed variation in shininess perception can be related to the image features of the highlights available in the different crops.

CORRELATION WITH HIGHLIGHTS' FEATURES

We performed correlation analysis to evaluate the relationships between the mean ratings of shininess and softness of the crops and the features calculated from the images, namely the mean luminance of the crops, and the coverage and contrast of the highlights. We only performed the correlations for the sets of crops in which we found significant differences with

the one-way ANOVA, i.e., fifteen sets for shininess and three for softness. In Fig. 5.12 we reported the correlation coefficients of the image features with shininess (top) and softness (bottom). Only the values significant at $p < .05$ were reported. The stimuli corresponding to the crop sets are reported in Fig. 5.35 and onward in the supplementary material. Note that the crop sets 1-3 for softness do not correspond to the crop sets 1-3 for shininess.

The top of Fig. 5.12 shows that for fourteen out of fifteen significantly different sets, shininess was positively and significantly correlated with the mean luminance of the crops. For eleven crop sets, shininess was also positively and significantly correlated with the coverage of the highlights. Three of the sets showed a significant positive correlation with the contrast of the highlights, whereas for one set the correlation with contrast was negative and significant.

The three sets with crops significantly different in softness reported in Fig. 5.12, were all positively and significantly correlated with the mean luminance. Two of them were also significantly and positively correlated to the coverage of the highlights. None of them was related to the contrast of the highlights.

5.4. GENERAL DISCUSSION

In Experiment 1, we aimed to determine which material attributes belong to the signatures of velvet and satin depicted in 17th century paintings. We further tested if removing shape and context information by only presenting crops of the fabric, caused a change in perception. We found that velvet and satin were judged to have different material attributes, as indicated by the two-way MANOVA (Fig. 5.4) and the PCAs (Fig. 5.6 and Fig. 5.7), and that the commonalities in the judgments were based on robust material signatures that are specific for velvet and satin. In the full figure condition, velvet was judged to be warmer, hairier, softer, heavier, and rougher, while satin was perceived to be shinier. In the PCAs for both conditions, shininess appears to be directed towards the satin cluster while the remaining attributes point more towards the velvet cluster. When we look at the velvet and satin clusters in the PCA for the full figure condition (Fig. 5.6) we also see that the materials are separated. In the crop condition (Fig. 5.7), this separation became less, implying that the distinction between satin and velvet decreases in the crop condition relative to the full figure condition. This is also shown in our finding that all material attributes were significantly different between satin and velvet in the full figure condition, but only part of them in the crop condition. This leads to the following result: satin and velvet depicted in the 17th century are perceptually distinct, but the distinction decreases when only viewing local information. But what is this perceptual distinction between satin and velvet based on?

In the rating tasks, participants were consistent in both conditions but less so in the crop condition. The agreement between participants varied depending on the perceptual attribute, which has been reported before (Fleming et al., 2013; van Zuijlen et al., 2020). Within the domain of computer vision Schwartz and Nishino (2019) argued that (for computer vision) visual material properties, such as shininess and hairiness, should be inherently local. However, Geirhos et al. (2018) showed that CNNs are strongly biased towards texture, i.e., local image features. This implies that computer vision algorithms currently rely on local information. Furthermore, Geirhos et al. (2018) showed that CNNs trained to learn a shape-based representation (i.e., a bias for global information) improve on accuracy and robustness. Similarly, providing global and context information decreases the idiosyn-

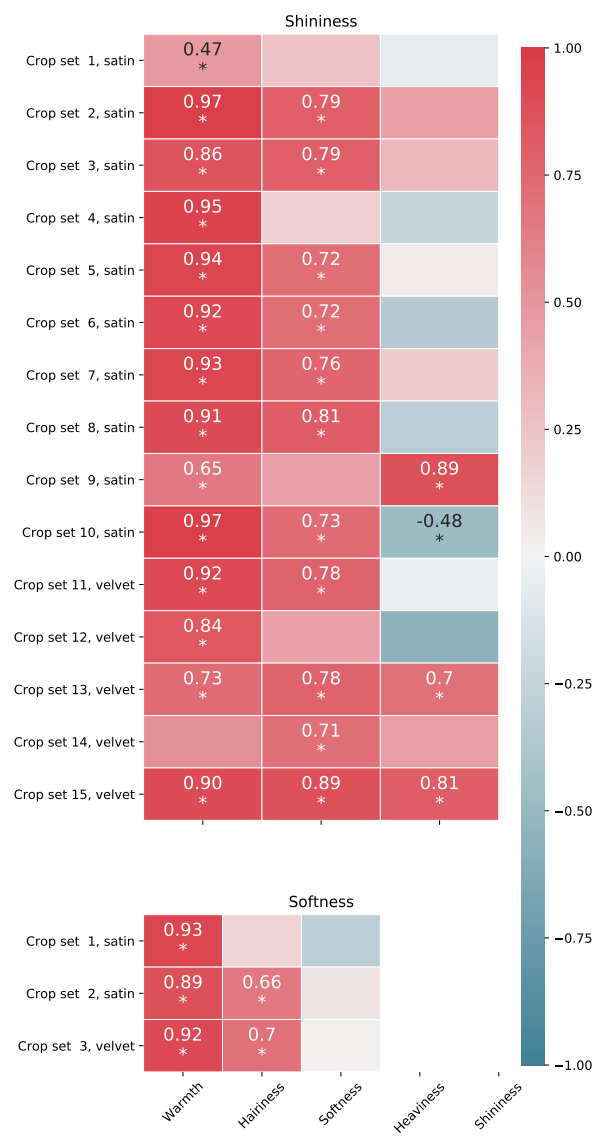


Figure 5.12: Correlation coefficients of shininess (top) and softness (bottom) with the image features highlights' contrast, highlights' coverage, and mean luminance of the crops. The values reported are significant at $p < .05$



Figure 5.13: One crop that was identified as a possible outlier. With this stimulus included, a strong negative correlation was found between softness and roughness, which is surprising based on the literature. With this crop removed, the correlation is no longer significant. This might be due to the visibility of the individual brushstrokes, which gave rise to a perceptual ambiguity.

crasy for our human data in our experiment. This implies that while both computer and human vision can form a clear or robust response from local information, the responses' robustness can be improved by providing global information.

The correlation matrices in Fig. 5.5 showed that roughness was negatively correlated to shininess in both viewing conditions. This is in agreement with many reflectance distribution models such as for instance the microfacets model (Cook and Torrance, 1982), in which rough surfaces are modeled as a distribution of specular microfacets, which orientation distribution determines the surface roughness and resulting width of the reflectance lobe (the rougher, the less glossy, see also for instance (Honson et al., 2020; Wendt and Faul, 2017)). We also see this negative correlation in the two PCA biplots (Fig. 5.6 and Fig. 5.7).

For roughness we found no correlation with softness for the full figure condition, which is in agreement with several studies that have shown that the main perceptual dimensions of tactile and visual perception of texture are roughness/smoothness and hardness/softness (Hollins et al., 1993; Nagano et al., 2013; Zhang et al., 2019). However, in the crop condition, we found a negative correlation between roughness and softness. This negative correlation might be ascribed to one outlier: a crop with clearly visible rough brushstrokes (see Fig. 5.13) which was on average perceived to be the second roughest fabric and the least soft. Indeed, removing this crop from the data made the correlation no longer significant. Possibly the roughness of the brushstrokes for this specific stimulus introduced an element of ambiguity in the judgment of the surface roughness of the fabric.

Heaviness was significantly negatively correlated with shininess in both viewing conditions. In the crop condition, no size information was available. If participants were able to retrieve the material identity, heaviness could have been inferred through an “associative approach” (Schmidt et al., 2017). One possible association could have been that darker objects are perceived to be heavier than brighter ones (Vicovaro et al., 2019; Walker et al., 2017). From additional analysis, we found that ratings of heaviness in the crop condition

were indeed highly negatively correlated with the mean luminance of the stimuli ($r = -0.73$, $p < .001$). Shininess, on the other hand, was highly and positively correlated with the mean luminance ($r = 0.75$, $p < .001$).

Softness, a material property relying on haptic information, is physically independent from the visual property of glossiness. However, they can be perceptually related since a perceptual association can be learned when intentionally induced (Ernst, 2007; Wismeijer et al., 2012), or from prior experience, since glossy materials tend to be hard (Ingvarsdóttir and Balkenius, 2020). In the full figure condition, there was indeed a high and significant negative correlation between shininess and softness, likely due to the identification of the objects and of the materials they were made of. Paulun et al. (2017) showed that the optical appearance of familiar materials creates expectations and influence stiffness perception. This might explain the lack of correlation between softness and shininess in the crop condition, where participants knew they were judging fabrics, but they were missing contextual information to recognize the fabrics' material, and thus were unable to draw from expectations.

In Fig. 5.4 (top), we reported the attributes that were perceived to be significantly different between the two viewing conditions, per material. If we considered a single fabric, we observed additional variations of attributes between conditions (see Fig. 5.8). This raised the question whether such variation in perception was due to our choice of the area to crop in the fabrics. Thus, in Experiment 2 we tested the relationship between the perception of shininess and softness and different areas cropped within the same fabric, spanning the whole fabric as much as possible. If different perceptions were triggered, they might be the result of the presence or absence of local image features within the crop. On the other hand, if all crops were perceived similarly, we might argue that local image features tend to be stable across the entire surface of the materials, at least within our set of stimuli.

The consistency within participants fluctuated greatly for different sets of crops, from 0.16 to 0.81 and from 0.18 to 0.77, for shininess and softness respectively. The consistency between participants showed similar fluctuations, from 0.13 to 0.87 for shininess, and from 0.08 to 0.78 for softness. The high agreement found for some sets of crops indicates that these crops evoked a clear and consistent perception. Simultaneously, the low agreement on other sets showed the opposite, namely that these crops were perceptually ambiguous. In the first experiment, stimuli presented with context and shape information, evoked a more consistent perception. It appears that cropping stimuli reduces the uniqueness of the evoked perception in some, but not all stimuli. The size, aspect ratio and area relative to the original image was kept constant within each set of crops, and can thus not explain the differences found. The local content of the crops within sets of crops must have caused the variety: the presence (or absence) of local image features in the crops of each set might be (in)sufficient to elicit a clear, consistent perception.

The results from the ANOVAs showed that crops were perceived to vary significantly in shininess within most of the crop sets. The presence of highlights on a surface is a well-known image feature for the perception of glossiness. According to Beck and Prazdny (1981), glossiness perception depends on the local presence of highlights, meaning that the direct area surrounding the highlight is perceived to be glossy, but not the whole surface per se. That is, they argue that glossiness perception is the direct response to local visual information, and not the result of some perceptual inference about the reflectance properties

of the whole surface. Similar results are discussed by (Berzhanskaya et al., 2005). They found that perceived gloss decreases as a function of the distance from the highlight. Thus when different parts of an object are considered, gloss perception will differ among the different parts depending on their vicinity to the highlights. This local quality of glossiness is in agreement with our results on the perception of shininess differing between the crops of a fabric.

We used three image features (mean luminance, coverage of the highlight and contrast of the highlight) to further analyze this relationship between the local image content and the evoked perception. In Fig. 5.12 (top), we showed that the mean luminance of the crops was highly and positively correlated with almost all the crop sets for shininess. This finding is in line with (Wiebel et al., 2015), who found that the mean luminance of photographs of real materials was a high-performance predictor, followed by the standard deviation of luminance, to differentiate between glossy and matte materials. Highlights are high-luminance regions of the surface, explaining the high correlation we observed between the mean luminance of the crops and the perceived shininess. Coverage of the highlights was also highly correlated with the perceived shininess for most of the crop sets. Coverage of highlights has been shown to be strongly associated with glossiness perception (Marlow and Anderson, 2013; Marlow et al., 2012), especially when coverage is the most reliable cue for the judgement of glossiness. This happens with objects whose shapes create higher variability in highlights' coverage rather than contrast or sharpness, under the same illumination. For our stimuli, within the same fabric, the folding configuration caused high variations of coverage that we found to be related to significant variations in shininess perception between the different crops of a fabric. High highlights' coverage is also related to higher mean luminance, given that the area of the surface covered with highlights, i.e., the high-luminance regions, increases. We indeed found the correlation between the mean luminance and the coverage averaged over all the crop sets, to be high and significant ($r = 0.78$ $p < .001$). The third image feature that we measured, the highlights' contrast, overall, was not strongly correlated with perceived shininess. In the three cases in which high and significant positive correlations were found, the contrast was also positively correlated with coverage. The opposite occurred for the only crop set that showed a significant negative correlation between contrast and shininess, i.e., the high contrast highlights covered the smallest regions of the fabrics' surface.

For softness perception, the ANOVAs showed no significant differences for most of the crop sets. So, while the perception of shininess might depend on local image features, this might not hold for softness. A possible explanation for this finding could be that softness is a mechanical property, rather than an optical property and therefore less associated to the image features. Another related possibility is that the image features that were analyzed are simply not the key triggers for softness. The question then arises whether other local features might explain the data, or whether mechanical attributes such as softness requires global features to explain the judgments.

In the bottom section of Fig. 5.12 we reported the three crop sets that were significantly different for softness perception. They all showed a high and significant positive correlation with the mean luminance of the crops. Two sets were also significantly positively correlated with the coverage of the highlights and one of these sets (crop set 2) is shown in Fig. 5.14. The crops in the top row were perceived to be significantly softer than the crops in the



Figure 5.14: Visualization of the crop set 2 from the bottom of Fig. 5.12. The image on top shows the locations where the crops were taken from the whole fabric. The crop in the top row were perceived to be significantly softer than the crops in the bottom row.

bottom row. What is apparent from these two rows of crops is that in the top row, the high luminance and the high coverage of the highlights allow to clearly see the folding shape of the fabric, in contradistinction to those in the bottom row. Local shape features, like textiles' folding, have been shown to play a role in the visual estimation of softness perception (Schmidt et al., 2020). For the stimuli shown in Fig. 5.14, the visibility of the shape deformation due to the folding could have been the driving cue for the perception of different levels of softness between the crops. This is in agreement with Xiao et al. (2016), who showed that the 3D folding configuration increases the accuracy of estimation of tactile material properties of fabrics.

Other cues, like the brightened contours', might be related to visual perception of softness via a cognitive association with velvet (Paulun et al., 2017; Schmidt et al., 2020; Zhang et al., 2019). Further research is needed to understand how local and global information contribute to and possibly interact in material perception, and whether such mechanisms are dependent on the material and property under consideration.

Since the 15th century, with the introduction of oil painting and a whole new range of possible visual effects, Netherlandish painters started to shift the attention from the rendering of space and volume to the rendering of materials, reaching their "golden age" in the 17th century. When the separation between diffuse and specular illumination started to be acknowledged and exploited (Gombrich, 1976), painters could visually differentiate velvet from satin, instead of rendering all the fabrics equally matte. This novel use of the highlights is what we quantified in Experiment 2 via image analysis, and related to the perception of shininess and softness. However, there is a stylistic aspect of the paintings

that we selected for our stimuli set, which we did not address here, in order to focus on the discussion on material perception. The paintings were made either with a neat, almost invisible brushwork (see Fig. 5.15, left) or with loose brushstrokes (see Fig. 5.15, right). These opposite pictorial manners were equally valued to produce a convincing effect (Gombrich, 1960), but their mechanisms are completely different. Paintings with the fine brushstrokes can be appreciated from a distance or from close by in a similar way, whereas paintings with coarse brushstrokes are unintelligible when one stands close or zooms in, but they make perfect sense and trigger a powerful convincing effect when seen in their entirety, at a proper distance. The different brushworks might have introduced an additional source of noise in our data, but they also raised further questions, like, how is the pictorial style (fine vs coarse) related to the use of local and global image cues for material depiction and perception? Future work in this direction could contribute to the emerging field of art and perception.

5.5. CONCLUSIONS

In this study, we found that warmth, heaviness, hairiness, and softness are key attributes of the material signatures of velvet, whereas shininess is a key attribute of the signature of satin, when studying the depiction of both fabrics in 17th century paintings. We further showed that the two fabrics, as depicted in 17th century paintings, and their material signatures were clearly perceptually distinct when the stimuli were presented in the full figure condition. On the other hand, the cropped condition, depriving the visual system of object shape and context information, caused higher ambiguity and made the distributions for the measured perceptual attributes of the two materials less distinct.

In Experiment 2, we showed that the perceived shininess is not stable across one single fabric. The perception of the optical property shininess based on a cropped area of the fabric was shown to be correlated to the presence of diagnostic image features in the crop, namely highlights. Moreover, shininess perception increased with the coverage of the highlights and their mean luminance.

The haptic property of softness, instead, did not differ significantly between crops of the same fabric. Further analysis of the softness data suggested that perception of this haptic property might be driven by local and global shape cues.

In conclusion, we have shown that velvet and satin were depicted with distinct perceptual material signatures, which painters started to employ around the 15th century, and highlights started to be exploited to render the characteristic appearance of different textiles (Gombrich, 1976). Highlights can be used to render the luster of satin and the softness of velvet, by indicating not only how the fabric reflects light but also by revealing the shape of the folds. Local image features of the highlights were found to be sufficient to trigger significant variations in shininess perception, but not for softness. This indicates that shininess is a local material property, whereas softness might require more global visual information relating to shape.



Figure 5.15: Examples from our stimuli set of a neat (left) and a loose use of brushstrokes (right), to illustrate the difference in these two styles which becomes apparent when moving closer to the physical painting – or when zooming in. The two crops (below) are from the crop-set that correspond to the paintings (above). Left: Self-portrait with the Portrait of his Wife, Margaretha van Rees, and their Daughter Maria by Adriaen van der Werff., 1699, Rijksmuseum, Amsterdam. Right: Portrait of a Man, Possibly Nicolaes Pietersz Duyst van Voorhout by Frans Hals, ca. 1636–38, The Metropolitan Museum of Art, New York.

BIBLIOGRAPHY

- Adelson, E. H. (2001). On Seeing Stuff: The Perception of Materials by Humans and Machines. *Proceedings of the SPIE*, 4299, 1–12. <https://doi.org/10.1117/12.429489>
- Akgun, M., Becerir, B., & Alpay, H. R. (2014). Effect of fabric layers on the relationship between fabric constructional parameters and percentage reflectance values of polyester fabrics. *Journal of Textiles*, 2014.
- Aliaga, C., O’Sullivan, C., Gutierrez, D., & Tamstorf, R. (2015). Sackcloth or silk? the impact of appearance vs dynamics on the perception of animated cloth. *Proceedings of the acm siggraph symposium on applied perception*, 41–46.
- Balas, B., Auen, A., Thrash, J., & Lammers, S. (2020). Children’s use of local and global visual features for material perception. *Journal of vision*, 20(2), 10–10.
- Barati, B., Karana, E., Sekulovski, D., & Pont, S. C. (2015). Retail lighting and textiles : Designing a lighting probe set, 1–22. <https://doi.org/10.1177/1477153515602953>
- Beck, J., & Prazdny, S. (1981). Highlights and the perception of glossiness. *Perception & Psychophysics*.
- Berzhanskaya, J., Swaminathan, G., Beck, J., & Mingolla, E. (2005). Remote effects of highlights on gloss perception. *Perception*, 34(5), 565–575.
- Beurs, W. (1692). *De groote waereld in ‘t kleen geschildert, of schilderagtig tafereel van ‘s weerelds schilderyen*. Amsterdam, the Netherlands: van Waesberge.
- Bouman, K. L., Xiao, B., Battaglia, P., & Freeman, W. T. (2013). Estimating the material properties of fabric from video. *Proceedings of the IEEE international conference on computer vision*, 1984–1991.
- Casati, R., & Cavanagh, P. (2019). *The visual world of shadows*. MIT Press.
- Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, 434(7031), 301–307.
- Cook, R. L., & Torrance, K. E. (1982). A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1), 7–24.
- Di Cicco, F., Wiersma, L., Wijntjes, M., & Pont, S. (2020). Material properties and image cues for convincing grapes: The know-how of the 17th-century pictorial recipe by willem beurs. *Art & Perception*, 8(3-4), 337–362.
- Di Cicco, F., Wijntjes, M. W., & Pont, S. C. (2019). Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *Journal of Vision*, 19(3). <https://doi.org/10.1167/19.3.7>
- Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of vision*, 7(5), 7–7.
- Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 3, 347–368. <https://doi.org/10.1037/t25352-000>
- Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8), 9. <https://doi.org/10.1167/13.8.9>

- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.
- Gombrich, E. H. (1976). *The heritage of apelles* (Vol. 3). Phaidon Oxford.
- Gombrich, H., E. (1960). *Art & Illusion. A study in the psychology of pictorial representation* (5th). Phaidon Press Limited.
- Gordenker, E. E. (1999). The rhetoric of dress in seventeenth-century dutch and flemish portraiture. *The Journal of the Walters Art Gallery*, 87–104.
- Ho, Y.-x., Landy, M. S., & Maloney, L. T. (2008). Conjoint Measurement of Gloss and Surface Texture. *19*(2), 196–204.
- Hollins, M., Aldowski, R. F., Rao, S., & Young, F. (1993). Perceptual dimensions of tactile surface texture : A multidimensional scaling analysis. *54*(6).
- Honson, V., Huynh-Thu, Q., Arnison, M., Monaghan, D., Isherwood, Z. J., & Kim, J. (2020). Effects of shape, roughness and gloss on the perceived reflectance of colored surfaces. *Frontiers in psychology*, 11.
- Ingvarsdóttir, K. Ó., & Balkenius, C. (2020). The visual perception of material properties affects motor planning in prehension: An analysis of temporal and spatial components of lifting cups. *Frontiers in psychology*, 11, 215.
- Koenderink, J., & Pont, S. (2003). The secret of velvety skin. *Machine vision and applications*, 14(4), 260–268.
- Koenderink, J. J., & van Doorn, A. J. (2001). Shading in the case of translucent objects. *Human vision and electronic imaging vi*, 4299, 312–320.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155–163.
- Lehmann, A.-S. (2015). Depicting the sacred through fabric. <http://blogs.getty.edu/iris/depicting-the-sacred-through-fabric/>
- Lehmann, A., & Stumpel, J. (In press). *Willem beurs – the big world painted small*, scholz, m. (transl.) The Getty Research Institute, Los Angeles, CA, USA.
- Liedtke, W. A. (2011). *Frans hals: Style and substance* (Vol. 61). Metropolitan Museum of Art.
- Lu, R., Koenderink, J. J., & Kappers, A. M. (1998). Optical properties (bidirectional reflection distribution functions) of velvet. *Applied Optics*, 37(25), 5974–5984.
- Marlow, & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision*, 13(14), 1–23. <https://doi.org/10.1167/13.14.2>
- Marlow, & Anderson, B. L. (2015). Material properties derived from three-dimensional shape representations. *Vision Research*, 115, 199–208. <https://doi.org/10.1016/j.visres.2015.05.003>
- Marlow, Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, 22(20), 1909–1913. <https://doi.org/10.1016/j.cub.2012.08.009>
- Marlow, P. J., Kim, J., & Anderson, B. L. (2017). Perception and misperception of surface opacity. *Proceedings of the National Academy of Sciences*, 114(52), 13840–13845.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1), 30.

- Nagano, H., Okamoto, S., & Yamada, Y. (2013). Visual and sensory properties of textures that appeal to human touch. *International Journal of Affective Engineering*, 12(3), 375–384.
- Parraman, C. (2014). The visual appearance and surface texture of materials according to the old masters. *Measuring, Modeling, and Reproducing Material Appearance*, 9018, 90180H.
- Paulun, V. C., Schmidt, F., van Assen, J. J. R., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of vision*, 17(1), 20–20.
- Pons-Moll, G., Pujades, S., Hu, S., & Black, M. J. (2017). Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4), 1–15.
- Pont, S. C. (2009). Ecological optics of natural materials and light fields. *Human vision and electronic imaging XIV*, 7240, 724009.
- Pont, S. C., & Koenderink, J. J. (2003). Split off-specular reflection and surface scattering from woven materials. *Applied optics*, 42(8), 1526–33. <https://doi.org/10.1364/AO.42.001526>
- Pont, S. C., & Te Pas, S. F. (2006). Material - Illumination ambiguities and the perception of solid objects. *Perception*, 35(10), 1331–1350. <https://doi.org/10.1068/p5440>
- Pottasch, C. (2020). Frans van mieris's painting technique as one of the possible sources for willem beurs's treatise on painting. *Art & Perception*, 1(aop), 1–17.
- Pottasch, C. (13 Jun. 2020). Frans van mieris's painting technique as one of the possible sources for willem beurs's treatise on painting. *Art & Perception*, 1–17. <https://doi.org/https://doi.org/10.1163/22134913-bja10013>
- Qi, L., Chantler, M. J., Siebert, J. P., & Dong, J. (2014). Why do rough surfaces appear glossy? *JOSA A*, 31(5), 935–943.
- Sayim, B., & Cavanagh, P. (2011). What Line Drawings Reveal About the Visual Brain. *Frontiers in Human Neuroscience*, 5(October), 1–4. <https://doi.org/10.3389/fnhum.2011.00118>
- Schmid, A. C., Barla, P., & Doerschner, K. (2020). Material category determined by specular reflection structure mediates the processing of image features for perceived gloss. *bioRxiv*, 1–45. <https://doi.org/10.1101/2019.12.31.892083>
- Schmidt, F. (2019). The art of shaping materials. *Art & Perception*, 8(3-4), 407–433.
- Schmidt, F., Fleming, R. W., & Valsecchi, M. (2020). Softness and weight from shape: Material properties inferred from local shape features. *Journal of vision*, 20(6), 2–2.
- Schmidt, F., Paulun, V. C., van Assen, J. J. R., & Fleming, R. W. (2017). Inferring the stiffness of unfamiliar objects from optical, shape, and motion cues. *Journal of Vision*, 17(3), 18–18.
- Schwartz, G., & Nishino, K. (2019). Recognizing material properties from images. *IEEE transactions on pattern analysis and machine intelligence*, 42(8), 1981–1995.
- Thomas, J. A. (1994). Fabric and dress in bronzino's portrait of eleanor of toledo and son giovanni. *Zeitschrift für Kunstgeschichte*, 57(H. 2), 262–267.
- Tuunainen, V. K., & Rossi, M. (2002). Ebusiness in apparel retailing industry-critical issues. *ECIS 2002 Proceedings*, 136.
- van Duijn, E., & Roeders, J. (2012). Gold-brocaded velvets in paintings by cornelis engebrechtsz. *Journal for Historians of Netherlandish Art*, 4, 4.

- van Eikema Hommes, M. H. et al. (2002). Discoloration in renaissance and baroque oil paintings. instructions for painters, theoretical concepts, and scientific data. *University of Amsterdam, PhD Thesis*, 16–46.
- van Mander, K. (1604). *Het schilder-boek*. Haarlem.
- van Zuijlen, M. J., Pont, S. C., & Wijntjes, M. W. (2020). Painterly depiction of material properties. *Journal of vision*, 20(7).
- Vicovaro, M., Ruta, K., & Vidotto, G. (2019). Influence of visually perceived shape and brightness on perceived size, expected weight, and perceived weight of 3d objects. *PloS one*, 14(8), e0220149.
- Walker, P., Scallan, G., & Francis, B. (2017). Cross-sensory correspondences: Heaviness is dark and low-pitched. *Perception*, 46(7), 772–792.
- Wendt, G., & Faul, F. (2017). Increasing the complexity of the illumination may reduce gloss constancy. *i-Perception*, 8(6), 2041669517740369.
- Wiebel, C. B., Toscani, M., & Gegenfurtner, K. R. (2015). Statistical correlates of perceived gloss in natural images. *Vision Research*, 115, 175–187. <https://doi.org/10.1016/j.visres.2015.04.010>
- Wijntjes, M. W. A., Doerschner, K., Kucukoglu, G., & Pont, S. C. (2012). Relative flattening between velvet and matte 3D shapes: Evidence for similar shape-from-shading computations. *Journal of Vision*, 12(1), 2–2. <https://doi.org/10.1167/12.1.2>
- Wijntjes, M. W. A., Spoiala, C., & de Ridder, H. (2020). Thurstonian scaling and the perception of painterly translucency. *Art & Perception*, 1–24. <https://doi.org/https://doi.org/10.1163/22134913-bja10021>
- Wijntjes, M. W., Xiao, B., & Volcic, R. (2019). Visual communication of how fabrics feel. *Journal of vision*, 19(2), 4–4.
- Wismeijer, D. A., Gegenfurtner, K. R., & Drewing, K. (2012). Learning from vision-to-touch is different than learning from touch-to-vision. *Frontiers in integrative neuroscience*, 6, 105.
- Xiao, B., Bi, W., Jia, X., Wei, H., & Adelson, E. H. (2016). Can you see what you feel? color and folding properties affect visual–tactile material discrimination of fabrics. *Journal of vision*, 16(3), 34–34.
- Zhang, F., De ridder, H., Barla, P., & Pont, S. (2019). A systematic approach to testing and predicting light-material interactions. *Journal of vision*, 19(4), 1–22.
- Zhao, S., Jakob, W., Marschner, S., & Bala, K. (2011). Building volumetric appearance models of fabric using micro ct imaging. *ACM Transactions on Graphics (TOG)*, 30(4), 1–10.
- Zhao, S., Luan, F., & Bala, K. (2016). Fitting procedural yarn models for realistic cloth rendering. *ACM Transactions on Graphics (TOG)*, 35(4), 1–11.

Table 5.3: ICC calculations for the inter-rater agreement in Experiment 1. The calculation was done using average rating, consistency agreement, two-way random effects model.

| | 95% CI | | | F test with true value 0 | | | |
|-------------|-------------|-------|-------------|--------------------------|-----|-----|-------|
| | Lower Bound | | Upper Bound | Value | df1 | df2 | Sig. |
| Full figure | | | | | | | |
| Warmth | 0.78 | 0.61 | 0.91 | 4.74 | 18 | 162 | <.001 |
| Hairiness | 0.92 | 0.85 | 0.96 | 12.3 | 18 | 162 | <.001 |
| Softness | 0.75 | 0.54 | 0.89 | 3.97 | 18 | 162 | <.001 |
| Heaviness | 0.89 | 0.79 | 0.95 | 9.05 | 18 | 162 | <.001 |
| Shininess | 0.92 | 0.86 | 0.96 | 12.8 | 18 | 162 | <.001 |
| Roughness | 0.61 | 0.27 | 0.82 | 2.54 | 18 | 162 | <.01 |
| Full figure | | | | | | | |
| Warmth | 0.5 | 0.08 | 0.77 | 1.99 | 18 | 162 | <.05 |
| Hairiness | -0.7 | -2.25 | 0.21 | 0.56 | 18 | 162 | >.05 |
| Softness | 0.49 | 0.06 | 0.77 | 1.97 | 18 | 162 | <.05 |
| Heaviness | 0.55 | 0.17 | 0.79 | 2.22 | 18 | 162 | <.01 |
| Shininess | 0.71 | 0.47 | 0.87 | 3.5 | 18 | 162 | <.001 |
| Roughness | 0.78 | 0.6 | 0.9 | 4.63 | 18 | 162 | <.001 |

5.6. SUPPLEMENTARY MATERIALS

First, in Table 5.3 present the the intra-class correlations for the inter rater agreement in Experiment 1.

Next, all the stimuli used in Experiment 1 are shown in Fig. 5.16 through Fig. 5.34 below. For stimuli we display the full figure condition on the left, and the crop condition on the right. Note that both images are resized to fit the page while in the experiment stimuli were presented at 600x600 and 200x200 pixels for the full figure and crop condition, respectively. The first 11 are the satin stimuli, the remaining 8 are the velvet stimuli.

Last, we visualize all the stimuli used in Experiment 2 in Fig. 5.35 and onward. The first 11 are satin stimuli, the remaining 8 are the velvet stimuli.

All images reproduced here are available under open access at a CC0 or CC BY 4.0 license. Two paintings (and the crops thereof) have not been reproduced here due to copy rights restraints.

5



Figure 5.16: Portrait of a Man, Possibly Nicolaes Pietersz Duyst van Voorhout, by Frans Hals. 1637, The Metropolitan Museum of Art

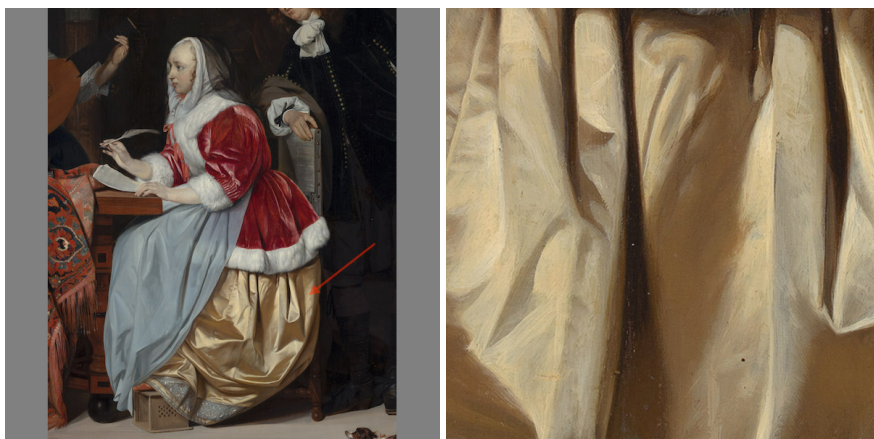


Figure 5.17: A Young Woman Composing a Piece of Music, by Gabriël Metsu. 1664, Mauritshuis, The Hague



5

Figure 5.18: Portrait of a Man, by Adriaen van der Werff. 1689, Mauritshuis, The Hague



Figure 5.19: The Oyster Meal, by Frans van Mieris the Elder. 1661, Mauritshuis, The Hague

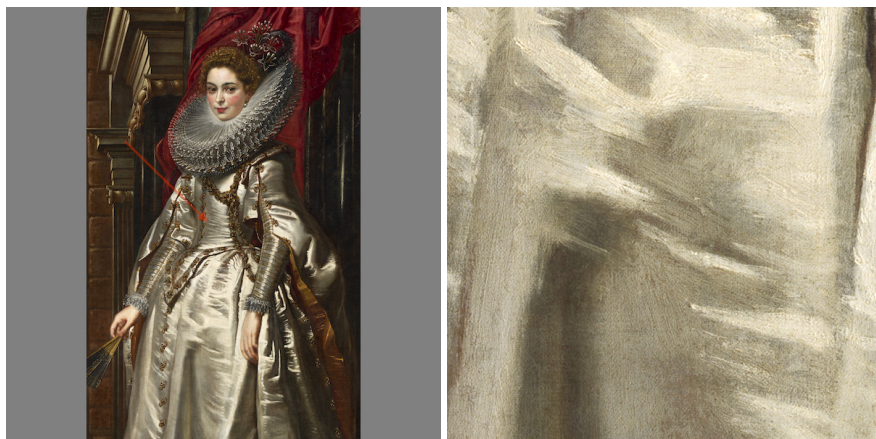
5



Figure 5.20: Philip, Lord Wharton, by Anthony van Dyck. 1632, National Gallery of Art



Figure 5.21: Catherine Howard, Lady d'Aubigny, by Anthony van Dyck. 1638, National Gallery of Art



5

Figure 5.22: Marchesa Brigida Spinola Doria, by Peter Paul Rubens. 1606, National Gallery of Art



Figure 5.23: Queen Henrietta Maria with Sir Jeffrey Hudson, by Anthony van Dyck. 1633, National Gallery of Art



Figure 5.24: William II, Prince of Orange, and his Bride, Mary Stuart, by Anthony van Dyck. 1641. The Rijksmuseum

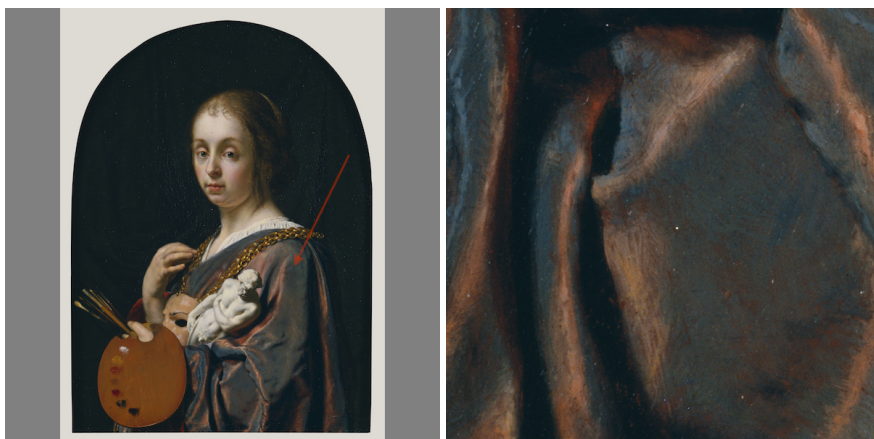


Figure 5.25: Pictura (An Allegory of Painting), by Frans van Mieris the Elder. 1661, J. Paul Getty Museum



5

Figure 5.26: The Letter Writer, by Frans van Mieris (I). 1680, The Rijksmuseum



Figure 5.27: Portrait of Agostino Pallavicini, by Anthony van Dyck. 1621, J. Paul Getty Museum

Figure 5.28: A Woman in a Red Jacket feeding a Parrot, by Frans van Mieris the Elder. 1663, The National Gallery of London. Images not reproduced due to copy-rights restrictions.



Figure 5.29: The Letter Writer, by Frans van Mieris (I). 1680, The Rijksmuseum

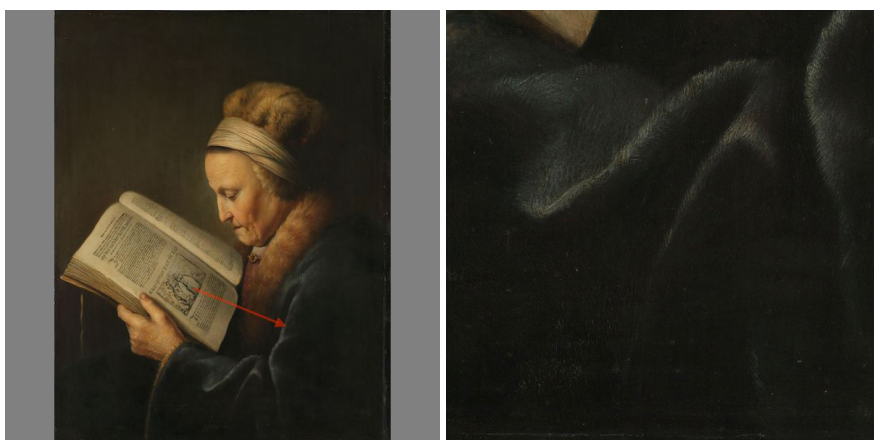


Figure 5.30: Old Woman Reading, by Gerard Dou. 1631, The Rijksmuseum

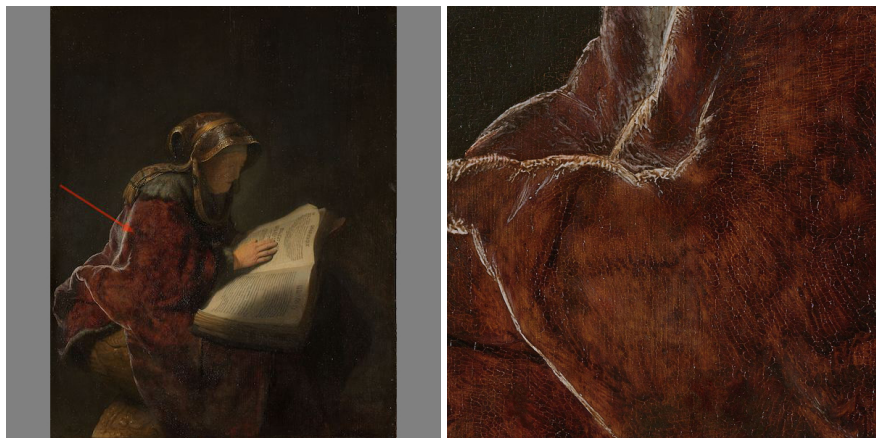


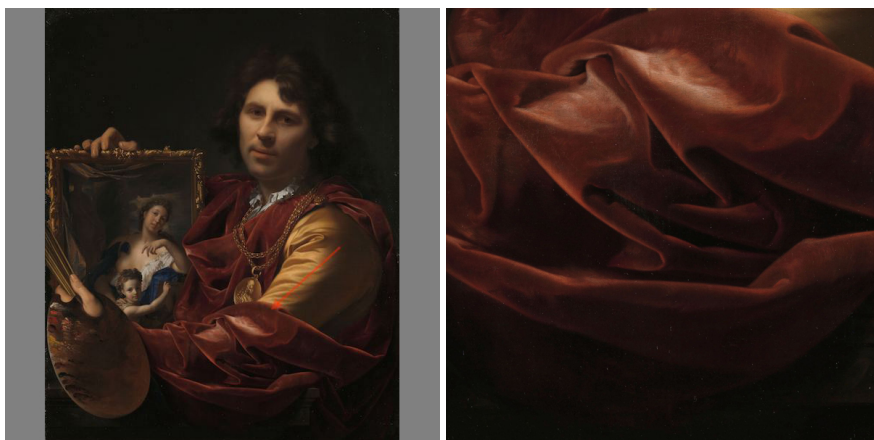
Figure 5.31: An Old Woman Reading, Probably the Prophetess Hannah, by Rembrandt van Rijn. 1631, The Rijksmuseum

5



Figure 5.32: Philip, Lord Wharton, by Anthony van Dyck. 1632, National Gallery of Art

Figure 5.33: Self Portrait of the Artist, with a Cittern, by Frans van Mieris the Elder. 1674, The National Gallery of London. Images not reproduced due to copy-rights restrictions.



5

Figure 5.34: Self-portrait with the Portrait of his Wife, Margaretha van Rees, and their Daughter Maria, by Adriaen van der Werff. 1699, The Rijksmuseum

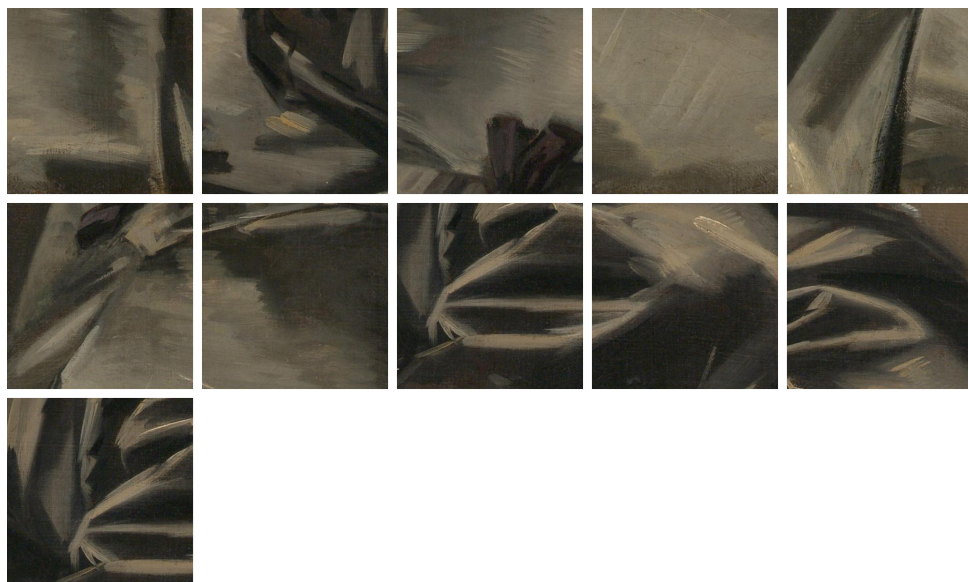


Figure 5.35: Stimulus 1 in Experiment 2. Crops taken from Portrait of a Man, Possibly Nicolaes Pietersz Duyst van Voorhout, by Frans Hals. 1637, The Metropolitan Museum of Art

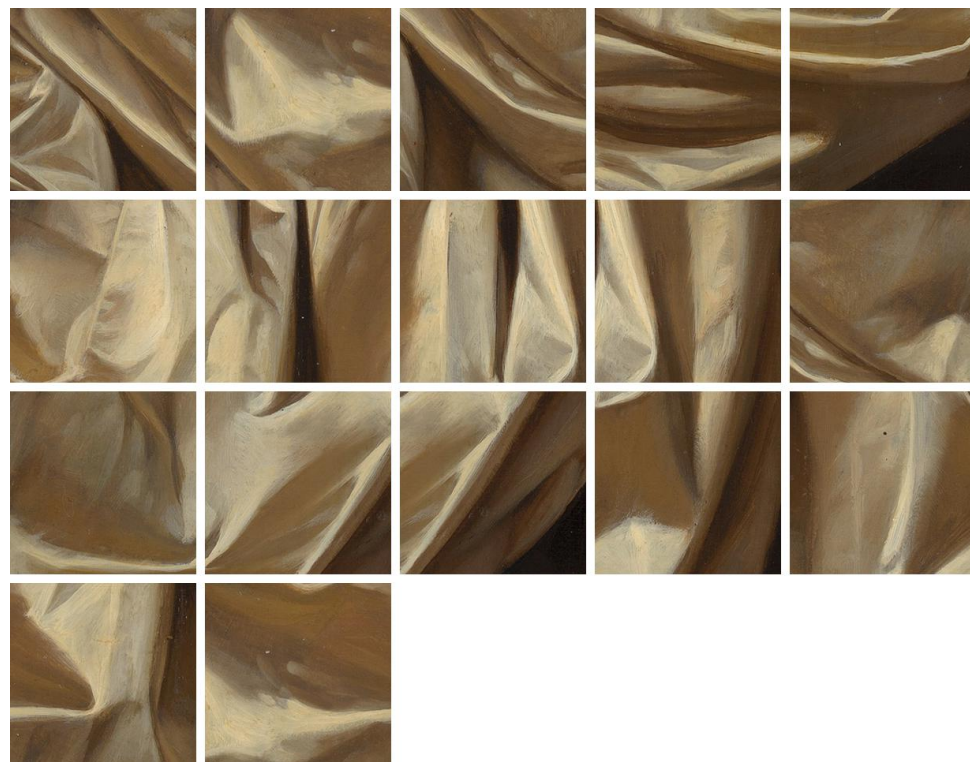


Figure 5.36: Stimulus 2 in Experiment 2. Crops taken from *A Young Woman Composing a Piece of Music*, by Gabriël Metsu. 1664, Mauritshuis, The Hague. Crop set 1 in Figure 12 shininess



Figure 5.37: Stimulus 3 in Experiment 2. Crops taken from *Portrait of a Man*, by Adriaen van der Werff. 1689, Mauritshuis, The Hague. Crop set 2 in Figure 12 shininess

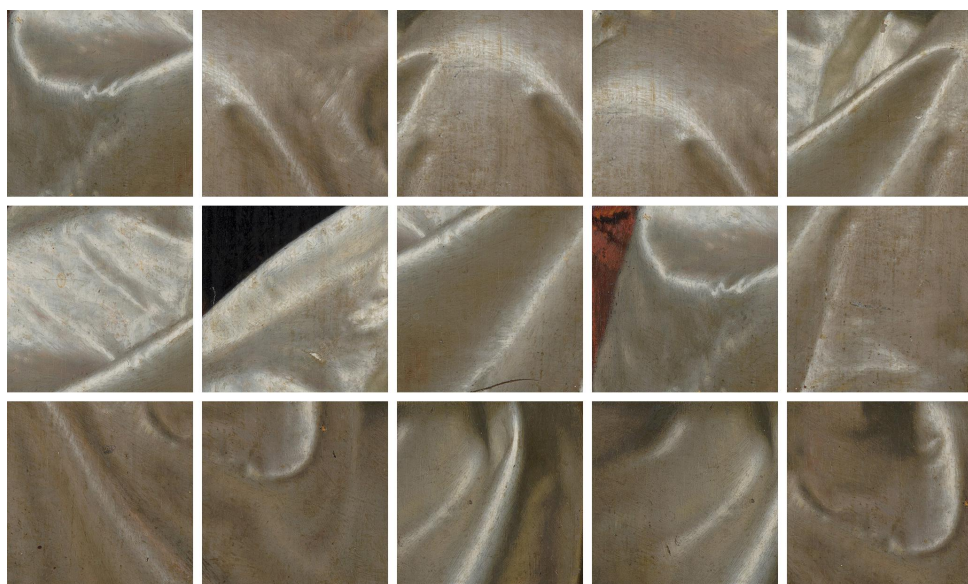


Figure 5.38: Stimulus 4 in Experiment 2. Crops taken from *The Oyster Meal*, by Frans van Mieris the Elder. 1661, Mauritshuis, The Hague. Crop set 3 in Figure 12 shine

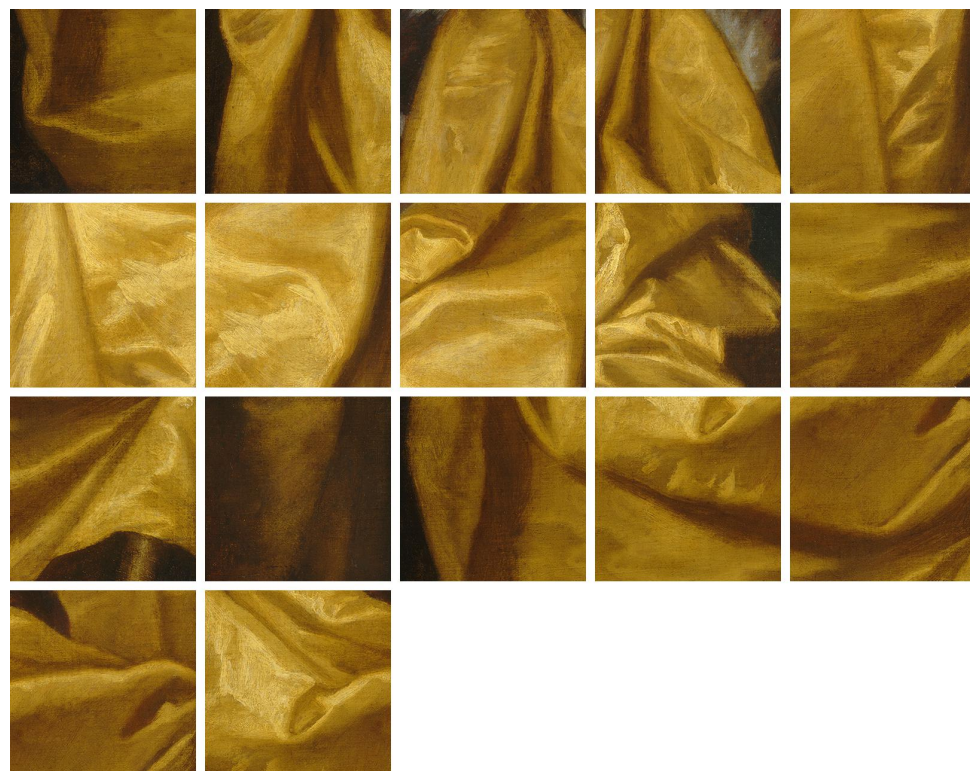


Figure 5.39: Stimulus 5 in Experiment 2. Crops taken from Philip, Lord Wharton, by Anthony van Dyck. 1632, National Gallery of Art. Crop set 4 in Figure 12 shininess. Crop set 1 in Figure 12 softness

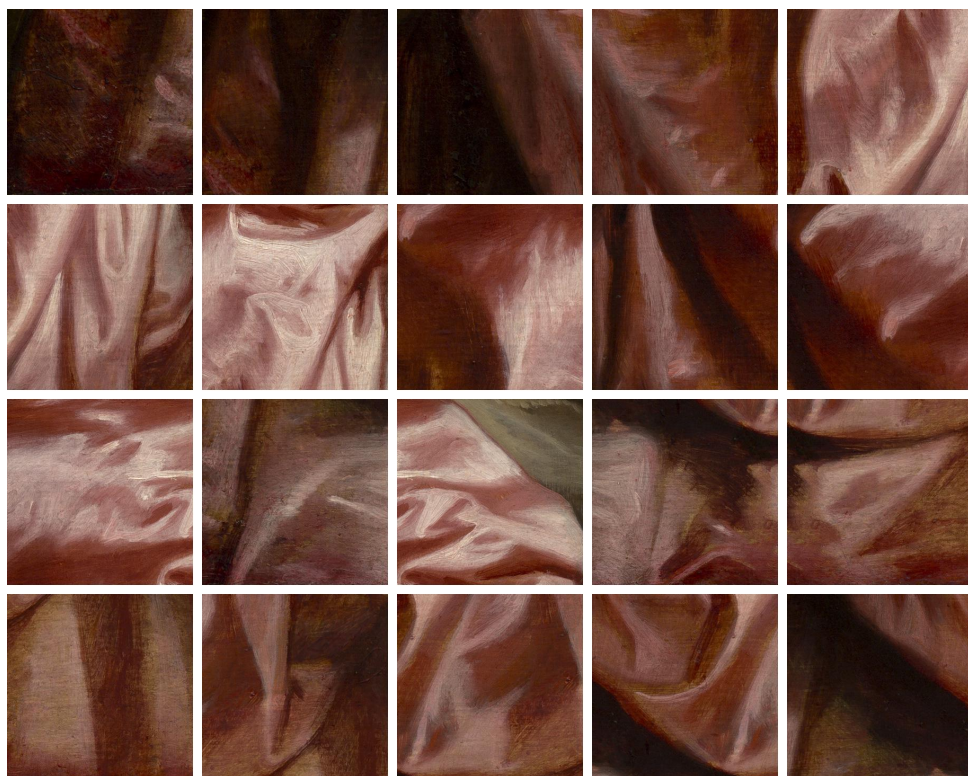


Figure 5.40: Stimulus 6 in Experiment 2. Crops taken from Catherine Howard, Lady d'Aubigny, by Anthony van Dyck. 1638, National Gallery of Art. Crop set 5 in Figure 12 shininess. Crop set 2 in Figure 12 softness

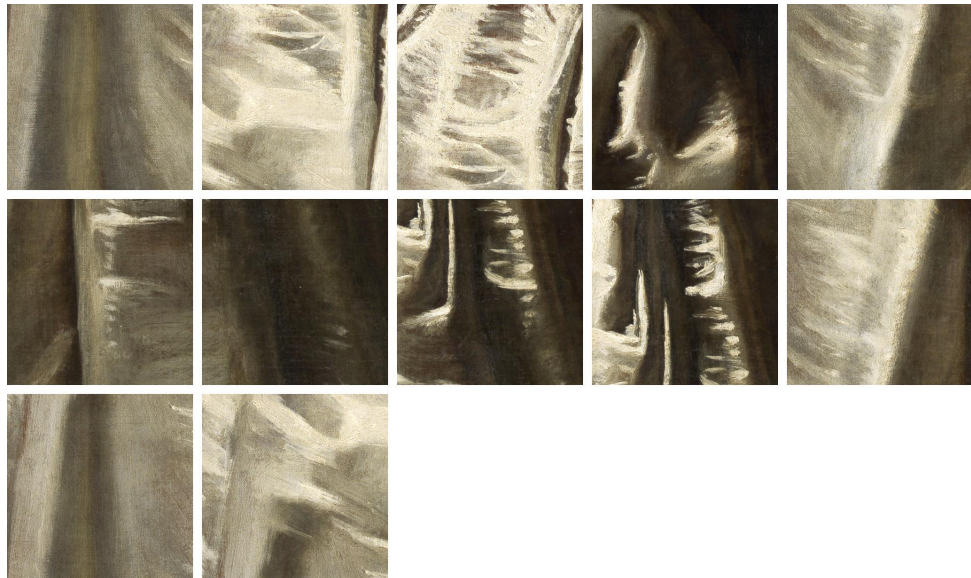


Figure 5.41: Stimulus 7 in Experiment 2. Crops taken from *Marchesa Brigida Spinola Doria*, by Peter Paul Rubens. 1606, National Gallery of Art. Crop set 6 in Figure 12 shininess

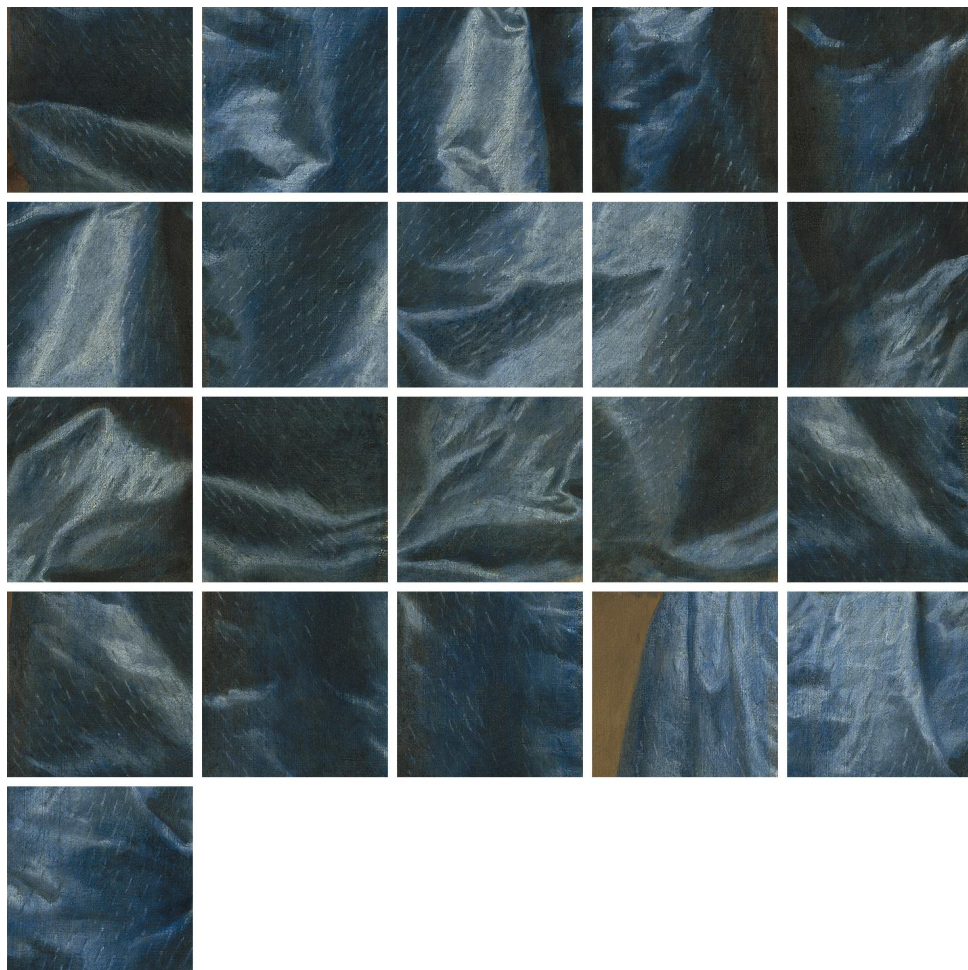
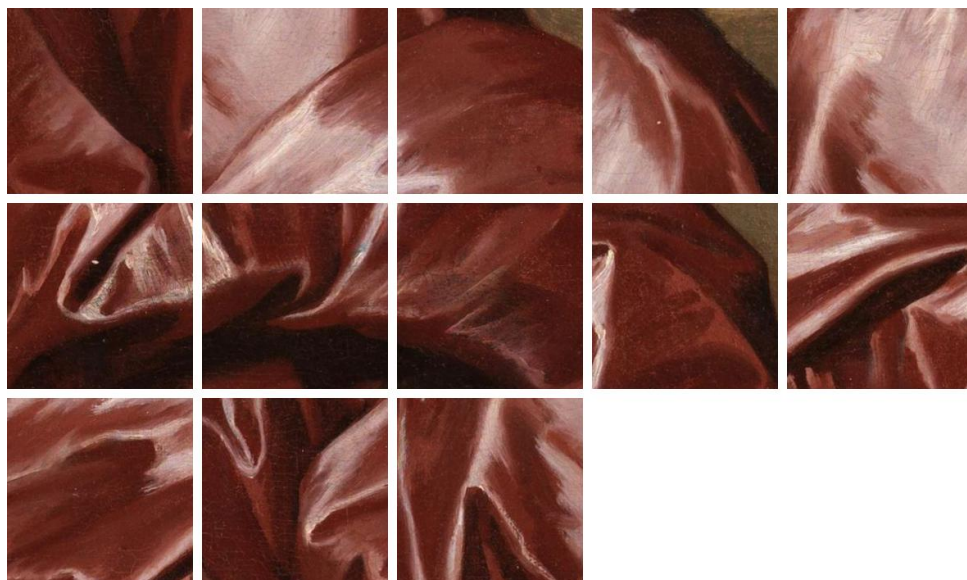


Figure 5.42: Stimulus 8 in Experiment 2. Crops taken from *Queen Henrietta Maria with Sir Jeffrey Hudson*, by Anthony van Dyck. 1633, National Gallery of Art. Crop set 7 in Figure 12 shininess



5

Figure 5.43: Stimulus 9 in Experiment 2. Crops taken from *William II, Prince of Orange, and his Bride, Mary Stuart*, by Anthony van Dyck. 1641. The Rijksmuseum. Crop set 8 in Figure 12 shininess.

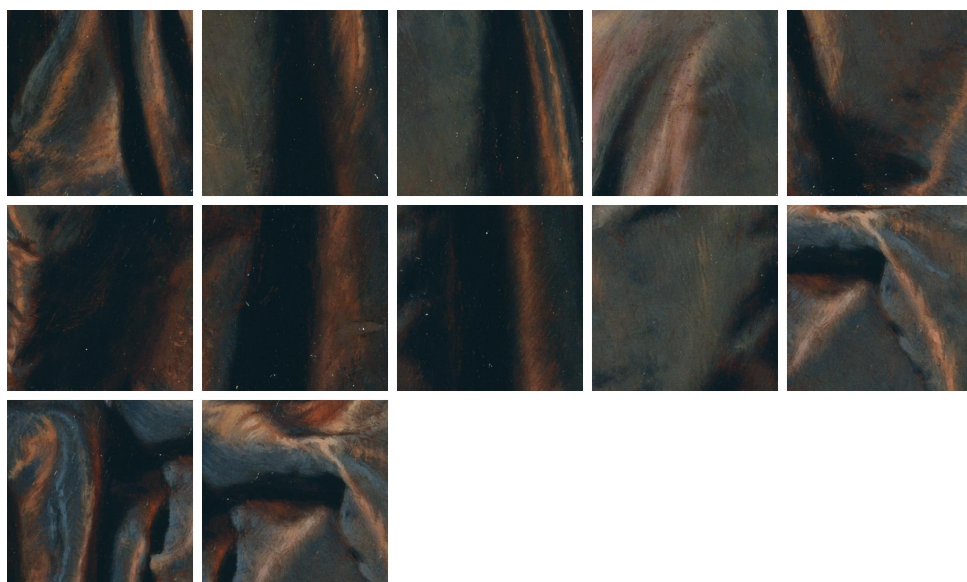


Figure 5.44: Stimulus 10 in Experiment 2. Crops taken from *Pictura (An Allegory of Painting)*, by Frans van Mieris the Elder. 1661, J. Paul Getty Museum. Crop set 9 in Figure 12 shininess.

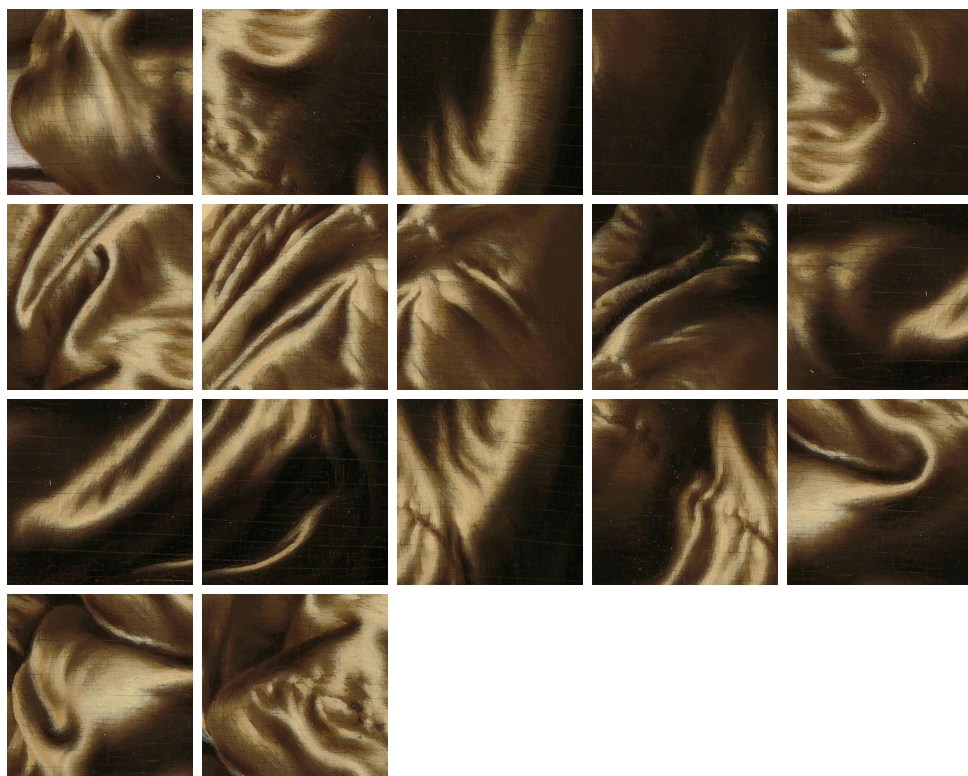


Figure 5.45: Stimulus 11 in Experiment 2 Crops taken from *The Letter Writer*, by Frans van Mieris (I). 1680, The Rijksmuseum. Crop set 10 in Figure 12 shininess.

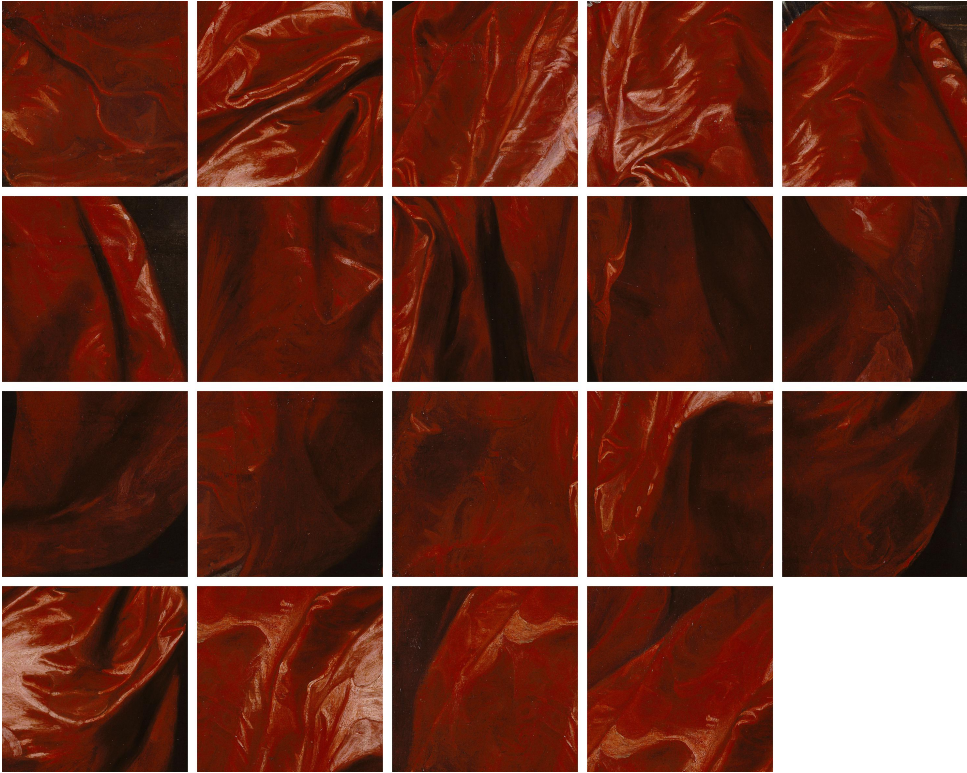


Figure 5.46: Stimulus 12 in Experiment 2. Crops taken from Portrait of Agostino Pallavicini, by Anthony van Dyck. 1621, J. Paul Getty Museum. Crop set 11 in Figure 12 shininess.

Figure 5.47: Stimulus 13 in Experiment 2. A Woman in a Red Jacket feeding a Parrot, by Frans van Mieris the Elder. 1663, The National Gallery of London. Crop set 12 in Figure 12 shininess. Images not reproduced due to copy-rights restrictions.

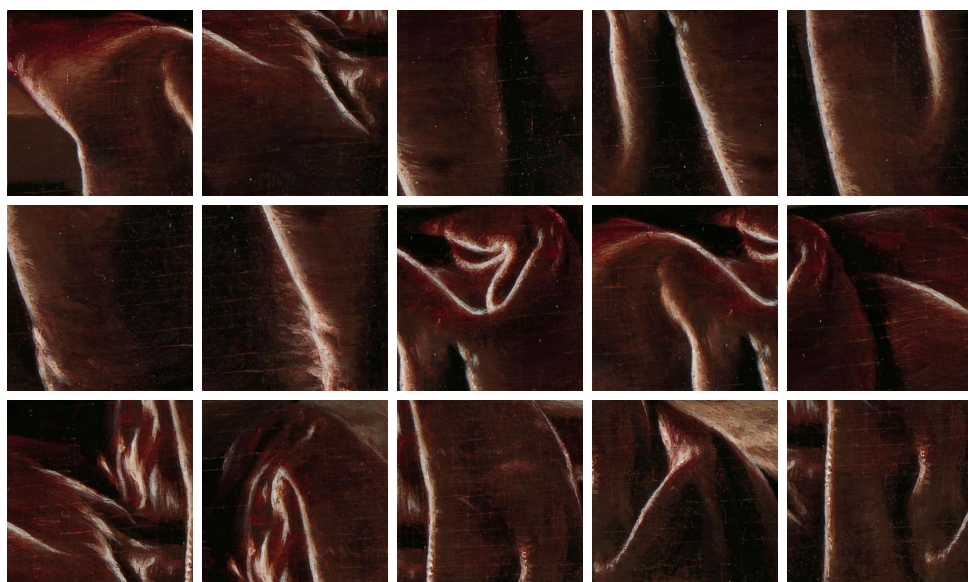


Figure 5.48: Stimulus 14 in Experiment 2. Crops taken from *The Letter Writer*, by Frans van Mieris (I). 1680, The Rijksmuseum. Crop set 13 in Figure 12 shininess.

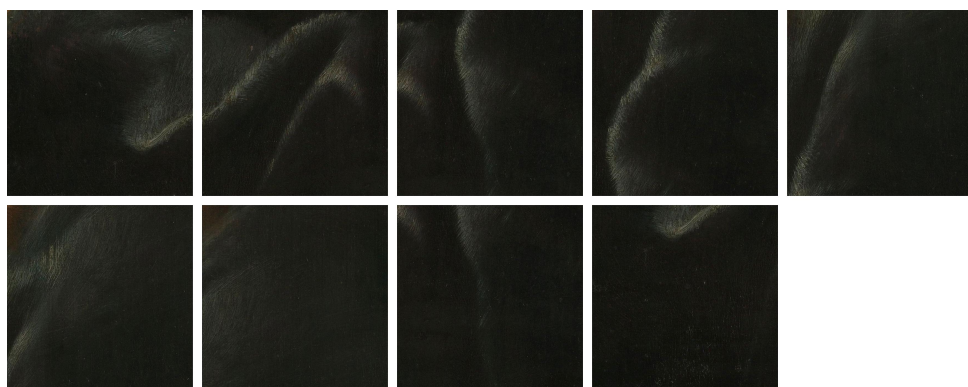


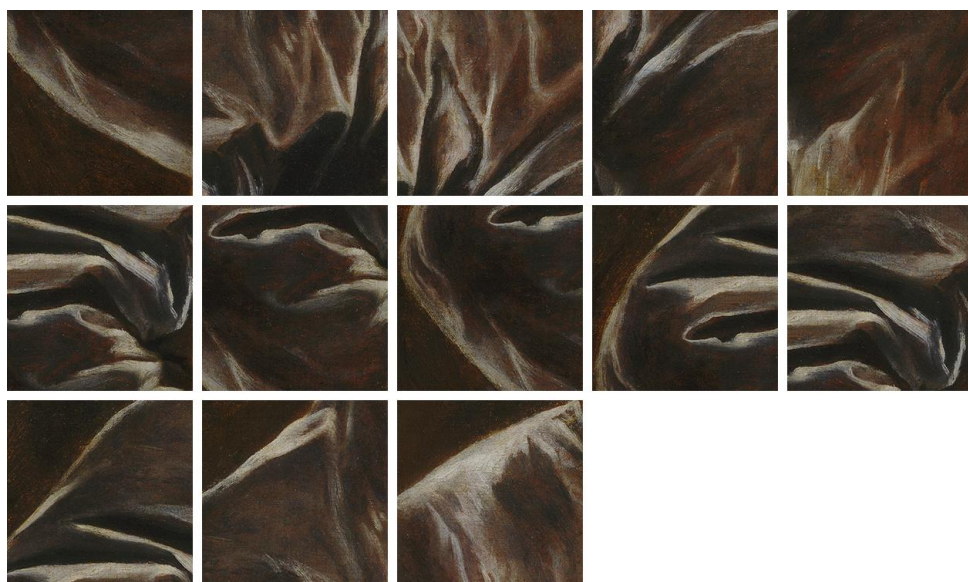
Figure 5.49: Stimulus 15 in Experiment 2. Crops taken from *Old Woman Reading*, by Gerard Dou. 1631, The Rijksmuseum. Crop set 14 in Figure 12 shininess.



5

Figure 5.50: Stimulus 16 in Experiment 2. Crops taken from *An Old Woman Reading, Probably the Prophetess Hannah*, by Rembrandt van Rijn. 1631, The Rijksmuseum. Crop set 15 in Figure 12 shininess.

Figure 5.51: Stimulus 17 in Experiment 2. Crops taken from *Self Portrait of the Artist, with a Cittern*, by Frans van Mieris the Elder. 1674, The National Gallery of London. Crop set 3 in Figure 12 softness. Images not reproduced due to copy-rights restrictions.



5

Figure 5.52: Stimulus 18 in Experiment 2. Crops taken from Philip, Lord Wharton, by Anthony van Dyck. 1632, National Gallery of Art

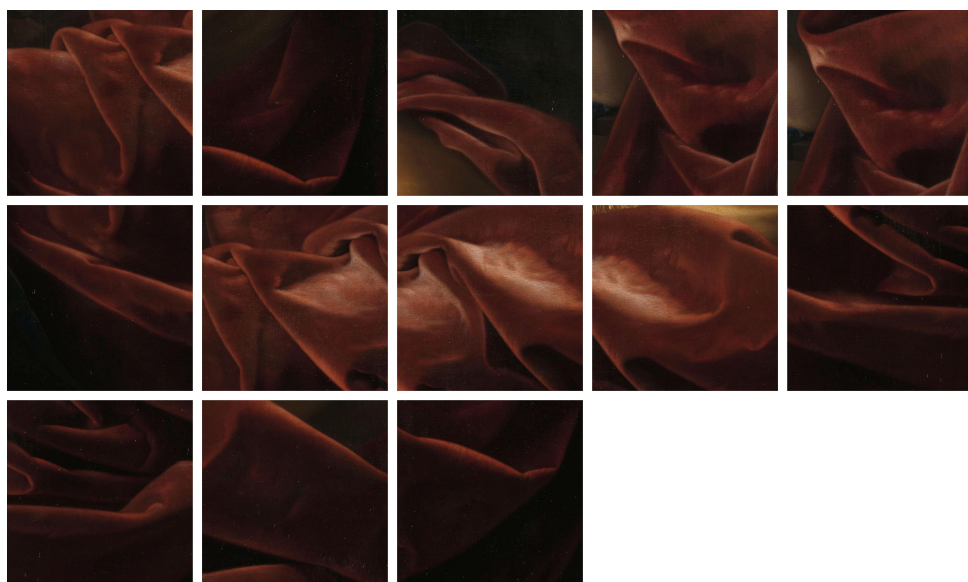


Figure 5.53: Stimulus 19 in Experiment 2. Crops taken from Self-portrait with the Portrait of his Wife, Margaretha van Rees, and their Daughter Maria, by Adriaen van der Werff. 1699, The Rijksmuseum

6

ON PERCEIVING THE LUSTER OF PEARLS. WHEN LESS IS MORE.

The optics of pearls is relatively well understood, however the perception of pearls has received less attention. Here we studied which image features make an object look pearl-like. We identified three image features: 1) a highlight, (2) a dark halo surrounding the highlight and (3) edge reflections. We hypothesized that these are used by the visual system as perceptual cues to estimate the appearance of pearls. We used pearl earrings depicted in paintings because painters identified the image features perceptually-relevant to render materials, including pearls, in a convincing manner. In the first of a series of experiments, we showed that there is a perceptual variability within these pearl stimuli, i.e., some objects appear more pearly relative to others. Next, we digitally enhanced the three image features mentioned above, and showed that this leads to an increase in the perception of pearliness. Next, we isolated each image feature - by digitally removing the other two features - to test the extent to which each isolated feature is responsible for the perception of pearliness both for novice participants and for experts of pearls. Surprisingly, we found that novice participants had a preference for the manipulated version in which the edge reflection and the dark halo were removed, thus with only the highlight, over the original depiction with all three image features. This indicates that the presence of a highlight alone satisfies the perceptual systems requirements to evoke the perception of pearliness for our stimuli, and that the other two optical features appear to interfere with the perception of pearliness. However, experts had a significant preference for depictions containing all three image features, demonstrating that a strong expert effect exists in the perception of pearls.

Submitted as: Mitchell J. P. van Zuijlen*, Francesca Di Cicco*, Pascal Barla, Maarten W. A. Wijntjes, Sylvia C. Pont; On perceiving the luster of pearls. When less is more. * Authors contributed equally

6.1. INTRODUCTION

Unlike most gemstones, pearls do not require human artifice to enhance their beauty, and as such they were likely the earliest gems known to mankind (Kunz and Stevenson, 1908a). The appraisal of the value and quality of pearls depends on a set of factors, like shape, size, and color. But the characteristic for which they are most admired and highly priced, is their luster, a term indicating the different interactions with light that are characteristic of pearls, both on the surface and within the internal layers. Luster is also referred to as brilliance or glow.

The question we aim to answer in this paper is: what makes an object look pearl-like? As previously shown (Di Cicco et al., 2020; Di Cicco et al., 2019; Sayim and Cavanagh, 2011; van Zuijlen et al., 2020; Wijntjes et al., 2020), the key to answer this kind of questions can be found by analyzing paintings and pictorial practices. Painters can deliberately enhance the perceptually-relevant features of pearls' for the sake of rendering their luxurious and lustrous appearance (Cavanagh, 2005; Miller, 1998). By including only the significant visual information, and leaving out what, even though physically correct, would not add to the overall convincingness of its perception, paintings represent an instructive source of information about the mechanisms of material perception. For example, the most famous pearl in art history, depicted in the "Girl with a Pearl Earring" painted by Vermeer (Fig. 6.1), does not look pearl-like and it might not be a pearl at all. Curators at the Mauritshuis (The Netherlands) - where the painting is displayed - have argued that the size of the pearl is "improbably large", and deemed it more likely that the pearl was rather a commonly available and cheap, handcrafted imitation pearl, made of metal or glass and then lacquered ("The Girl Without a Pearl Earring?", 2017). Beside the shape, the appearance itself of the pearl is not convincing. For example, Icke (1987) pointed out that the look is too metallic to be a pearl. Understanding which cues determine the appearance of pearls can thus also help art historians and conservators with an objective and consistent method for the identification of the depicted materials.

Why are pearls lustrous? Like every other material property, luster can be treated both as a physical and a perceptual issue (Mondonneix et al., 2017). While the optics of pearls is well understood and has been thoroughly studied (Landman et al., 2001; Raman and Krishnamurti, 1954a, 1954b), their perception remains largely unexplored.

The physical explanation of the luster of pearls is to be found in their structure and optical behaviour. Natural pearls are formed by mollusks such as oysters and clams, which secrete shell material as a response to damages to their tissue, caused by the intrusion of an irritant (Taylor and Strack, 2008). The shell material is made mostly of calcium carbonate, also called nacre or mother of pearl, interlaced by an organic membrane of conchiolin, and deposited around the nucleus in concentric layers (Landman et al., 2001). The nacre consists of an ordered structure of thin, parallel, crystal sheets (Rouhana and Stommel, 2020). The thickness of the translucent nacre, and its arrangement with the opaque conchiolin, determine the lustrous and iridescent appearance of pearls (Landman et al., 2001). When light hits a pearl, it is partly reflected off the surface and partly absorbed and refracted via thin-film interference, through the translucent nacre layers (Raman and Krishnamurti, 1954a, 1954b). The scattering of light follows an anisotropic path along the lateral direction, parallel to the nacre plates rather than perpendicularly (Raman and Krishnamurti, 1954a, 1954b).



Figure 6.1: Johannes Vermeer, *Girl with a Pearl Earring*, c. 1665. Mauritshuis, The Hague, The Netherlands.

In their comprehensive discussion of the optical behaviour of pearls, Raman and Krishnamurti (1954a, 1954b) identified three main optical phenomena: the reflection and diffraction of light creating the focused image of the light source, a darker diffusion halo surrounding this reflection, and a general, lateral diffusion of light reaching the highest intensity along the periphery of the pearl, termed "whispering gallery effect".

The optical behaviour of pearls described by Raman and Krishnamurti (1954a, 1954b) was measured under collimated lighting conditions in a controlled lab environment. However, as for any other material, pearls' appearance depends on the combination of object shape, optical properties and illumination conditions (Fleming et al., 2003; Marlow et al., 2012; Olkkonen and Brainard, 2010; Pont and te Pas, 2006). Given that the shape of pearls is mostly spherical, the visibility of the optical phenomena identified by Raman and Krishnamurti (1954a, 1954b) will depend on the illumination environment under which the pearls are observed. More ecologically valid lighting conditions can cause the emergence of different features and thus a wider range of appearances of pearls.

While lighting conditions within paintings are often canonical, with strong conventions for lighting coming from the top-left (Carbon and Pastukhov, 2018; Di Cicco et al., 2019; Mamassian, 2008; Van Zuijlen et al., 2021; Wijntjes, 2021), they nevertheless contain more variability than a single collimated light source.

Through observation of a large selection of painted pearls we identified three distinct image features that appear consistently across paintings of pearls. These image features are the highlight, the dark halo around the highlight, and the edge reflection. A pearl depicted

within a painting that exemplifies all three image features is presented in Fig. 6.2. We hypothesized that these three image features are perceptual cues for the visual perception of pearls.



Figure 6.2: A section from the painting "Portret van Adriana Croes" by Johannes Cornelisz. Verspronck, 1644, The Rijksmuseum, which exemplifies all three image features that we hypothesized are used by the visual system as perceptual cues. First, the highlight in the top-left region of the pearl. Next, in the middle of the pearl, bordering the highlight is a dark halo. Lastly, the reflection along the edges of the pearl.

A number of perceptual studies have researched the first of these three image features, namely the highlight, but not in relation to the perception of pearls' appearance. The role and constraints of the highlights have been the object of thorough investigations, with the main aim of understanding gloss perception (Chadwick and Kentridge, 2015). The presence of even simple shaped highlights (e.g. from a single-point light source) on the surface, is enough to trigger the perception of glossiness (Beck and Prazdny, 1981), even though the estimation of reflectance properties improves under more complex, real-world illumination conditions (Fleming et al., 2003). This is probably due to the increased coverage of highlights, one of the image cues proposed to elicit gloss perception, together with contrast and sharpness (Marlow and Anderson, 2013; Marlow et al., 2012). For the bright spots

on the surface to be perceived as specular reflections, congruency constraints apply. Thus, their orientation and brightness need to be congruent with the 3D shape of the object and its shading profile (Kim et al., 2011; Koenderink and van Doorn, 1980; Marlow et al., 2011), unless one considers a cluster of objects oriented differently, like a bunch of grapes, then the rule of orientation congruence can be broken without hindering the perception of glossiness (Di Cicco et al., 2019).

The perceptual role of the dark halo surrounding the highlight, has never been researched to the best of our knowledge, probably because it is an optical phenomenon peculiar to pearls. Fu et al. (2014) conducted microstructural analysis of real pearls to determine the origin of this optical effect. They attributed it to the heterogeneous nature of the reflecting layers of the pearl, made of crystals of aragonite embedded in the organic matrix. While resulting from a different optical phenomenon, the locally darker luminance extrema of the dark halo might be visually compared to the lowlights, which have been shown to contribute and even trigger a convincing perception of glossiness in the absence of highlights (Kim et al., 2012).

Finally, the edge reflections are created by reflections of the surroundings along the edges of the pearl's surface.

To further illustrate these three features, we photographed two real pearls using studio lighting set up (Fig. 6.3). All the images were illuminated from above. In the two images on the left, the pearls were placed against a dark background. The lighting condition created a clear highlight and some glow along the edges. Note that this glow is due to the lateral scattering of light parallel to the nacre layers, an optical phenomenon also described by Raman and Krishnamurti (1954a, 1954b), and it is different from what we referred to here as edge reflection. In the images in the middle and in the right columns of Fig. 6.3, we added a white background, first only behind the pearl (middle) then also below (right). Such light background could be clothing or human skin within paintings (such as in Fig. 6.2), and it is reflected along the edges of the pearl, creating the edge reflection. The addition of the white background below the pearl causing the bottom reflection, also delineates the dark region around the highlight, which we referred to as dark halo.

It has been shown that humans are incredibly good at categorizing materials into different classes with high accuracy (Fleming et al., 2013), and at telling apart real from fake materials, even within a brief presentation time (Sharan et al., 2009, 2014). To understand how humans can categorize materials, the next step in research is to focus on the cues that trigger the perception of material classes. For example, Todd and Norman have studied the role of reflections on the perception of glass (Todd and Norman, 2019, 2020) and on metal (Todd and Norman, 2018). Schmid et al. (2020) hypothesized that discriminating between material classes is the primary aim of our ability to distinguish between gloss levels, based on the image structure of the specular reflections. They identified a set of image features of the specular reflections, i.e. both highlights and lowlights, and showed that these can be manipulated to evoke robust perception of different material classes, including pearlescent materials.

Understanding the image features employed by the visual system to judge the appearance of pearls is not only a fundamental question about the functioning of the visual system, it also has practical applications. For example, it would help to clarify and optimize the evaluation criteria used by experts during quality control through computer graphics



Figure 6.3: Two sets of photographs of real pearls showing the three image features that we hypothesized are used by the visual system as perceptual cues for the visual perception of pearliness. The pearls in all images are illuminated from above. The images on the left were taken against a black background to capture the highlight in isolation. In the images in the middle and on the right, a white background was added only behind (middle) and both behind and below (right) the pearls. The reflection of the white background at the bottom of the pearls causes the edge reflection and delineates the dark halo around the highlight.

simulations (Nagata et al., 1997), or switch from the time-consuming judgement prone to subjective biases of human experts to computer vision assessment of pearls' quality (Mon-donneix et al., 2017).

In this study, we are interested in the image features that serve as perceptual cues to identify pearls depicted in paintings. As stimuli for this study, we exclusively used pearls earrings, which are commonly depicted within paintings. In this way, each stimulus has a relatively similar context, i.e., a human face. We first experimentally demonstrated the intuitive finding that the perception of pearls is not simply uniform: some pearls are more pearl-like than others. This of course leads to the main question: What makes a pearl look more pearl-like? Therefore, we manipulated the three pearl image features - highlights, dark halo, edge reflections - first by digitally enhancing them, and then by deleting two of the three features in turn, to study the extent of their role as perceptual cues for the perception of pearliness.

6.1.1. OVERVIEW OF THE EXPERIMENTS

In Experiment 1, we tested the extent to which participants perceived the pearls depicted in earrings to look pearl-like. The hypothesis behind this experiment was that not all depictions were equally successful in rendering a convincing appearance of pearls. Thus, the aim of the experiment was to test whether the painted pearls indeed varied in how much they looked like a pearl.

In Experiment 2, we digitally enhanced the three image features under investigation in this study - highlight, dark halo and edge reflection - using a self-developed image editing process. Participants were asked to choose the most pearl-like between the original and the enhanced version of each pearl, with a two-alternative forced choice experiment. In this experiment we tested whether the pearls would be consistently chosen to look more pearl-like, by enhancing the image features that we expected to trigger the appearance of pearls.

In Experiment 3, we tested the causal relationship between each of the three image features and the appearance of pearls, by digitally deleting two of the three features. Participants chose with a categorization task, to what extent the three manipulated versions of the stimuli as well as the original stimuli looked like pearl, glass, metal or stone.

In Experiment 4, we performed a follow-up of the results in Experiment 3. Two groups of participants, experts of pearls and novices, performed a two-alternative forced choice experiment to choose the most pearl-like between the original version and the highlight-only version. We found that novices and experts attended to different features in their judgments, and that experts chose consistently the original stimuli, whereas novices preferred the highlight-only manipulations.

6.2. EXPERIMENT 1

6.2.1. PARTICIPANTS AND STIMULI

All participants in the studies reported within this paper were naive to the purpose of the experiments. Each participant agreed with the informed consent prior to the experiment and received compensation for their participation. The experiments adhered to the tenets of the Declaration of Helsinki and were approved by the Human Research Ethics Committee of the Delft University of Technology.

For all the experiments (except for a part of Experiment 4), participants were recruited through Amazon Mechanical Turk (AMT). AMT is an online platform for crowd-sourced data gathering that can be used to rapidly obtain high-quality data (Buhrmester et al., 2011). Using the AMT system of qualifications, we restrict the tasks to participants that have completed at least 1000 tasks of which at least 95% were accepted (Peer et al., 2014). Furthermore, we added the qualification that required participants to be located within the USA, with the aim that participants would have a sufficient knowledge of the English language to understand the instructions. These three qualifications apply to every experiment reported in this paper that was performed through AMT.

Besides the 'qualifications' that restrict participants from starting a task, we also defined a number of exclusion criteria for participants based on their performance in these tasks. For the first experiment we excluded participants that either (1) failed an attention check (Aust et al., 2013; Oppenheimer et al., 2009) i.e., in the instructions participants were explicitly

instructed to write the word ‘attention’ in the comment section at the end of the experiment, or (2) participants that responded too fast by having a median trial time under 1 seconds or (3) by having an intra-rater agreement below a cut-off value of 0.3. Here, the intra-rater agreement was calculated as the averaged Pearson correlation across the three repetitions of the trials. The cut-off value of 0.3 was determined using a look-up table for Pearson correlation’s critical values for a one-tailed distribution at an alpha of .005, which is 0.267 when n equals 90. We rounded the critical value (i.e., 0.267) up to 0.3, which makes our cut-off slightly more conservative.

In Experiment 1, we collected the data from 40 participants. Using the exclusion criteria described above, we removed a total of 31 participants, the majority of which ($n = 29$) were removed due to failing the attention check and a further 2 were removed due to failing the intra-rater agreement criterion. Of the participants that failed the attention-check, 14 also failed the time criterion, i.e., they responded too fast.

We analyzed the data from the 9 remaining participants, who rated 91 sections of paintings that contained a pearl earring. Stimuli were selected from the Materials In Paintings (MIP) database (Van Zuijlen et al., 2021, <https://materialsinpaintings.io.tudelft.nl/>). The size of the depicted pearls varied between paintings. A simple, but limited way to control for this would be to standardize the size of the pearls, i.e., to resize each painting so that the pearls have the same size across all paintings. However, real-world pearls are not all the same size and therefore standardizing in this way would also remove a portion of the variability which the painter might have explicitly chosen to include. This might furthermore impact the scale and content of the segmented context drastically, which in turn might affect the data as a confounding factor. Instead, we chose to standardize the size of pearls relative to something that is consistent in size in the real world and that is present across all the paintings in our set. The iris of the eye of the figures wearing the pearl earrings fits these requirements. One of the researchers annotated the size of the iris for the figures wearing the pearls, which was then used to resize the paintings. Next, we manually created a segment of 200 x 200 pixels for each painting, with the pearl roughly in the middle. A few examples can be seen in Fig. 6.4 and the full set can be found in the supplementary materials.



Figure 6.4: Examples of pearl earring stimuli used in Experiment 1. Note the visible variability in pearl sizes that was maintained by standardizing the size of the original painting to the size of the iris. The attribution for each painting can be found in the section.

6.2.2. PROCEDURE

Before starting the rating task, we asked participants to calibrate the presentation size of the stimuli using the method described in (Di Cicco et al., 2021, *in press.*), where users matched the size of a digital credit card on their screen to a physical credit card in their possession. In this way we could control for the varying display sizes, and therefore ensure a consistent presentation size across participants. After this calibration, participants performed a free-viewing task of 10 seconds of 18 stimuli randomly selected from the stimulus set, to get an idea of the range of the stimuli. Finally, participants performed five randomly selected practice trials before starting the real task.

Participants were explicitly instructed that each stimulus would contain a pearl earring. They were asked to rate “to what extent” the pearl “looks like a pearl”, using a continuous scale ranging from “Does not look pearl-like” to “Looks very pearl-like”. At the end of the experiment, participants were asked to answer the question “What did you base your judgements on?”. Each stimulus was rated three times, for a total of 273 trials per observer. The trials were randomized across participants.

6.2.3. RESULTS AND DISCUSSION

In Experiment 1, participants rated the perceived pearliness of 91 stimuli, three times per stimulus. The intra-rater agreement, calculated as the average Pearson rank order correlation across the three repetitions of each participant, was 0.54. Note however that this value is slightly inflated due to the exclusion of participants scoring below 0.3. Next, we looked at the consistency between participants by calculating the intraclass correlation coefficient (ICC). ICC estimates and their 95% confidence intervals were calculated using the pingouin package for Python, based on a mean-rating ($k = 3$), consistency, two-way random-effects model (ICC3k; (Koo and Li, 2016; McManus and Humphrey, 1973; Shrout and Fleiss, 1979). The ICC was estimated to be 0.78, in a 95% confidence interval between 0.7 and 0.84, which was found to be significantly different from 0 ($F(90,810) = 4.49, p < .0001$). The mean perceived pearliness spans a perceptually distinguishable range from 0.11 to 0.82.

Taken together, these results show that participants had a robust and non-random perception of pearliness that was agreed upon between participants. In Fig. 6.5 we visualized the mean ratings across all participants, where we took the median of each participants over the three repetitions, and we showed the two least and the two most “pearly” pearls in our stimuli set. The full range of stimuli ordered on pearliness can be found in the supplementary materials Fig. 6.12.

The results of Experiment 1 indicate that the perception of pearliness can vary across painted depictions of pearls. Such variation in the appearance of pearliness may be caused by differences in the features of the painted pearls. This raised the follow-up question: which are these features? As discussed earlier, we hypothesized that three perceptual features are used by the visual system to trigger the perception of pearliness - highlight, dark halo and edge reflection. On observation of our stimuli, we found these three image features to be present in all the stimuli of our set, but with different weight combinations depending on the artist. We tested our hypothesis in the following experiments.

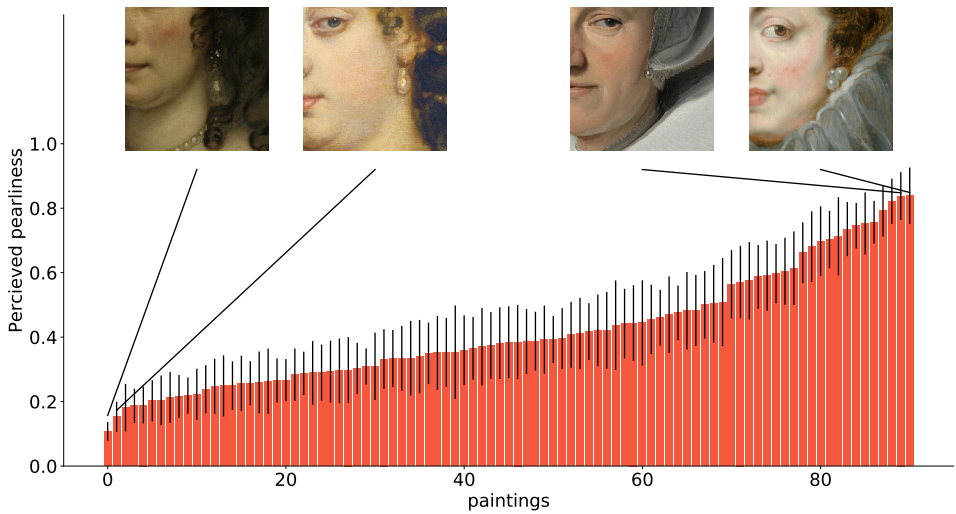


Figure 6.5: The average ratings across all participants, where we take the median rating per participant, for each painting. Error bars indicate the standard error. Higher ratings indicate that the pearl earring was rated as perceptually more pearly, and the opposite for lower values. The two least pearly, as well as the two most pearly pearls are visualized. The attribution for each painting can be found in the section.

6.3. EXPERIMENT 2

In Experiment 2, we tested our hypothesis that the three images features (highlights, dark halo and edge reflections) are cues that trigger the perception of pearliness. We enhanced the intensity of these three image cues in the 91 stimuli used in Experiment 1, employing a self-developed image-editing process. The manipulated stimuli, together with their original version, were used in a two-alternative forced choice test. According to our hypothesis, the manipulated depiction should be perceived as more pearly relative to the original stimuli.

6.3.1. PARTICIPANTS AND STIMULI

In Experiment 2 we compared the perceptions of the original paintings to the perceptions of their Pearliness Enhancement Transform (PET) version. The painting segments we tested and manipulated were the same 91 stimuli used in Experiment 1. The self-developed PET algorithm, inspired by Wetness enhancing transformation (WET) (Sawayama et al., 2017), consists of three mostly automatic image manipulations. First, we increased the luminance of the whole pearl (both light and dark tones). In the second manipulation, the visibility of the dark halo and the edge reflection are enhanced by creating a mask that adds dark tones to the dark halo and bright tones to the edge reflection; the mask is then blurred to create smooth transitions between these added image features. Finally, a glow filter is applied to the pearl image region in order to reproduce the visual effect of pearl translucency. Details of the step-by-step process of PET manipulation can be found in section 6.9 in supplementary material. Examples of the original version of the pearls in the earrings and their PET version, are shown in Fig. 6.6. The complete list of PET manipulated stimuli can be found in section 6.9 in the supplementary material.

The task was originally performed by 20 participants, recruited on AMT. Identical to the previous experiment, we used an attention check and data from participants that failed the attention check were not analyzed. A total of 10 participants were removed from the task due to the attention check, and the data for the remaining 10 participants were analyzed.

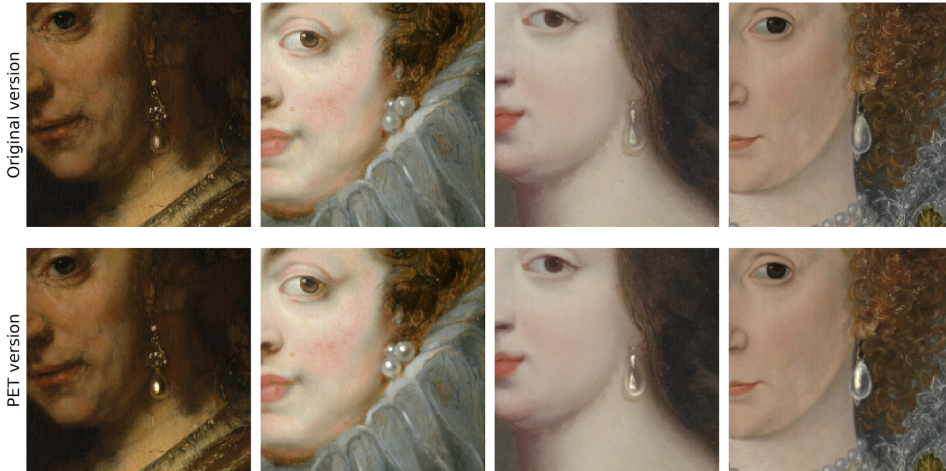


Figure 6.6: Examples of the original stimuli (top row) and their manipulated PET versions (bottom row). The attribution for each painting can be found in the section.

6.3.2. PROCEDURE

Identical to Experiment 1, participants performed a size-calibration task before starting the experiment. In Experiment 2 we conducted a two-alternative forced choice (2AFC) test. The original and the manipulated versions of the stimuli were presented side-by-side. The location of each version (i.e., left or right presentation) was randomized across trials. Participants were informed that they would be presented with “two versions of a painting”, and they were asked to use the mouse to click the image that they perceived to “contain the most pearl-like pearl”. Each of the 91 pairs of stimuli was presented three times, for a total of 273 trials.

6.3.3. RESULTS AND DISCUSSION

Fig. 6.7 shows the proportion of which version of the stimuli was perceived to be more pearly, averaged across participants for each stimulus, for which we took the median over the three repetitions. Across participants, there was a strong tendency to perceive the manipulated version as more pearly. A 1-sample t-test confirms that the difference between the PET and the original version was significant ($t(90) = 37.6, p < .0001$).

We hypothesised that highlight, edge reflection and the dark halo are the perceptual cues that trigger the perception of pearliness. Our findings show that enhancing the visibility of these image features leads to an increase in the perception of pearliness. We have shown that a causal relationship exists between at least one of these features and the in-

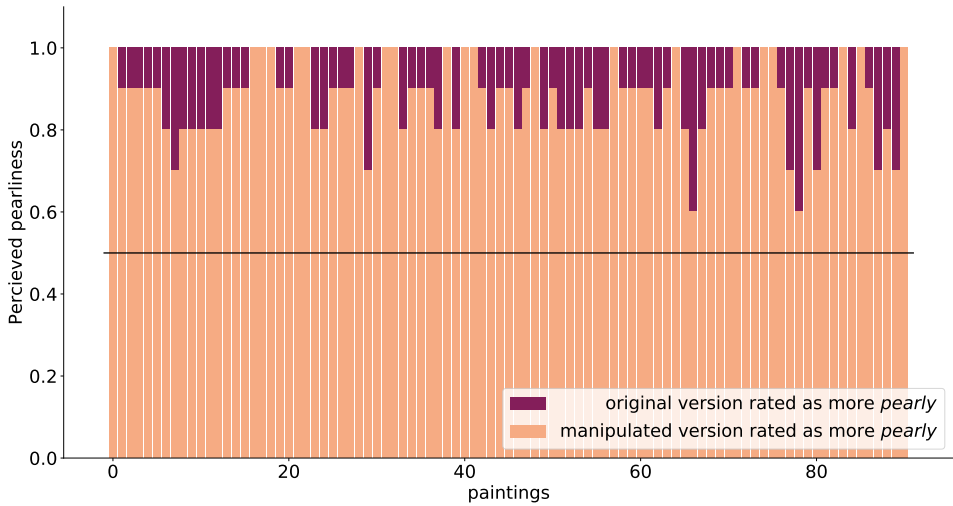


Figure 6.7: The proportion of which version of the stimuli was chosen to look more pearly by participants. A value of 1, which makes a full orange bar, such as for the most left-wards stimuli, implies that all the participants rated the manipulated version as being more pearly. For each participant, the median of the three repetitions was used.

6

creased perception of pearlyness. However, in Experiment 2 the contributions of the three image features were conflated. Thus, in order to measure the relative contribution of each feature to the perception of pearlyness, in the next experiment we isolated them via image manipulation of the original versions.

6.4. EXPERIMENT 3

In Experiment 3, we conducted a material categorization task inspired by Todd and Norman (2019), to test if the image features identified in this study triggered a convincing perception of pearlyness when seen in isolation, or if they rather elicited the appearance of different materials, like metal, glass, and stone.

6.4.1. PARTICIPANTS AND STIMULI

To create the stimuli for Experiment 3, we selected the 15 images that were perceived to be most pearly on average in Experiment 1. Next, we created three alternative versions for each image, where each version only contained one of the three image features. The aim of this was to test the causal relationship between the individual image features and the pearly appearance. That is why we tested only the images that were judged as perceptually most pearly, and left out the ones that perhaps were already perceived to look more like something else.

The image manipulations were done in Gimp version 2.10.22, using the Resynthesizer plugin. We first upscaled the image from 200 x 200 pixels to 800 x 800 pixels. Then, we applied the Heal Selection tool, which is part of the plugin, using a sampling width of 20 pixels in order to erase two of the three image features in turn, and create versions of the

stimuli that contained either only the highlight, only the edge reflection or only the dark halo (see Fig. 6.8). In the last step, we downscaled the images back to their original size. The 15 original stimuli and their three manipulated versions can be found in section 6.9 in the supplementary material. Note that for the edge reflection only version of some stimuli, the edge reflection might appear attenuated compared to the original version. The reason of such perceptual effect is the lack of contrast created by the absence of the dark halo.

A total of 30 participants were recruited on AMT to perform this experiment. The exclusion criterion was the attention check that was also used in the previous two experiments, which led to the exclusion of 13 participants. The data from the remaining 17 participants was analyzed and is shown below.



Figure 6.8: Examples of the four different versions of the stimuli used in Experiment 3. On the left, two examples of the original stimuli are presented. These are identical to the stimuli used in Experiment 1 and 2. In the second column, the highlight only version is shown, in which the dark halo and edge reflection were manually removed. In the third column, only the dark halo remains visible, and in the last column only the edge reflection is visible. The attribution for each painting can be found in the section.

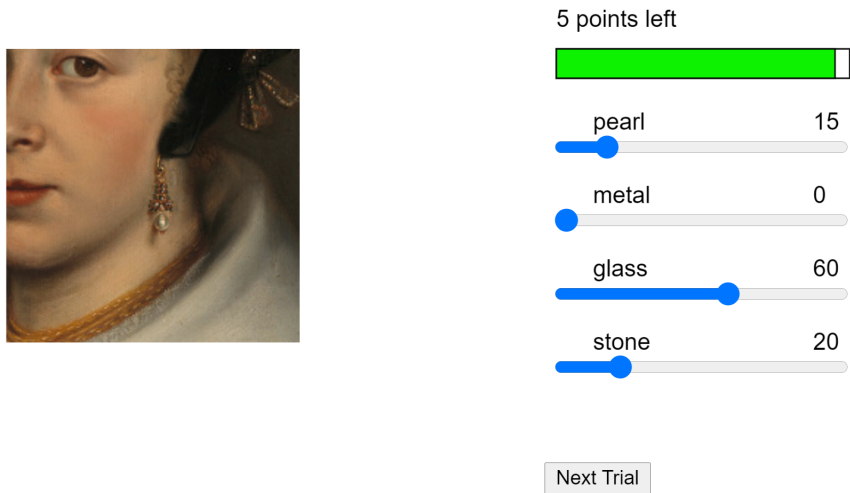
6.4.2. PROCEDURE

Before starting the experiment, participants performed the size-calibration and the free-viewing tasks as described for Experiment 1 and 2. Participants were instructed that they would be presented with sections of paintings containing an earring. They were further instructed that “because the earrings come from paintings, it is not always clear what the earring is made out of”, and that the earrings “can look like a combination of” glass, metal or stone. Participants were then asked to categorize the appearance of each earring by assigning 100 points across four material categories: pearl, metal, glass, and stone. For example, if an earring looked completely pearl-like, a participant could assign all 100 points to the pearl attribute. Participants could continue to the next trial only after all 100 points were assigned. Points were always assigned in multiples of 5, which we intended to make the task less tedious. This is mathematically identical to asking participants to assign 20

points in multiples of 1, but the authors found that a total of 100 points was conceptually easier. See Fig. 6.9 for an example of the interface used in this task. Observers were presented with one stimulus at a time to categorize. There were three repetitions for each stimulus, for a total of 180 trials. The trials were randomized across participants.

To what extent does the earring look like the four materials?

Assign the points to the materials relative to how much the earring represents that category



Trial 0 out of 5 practice trials

Figure 6.9: Example of the task in Experiment 5. Participants were required to assign all points (in multiples of 5) to the extent in which they perceived the earring to look like the materials reported above the sliders. The participants were only able to progress once all 100 points were assigned. Therefore in this case of this example, the participants is required to assign 5 more points. The green bar at the top serves as a visual indication of the remaining points and displayed a linearly interpolated color between red (100 points left) and green (0 points left). The attribution for the painting can be found below the references.

6.4.3. RESULTS AND DISCUSSION

Participants categorized four versions of 15 pearl earrings, and for each version they indicated to what extent they perceived the earring to be made out of pearl, metal, glass, and stone. The mean ratings across participants and the significant differences between categories per stimuli version are shown in Fig. 6.10. The original version and the highlight only version of the stimuli, were categorized to look significantly more pearl-like than metal, glass or stone-like. On the contrary, the versions of the stimuli that only showed the dark halo or the edge reflection, were categorized similarly as pearl or stone, and pearl, stone or glass respectively.

These results appear to indicate that the highlight is the most salient feature to trigger pearliness perception. When comparing between versions, it is somewhat surprising to see that the categories distribution for the highlight-only version is so similar to the original version, which seems to imply that the dark halo and edge reflection do not contribute to

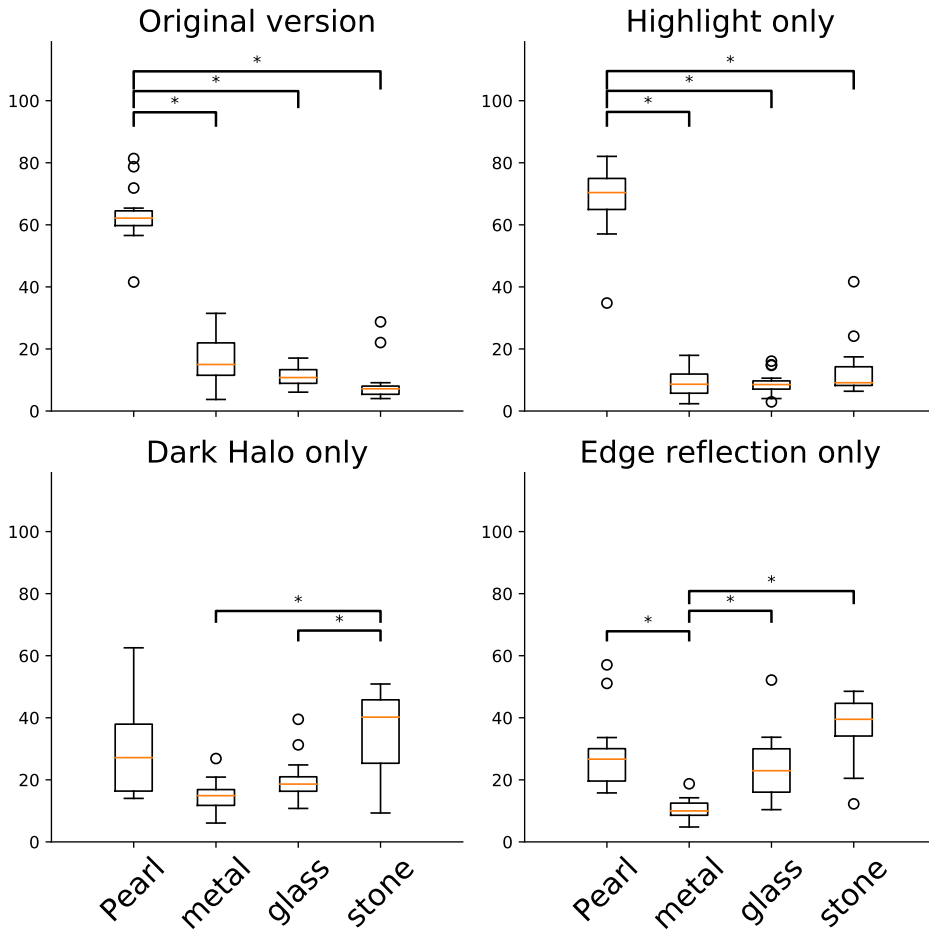


Figure 6.10: Boxplot distributions for the ratings for each of the four versions of the pearl earrings. The box extends from the upper and lower quartiles and the middle lines indicate the median. Outliers more than 1.5 times the interquartile range are plotted as points. In the original version (top-left) the stimuli were not manipulated. In the other three versions, only one of the image features remained, while the other two image features were digitally removed. Significant differences at Holm–Šidák adjusted p-values of 0.05 are indicated with an asterisks (*).

the perception of pearliness. Note that ‘highlight-only’ refers to the manipulation in which the other two images features were removed. When directly comparing the original and highlight-only version, it appears that the highlight-only version was rated as more pearly relative to the original version, which contains all image features. As Fig. 6.10 only contains significance tests per version, we performed one extra dependent t-test for paired samples which showed that indeed, the highlight only version was rated as significantly more pearly, $t = -3.89$, $p < .005$.

The importance of the highlight for pearls and the resulting change in material cate-

gorization when the highlight was removed, were not surprising results. What we did not expect though, was that the version of the stimuli in which we isolated the highlight was judged as pearl significantly more than the original version having all three image features. One influencing factor could be the level of experience with the appearance of real pearls of our participants. It is probable that the vast majority of our participants were no experts of pearls nor were they often exposed to real pearls. In Tani et al. (2014), it was shown that when asked to visually rank real pearls based on quality, novices and experts attend to different features to judge pearls' quality (Tani et al., 2014). One can assume that this difference between novice and experts extends to the judgement of pearliness for pearls depicted within paintings. Possibly pearl experts would disagree with our novice participants that judged the highlight-only version as more pearl-like than the original version. Therefore we conducted the next experiment, in which we compared the perception of the pearl-like appearance for the original and the highlight-only versions, by expert and novice observers.

6.5. EXPERIMENT 4

The previous experiment showed that the stimuli having only the highlight were categorized as pearls more often than the original stimuli, suggesting that the dark halo and the edge reflection did not contribute to the perception of pearliness, or even suppressed the perception thereof when shown in isolation. In Experiment 4, we tested whether experience with real pearls affects the perception of the pearls. We used a two-alternative forced choice paradigm, identical to Experiment 2, to check the perceptual difference between the original version containing all image features and the highlight-only version, i.e., the version in which the black halo and the edge reflect were removed.

6.5.1. PARTICIPANTS AND STIMULI

We collected data from two groups of participants: novices and experts. The novice participants ($n = 30$) were recruited on the AMT platform, identical to the previous experiments. The expert participants ($n = 5$) consisted of professional pearl appraisers and art historians. The experts were approached by email and performed the experiment in identical settings as the AMT participants. All participants were naive to the purpose of the experiment, they were only informed that we were “interested in the perception of pearls.”

For the experts, we did not use any exclusion criteria. For the novice participants, we used the same exclusion criterion as reported for Experiment 2, i.e., the attention check. In this case it led to the removal of 21 novice participants.

The stimuli were the 15 original versions of the paintings and the 15 highlight-only versions used in Experiment 3. Two examples of the stimuli are shown in the first and second columns of Fig. 6.8.

6.5.2. PROCEDURE

The task was identical to the two-alternative forced choice experiment described in Experiment 2, executed with the same procedure. Each trial was repeated three times, for a total of 45 trials.

6.5.3. RESULTS AND DISCUSSION

Fig. 6.11 shows the number of times that each version of the stimuli was chosen to appear more pearly by novices (top) and experts (bottom). For the novice participants, it shows that the highlight-only version was consistently chosen as looking more pearl-like than the original, which was confirmed by a 1-sample t-test ($t(14) = 9.8, p < .0001$). Experts, on the other hand, had a strong preference for the original depiction of the pearl, $t(14) = -9.99, p < .0001$. It is interesting to note that one of the experts spontaneously mentioned that without the “dark area surrounding the highlight” the pearl looks less realistic, which is in agreement with our initial hypothesis.

6.6. GENERAL DISCUSSION

In this study, we identified the image features used by painters to depict pearls (highlight, dark halo, edge reflection). The aim of this study was to determine whether these image features are used as perceptual cues by the visual system to trigger the perception of pearls.

In Experiment 1, we measured to what extent pearl earrings depicted in paintings were perceived pearl-like. We found that the stimuli in our set covered a wide range of pearliness, indicating that certain features of their appearance allowed participants to tell apart what does and what does not look like a pearl. Real pearls do not all have the same appearance, otherwise quality controls would not be necessary. Shape, size, and color of pearls are important quality parameters for pearls' evaluation (Klein, 2014; Kunz and Stevenson, 1908b) which likely influenced the ratings of this experiment. The two most pearly pearls look small and round while the least two are bigger and drop-like in shape (Fig. 6.5). When participants reported what they based their judgements on, several indeed referred to shape and size and a single participant mentioned the color. But most of all, participants referred to the optical appearances of the objects, more than they mentioned shape, size, and color combined. This is probably because the most important parameter to assess pearls is their luster, the way they interact with light (Mondonneix et al., 2017). Words like “shiny”, “glossiness”, “luster” and “gleam” were used by the majority of participants to describe the objects that looked more pearl-like. But remarks were not limited to specular reflections. Participants also reported to have judged the “opacity”, “whiteness” and “solidity”, and stated that more transparent objects did not look like pearls.

We hypothesized that three image features (highlight, edge reflection and a dark halo) constitute the perceptual cues for pearls' recognition. Previous studies have classified the image features used for luster assessment with the main aim of automatizing the quality control process. Mondonneix et al. (2017) related a number of image features, including the six types of gloss identified by Hunter et al. (1937) to the perceptual luster assessment of Tahitian pearls, a type of pearl showing a black body color due to the pigmentation of the organic material. The majority of features used by Mondonneix et al. (2017) are related to highlight features, such as specular gloss, gloss contrast and directness-of-reflected-image gloss. In this study we merely considered the presence (or absence) of highlights with our highlight feature. Our next feature, edge reflections, was however not considered in Mondonneix et al. (2017). Note here that sheen, i.e., specular reflections at a grazing angle, and our edge reflection feature might in some conditions appear perceptually similar but are caused by separate optical processes. Lastly, in Mondonneix et al. (2017) there is no mention of a dark halo. This is most likely because the body color of the pearls under

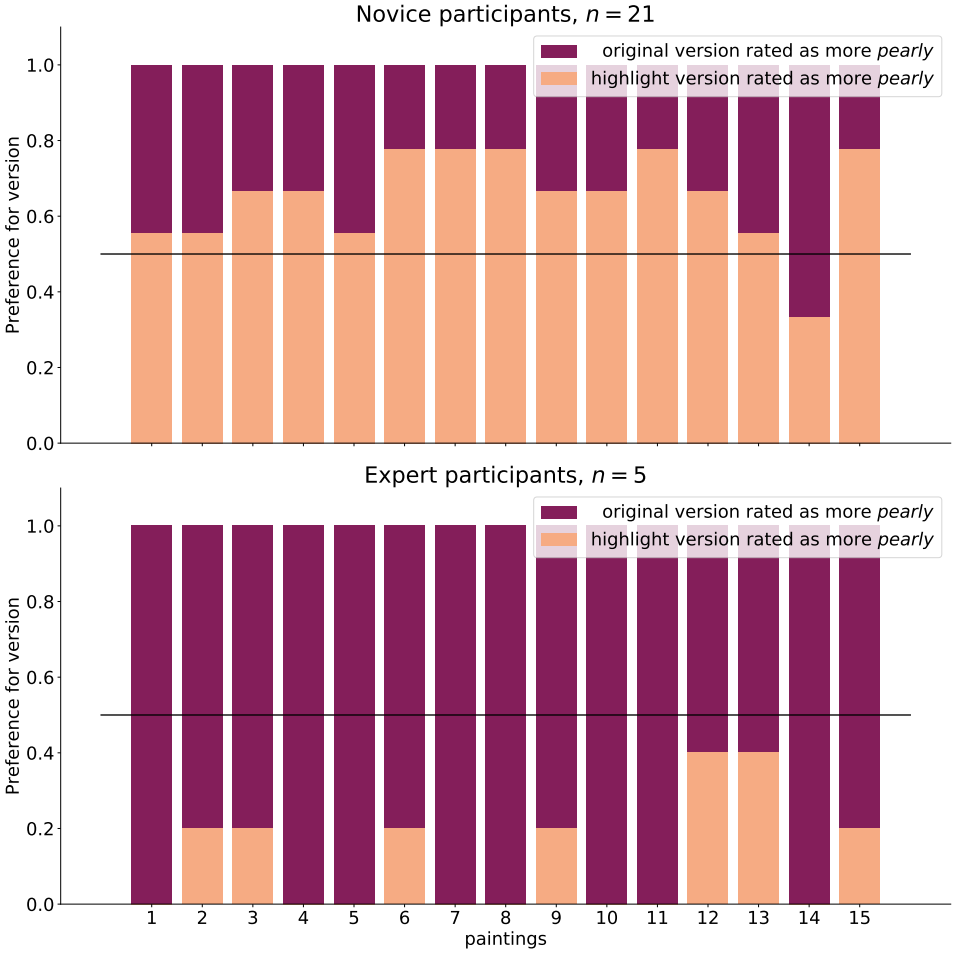


Figure 6.11: The preference across novice (top) and expert (bottom) participants for the original or highlight-only version of the stimuli. A value of 0 (a full crimson bar) would indicate that participants always choose the original version as more pearly, and a value of 1 (a full salmon bar) would mean the participants always choose the original version as the most perceptually pearly.

investigation was already dark. Nagata et al. (1997) tested the visual judgements by novice observers of the pearl-like quality of different sections of a pearl for both photographs and synthesized images of pearls. They found that the section of the image that presented both the highlight feature and the interference color, was the most characteristic of pearls. This confirms the relevance of the highlight feature in pearls' judgement.

It is interesting to note that both studies mentioned above also tested iridescence in addition to specular reflections. Iridescence here is an optical phenomenon typical of pearls, caused by the combined effects of dispersion, diffraction, and thin-film interference due to the layers of nacre. However, within our stimuli set of paintings, iridescence was never rendered in the pearl earrings. We can speculate that leaving out the iridescence was done deliberately by painters, who knew the key image cues for a convincing depiction of materials, and iridescence might not have been necessary for the identification of pearls.

To test the effect of our set of image features as perceptual cues for pearliness perception, in Experiment 2 we enhanced the visibility of these image features using our Pearliness Enhancement Transform (PET) to check if this would increase the pearl-like appearance of the stimuli. We found a significant effect where the PET versions were consistently chosen to look more pearly than the originals. The brightness of the pearl was indicated by participants as the main trait they based their judgements on, confirming that the most characteristic property of pearls is indeed their luster, due to the complex scattering of light that results in a specular reflection "whiter than white" (Klein, 2014). High contrast highlights were found to be highly correlated with the perception of glossiness of pearlescent materials (Schmid et al., 2020). However, with the PET manipulation we enhanced the brightness not only of the highlight but also of the edge reflection, and we also increased the darkness of the dark halo, which in turn increased the contrast of the highlight even more. As such, it was difficult to disentangle the separate effect of increased pearl-perception for each of the three image features.

To test the relative contribution of each of these three image feature to the perception of pearliness, we ran Experiment 3. Surprisingly, when comparing the ratings between versions, the highlight-only stimuli were perceived to be significantly more pearl-like than the original version. Our finding that the version with only the edge reflection was categorized as glass much more often than the other versions is in agreement with Todd and Norman (2019), who reported that glass reflections are mostly visible along occluding object boundaries.

In Experiment 4, we further investigated this finding with a 2AFC paradigm, presenting the original stimuli and the highlight-only stimuli to novices and experts of pearls. Our findings showed that novice participants had a significant preference for the highlight-only version, whereas experts had a strong and significant preference for the version with all three image features. From this we can conclude there is an effect of visual expertise on the perception of pearls.

It has been shown in literature that experts of pearls and novices evaluate different features when judging the quality of real pearls (Tani et al., 2014), and that novice observers improve their ability to distinguish between key image features of photographed pearls when they are exposed to real pearls before the judgement (Nagata et al., 1997). Tani et al. (2014) had novices and pearl experts to visually rank pearls based on quality, and tested the contribution of glossiness and interference color, both in isolation and combined.

They found that glossiness could explain the judgements of the novices four times more than it could explain the expert judgement, leading them to the conclusion that experts used features that have not been considered in their study. Their finding of the increased relevance of glossiness for novices is in agreement with our results. A possible explanation could be that people who work with real pearls, i.e., pearl experts, are more familiar with real pearls, and they might build a more complete *schemata* (Gombrich, 1960) in their mind of how a pearl is supposed to look-like, rather than simply attending to glossiness as novice users. We speculate that, as a result of visually studying pearls with the goal of depicting them, painters could also develop an expertise in pearl appearance. This has likely led them to reproduce some or all of the three image features through painting. Furthermore, fake pearls, being made out of plastic or glass, do not have the layered structure found in real pearls and as a result only show a white spot as specular reflection, where real pearls present a more complex combination of features (Mondonneix et al., 2017).

6.7. CONCLUSION

The present article described the perception of pearls depicted in paintings. We identified three image features (highlight, edge reflection and the dark halo surrounding the highlight) that might be used by the visual system as cues for the perception of pearliness. In a series of experiments, we first demonstrated the intuitive result that pearls depicted within paintings can vary on perceived pearliness. We further demonstrated that increasing the intensity of the three image features led to an increase in the perceived pearliness for our stimuli, implying that at least one of these features is used by the human visual system for the perception of pearliness. When attempting to disentangle the contribution to pearliness for these three features, we found that removing the edge reflections and the dark halo increased the perception of pearliness for novice participants, while it decreased for expert viewers. This implies that there is a strong effect of exposure or expertise on the perception of pearliness. We can conclude that from our set of image features, highlights are most important for the perception of pearliness, while the benefit of edge reflection and the dark halo depends strongly on familiarity with high-quality pearls.

6.8. ATTRIBUTION

All paintings used and reproduced within this paper are available under open access at a CC0 or CC BY 4.0 licences. Paintings are listed in order of appearance, from left to right.

- **Figure 1.** Johannes Vermeer, 1635. *Girl with the pearl earring*, Mauritshuis.
- **Figure 2.** Johannes Cornelisz Verspronck, 1644. *Portret van Adriana Croes*, The Rijksmuseum.
- **Figure 4.** Frans II Pourbus, 1600. *Portrait of Margaret of Austria, Consort of Philip III*, The Rijksmuseum.
- **Figure 4.** Pieter Dubordieu, 1638. *Portrait of a Woman*, The Rijksmuseum.
- **Figure 4.** Nicolaes Maes, 1665. *Portrait of a young Woman*, The Rijksmuseum.
- **Figure 4.** Isaack Luttichuys, 1656. *Portrait of a Young Lady*, The Rijksmuseum.

- **Figure 5.** Jan Mijtens, 1661. *Maria de Witte Franoisdr (b 1616). Wife of Johan van Beaumont*, The Rijksmuseum.
- **Figure 5.** Probably chiefly studio of Sir Peter Lely, 1663. *Barbara Villiers, Duchess of Cleveland*, National Gallery of Art.
- **Figure 5.** Johannes Cornelisz Verspronck, 1644. *Portret van Adriana Croes*, The Rijksmuseum.
- **Figure 5.** Peter Paul Rubens, 1630. *Portrait of Isabella of Bourbon*, The Art Institute of Chicago.
- **Figure 6.** Rembrandt van Rijn, 1660. *Woman with a Pink*, The Metropolitan Museum of Art.
- **Figure 6.** Peter Paul Rubens, 1630. *Portrait of Isabella of Bourbon*, The Art Institute of Chicago.
- **Figure 6.** Missing value, 1675. *Okänd kvinna*, Nationalmuseum.
- **Figure 6.** British Painter, 1600. *Portrait of a Woman*, The Metropolitan Museum of Art.
- **Figure 8.** Frans van Mieris (I), 1678. *A Man and a Woman*, The Rijksmuseum.
- **Figure 8.** Van Dyck, Anton, 1638. *Diana Cecil, Countess of Oxford*, Museo Nacional del Prado.
- **Figure 9.** Dirck Craey, 1650. *Portrait of a Woman, thought to be Catharina Kettingh (1626/27-73), Wife of Bartholomeus Vermuyden*, The Rijksmuseum.

BIBLIOGRAPHY

- Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior research methods*, 45(2), 527–535.
- Beck, J., & Prazdny, S. (1981). Highlights and the perception of glossiness. *Perception & Psychophysics*.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5.
- Carbon, C.-C., & Pastukhov, A. (2018). Reliable top-left light convention starts with early renaissance: An extensive approach comprising 10k artworks. *Frontiers in psychology*, 9, 454.
- Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, 434(7031), 301–307.
- Chadwick, A. C., & Kentridge, R. W. (2015). The perception of gloss: A review. *Vision Research*, 109(PB), 221–235. <https://doi.org/10.1016/j.visres.2014.10.026>
- Di Cicco, F., van Zuijlen, M. J., Wijntjes, M. W., & Pont, S. C. (2021, in press.). Soft like velvet and shiny like satin: Perceptual material signatures of fabrics depicted in 17th century paintings. *Journal of Vision*.
- Di Cicco, F., Wiersma, L., Wijntjes, M., & Pont, S. (2020). Material properties and image cues for convincing grapes: The know-how of the 17th-century pictorial recipe by willem beurs. *Art & Perception*, 8(3-4), 337–362.
- Di Cicco, F., Wijntjes, M. W., & Pont, S. C. (2019). Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *Journal of Vision*, 19(3). <https://doi.org/10.1167/19.3.7>
- Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 3, 347–368. <https://doi.org/10.1037/t25352-000>
- Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8), 9. <https://doi.org/10.1167/13.8.9>
- Fu, F., Tian, L., Xu, X., Xu, S., & Hu, X. (2014). Influence of microstructure on optical behaviour of freshwater cultured pearls. *Materials Research Innovations*, 18(4), 245–250.
- The girl without a pearl earring? (2017). <https://www.mauritshuis.nl/en/discover/news-archive/2014/draagt-het-meisje-wel-een-parel>
- Gombrich, E. H. (1960). *A study in the psychology of pictorial representation*. Pantheon Books.
- Hunter, R. S. et al. (1937). Methods of determining gloss. *NBS Research paper RP*, 958.
- Icke, V. (1987). "MEISJE MET GEEN PAREL"[GIRL WITH NO PEARL EARRING]. *Nederlands Tijdschrift voor Natuurkunde*, 80(12), 418–419.

- Kim, J., Marlow, P., & Anderson, B. L. (2011). The perception of gloss depends on highlight congruence with surface shading. *Journal of Vision*, 11(9), 4–4.
- Kim, J., Marlow, P. J., & Anderson, B. L. (2012). The dark side of gloss. *Nature Neuroscience*, 15(11), 1590–1595. <https://doi.org/10.1038/nn.3221>
- Klein, T. (2014). The lustre of pearls. *Europhysics News*, 45(3), 15–16.
- Koenderink, J. J., & van Doorn, A. J. (1980). Photometric invariants related to solid shape. *Optica Acta: International Journal of Optics*, 27(7), 981–996.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155–163.
- Kunz, G. F., & Stevenson, C. H. (1908a). *The book of the pearl: The history, art, science, and industry of the queen of gems*. Century Company.
- Kunz, G. F., & Stevenson, C. H. (1908b). *The book of the pearl: The history, art, science, and industry of the queen of gems*. Century Company.
- Landman, N., Mikkelsen, P., Bieler, R., & Bronson, B. (2001). *Pearls: A natural history. harry and abrams*. Harry N. Abrams Inc., New York, USA.
- Mamassian, P. (2008). Ambiguities and conventions in the perception of visual art [Vision Research Reviews]. *Vision Research*, 48(20), 2143–2153. <https://doi.org/https://doi.org/10.1016/j.visres.2008.06.010>
- Marlow, & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision*, 13(14), 1–23. <https://doi.org/10.1167/13.14.2>
- Marlow, Kim, J., & Anderson, B. L. (2011). The role of brightness and orientation congruence in the perception of surface gloss. *Journal of Vision*, 11(9), 16–16.
- Marlow, Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, 22(20), 1909–1913. <https://doi.org/10.1016/j.cub.2012.08.009>
- McManus, I. C., & Humphrey, N. K. (1973). Turning the left cheek. *Nature*, 243(June 1973), 271–272. <https://doi.org/10.1038/243271a0>
- Miller, J. (1998). *On reflection*. National Gallery Publications London.
- Mondonneix, G., Chabrier, S., Mari, J.-M., & Gabillon, A. (2017). Tahitian pearls' luster assessment automation. *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–9.
- Nagata, N., Dobashi, T., Manabe, Y., Usami, T., & Inokuchi, S. (1997). Modeling and visualization for a pearl-quality evaluation simulator. *IEEE Transactions on Visualization and Computer Graphics*, 3(4), 307–315.
- Olkkonen, M., & Brainard, D. H. (2010). Perceived glossiness and lightness under real-world illumination. *Journal of vision*, 10(9), 5–5.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4), 867–872.
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4), 1023–1031.
- Pont, S. C., & te Pas, S. F. (2006). Material—illumination ambiguities and the perception of solid objects. *Perception*, 35(10), 1331–1350.

- Raman, C., & Krishnamurti, D. (1954a). On the chromatic diffusion halo and other optical effects exhibited by pearls. *Proceedings of the Indian Academy of Sciences-Section A*, 39(6), 265–271.
- Raman, C., & Krishnamurti, D. (1954b). The structure and optical behaviour of pearls. *Proceedings of the Indian Academy of Sciences-Section A*, 39(5), 215–222.
- Rouhana, R., & Stommel, M. (2020). Modelling and experimental investigation of hexagonal nacre-like structure stiffness. *Journal of Composites Science*, 4(3), 91.
- Sawayama, M., Adelson, E. H., & Nishida, S. (2017). Visual wetness perception based on image color statistics. *Journal of Vision*, 17(5), 7–7.
- Sayim, B., & Cavanagh, P. (2011). What Line Drawings Reveal About the Visual Brain. *Frontiers in Human Neuroscience*, 5(October), 1–4. <https://doi.org/10.3389/fnhum.2011.00118>
- Schmid, A. C., Barla, P., & Doerschner, K. (2020). Material category determined by specular reflection structure mediates the processing of image features for perceived gloss. *bioRxiv*, 1–45. <https://doi.org/10.1101/2019.12.31.892083>
- Sharan, L., Rosenholtz, R., & Adelson, E. H. (2009). Material perception : What can you see in a brief glance? *Vision Sciences Society Annual Meeting Abstract*, 9(8), 2009.
- Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision*, 14(9), 1–24. <https://doi.org/10.1167/14.9.12>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, 86(2), 420.
- Tani, Y., Nagai, T., Koida, K., Kitazaki, M., & Nakauchi, S. (2014). Experts and novices use the same factors—but differently—to evaluate pearl quality. *PloS one*, 9(1), e86400.
- Taylor, J., & Strack, E. (2008). Pearl production. In P. S. J. Lucas (Ed.), *The pearl oyster* (pp. 273–302). Elsevier Science.
- Todd, J. T., & Norman, J. F. (2018). The visual perception of metal. *Journal of Vision*, 18(3), 9–9.
- Todd, J. T., & Norman, J. F. (2019). Reflections on glass. *Journal of vision*, 19(4), 26–26.
- Todd, J. T., & Norman, J. F. (2020). Contours produced by internal specular interreflections provide visual information for the perception of glass materials. *Journal of Vision*, 20(10), 12–12.
- Van Zuijlen, M. J., Lin, H., Bala, K., Pont, S. C., & Wijntjes, M. W. (2021). Materials in paintings (mip): An interdisciplinary dataset for perception, art history, and computer vision. *Plos one*, 16(8), e0255109.
- van Zuijlen, M. J., Pont, S. C., & Wijntjes, M. W. (2020). Painterly depiction of material properties. *Journal of vision*, 20(7).
- Wijntjes, M. W. A., Spoiala, C., & de Ridder, H. (2020). Thurstonian scaling and the perception of painterly translucency. *Art & Perception*, 1–24. <https://doi.org/https://doi.org/10.1163/22134913-bja10021>
- Wijntjes, M. (2021). Shadows, highlights and faces: The contribution of a ‘human in the loop’ to digital art history. *Art & Perception*, 9(1), 66–89.

6.9. SUPPLEMENTARY MATERIALS

ALL ORDERED PEARL STIMULI AND ATTRIBUTION

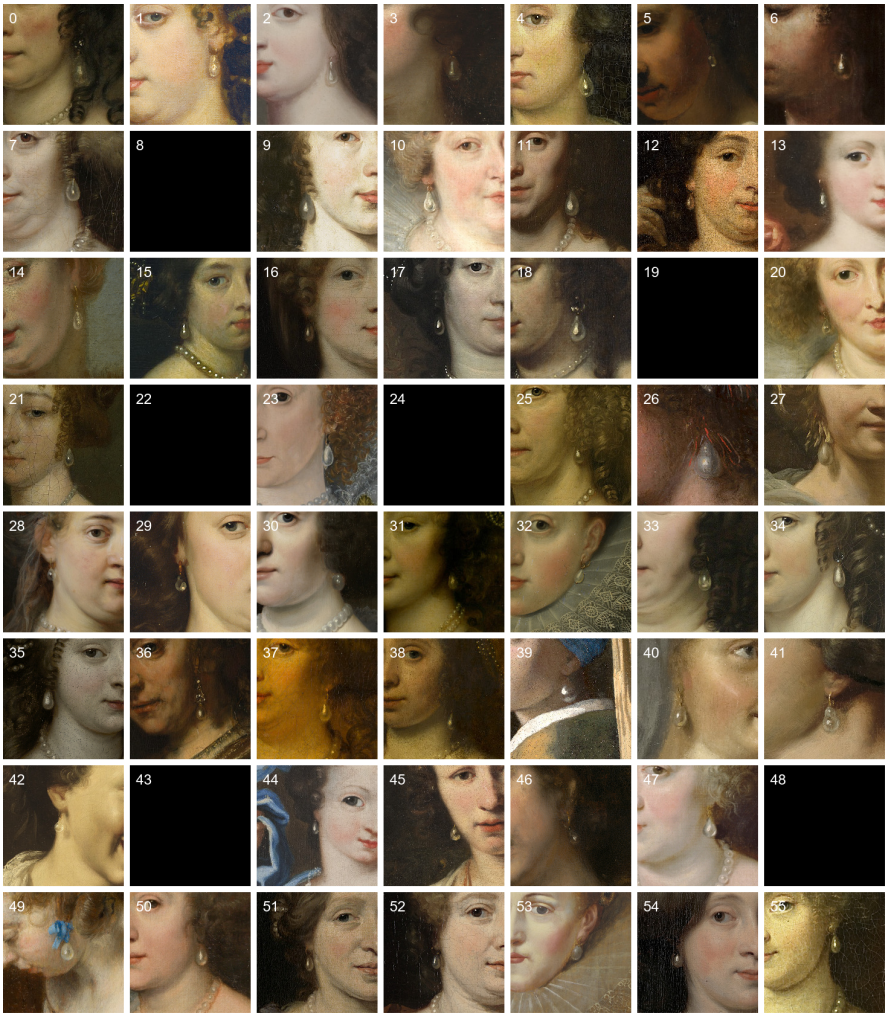


Figure 6.12: Figure continued on next page. All the pearl stimuli ordered from least pearly to most pearly, as rated by participants in Experiment 1. The ordering goes from left to right, top to bottom, where the top-left is least and bottom-right most pearly. 9 of the pearl stimuli are from images from the National Gallery of London, which can not be reproduced here due to copyright and have therefore been replaced by a black square. The attribution for each of the paintings can be found in the Attribution section. Note that the number in the top-left of each image here is only used to link images to its attribution and was not visible to the participants.

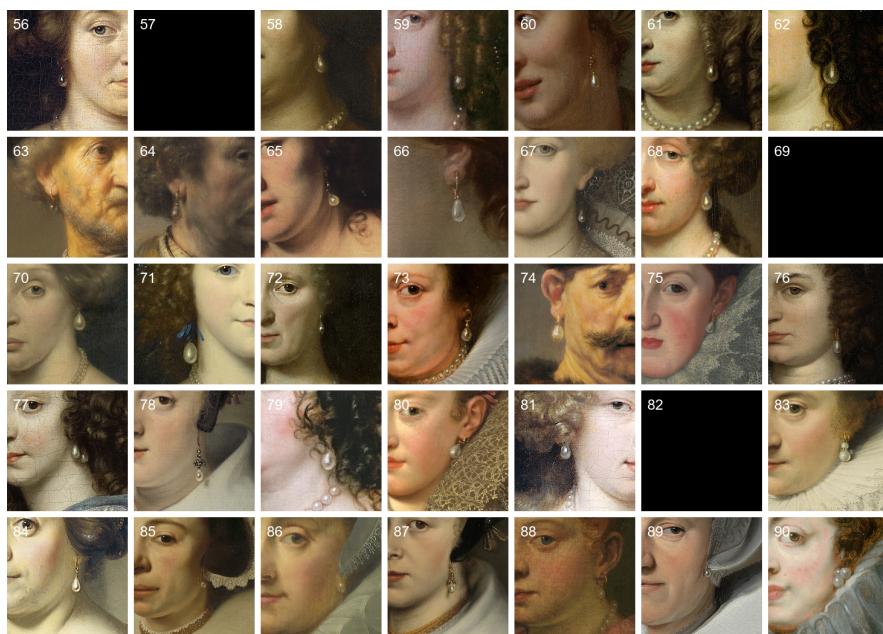


Figure 6.13: Figure continued from previous page.

ATTRIBUTION

Below is the list of all paintings used within this study. For each painting, we only used a small section of the painting that depicted the pearl earring. All images reproduced in the main document or in this supplementary document are reproduced under CC BY 4.0 or CC 0.0.

1. *Maria de Witte Françoisdr (b 1616). Wife of Johan van Beaumont* by Jan Mijtnens. 1661. The Rijksmuseum
2. *Barbara Villiers, Duchess of Cleveland* by Probably chiefly studio of Sir Peter Lely. 1663. National Gallery of Art
3. *Okänd kvinna* by Missing value. 1675. Nationalmuseum
4. *Ulrika Eleonora d.ä. 1656-1693, drottning av Sverige prinsessa av Danmark* by Jacques D' Agar. 1677. Nationalmuseum
5. *Portrait of a young woman* by Eglon van der Neer. 1660. The Rijksmuseum
6. *Young Woman with a Pearl Necklace* by Copy after Willem Drost. None. The Metropolitan Museum of Art
7. *Company Making Music* by Missing value. 1665. Mauritshuis
8. *Margaretha Munter (1639-1711), second Wife of Jacobus Trip* by Lambertus Jansz. de Hue. 1668. The Rijksmuseum
9. *Portrait of a Young Woman* by Gerrit Dou. 1655. The National Gallery of London

10. *Portrait of a young Woman by Nicolaes Maes*. 1665. The Rijksmuseum
11. *Marie de Medici, Queen of France by Rubens, Pieter Paul*. 1622. Museo Nacional del Prado
12. *Portrait of Cornelia Craen van Haeften (1622-1678) by Cornelis Janssens van Ceulen (II)*. 1670. The Rijksmuseum
13. *Portrait of Catharina Dierquens (1664-1715) by Nicolaes Maes*. 1682. Mauritshuis
14. *Karl XII, 1682-1718, konung av Sverige, pfalzgreve av Zweibrücken och Hedvig Sofia, 1681-1708, prinsessa av Sverige, hertiginna av Holstein-Gottorp by David Klöcker Ehrenstrahl*. 1687. Nationalmuseum
15. *Diana and Her Nymphs on the Hunt by Workshop of Peter Paul Rubens*. 1627. J. Paul Getty Museum
16. *Girl Standing before a Mirror by Caspar Netscher*. 1668. The Art Institute of Chicago
17. *Ingena Rotterdam (died 1704), Betrothed of Admiral Jacob Binkes by Nicolaes Maes*. 1676. The Metropolitan Museum of Art
18. *Johanna le Gillon, Wife of Hieronymus van Beverningk by Jan de Baen*. 1670. The Rijksmuseum
19. *A couple with six children by Jürgen Ovens*. 1664. The Rijksmuseum
20. *Portrait of a Lady and a Girl by Caspar Netscher*. 1679. The National Gallery of London
21. *Portret van Helena Fourment (1614-1673) by Peter Paul Rubens*. 1635. The Rijksmuseum
22. *Aletta Pancras (1649-1707) Wife of François de Vic by Gerard ter Borch (II)*. 1670. The Rijksmuseum
23. *Judith by Eglon Hendrik van der Neer*. 1678. The National Gallery of London
24. *Portrait of a Woman by British Painter*. 1600. The Metropolitan Museum of Art
25. *Portrait of the Infanta Isabella by Studio of Peter Paul Rubens*. 1615. The National Gallery of London
26. *Margaretha van Raephorst (d 1690). Wife of Cornelis Tromp by Jan Mijntens*. 1668. The Rijksmuseum
27. *A Fishmonger at the Door by Missing value*. 1663. Mauritshuis
28. *A Musical Company by Gerard van Kuijl*. 1651. The Rijksmuseum
29. *Young Woman in Fantasy Costume by Rembrandt van Rijn*. 1633. The Rijksmuseum
30. *Portrait of a Woman by Jacob Adriaensz Backer*. 1647. J. Paul Getty Museum
31. *Portrait of Amalia van Solms (1602-75) by Michiel Jansz van Mierevelt*. 1632. The Rijksmuseum
32. *Portrait of a Woman by Adriaen Hanneman*. 1653. The Metropolitan Museum of Art

33. *Portrait of Margaret of Austria, Consort of Philip III* by *Frans II Pourbus*. 1600. The Rijksmuseum
34. *Portrait of Catharina van der Voort* by *Abraham van den Tempel*. 1667. The Rijksmuseum
35. *Portrait of Jacoba van Orliens, Wife of Jacob de Witte of Haamstede* by *Jan Mijntens*. 1660. The Rijksmuseum
36. *Portrait of a Woman* by *Jan Mytens*. 1660. J. Paul Getty Museum
37. *Woman with a Pink* by *Rembrandt van Rijn*. 1660. The Metropolitan Museum of Art
38. *Portrait of Elisabeth van Bebbber (1643-1704)* by *Caspar Netscher*. 1677. Mauritshuis
39. *Portrait of a Woman* by *Bartholomeus van der Helst*. 1659. Mauritshuis
40. *Girl with a Pearl Earring* by *Johannes Vermeer*. 1665. Mauritshuis
41. *Konversationsstykke* by *Bartholomeus Maton*. 1679. Nationalmuseum
42. *Samson and Delilah* by *Jan Lievens*. 1632. The Rijksmuseum
43. *A Girl Holding a Glass ("Taste", One of a Series of the Five Senses)* by *Hendrick ter Brugghen*. 1620. Nationalmuseum
44. *Portrait of a Lady* by *Caspar Netscher*. 1683. The National Gallery of London
45. *Hedvig Sophia of Sweden (1681–1708), Swedish princess and a Duchess Consort of Holstein-Gottorp, Spouse Frederick IV, Duke of Holstein-Gottorp* by *Anna Maria Ehrenstrahl*. 1684. Nationalmuseum
46. *Women with Pearls in her Hair* by *Ferdinand Bol*. 1653. Nationalmuseum
47. *Lucretia* by *Rembrandt van Rijn*. 1664. National Gallery of Art
48. *Anne of Austria, Queen of France* by *Rubens, Pieter Paul*. 1622. Museo Nacional del Prado
49. *Portrait of a Woman* by *Style of Anthony van Dyck*. 1635. The National Gallery of London
50. *Woman Writing a Letter* by *Gerard ter Borch*. 1655. Mauritshuis
51. *Amalia de Solms-Braunfels* by *Van Dyck, Anton*. 1631. Museo Nacional del Prado
52. *Portrait of Catharina Pottey, Sister of Willem and Sara Pottey* by *Nicolaes Maes*. 1677. The Rijksmuseum
53. *Portrait of Petronella Dunois* by *Nicolaes Maes*. 1681. The Rijksmuseum
54. *Isabella Klara Eugenia, 1566-1633, prinsessa av Spanien ärkehertiginna av Österrike* by *Peter Paul Rubens*. None. Nationalmuseum
55. *Gerard Röver, Merchant and Shipowner in Amsterdam* by *Nicolaes Maes*. 1684. The Rijksmuseum
56. *Portrait of a woman, possibly a member of the van Citters family* by *Caspar Netscher*. 1674. The Rijksmuseum

57. *Sara Pottey (1651-1705), Wife of Johan van Bochoven* by Daniël Haringh. 1680. The Rijksmuseum
58. *Lady Elizabeth Thimbelby and her Sister* by Anthony van Dyck. 1637. The National Gallery of London
59. *Cornelia Teding van Berkhout (1614-80), Wife of Maerten Harpertsz Tromp* by Jan Lievens. 1646. The Rijksmuseum
60. *Alida de Lange, Wife of Johan Rammelman* by Netscher, Caspar. 1679. Museo Nacional del Prado
61. *Woman Selling Herrings* by Godfried Schalcken. 1677. The Rijksmuseum
62. *Portrait of Philippina Staunton, Wife of Roelof van Arkel (1632-1709), lord of Broeckhuijsen* by Caspar Netscher. 1668. The Rijksmuseum
63. *Portrait of Amalia van Solms (1602-75)* by Gerard van Honthorst. 1650. The Rijksmuseum
64. *Old Man with a Gold Chain* by Rembrandt Harmenszoon van Rijn. 1631. The Art Institute of Chicago
65. *Portrait of Rembrandt* by Rembrandt van Rijn. 1650. National Gallery of Art
66. *A Lady Playing the Lute* by Ferdinand Bol. 1654. Nationalmuseum
67. *Joseph and Potiphar's Wife* by Guido Reni. 1630. J. Paul Getty Museum
68. *Maria Eleonora (1599-1655), Princess of Brandenburg, Queen of Sweden* by Michiel van Mierevelt. 1619. Nationalmuseum
69. *Portrait of Maria Timmers (1658-1753)* by Missing value. 1683. Mauritshuis
70. *Portrait of a Young Woman* by Attributed to Willem Drost. 1679. The National Gallery of London
71. *Portrait of Petronella van der Burcht (1657-1682)* by Zacharias Blijhooft. 1674. The Rijksmuseum
72. *Portrait of a Young Lady* by Isaack Luttichuys. 1656. The Rijksmuseum
73. *Portrait of a Woman* by Jan Verkolje (I). 1691. The Rijksmuseum
74. *Portrait of a Woman, possibly Clara Fourment (1593-1643)* by Peter Paul Rubens. 1630. Mauritshuis
75. *A Polish Nobleman* by Rembrandt van Rijn. 1637. National Gallery of Art
76. *nan* by González, Bartolomé. 1600. Museo Nacional del Prado
77. *Portrait of Dina Lems, Wife of Jan Valckenburgh* by Daniel Vertangen. 1660. The Rijksmuseum
78. *Matthijs Pompe van Slingelandt and Family* by Johannes Mijtens. 1655. Nationalmuseum
79. *Maria Rey (1630/31-1703). Wife of Roelof Meulenaer* by Ferdinand Bol. 1650. The Rijksmuseum

80. *Diana Cecil, Countess of Oxford* by *Van Dyck, Anton*. 1638. Museo Nacional del Prado
81. *Margherita Gonzaga (1591-1632), Princess of Mantua* by *Frans Pourbus the Younger*. None. The Metropolitan Museum of Art
82. *Portrait of Helena Ctaharina de Witte 91661-95), wife of Iman mogge, lord of Haamstede* by *Caspar Netscher*. 1678. The Rijksmuseum
83. *Portrait of a Woman with a Black Cap* by *Jan de Braij*. 1657. The National Gallery of London
84. *Portrait of Ernestine Yolande (1594-1663), Princess of Ligne* by *Missing value*. 1618. Mauritshuis
85. *A Man and a Woman* by *Frans van Mieris (I)*. 1678. The Rijksmuseum
86. *Portrait of a Woman* by *Pieter Dubordieu*. 1638. The Rijksmuseum
87. *Lucretia del Prado, Wife of Jeremias Boudinois* by *Gortzius Geldorp*. 1610. The Rijksmuseum
88. *Portrait of a Woman, thought to be Catharina Kettingh (1626/27-73), Wife of Bartholomeus Vermuyden* by *Dirck Craey*. 1650. The Rijksmuseum
89. *Young Venetian Woman* by *Tintoretto, Domenico*. 1600. Museo Nacional del Prado
90. *Portrait of Adriana Croes* by *Johannes Cornelisz. Verspronck*. 1644. The Rijksmuseum
91. *Portrait of Isabella of Bourbon* by *Peter Paul Rubens*. 1630. The Art Institute of Chicago

PEARLINESS ENHANCEMENT TRANSFORM (PET)

To manipulate the stimuli we developed the Pearliness Enhancement Transform (PET). PET consists of three image manipulations applied to the pearl linked to the three image features we have previously discussed. As a prerequisite, PET requires the pearl's location and an approximation of the pearl's shape. Below, we first discuss how the pearls location and shape were identified, after which we discuss the three image manipulations in depth. All computations are GPU based with a GLSL implementation in the Gratin software ¹. As the input images are rather small (200×200 pixels) they are first upscaled to 800×800 pixels. Then the PET is applied, after which they are again downscaled back to their original resolution.

Pearl Shape The image region corresponding to the earring is determined by 5 parameters: its central position \vec{p} in pixels, its size s in pixels, its aspect ratio $r \in (0, 1]$ (assuming that a pearl is always elongated in the vertical direction), its orientation θ in degrees and its vertical asymmetry $a \in \mathbb{R}^+$ (assuming a pearl may be narrower at its top than at its base).

All shape parameters are set manually to match the estimated contour of the earring in the image. This process exports two images: a gray-level *mask* m of the estimated image region, and a *normal map* \vec{n} for the corresponding shape.

6

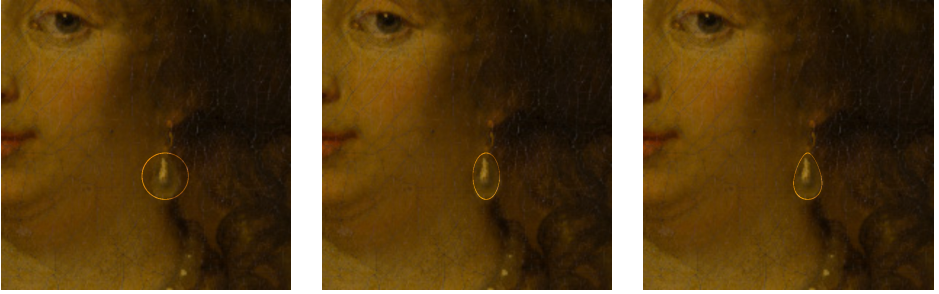


Figure 6.14: The outline of the pearl earring is first defined as a circle of center \vec{p} and radius s (left). It is then deformed to an elliptic shape via the ratio r (middle) and to a “drop-like” shape via the vertical asymmetry a (right). For attribution see #37 in the Attribution section.

The mask is determined implicitly: we assign a gray level value $m(\vec{q})$ to a pixel \vec{q} by imposing that the distance $d(\vec{q})$ to the contour is smaller than or equal to 1. Here $\tilde{\vec{q}} = \mathcal{R} [\vec{q} - \vec{p}] + \vec{p}$ corresponds to \vec{q} transformed by a rotation matrix \mathcal{R} of $-\theta$ radians around \vec{p} . The distance is given by $d(\vec{q}) = \|S [\vec{q} - \vec{p}]\|$, where S is a non-uniform scaling (i.e., diagonal) matrix given by $S = \text{diag}(\frac{1}{r(1-a\tilde{q}_1)s}, \frac{1}{s})$, where \tilde{q}_1 is the vertical coordinate of $\tilde{\vec{q}}$. When $a = 0$, $S = \text{diag}(\frac{1}{rs}, \frac{1}{s})$ is independent of $\tilde{\vec{q}}$, and yields an ellipse in general, or a circle when $r = 1$. When $a \neq 0$, the scaling depends on the position \tilde{q}_1 along the vertical axis, pinching the shape at the top and inflating it at the bottom, yielding a “drop-like” shape. For the final mask, we allow for a narrow transition around $d(\vec{q}) = 1$ by setting $m(\vec{q}) = 1 - \text{smoothstep}(1 - \epsilon, 1 + \epsilon, d(\vec{q}))$, where we use the GLSL smoothstep function, and

¹<http://gratin.gforge.inria.fr/>

systematically set $\epsilon = 0.12$.

The normal map \tilde{n} is also computed implicitly, starting from the distance function. Indeed, since $d(\tilde{q}) \leq 1$ inside the mask, $S[\tilde{q} - \tilde{p}]$ can be considered as the projection of \tilde{n} onto the picture plane. The normal is then recovered by lifting its projection so that it has unit length: $\tilde{n}(\tilde{q}) = (S[\tilde{q} - \tilde{p}], \sqrt{1 - \|d(\tilde{q})\|^2})$.

Luminance Adjustment The first image manipulation simply consists in adjusting the luminance of pixels in order to make both darkest and brightest tones lighter. This should be done in a way that depends on the range of luminances in the input image. For that reason, we first compute the mean μ and standard deviation σ of the luminance histogram of whole the input image.



Figure 6.15: The earring region in the input image (left) is first modified to slightly increase the luminance of the dark tones via α_0 (middle), then modified to strongly increase the bright tones via α_1 (right). For attribution see #37 in the Attribution section.

The color $C(\tilde{q})$ at a pixel \tilde{q} inside the mask m is then modified using

$$C(\tilde{q}) \leftarrow [1 + \alpha_0(L(\tilde{q}))][1 + \alpha_1(L(\tilde{q}))] C(\tilde{q}), \quad (6.1)$$

where $L(\tilde{q})$ is the luminance intensity of pixel \tilde{q} , and α_0 and α_1 increase the luminance of the darkest and brightest tones respectively and are given by:

$$\alpha_0(l) = \text{smoothstep}(1 - (\mu - \sigma), 1, 1 - l), \quad (6.2)$$

$$\alpha_1(l) = \text{smoothstep}(\mu - \sigma, \mu + 4\sigma, l). \quad (6.3)$$

The boundaries (first two parameters) of the GLSL smoothstep functions could be adjusted to increase or decrease the strength of the effect; here they have been chosen manually to produce consistent results for the full set of 91 input earring images.

Additional reflections The next image manipulation adds two types of reflections to the pearl earring image region: a dark spot around the main highlight, and an elongated bright reflection close to its contour, following observations from paintings as well as recommendations from painting tutorials.

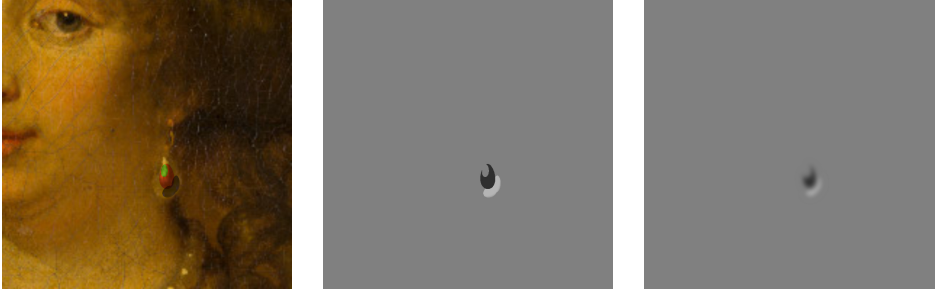


Figure 6.16: Using the normal map \vec{n} , we create three regions (left): inside the highlight in green, in the dark spot in red, and in the bright elongated region in dark gray. We assign a dark tone to pixels belonging to the dark spot but not to the highlight, and a bright tone to pixels belonging to the bright reflection but not to the dark spot (middle). The resulting gray level map R is blurred (right). For attribution see #37 in the Attribution section.

As with the pearl shape estimation, this step involves manual interaction for locating the position of the pixel \vec{q}_h at the center of the main highlight and for specifying its size s_h . We use the normal map to assign a normal $\vec{n}_h = \vec{n}(\vec{q}_h)$ to the highlight.

We then define three distance functions d_h , d_0 and d_1 , which locate the distance to the border of the main highlight, of the dark spot and of the bright elongated reflection respectively. All three functions are based on the normal map \vec{n} , and hence automatically conform to the previously estimated pearl shape from Section 6.9. They are computed as follows:

$$d_{h,0,1}(\vec{q}) = \frac{1 - \vec{n}_{h,0,1} \cdot \vec{n}(\vec{q})}{s_{h,0,1}}. \quad (6.4)$$

When \vec{n}_q is aligned with either \vec{n}_h , \vec{n}_0 or \vec{n}_1 , the corresponding distance is equal to 0. As the normals start to spread apart, the distance increases until it reaches 1 for a reflection of size s_h , s_0 or s_1 respectively.

The dark spot is positioned in a direction opposite to the highlight with respect to the center of the earring. To this end, we start from \vec{n}_h , the 2D projection of \vec{n}_h onto the picture plane; we then lift the symmetric of this 2D vector to a unit 3D vector using $\vec{n}_0 = \frac{(-\vec{n}_h, 3)}{\|(-\vec{n}_h, 3)\|}$. This has the effect of placing the dark spot closer to the earring center so that it does not overlap too much with the second, bright elongated reflection. We systematically set the extent of the dark spot to $s_0 = 4s_h$.

The bright elongated reflection is also positioned in a direction opposite to the highlight, but farther from it. As before, we start from \vec{n}_h but modify it to push the reflection closer to the boundary using $\vec{n}_1 = -\frac{0.7+0.3\|\vec{n}_h\|}{\|\vec{n}_h\|}\vec{n}_h$. The resulting 2D vector is then lifted to a unit 3D vector: $\vec{n}_1 = (\vec{n}_1, \sqrt{1 - \|\vec{n}_1\|^2})$. Elongation is obtained by computing the extent in a way dependent on \vec{q} . We use $s_1 = (1 + 8|\vec{n}(\vec{q}) \cdot \vec{t}_1|)s_h$, where $\vec{t}_1 = \frac{\vec{n}_h \times \vec{n}_1}{\|\vec{n}_h \times \vec{n}_1\|}$; in effect, the bright reflection is thus 8 times wider in the direction running along the contour of the earring.

We next create a reflection image R based on these three distance functions, using the minimal and maximum luminances m and M of the original input image. We assign the dark spot a luminance that is slightly less dark than m (e.g., to avoid having it as dark as

the eye pupil), but we use the maximal luminance for the elongated bright reflection:

$$R(\vec{q}) = \begin{cases} 0.8m + 0.2M & \text{if } d_0(\vec{q}) \leq 1 \text{ and } d_h(\vec{q}) > 1 \\ M & \text{if } d_0(\vec{q}) > 1 \text{ and } d_1(\vec{q}) \leq 1 \\ 0.5 & \text{otherwise.} \end{cases} \quad (6.5)$$

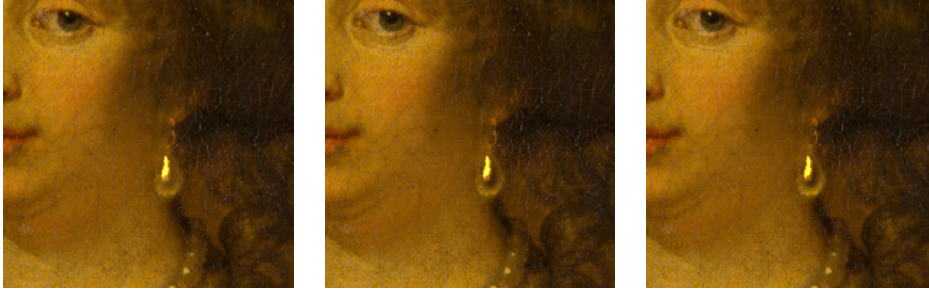


Figure 6.17: Starting from the previously modified image (left), we first blend it with the dark spot (middle), then with both the dark spot and the bright elongated reflection (right), using a soft light mode (see text). For attribution see #37 in the Attribution section.

The reflection image R is then blurred with a large kernel to smooth out boundaries across the added reflections, and finally combined with the image output from Equation 6.1 using a soft light mode:

$$C(\vec{q}) \leftarrow \text{softlight}(C(\vec{q}), G * R(\vec{q})), \quad (6.6)$$

where G is a 40×40 Gaussian blur kernel with $*$ the convolution operator and where softlight is a classic blending mode in image manipulation software that makes dark pixels darker and bright pixels brighter. We use the softlight formula employed in Gimp:

$$\text{softlight}(A, B) = ((1 - A)B + C)A \quad (6.7)$$

where

$$C = 1 - (1 - A)(1 - B) \quad (6.8)$$

Glow filter Shape The last manipulation adds a slight glow to the pearl. It is obtained by blurring the pixels in the mask m , and combining the result with the unblurred pixels:

$$C(\vec{q}) \leftarrow \max(C(\vec{q}), G * C(\vec{q})), \quad (6.9)$$

where G is a 10×10 Gaussian blur kernel and $*$ denotes the convolution operator.

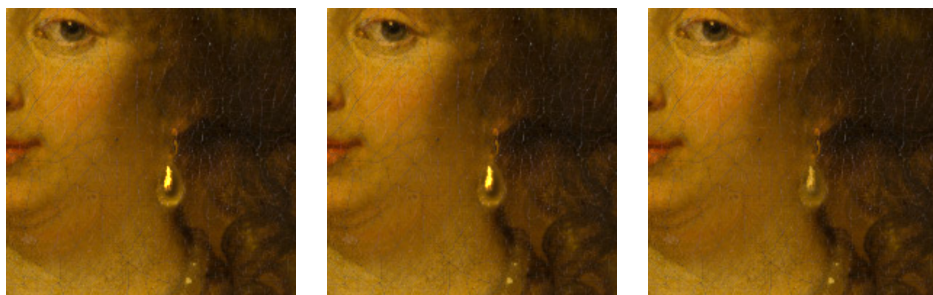


Figure 6.18: Starting from the previously modified image (left), we add a slight glow effect (middle). We compare this final result with the original input image (right). For attribution see #37 in the Attribution section.

ALL PET MANIPULATED STIMULI

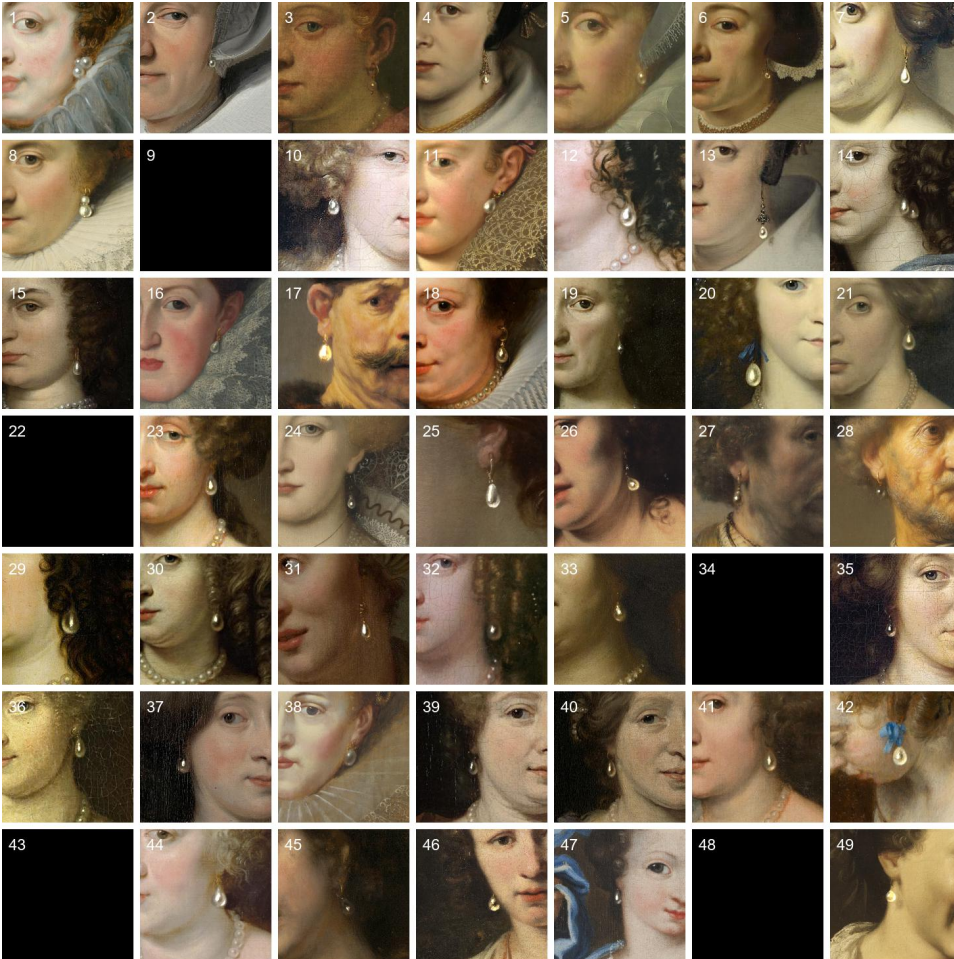


Figure 6.19: Figure continued on next page. All the PET stimuli, in the same order as Fig. 6.12. 9 of the pearl stimuli are from images from the National Gallery of London, which can not be reproduced here due to copyright and have therefore been replaced by a black square. Note that the number in the top-left of each image here is only used to link images to its attribution and was not visible to the participants. For attribution see the corresponding numbers in the Attribution section.

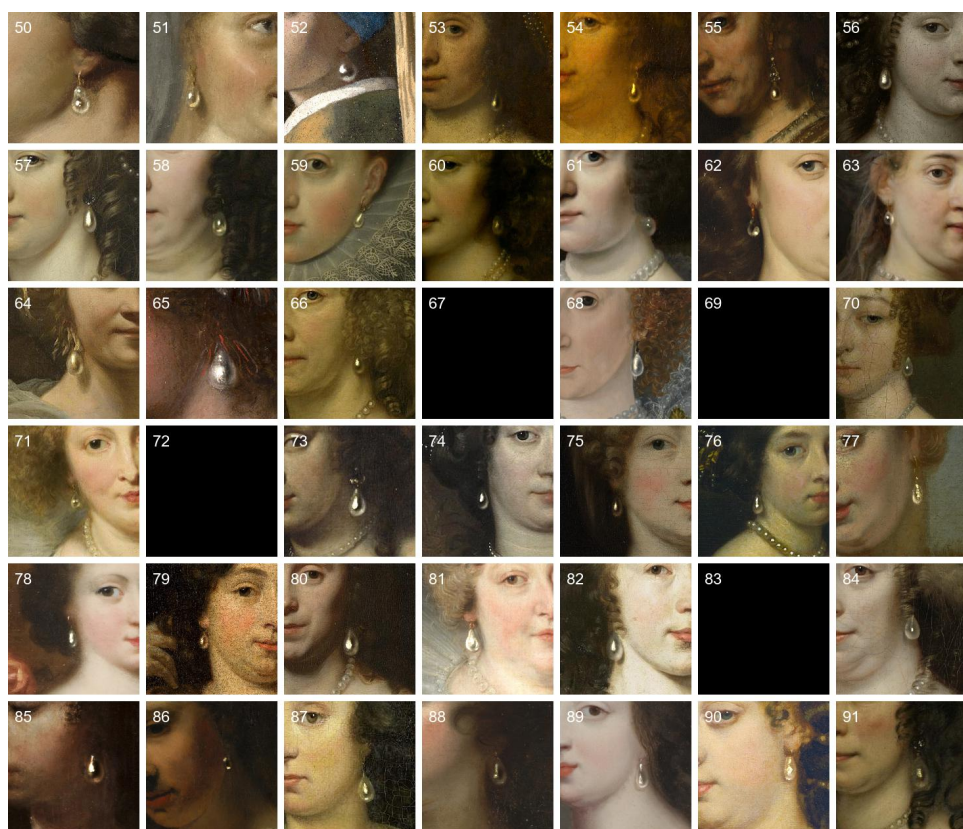


Figure 6.20: Figure continued from previous page.

ORIGINAL AND MANIPULATED VERSIONS

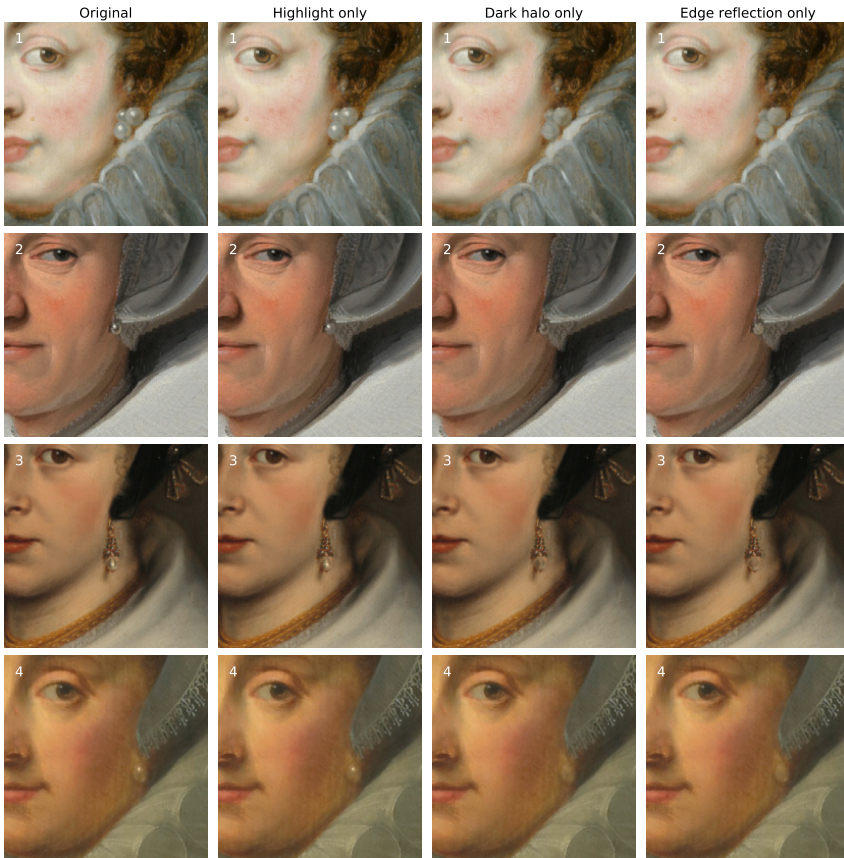


Figure 6.21: Figure continued on the next page. The original and the 3 manipulated versions of the stimuli used in experiment 3, following the same order as in Fig. 6.12. One image (#3 in Fig. 6.12) did not show all the image features and thus could not be manipulated consistently in relation to the other stimuli and has therefore been skipped. As such, instead of the 15 most pearly stimuli, we actually used the 16 most pearly stimuli minus image #3 in Fig. 6.12. A second image (#9 in Fig. 6.12, which would be #8 here) has not been reproduced here due to copyright. The numbers here refer to the PCA bi-plot in figure 9 of the main document and have not been shown to participants. Note that in some cases it can be difficult to see the edge reflections due to the reduced contrast as a result of the removal of the black halo. For attribution see the corresponding images in the Attribution section.

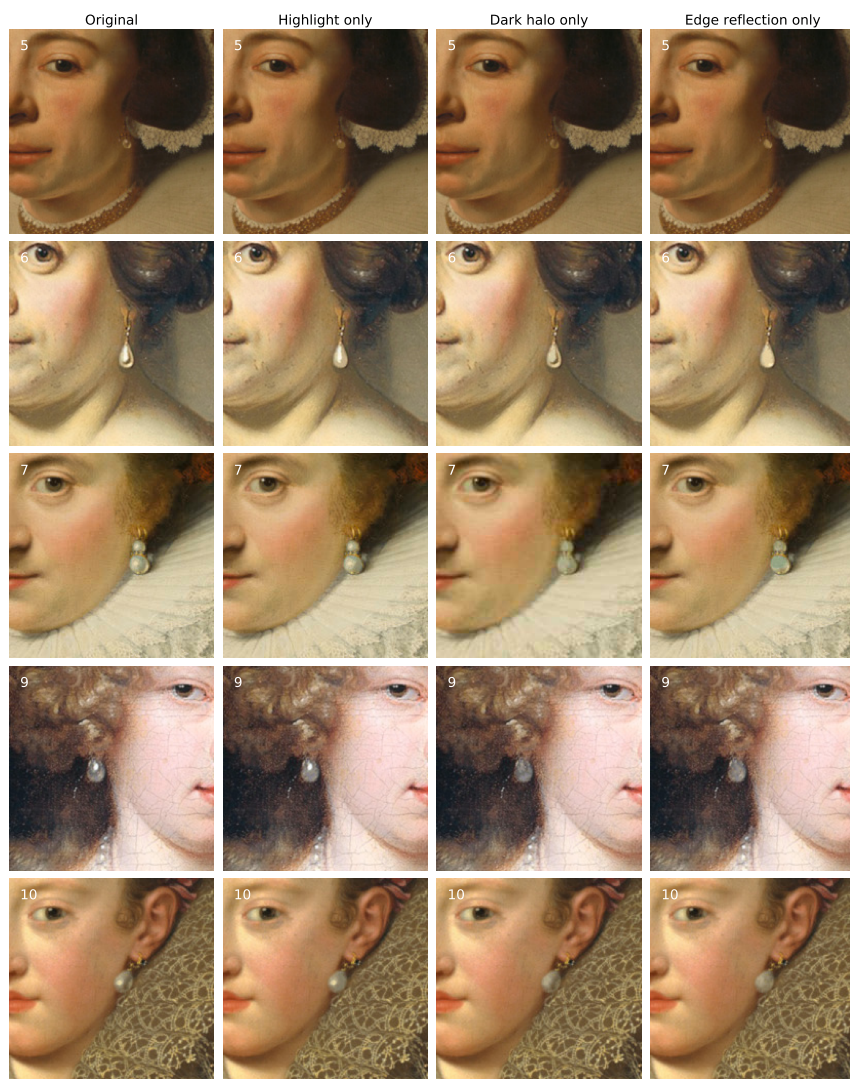


Figure 6.22: Figure continued from previous page.

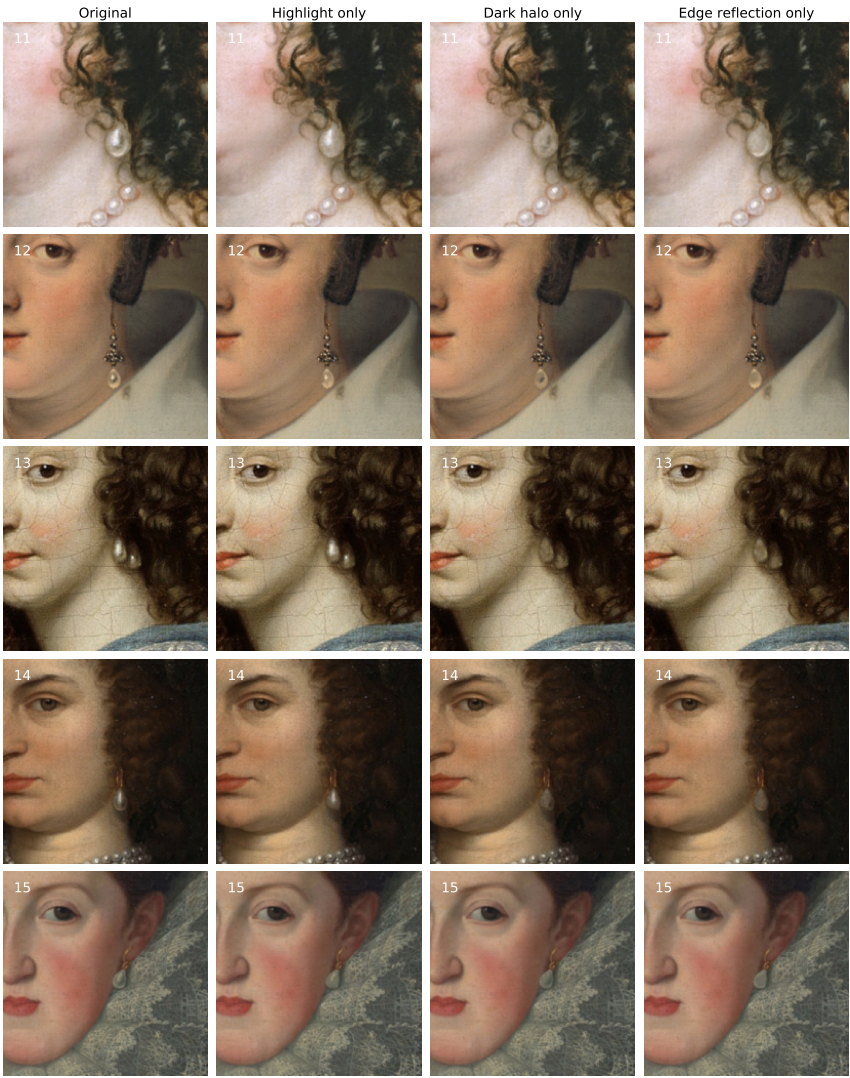


Figure 6.23: Figure continued from previous page.

7

CONCLUSION

Material perception is a vital ability that enables us to successfully interact with the world. This ability extends to the painterly depictions of materials found in paintings. The main aim of this thesis was to study the perception of materials and material attributes depicted within paintings and to enable other researchers to study material depiction within paintings. This aim has been accomplished in several scientific articles that study material perception in paintings and with the release of the Materials In Paintings dataset. Detailed conclusions for each of the studies can be found within the corresponding chapters; here we discuss the implications and contributions of this thesis.

The first major contribution of this is the Materials In Paintings (MIP) dataset. The MIP dataset contains a large set of high-quality, digital reproductions of paintings, representative of western art history. For each of these paintings we have collected a number of annotations relating to material identity and location. We have demonstrated the multidisciplinary utility of this dataset for perception, art history and computer vision.

The annotational paradigm we used to collect the majority of our data relies on human vision to provide insights that may be difficult, if not impossible, to collect using automated, computational methods (Wijntjes, 2021). Often, only *after* collecting human annotations can computer vision leverage these annotations into a computational approach. By using large samples of participants we can use the human visual system to measure or judge image features. For example, while it is trivial to extract color, contrast, or edges from an image computationally, the same can not be said of extracting subjective qualities such as beauty or perceptual judgements of material attributes.

Using this annotation method we collected perceptual judgements of material attributes, such as glossiness, hardness, translucency, etc., in chapter 2 for 15 materials. Using these annotations we reaffirmed the finding of “material signatures” by Fleming et al. (2013), i.e., distributions of material attributes that are related to material categories in distinct and well defined patterns. We showed that similar distributions can be found for materials depicted within paintings as can be found for materials captured within photographs. The implication of this finding is that insights generated by studying painterly material depictions and the perception thereof could transfer to material perception as a whole, i.e., material perception

independently of the medium of depiction.

We found that the averaged “material signatures” distributions are distinct for the coarse-grained material categories (e.g., stone, fabrics, etc.) that we measured. This raised the question if equally distinct and well-defined material signatures could be found for fine-grained material categories (e.g., marble and limestone or velvet and satin). In chapter 5 we measured a number of perceptual attributes for velvet and satin, which are two subordinal material categories of the superordinal category of fabrics. We found that the distributions of measured material attributes for these two materials were also distinct. We speculate that experiments will likely continue to find distinct distributions for material categories as long as we can visually distinguish between those materials. In other words, when we can not visually distinguish between two materials categories the material signatures will likewise not be distinct. While we did not explicitly test for material recognition or distinguishability, we did find that only presenting local information, which we assumed reduced material recognizability, resulted in less distinct distributions of material attributes for velvet and satin.

While the material signatures averaged across participants are distinct and well defined for materials we also found that variation exists between participants when rating perceptual attributes. Specifically, we found that the consistency within and the agreement between participants, while both quite high, displayed some variation. This can be seen in figure 2.4 from chapter 2 which shows the intra- and inter-rater agreements for perceptual attributes averaged over 15 materials. Similar results for within participants consistency have been found for velvet and satin, in figure 5.3 from chapter 5¹. Some variation within the intra- and inter-rater agreement is found in virtually any perceptual experiment and is therefore not very surprising. However, what is interesting is that this variation appears to be dependent on the perceptual attribute being rated. This is interesting as this might explain some instances of misperception, where one material is wrongfully interpreted as another material.

Above we have only discussed the *averaged* distributions of perceived material attributes, which are robust. However, different instances of a material can of course evoke different perceptions. This can for example be seen in the PCA bi-plot shown in figure 2.9: if all instances of a material were perceptually identical in terms of perceptual attributes, than their should be no spread within material clusters within the PCA space. One could argue that this spread is merely noise, but then why do different clusters display different patterns? I would argue that it is the variation in the materials appearance that causes this perceptual variation - and not just noise. However, one might continue to argue that the stimuli displayed a variety of shapes and were depicted within varying contexts, and that might explain the perceptual deviation from the “average” material signature. In chapter 6 we used painterly depictions of pearl earrings, which all had a roughly comparable shape and context, and nevertheless found perceptual variations². We furthermore showed that manipulating a number of perceptual image features related to the material appearance, without manipulating shape or context, caused participants’ perception of the pearls

¹For agreement between participants a different method of calculation was applied that invalidates the comparison.

²It might be worth noting that in chapter 2 we measured perceptual attributes, such as glossiness, roughness, etc., while in chapter 6 we measured pearliness, i.e., the extend to which the stimuli looked like a pearl. As such, there is some asymmetry as in the former we studied material estimation and in the latter material categorization. Nevertheless, we consider both to be expressions of *perceptual variation* in this case.

to change. Taken together this implies that it is variations within the appearance of the material itself that account for these perceptual variations.

In addition to variations within the appearance of materials leading to perceptual variations, for pearl perception we also found striking perceptual variation between two sets of participants; pearl novices and pearl experts. Despite being presented with identical stimuli, i.e., painterly depictions of pearls, these two sets of participants appeared to have significantly diverging perceptions. Specifically, novice participants had a significant preference for pearls in which two image features were digitally removed, while pearl experts indicated that they perceived the original pearls, containing all image features, as more pearly. Note here that from an optical point of view, the original pearl should indeed be considered as more pearly. While it is possible that the reduced pearliness perception for pearl experts was induced by noticing small artefacts of the digital removal of two image features, we speculate that instead the pearl experts displayed a visual expertise, i.e., a perceptual familiarity with pearls from previous interactions. Within vision science the study of visual expertise has largely focused on face perception (Diamond and Carey, 1986; Gauthier and Tarr, 1997; Harel, 2016; Young and Burton, 2018) and to our knowledge little work exists that explicitly considered visual expertise for material perception. In a study by Tani et al. (2014), where novices and pearl experts were asked to order pearls based on quality, it was shown that experts use a method for evaluating quality that was different from what was used by novices. Biederman and Shiffrar (1987) proposed that visual expertise might be related to acquired knowledge of distinguishing features. Indeed, in a recent eye tracking study it was shown that novices learned to attend to critical locations that contained the distinguishing features required for a categorization task (Dickter and Baker, 2019). Without additional studies, we can only speculate further on the existence of such an effect of expertise on material perception. Might expertise with certain materials - or material attributes - explain some of the differences found between participant discussed previously?

FUTURE WORK

The MIP dataset currently contains the paintings from 9 galleries, which we believe gives us a reasonably representational sample of western art. However, by now there are many more digital reproductions of paintings available. Collecting additional paintings would greatly improve the generalizability for any art historical findings. Likewise, paintings could be collected that cover non-western art to improve generalizability.

Besides the collection of additional paintings, the density of the bounding boxes for the existing paintings could be improved. For each painting, we have exhaustively labelled material presence, i.e., we know if a material is present or not. However, we do not always know how often a material is present as material segmentations were not always done exhaustively. Instead, for each painting/material combination we have collected at most five bounding boxes. Increasing this maximum, or removing this upper-limit entirely, would allow for more detailed ecological inquiries. For example, we found that the presence of the coarse grained materials (fabrics, skin, metal, etc.,) within a painting remained relatively stable over time, however it might be that the quantity with which materials occur within a painting does fluctuate.

Furthermore, the existing bounding boxes have not all been assigned with a fine-grained material label because in many cases no consensus was reached between labelling annota-

tors. We can assume that failing to reach consensus is at least partially due to a more difficult stimulus. On the one hand, this implies that the fine-grained labels in our dataset, i.e., the boxes that have reached consensus are robust and clear stimuli. On the other hand, it means that difficult, ambiguous or unclear stimuli are under-represented in our dataset. These difficult stimuli would be very interesting to study. What makes them unclear? Do these stimuli evoke robust but idiosyncratic responses or are they not robust at all? Perhaps alternative methods could be used to reach consensus for difficult images, such as the *consensus game* by Upchurch et al. (2016).

In addition to improving the existing annotations for our set of paintings, there are also novel annotations of which collection could be beneficial. For example, we have demonstrated that annotating image features such as highlights can reveal painterly conventions or the stylized methods of depictions that evoke a robust perceptual response. This demonstrates that simple annotational interfaces can be used to simplify difficult, lengthy, or complex tasks by distributing the labor across a variety of simple and straightforward tasks, or across a large selection of annotators. By for example directly annotating a set of image features we might be able to describe certain materials by the typical image features they depict. More broadly speaking, using annotation methods we could link image features to perceptual responses. The difficulty herein however lies with knowing what to annotate. For example, image features such as the contrast, coverage, sharpness (Marlow and Anderson, 2013) and motion of highlights (Doerschner et al., 2011), as well as the presence of lowlights (Marlow et al., 2012) have been proposed to trigger the perception of shininess. Recently, other perceptual attributes have received increasing attention, such as translucency (Di Cicco, Wiersma, et al., 2020; Di Cicco, Wijntjes, et al., 2020; Fleming and Bülthoff, 2005; Gkioulekas et al., 2015; Gkioulekas et al., 2013) and viscosity (Van Assen et al., 2018; van Assen et al., 2020). For example, it was found by Paulun et al. (2017) that the magnitude of object deformation in response to external forces was strongly related to perceived rigidity. However, for many other attributes it is not always clear what features are related. Visualizing the image regions used by neural networks for material categorization, as we have demonstrated in chapter 5, could help reveal image features or novel cues that trigger the perception of material categories or material attributes.

As we have discussed in previous sections, individual differences exists between human annotators. By discarding individual differences and averaging across annotators we can reveal information encapsulated within the image itself. On the other hand, analysis of these individual differences can reveal insights into the perceptual system. The former, averaging across annotation, could be of direct interest to digital art history. However, to our knowledge human annotations are only rarely used within this field. One example from the literature is given by Carbon and Pastukhov (2018), who collected annotations of the lighting direction for a large set of paintings. Doing so, they were able to confirm the top-left lighting convention, i.e., the convention that lighting predominately originates from the top-left of the scene. While computer vision often replaces human annotations with ground truth, this option is not available for digital art history as ground truth is usually impossible to retrieve or might not exist in the first place. The latter, analyzing individual differences between annotators, can be of value to perception science by providing insights into the human perceptual system. For example, in chapter 2 we use annotators to judge material attributes, such as glossiness and roughness, and found that the (dis-)agreement between

participants depended on the material being judged. Another example can be found in chapter 3 where we had groups of at least 5 annotators label material identities. In this way a piece of fabric would be labelled as silk or velvet, etc. In many cases participants disagreed and no consensus was reached. What leads some participants to interpret a fabric stimulus as silk, while others interpret it as velvet? Do different annotators attend to different image features? Or are image features interpreted differently amongst annotators? We ask these questions explicitly for painting stimuli, but these questions are equally valid for photographs or renderings.

An added benefit of such annotations is that they can be applied for a wide variety of research questions, assuming that the data is publicly available. For example, in chapter 3 we used material bounding box annotations and found that glass windows have a strong spatial bias to be depicted in the top-left of the painting, which appears strongly related to the previously mentioned top-left lighting convention confirmed by Carbon and Pastukhov (2018). The finding that two very dissimilar types of annotations can both point towards a very similar result, i.e., a top-left lighting convention, is indicative not just of a high external validity, but also that annotations might be useful besides their primary purpose. Thus the collection, and especially the publicly sharing, of such annotations could lead to novel insights.

LIMITATIONS

There is one “limitation” that is especially worth discussing, which concerns all chapters of this thesis: the usage of paintings itself. Here we used quotation marks around the word limitations, as we believe the usage of paintings is one of the primary assets of this thesis. Nevertheless, while paintings can provide many benefits, there can also be several drawbacks that warrant discussion.

The first such drawback is the underlying assumption of this thesis related to the artist’s skills of material depiction. As argued several times throughout this thesis, we assume that painters are skilled at depicting material and their attributes. This assumption is based on the observation that, when looking at materials depicted by old master painters such as Van Eyck, Velazquez, or Caravaggio, we experience a rich, convincing perception of these materials. Thus, based on this we feel warranted to assume that, at the least, some painters have the expertise, the skill, to depict materials in a way that robustly triggers the human visual system by capturing the required perceptual cues to evoke material perceptions. By collecting paintings from curated art galleries we aimed to collect paintings that were primarily made by artists who possess this skill of material depiction. However, it is entirely possible that at least some of the paintings in our collection were made by painters without this material expertise. An added complication here is the subjective, unclear nature of this skill or expertise of material depiction.

The history of art has seen countless varying, changing, and evolving methods of depictions. Rarely can clear *before* and *after* periods be distinguished, such as before and after the discovery of perspective. Usually, variations in style appear and disappear gradually across time and regions. Throughout all these changes in depiction, possibly, if not likely, entirely unbeknownst or unintended by the painters, the material depictions have also changed. This too adds to the complexity of any attempt trying to extract or capture this painterly expertise of material depiction. On the other hand however, it makes this material

depiction expertise all the more interesting. How, for example, did painters across time and different styles depict some materials? Furthermore, this material depiction expertise is not only limited to paintings. Food photographers, for example, are known to use many “tricks” to make their photographic results look more appealing to the visual perceptual system, such as for example substituting milk with glue or hair conditioner in an attempt to create a perfect photograph of bowl of cereal (Campbell, 2012). For a more complete understanding of material perception we need a broader study that encompasses a wider variety of artistic material depictions, such as food photographers and digital artists.

In addition to this complex notion of a painters expertise of material depiction, there is also a more physical, objective issue that arises when working with paintings, namely the degradation of paintings. This can lead to unintended perceptual effects. For example, in one unpublished pilot experiment from chapter 2 a piece of porcelain was perceived as very fragile. On visual inspection we noticed that the white paint used to depict this porcelain had deteriorated and had become transparent. Consequently, the porcelain itself appeared so thin as to be transparent. Here, the transparency likely evoked the perception of fragility in a way that was, likely, not intended by the painter. In our current studies we did not take degradation into account and did not account for how degradation might alter the material depictions in ways that were not intended by the painters.

While the evoked perception of a painting can be altered by deterioration, and potentially by restoration, it is in practise not possible to alter the evoked perception in a controlled manner. This results in a fundamental difference between the use of paintings as stimuli and the use of more conventional vision science stimuli, such as photos and renderings. For the latter, experimenters can manipulate and control parameters such as for example the perspective, lighting, size, and the properties of the materials presented. For paintings, comparable experimental control is only possible during the painterly process, which means that researchers would need to create novel painterly stimuli (e.g., Di Cicco et al., 2019) if controlled stimuli are required. Despite this fundamental difference during the stimuli creation, the stimuli and the perceptions these evoke are comparable. All three types of stimuli discussed here (paintings, photos, and renderings) could in theory result in perceptually identical stimuli³, despite the fundamental differences in experimental control discussed here. This lack of control over experimental parameters within the stimulus can be overcome by relating perceptual data to independent variables quantified and measured within the stimuli itself. Furthermore, instead of setting experimental parameters themselves, paintings allow researchers to rely on painters, who have carefully chosen the parameters of depiction in order to achieve an optimal visual representation.

All in all, we conclude that studying painterly depictions for perception introduces novel challenges and complexities, but that these are far out-weighted by the benefits, as we have shown that painters’ perceptual expertise can be leveraged into insights that can be revealed by the perceptual study of painterly depictions.

³The most simple being a monochrome plane, but more complex scenes could in theory be captured, too.

BIBLIOGRAPHY

- Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, memory, and cognition*, 13(4), 640.
- Campbell, T. (2012). *Food photography & lighting: A commercial photographer's guide to creating irresistible images*. New Riders.
- Carbon, C.-C., & Pastukhov, A. (2018). Reliable top-left light convention starts with early renaissance: An extensive approach comprising 10k artworks. *Frontiers in psychology*, 9, 454.
- Di Cicco, F., Wiersma, L., Wijntjes, M., & Pont, S. (2020). Material properties and image cues for convincing grapes: The know-how of the 17th-century pictorial recipe by willem beurs. *Art & Perception*, 8(3-4), 337–362.
- Di Cicco, F., Wijntjes, M. W., & Pont, S. C. (2019). Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *Journal of Vision*, 19(3). <https://doi.org/10.1167/19.3.7>
- Di Cicco, F., Wijntjes, M. W., & Pont, S. C. (2020). If painters give you lemons, squeeze the knowledge out of them. a study on the visual perception of the translucent and juicy appearance of citrus fruits in paintings. *Journal of vision*, 20(13), 12–12.
- Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of experimental psychology: general*, 115(2), 107.
- Dickter, A. H., & Baker, C. I. (2019). Impact of developing perceptual expertise on eye fixations. *Journal of Vision*, 19(10), 31b–31b.
- Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., & Kersten, D. (2011). Visual motion and the perception of surface material. *Current Biology*, 21(23), 2010–2016.
- Fleming, R. W., & Bühlhoff, H. H. (2005). Low-Level Image Cues in the Perception of Translucent Materials. *ACM Transactions on Applied Perception*, 2(3), 346–382. <https://doi.org/10.1145/1077399.1077409>
- Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8), 9. <https://doi.org/10.1167/13.8.9>
- Gauthier, I., & Tarr, M. J. (1997). Becoming a “greeble” expert: Exploring mechanisms for face recognition. *Vision research*, 37(12), 1673–1682.
- Gkioulekas, I., Walter, B., Adelson, E. H., Bala, K., & Zickler, T. (2015). On the appearance of translucent edges. *Proceedings of the ieee conference on computer vision and pattern recognition*, 5528–5536.
- Gkioulekas, I., Xiao, B., Zhao, S., Adelson, E. H., Zickler, T., & Bala, K. (2013). Understanding the role of phase function in translucent appearance. *ACM Transactions on graphics (TOG)*, 32(5), 1–19.
- Harel, A. (2016). What is special about expertise? visual expertise reveals the interactive nature of real-world object recognition. *Neuropsychologia*, 83, 88–99.

- Marlow, & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision*, 13(14), 1–23. <https://doi.org/10.1167/13.14.2>
- Marlow, Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, 22(20), 1909–1913. <https://doi.org/10.1016/j.cub.2012.08.009>
- Paulun, V. C., Schmidt, F., van Assen, J. J. R., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of vision*, 17(1), 20–20.
- Tani, Y., Nagai, T., Koida, K., Kitazaki, M., & Nakauchi, S. (2014). Experts and novices use the same factors—but differently—to evaluate pearl quality. *PloS one*, 9(1), e86400.
- Upchurch, P., Sedra, D., Mullen, A., Hirsh, H., & Bala, K. (2016). Interactive consensus agreement games for labeling images. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 4(1).
- Van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current biology*, 28(3), 452–458.
- van Assen, J. J. R., Nishida, S., & Fleming, R. W. (2020). Visual perception of liquids: Insights from deep neural networks. *PLoS computational biology*, 16(8), e1008018.
- Wijntjes, M. (2021). Shadows, highlights and faces: The contribution of a ‘human in the loop’ to digital art history. *Art & Perception*, 9(1), 66–89.
- Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in cognitive sciences*, 22(2), 100–110.

ACKNOWLEDGEMENTS

If you are reading this, that means that I have finished my thesis and my PhD project has come to an end. In the past four years, I have worked hard to bring this moment forth. Of course, I did not do this alone. Many people have supported me during this journey.

First, I want to thank my supervisors. Maarten, without you none of this would have been possible. You made me feel welcome and comfortable before my first job interview even started. Thank you for your patience when I disagreed with you ten times in a row (when you were usually right after all), or with my countless questions. I have learned a lot in these past four years and much of this is because of you! Thank you. Sylvia, thank you for your endless council and feedback. Your critical feedback has helped me improve greatly while your positive feedback has greatly improved my confidence in my skills and abilities. Once, a few months after I started my PhD, you jokingly mentioned our academic family tree - something along the lines of your PhD promotor being my PhD grandfather-promotor. You truly did make it feel like a family, and I will always be grateful for that.

Next, my fellow PhD students and postdocs. Francesca, buongiorno, thank you for being there during my entire PhD journey. You always listened to me complain and even let me self-proclaim myself as Sofia's favorite uncle. Thanks for all the drinks, dinners, laughs, and collaborations! I hope we can have more fun conferences in the future. Cristina, thank you for the energy you introduced into the lab. It was a pleasure working with you and answering al jou vragen over Nederlands! Karina, thank you for the drinks and chats, there was never a boring moment when you were in the office. Yuguang, thank you for continuing the project once I depart. It's a shame that we were rarely in the office together. I hope we will get plenty of chances to get those drinks at the next VSS! Cehao, thank you for making the lab a little livelier. We never did manage to get a group outing to Spicy City. Jan Jaap, we were only in the lab for a short time together and now it seems like I will follow in your footsteps. Jess, thank you for the fun lunch breaks! Each time I asked you about anything, you could give me numerous references and a great answer. You taught me a lot about things I didn't even yet know I was interested in! Fan and Tatiana, thank you for going before, and for showing me what to expect. I would also like to thank Kavita, from Cornell University, for inviting me to work from the CS department. Paul and Hubert, you helped me a lot with many of the technical details of the project when I was starting out. This was super helpful and has taught me a lot!

Besides my colleagues, there was also my real family. The next part will be written in Dutch, for them. Mam, pap, Maikel, bedankt voor alles. Zonder jullie was ik hier niet. Ik heb misschien nooit echt duidelijk gemaakt wat ik nou eigenlijk heb uitgespookt de afgelopen jaren, maar jullie waren altijd vol trots. Bedankt voor alle liefde en support! Maik, vooral je statistiek vragen de afgelopen maanden hebben mij zelf ook weer veel geleerd.

For Sebastiaan and Lucas, and all others from *vulululu*, thank you for showing me there is more to life than work. Thank you Rick, for wagering that you would publish a book

before I would: I win. Thank you to Ada, for showing me “it’s just a job”. Thank you to Jorrit, for showing me the difference between Java and JavaScript.

Last, thank you to my loving partner, Bárbara, who insisted I shouldn’t mention her here. It is a good thing I don’t always listen to you. Thank you for all your support and love, and for listening to me complain for countless hours. To the moon and back.

For all those not mentioned by name, who are still reading, know that I am grateful for the invaluable support freely given over the past years. For all those reading this: thank you!

SUMMARY

The world around us is filled with materials. Our ability of visual material perception informs us how to navigate and interact with our environment. It tells us, for example, whether food is fresh, if a chair is strong enough to sit on, how much force to use to pick up a glass, etc. Painters have studied how to depict the world and the materials therein for thousands of years. We believe that the material depictions within paintings can be leveraged into insights for the scientific understanding of material perception. In this thesis, we studied the perception of painterly depictions of materials and aimed to make the study thereof accessible to other researchers with the release of the Materials In Paintings dataset. We collected a large set of paintings from museums and galleries. Then, we used an online crowd-sourcing approach to annotate material identity (fabrics, stone, etc.) and gather spatial material segmentations (i.e., “cutting out” piece of the painting that depict the material). In the first study, we measured the perception of material attributes (soft, rough, fragile, etc.) across a range of materials and found that painterly materials trigger distinct distributions of perceived attributes and we furthermore compared these distributions to those for photographic materials. In the second study, we continued crowd-sourcing annotations on material identity and material segmentations and combined these into the Materials In Paintings dataset. In a number of cross-disciplinary demonstrations we presented novel findings across art history, human perception, and computer vision. While these demonstrations are useful in their own right, the main focus here was the release of the dataset. Next, we used the dataset as a source of stimuli for two studies into specific materials. First, for fabrics, we studied the perception of satin and velvet and the effect of presenting only local or, both local and global information, and found that the perceptual distinction between these two fabrics becomes more ambiguous when removing global information. Furthermore, we showed that local image cues can affect perceptual responses for shininess but not for softness. Lastly, we studied the perception and depiction of pearls by identifying three image features that might trigger the perception of pearliness. In a series of experiments, we confirm the role of these image features but find that the presence of only one of these image features, highlights, is already sufficient for naive participants to trigger the perception of pearliness. Conversely, expert participants (art historians or pearl experts) perceive depictions with all three features as more pearly, which implies the existence of visual expertise for pearl perception. All in all, in this thesis we show the benefits of studying material perception through painterly depictions of materials and enable further study with the release of the MIP dataset.

SAMENVATTING

De wereld om ons heen is vol met materialen. Ons vermogen om deze materialen visueel waar te nemen helpt ons om te navigeren en interacteren met onze omgeving. Het vertelt ons bijvoorbeeld of voedsel vers is, of een stoel sterk genoeg is om op te zitten, hoeveel kracht we moeten gebruiken om een glas op te pakken, of de vloer glad is enzovoort. Schilders hebben duizenden jaren bestudeerd hoe zij de wereld en de materialen daarin kunnen weer-geven. Wij geloven dat de schilderkunstige weergave van materialen kan worden benut om inzicht te creëren in materiaal perceptie. In dit proefschrift bestuderen we de perceptie van schilderkunstige afbeeldingen van materialen en probeerden we de studie daarvan toegankelijk te maken voor andere onderzoekers met de publicatie van de *Materials In Paintings* dataset. We verzamelden een groot aantal schilderijen van musea en galerieën. Vervolgens gebruiken we een online crowd-sourcing benadering om materiaalidentiteit (hout, steen, metaal, enz.) te annoteren en materiaalsegmentaties te verzamelen (d.w.z. stukken digitaal “uitsnijden” van het schilderij, waarin alleen het materiaal in kwestie is afgebeeld). In de eerste studie meten we de perceptie van materiaal attributen (zacht, ruw, breekbaar, enz.) in een reeks materialen en vinden dat verschillende schilderkunstige materialen verschillende distributies van materiaal attributen vertonen. Verder vergelijken we deze distributies met die voor fotografische materialen. In de tweede studie gaan we verder met crowd-sourcing-annotaties over materiaalidentiteit en materiaalsegmentaties en combineerden we deze in de *Materials In Paintings* dataset. In een aantal multidisciplinaire demonstraties werden nieuwe bevindingen voor de kunstgeschiedenis, visuele perceptie en computer-visie gepresenteerd. Hoewel deze demonstraties op zichzelf al nuttig zijn, lag de focus hier op het publiceren en introduceren van de dataset. Vervolgens gebruikten we de dataset als bron voor twee onderzoeken naar specifieke materialen. Ten eerste bestudeerden we voor textiel de perceptie van satijn en fluweel en het effect van het presenteren van alleen lokale of, zowel lokale als globale informatie. Hierbij vonden we dat het perceptuele onderscheid tussen deze twee stoffen minder duidelijker wordt bij het verwijderen van globale informatie. Bovendien lieten we zien dat lokale beeldsignalen de perceptuele reacties voor *glans* kunnen beïnvloeden, maar niet op *zachtheid*. Ten slotte bestudeerden we de perceptie en afbeelding van parels. Hierbij identificeerden we drie beeldkenmerken waarvan wij stellen dat deze de perceptie van parelachtigheid kunnen opwekken. In een reeks experimenten bevestigden we de rol van deze beeldkenmerken en vonden we dat de aanwezigheid van slechts één van deze beeldkenmerken, highlights, al voldoende is voor naïeve deelnemers om de perceptie van parelachtigheid op te wekken. Echter, deskundige deelnemers (kunst-historici of parelexperts) geven aan dat zij de afbeeldingen met alle drie de kenmerken als meer parelachtig ervaren, wat impliceert dat er visuele expertise bestaat voor de perceptie van de parel. Al met al lieten we in dit proefschrift de voordelen zien van het bestuderen van materiaal perceptie door middel van schilderkunstige afbeeldingen van materialen en maken we verder onderzoek mogelijk met de publicatie van de MIP-dataset.

LIST OF PUBLICATIONS

Papers

8. **van Zuijlen, M.J.P.**, Lin, H., Bala, K., Pont, S.C., Wijntjes, M.W.A. (2021). Materials In Paintings (MIP): An interdisciplinary dataset for perception, art history, and computer vision. *Plos one*, 16(8), e0255109.
7. **van Zuijlen***, **M.J.P.**, Di Cicco*, F., Wijntjes, M.W.A., Pont, S.C. (2021). Soft like velvet and shiny like satin: perceptual material signatures of fabrics depicted in 17th century paintings. *Journal Of Vision.*, 2021(5) 1 - 22. * authors contributed equally
6. **van Zuijlen***, **M.J.P.**, Di Cicco*, F., Wijntjes, M.W.A., Pont, S.C. (2021). On perceiving the lustre of pearls. When less is more. *Submitted*. * authors contributed equally
5. Lin, H., **van Zuijlen, M.J.P.**, Wijntjes, M.W.A., Pont, S.C., Bala, K (2020). Insights From A Large-Scale Database of Material Depictions In Paintings. *International Conference on Pattern Recognition*, 2021(531-545).
4. Lin, H., **van Zuijlen, M.J.P.**, Wijntjes, M. W., Bala, K. (2021). What Can Style Transfer and Paintings Do For Model Robustness? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
3. Wijntjes, M.W.A. **van Zuijlen, M.J.P.** (2021). Sketch and test: visual crowd research using p5.js. *Submitted*.
2. **van Zuijlen, M. J.P.**, Pont, S. C., Wijntjes, M. W.A. (2020). Painterly depiction of material properties., *Journal of Vision*. 20(7), 1-17.
1. **van Zuijlen, M.J.P.**, Pont, S.C., Wijntjes, M.W.A. (2020). Conventions and temporal differences in painted faces: A study of posture and color distribution. *Human Vision and Electronic Imaging proceedings*, 2020(11), 1-7.

Abstracts

12. **van Zuijlen, M.J.P.**, Pont, S.C., Wijntjes, M.W.A. (2017). Depicted material categories in online museum collections. *Visual Science of Art Conference (VSAC), Berlin. Art Perception*, 5(2017) 337-426.
11. **van Zuijlen, M.J.P.**, Pont, S.C., Wijntjes, M.W.A. (2017). Human classification of depicted materials in paintings. *European Conference on Visual Perception (ECVP), Berlin, 2017. Perception*.
10. **van Zuijlen, M.J.P.**, Pont, S.C., Wijntjes, M.W.A. (2018). Has the depiction of the human face changed over the centuries? *Visual Science of Art Conference (VSAC), Trieste. Art Perception*, 6(2018) 151-207.

9. Hulbert, A., **van Zuijlen, M.J.P.**, Spoiala, C., Wijntjes, M.W.A. (2018). Painting the time of day: Colour determines perceived circadian phase in visual art *Visual Science of Art Conference (VSAC), Trieste. Art Perception*, 6(2018) 151-207.
8. **van Zuijlen, M.J.P.**, di Cicco, F., Pont, S.C., Wijntjes, M.W.A. (2018). Material cues from the past: Experimentally testing a historical description of material rendering. *European Conference on Visual Perception (ECVP), Trieste. Perception*, 48(s1). 1 - 233
7. **van Zuijlen, M.J.P.**, Upchurch, P., Pont, S.C., Wijntjes, M.W.A. (2019). Material property space analysis for depicted materials. *Vision Sciences Society (VSS). Journal of Vision*, 19(10), 251a-251a..
6. **van Zuijlen, M.J.P.**, Pont, S.C., Wijntjes, M.W.A. (2019). Relations Between Material Classes, Material Attributes, and Image Statistics *European Conference on Visual Perception (ECVP), Trieste. Perception*, 48(2s). 1 - 236
5. **van Zuijlen, M.J.P.**, Pont, S.C., Wijntjes, M.W.A. (2018). Material Incidence and spatial analysis of materials in Visual Pre-Modern Art. *Visual Science of Art Conference (VSAC), Leuven. Art Perception*, 7(2019) 251-334.
4. di Cicco, F., **van Zuijlen, M.J.P.**, Pont, S.C., Wijntjes, M.W.A. (2020). Perceptual signatures of velvet and satin depicted in 17th century paintings. *Vision Sciences Society (VSS). Journal of Vision*, 20(11), 481-481..
3. **van Zuijlen, M.J.P.**, Lin, H., Bala, K., Pont, S.C., Wijntjes, M.W.A. (2020). A Database of Painterly Material Depictions. *Vision Sciences Society (VSS). Journal of Vision*, 20(11), 1127-1127..
2. Louwers, G.L.M., Pont, S.C., Havranek, M., **Van Zuijlen, M.J.P.** (2020). Designing with Dynamic Light Textures - Enlightening Designers *LED Professional Review: Trends & Technologie for Future Lighting Solutions.*, (80), 22.
1. **van Zuijlen, M.J.P.**, Lin, H., Bala, K., Pont, S.C., Wijntjes, M.W.A. (2021). Painterly depictions of glasses display a stylized pattern of highlights *Vision Sciences Society (VSS), accepted*.