Delft University of Technology

Appearance rendering by painters, engravers and generative AIs
Material perception and depiction across different styles and media

Zhao, Y.

DOI
10.4233/uuid:b105d4d4-c85c-4a02-8aad-1f0067fd853b

Publication date
2025

Document Version
Final published version

Citation (APA)
Zhao, Y. (2025). *Appearance rendering by painters, engravers and generative AIs: Material perception and depiction across different styles and media*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b105d4d4-c85c-4a02-8aad-1f0067fd853b

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.

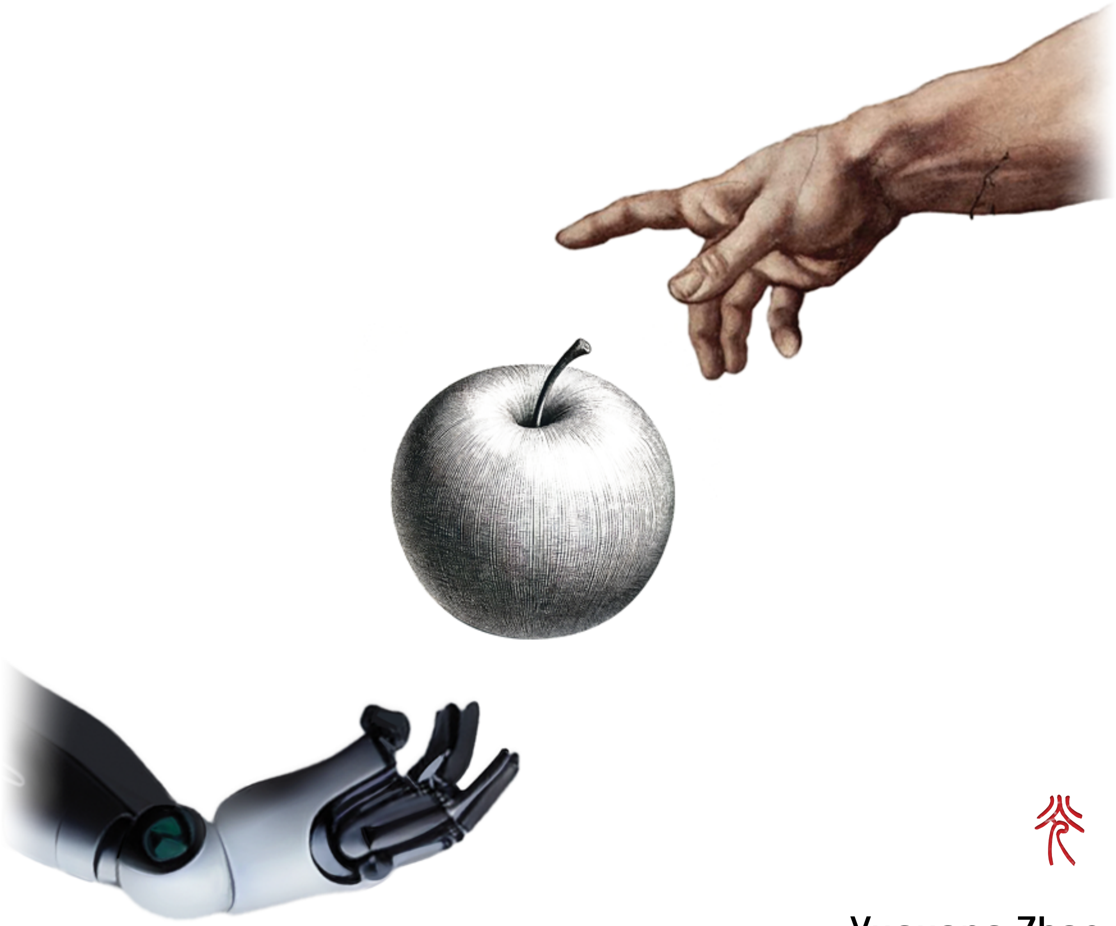
Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

APPEARANCE RENDERING

by Painters, Engravers and Generative AIs

Material Perception and Depiction
Across Different Styles and Media



Yuguang Zhao

Appearance rendering by painters, engravers and generative AIs

Material perception and depiction across different styles and media

APPEARANCE RENDERING BY PAINTERS, ENGRAVERS AND GENERATIVE AIS

MATERIAL PERCEPTION AND DEPICTION ACROSS DIFFERENT STYLES AND MEDIA

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Wednesday 16 April 2025 at 12:30 o'clock

by

Yuguang ZHAO

Master of Science in Human-Technology Interaction,
Eindhoven University of Technology, the Netherlands,
born in Taiyuan, China.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Dr. M.W.A. Wijntjes	Delft University of Technology, promotor
Prof. dr. H. de Ridder	Delft University of Technology, promotor
Independent members:	
Prof. dr. J. Dik	Delft University of Technology
Prof. dr. J. Wagemans	KU Leuven
Prof. dr. A.C. Hurlbert	Newcastle University
Dr. N.J.E. van Noord	University of Amsterdam
Prof. dr. ir. R.H.M. Goossens	Delft University of Technology, reserve member
Non-independent member:	
Prof. dr. J.F.H.J. Stumpel	Utrecht University

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, art and perception, depiction style, pictorial media, generative AI

Printed by: Proefschrift All-in-One

Front & Back: Cover designed and created by Yuguang Zhao, incorporating a detail from: The Creation of Adam by *Michelangelo di Lodovico Buonarroti Simoni*, 1508-1512.

Copyright © 2024 by Y. Zhao

ISBN 978-94-93406-47-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

CONTENTS

SUMMARY

In contemporary society, we are surrounded by not only physical materials, but also images of them. We are capable of judging materials and their properties with only visual information. For instance, if an object looks solid and glossy or soft and fluffy. This ability is called material perception. As for images, there are various ways of image making, such as photography, painting, computer rendering, etc. And a new method has emerged recently: generative AI. All these image generation methods can produce different appearances of the same object or material. In this thesis, we studied human visual perception of two types of appearances: appearance as material property, appearance as pictorial style and the interaction between them.

In Chapter 2, we investigated depiction style by zooming in on a single motif, an apple. By using the fragments instead of the whole painting, we were able to keep the subject matter relatively constant, and isolate style from composition as well as other contextual information. We first constructed a perceptual space of style using similarity judgements from online participants. Then we fitted perceived attributes to this space to understand its dimensions. The data resulted in a three-dimensional space. Dimension 1 is associated with smoothness and brushstroke coarseness. Dimensions 2 and 3 are related to hue and chroma. Surprisingly, we also found a rotational relation between creation year and the first two dimensions, revealing a certain cyclic, repetitive pattern of style. The results suggest style can already be perceived in fragments of paintings.

In Chapter 3, we studied the influence of medium on appearance. For example, imagine an oil-painted apple and a pencil-sketched apple: they can have different appearances. The comparison between different media has rarely been studied. One possible reason is the difficulty to isolate medium from its confounding factor, subject matter. We found a solution by comparing oil paintings and their engraved reproductions. The identical content gave us a perfect opportunity to compare material perception from two distinct media. We collected 15 pairs, consisting of 88 fragments depicting different materials like fabric, skin, wood and metal. We also created three manipulations to understand the effect of color (a grayscale version) and contrast (equalized histograms towards both painting and engraving). We collected ratings on five attributes: three-dimensionality, glossiness, convincingness, smoothness and softness. Paintings showed a broader perceived range than engravings, with contrast equalization having a greater impact on perception than color removal. Possibly engravers used local contrast to compensate the absence of color.

In Chapter 4, we analyzed an emerging medium from a non-human creator, generative AI. In two experiments, we explored human material perception using generative AI stimuli and compared the perceptual spaces of three generative AI models, as well as a computer-generated BRDF stimulus set, the MERL dataset. In Experiment 1, we used text descriptions of 32 materials from MERL (e.g. ‘green fabric’) as prompts for DALL-E 2 and Midjourney v2. Both AI models resulted in a 2D space while MERL resulted in a 1D one. The three spaces showed low similarity, suggesting the AI models generated unique and different images of materials from identical text prompts. In Experiment 2, we explored another text-to-image model Stable Diffusion v1.5 with an add-on, ControlNet. ControlNet allowed us to add additional graphical constraints besides text input. In this way we could inspect more complex shapes. We kept the same 32 descriptions and generated material blobs in three shapes, from simple to more complex geometry. The three perceptual spaces from the three shapes showed high similarity, indicating both robust structure and minor influence of object shape on material perception. Interestingly, the perceptual spaces from Experiment 2 also shared similar structure as perceptual spaces from other material studies using real-world photos, computer renderings and depictions. In sum, we investigated visual perception through the lens of art by examining appearances rendered by painters, engravers and generative AIs.

Samenvatting

In de hedendaagse maatschappij worden we niet alleen omringd door fysieke materialen maar ook door afbeeldingen ervan. We zijn in staat om materialen en hun eigenschappen te beoordelen met behulp van alleen maar visuele informatie. Bijvoorbeeld, of een object er solide en glanzend uitziet of zacht en pluizig. Dit vermogen wordt materiële perceptie genoemd. Wat betreft afbeeldingen zijn er verschillende manieren om afbeeldingen te maken, zoals fotografie, schilderen, computer rendering, etc. En recent is er een nieuwe methode bijgekomen: generatieve AI. Al deze methoden voor het genereren van afbeeldingen kunnen verschillende verschijningsvormen van hetzelfde object of materiaal produceren. In dit proefschrift bestudeerden we de menselijke visuele perceptie van twee verschijningsvormen: niet alleen als materiële eigenschap maar ook als picturale stijl plus de interactie daar tussen.

In Hoofdstuk 2 onderzochten we de weergavestijl door in te zoomen op een enkel motief, een appel. Door de fragmenten te gebruiken in plaats van het hele schilderij, konden we het onderwerp relatief constant houden en stijl isoleren van compositie en andere contextuele informatie. We construeerden eerst een perceptuele ruimte voor stijl met behulp van gelijkenisoordelen van online deelnemers. Vervolgens pasten we waargenomen kenmerken in deze ruimte om de dimensies te begrijpen. De gegevens resulteerden in een driedimensionale ruimte. Dimensie 1 wordt geassocieerd met gladheid en grofheid van de penseelstreek. Dimensies 2 en 3 zijn gerelateerd aan tint en chroma. Verrassend genoeg vonden we ook een circulaire relatie tussen het scheppingsjaar en de eerste twee dimensies, wat een bepaald cyclisch, repetitief patroon van stijl onthulde. De resultaten suggereren dat stijl al kan worden waargenomen in fragmenten van schilderijen.

In Hoofdstuk 3 bestudeerden we de invloed van medium op de verschijningsvorm. Stel je bijvoorbeeld een met olieverf geschilderde appel voor en een met potlood geschetste appel: ze kunnen er anders uitzien. De vergelijkingen tussen verschillende media zijn zelden bestudeerd. Een mogelijke reden is de moeilijkheid om medium te isoleren van het onderwerp zijnde een verstorende factor. We vonden een oplossing door olieverfschilderijen en hun gegraveerde reproducties te vergelijken. De identieke inhoud gaf ons een perfecte gelegenheid om de materiële perceptie van twee verschillende media te vergelijken. We verzamelden 15 paren, bestaande uit 88 fragmenten die verschillende materialen afbeelden, zoals stof, huid, hout en metaal. We hebben ook drie manipulaties gecreëerd om het effect van kleur (via een grijstintenversie) en contrast (via gelijkgemaakte histogrammen voor zowel schilderij als gravure) te begrijpen. We hebben beoordelingen verzameld voor vijf kenmerken: driedimensionaliteit, glans, overtuigingskracht, gladheid en zachtheid. Schilderijen lieten een breder waargenomen bereik zien dan gravures, waarbij het gelijktrekken van contrast een grotere impact had op de perceptie dan het verwijderen van kleuren. Mogelijk gebruikten graveurs lokaal contrast om de afwezigheid van kleur te compenseren.

In Hoofdstuk 4 hebben we een opkomend medium van een niet-menselijke maker geanalyseerd, te weten generatieve AI. In twee experimenten hebben we menselijke materiële perceptie onderzocht met behulp van generatieve AI-stimuli en de perceptuele ruimtes van drie generatieve AI-modellen door ze niet alleen met elkaar te vergelijken maar ook met een door de computer gegenereerde BRDF-stimulusset, de MERL-dataset. In Experiment 1 hebben we tekstbeschrijvingen van 32 materialen van MERL (bijv. 'groene stof') gebruikt als prompts voor DALL-E 2 en Midjourney v2. Beide AI-modellen resulteerden in een 2D-ruimte, terwijl MERL resulteerde in een 1D-ruimte. De drie ruimtes vertoonden weinig gelijkenis, wat suggereert dat de AI-modellen unieke en verschillende afbeeldingen van materialen genereerden uit identieke tekstprompts. In Experiment 2 onderzochten we een ander tekst-naar-afbeeldingsmodel, te weten Stable Diffusion v1.5 met een add-on, ControlNet. Met ControlNet konden we extra grafische beperkingen toevoegen naast tekstinvoer. Op deze manier konden we complexere vormen onderzoeken. We behielden dezelfde 32 beschrijvingen en genereerden per materiaal blobs in drie vormen, met de geometrie variërend van eenvoudig tot meer complex. De drie perceptuele ruimtes van de drie vormen vertoonden veel gelijkenis, wat duidt op zowel een robuuste structuur als een kleine invloed van de objectvorm op de materiële perceptie. Interessant genoeg deelden de perceptuele ruimtes van Experiment 2 ook een vergelijkbare structuur met perceptuele ruimtes van andere materiaalstudies verkregen met behulp van echte foto's, computerweergave en afbeeldingen. Samengevat, we onderzochten visuele perceptie door de lens van kunst via het nagaan van verschijningsvormen zoals gecreëerd door schilders, graveurs en generatieve AI's.

1

Introduction

1.1. Material Perception and Depiction

We are constantly surrounded by a variety of materials that make up our environment, from wood table tops to ceramic coffee mugs. We use our visual sense to understand and interact with these materials. Without touching, we can already recognize the world around us and even judge the properties of the materials we encounter. Humans have the ability to visually recognize materials with high speed and accuracy (Fleming, 2017; Sharan et al., 2009, 2014). This ability, which can be called material perception, helps us to interact with the world. For example, it can help us to judge if a fruit is ripe or if a surface is solid to step onto. In addition to physical materials, as someone who lives in the modern society, we also encounter enormous amounts of depicted materials. For example, magazines, photographs, computer rendered movie scenes and images from generative artificial intelligence (AI) models. Although in the end they are just printed ink on a piece of paper, or pixels on a screen instead of physical objects, we can still easily recognize different materials and infer their properties from the images. This underlines the powerful utility of images for humans.

For physical materials, recognizing them visually and inferring their properties rapidly is actually a complex task. The information received by the retina are complex light patterns, shaped directly by the objects' three-dimensional properties, reflectance, and transmittance (Anderson, 2011). Recognizing materials from images shares some commonalities, but there are also some differences. First, images are produced with various techniques. Photographs are rather direct captures of the real world. Instead of the retina, light is projected onto a film or a digital sensor. Computer-generated imagery (CGI) also takes a physics based approach. To generate a 2D image with CGI, a physics based rendering engine would require the 3D model of the object, its surroundings, their material properties, the light source and sometimes the environmental light map. It calculates how the light interacts with the scene from the corresponding viewpoint. Visual arts such as drawings and paintings, on the other hand, take a different approach. Usually created directly on a 2D surface, a painting can be seen as the artist's subjective interpretation of how objects are perceived. Paintings not necessarily fully obey the rules of physics (Cavanagh, 2005), even for those that are intended to be realistic rather than stylistic. For instance, the shadows of some objects might violate the shadow directions of the rest of the scene, but viewers might not notice the inconsistency since the human visual system is insensitive to illumination inconsistencies (Jacobson and Werner, 2004; Mamassian, 2004; Ostrovsky et al., 2005; Wijntjes and de Ridder, 2014). Or, as Figure 1.1 shows, different objects in the same scene might have inconsistent light sources, even though, at a glance, nothing seems to be wrong. Throughout history, artists have found shortcuts to depict convincing scenes without fully following the laws of physics. The shortcuts work for viewers since the viewers share a similar visual system as the artist. In other words, art is made from perception and for perception. This makes art a great material for understanding human's visual perception. In this thesis, we will study visual perception through the lens of art.

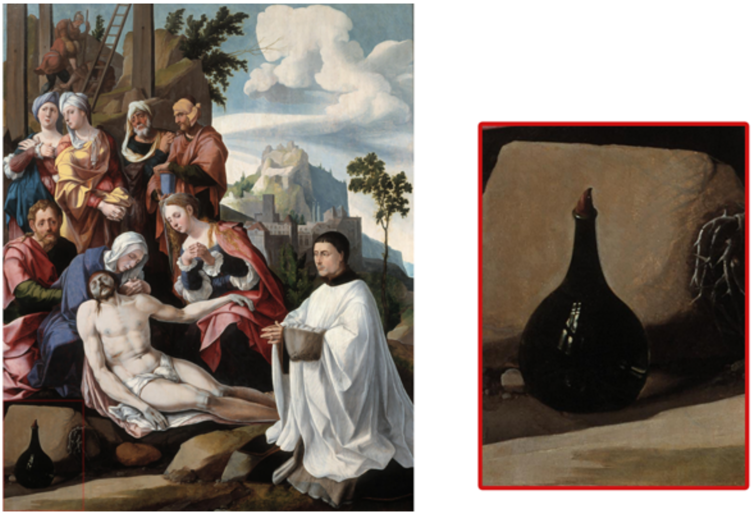


Figure 1.1: All the human figures appear to be outside. However, the glass bottle in the bottom left corner (enlarged on the right side) appears to be indoors. Its small highlight indicates a small light source possibly in the shape of a window. Jan van Scorel, *The Lamentation of Christ*, 1535. Downloaded from the online repository of Centraal Museum, Utrecht.

1.2. RATIONALE OF THE THESIS

In visual arts, the same object or material can be rendered in different ways, resulting in different appearances. For example, flowers depicted by van Gogh have a different appearance than flowers from de Heem; a painted dress can have a different appearance than an engraved one; even the latest AI generated images have, what may be called, an 'AI look'. In the example of depicting flowers, the two artists made their choices of how something is depicted, which can be called style. We will investigate visual style in Chapter 2. The comparison between painting or engraving the same object or material actually reflects the influence of different media on appearance, which will be discussed in Chapter 3. And Chapter 4 will dive into AI generated appearances.

IT STARTED FROM STYLE

The emerging technique of style transfer (Gatys et al., 2016) and the rapid development of generative AI have brought renewed attention to the topic of visual style. These tools make it easy nowadays to create desired images in different styles with text descriptions, prompts (e.g., a flower in the style of watercolor, or a flower in the style of Monet). Style is not a new topic in art history nor in vision science. Approximately one century ago, Heinrich Wölfflin used a top-down approach employing pre-defined concepts to differentiate the Renaissance from the Baroque (Wölfflin, 2012). To this end, he came up with five principles for describing style differences. However, the approach has its limitations when exploring unknown concepts we are interested in. Here, bottom-up methods such as multidimensional scaling (MDS) are to be preferred. Often used in perception studies (Agarwal et al., 2007; Di Cicco et al., 2020; Ferwerda et al., 2001; Graham et al., 2010; Hebart et al., 2020; Toscani et al., 2020), MDS first collects (dis)similarity judgements from observers among pairs of stimuli to obtain a measure for the perceptual distances. Then it constructs a low-dimensional space where the distance between points reflects as good as possible their perceptual distance (see Mead (1992) for a review). Note that the term perceptual space is often interchangeable with 'embedding' or 'MDS solution'. We found style studies using the MDS method, with no (Berlyne and Ogilvie, 1974) or little control (O'Hare, 1976; Ruth and Kolehmainen, 1974) on the content (subject matter). Content and composition were found to be part of the stylistic choices. In the current project, we are interested more in *how* something is depicted instead of *what*. Hence we will strive for isolating style from confounding factors such as subject matter. For example, Figure 1.2 illustrates two oil paintings that are both substantially different in style and subject matter. It can be challenging, or at least not straightforward to focus on style differences only.

We are also aware of other style research from different perspectives, such as image statistics (Rao et al., 1999; Sablatnig et al., 1998) or computational approaches (Elgammal et al., 2018; Graham et al., 2010). In this project we would like to focus on the angle of human perception, as I argued in section 1.1. In sum, we are interested in studying how humans perceive an object or material when rendered in different ways.

METHODOLOGY: A PILOT STUDY ON FLOWER STILL-LIFE PAINTINGS

Flower still life paintings first catched my attention. I noticed that they share relatively uniform content: a collection of flowers placed in a vase, the vase is often on a surface such as a table. So we conducted the very first experiment with flower still life oil paintings. In the first online experiment, Pilot Experiment 1, we collected pairwise similarity judgements on ten flower still life paintings from 20 online participants. The selection of paintings covered three art movements, Baroque, Impressionism and Realism. Their creation years ranged from early 18th century to late 19th century. We asked for stylistic similarity ratings between each pair of the paintings, without specifying the concept of style in too much detail, leaving room for participants to have their own interpretation and criteria. The data analysis resulted in a 2D perceptual space as suggested by the stress values. The stress value is a key indicator to determine the dimensionality of an MDS analysis (Kruskal, 1964). The results shown in Figure 1.3A seemed promising. The four Realism paintings on the right side are clearly separated from the rest. One assumption might be that this separation is due to a difference in composition. In the

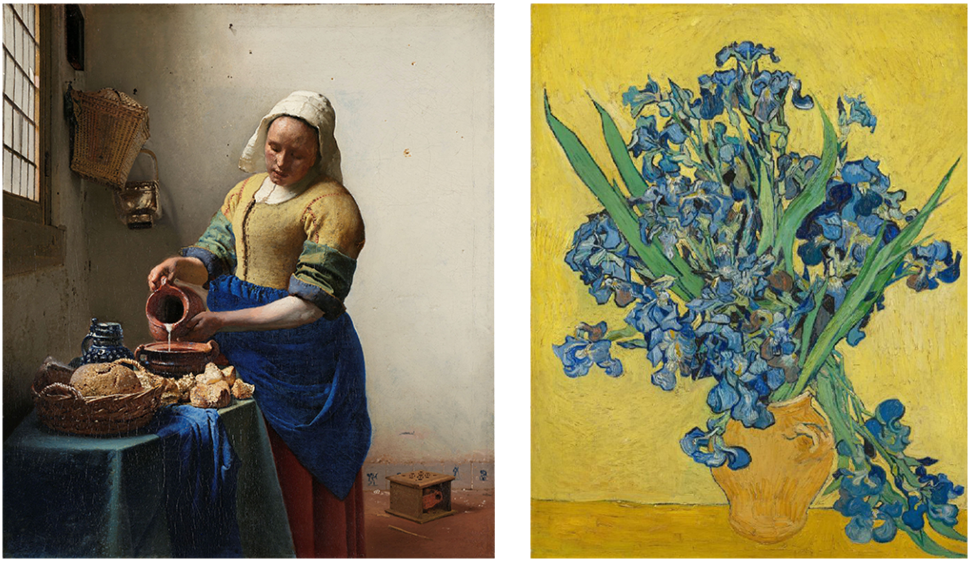


Figure 1.2: Subject matter can be a confounding factor for judging depiction style. Left: Johannes Vermeer, *The Milkmaid*, 1660. Downloaded from the online repository of the Rijksmuseum, Amsterdam. Right: Vincent van Gogh, *Irises*, 1890. Downloaded from the online repository of Van Gogh Museum, Amsterdam.

four Realism paintings, the background occupied a larger area than the rest. There was no clear separation between the Baroque cluster and the Impressionism cluster, possibly because: 1) they share similar composition; 2) they can both be described as colorful while the Realism paintings present less diversity in color and less saturated colors.

As mentioned above, we would like to study the *how* rather than the *what* of style, so the next step was trying to remove the influence of composition. To achieve that, we conducted the same experiment, but instead of using the whole paintings as stimuli, we used cut-outs of flowers from the same selection of paintings. Again we reached a 2D solution, as shown in Figure 1.3B. However, the structure of that space was obviously different from that in Figure 1.3A, indicating different criteria might have been used by the participants. More specifically, the Impressionism cluster separated from the Baroque one, and one of the Realism paintings moved towards the Baroque cluster. We assumed the brushstroke coarseness might be one of the judging criteria (which was later confirmed in Chapter 2).

At the same time, we explored another methodological direction. Two types of tasks are often used to collect (dis)similarity judgement data for MDS analysis. Both have their advantages and disadvantages. Pairwise rating, as we used in the first two experiments, presents two stimuli in each trial and asks participants to rate their (dis)similarity on a scale, from not similar to very similar. The advantage of pairwise rating is it requires a relatively small number of trials to collect the full (dis)similarity matrix for MDS analysis: $n(n-1)/2$ trials where n is the total number of stimuli. However, it has the disadvantage of being sensitive to individual differences: for example, scale range can vary considerably between observers and may also depend on preceding trials (Linde, 1975; O'Hare,

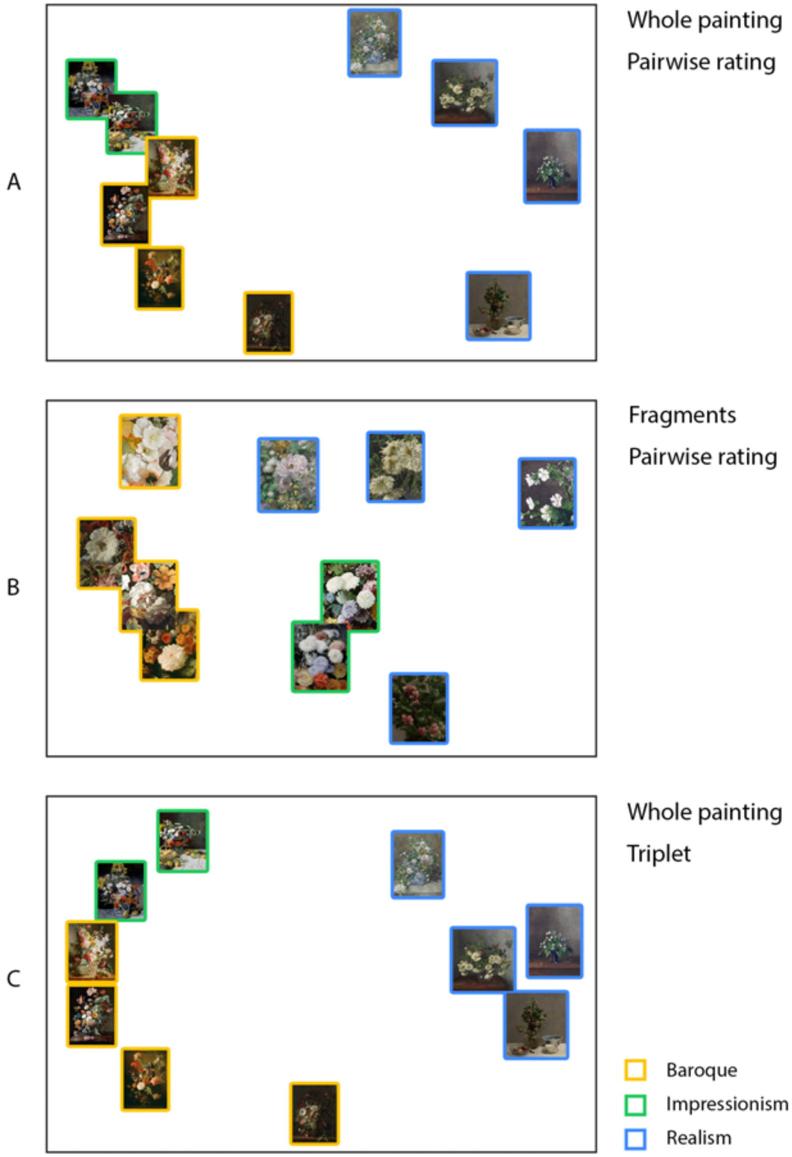


Figure 1.3: A) 2D MDS solution of a pairwise rating task using ten flower still-life oil paintings from three art periods: Baroque, Impressionism, Realism. B) 2D MDS solution of a pairwise rating task using ten fragments, one per painting as shown in Figure 1.3A. C) Same as Figure 1.3A, but now based on the results of a triplet judgement task.

1976). The alternative choice, a triplet comparison task, presents three stimuli in each trial and asks participants to choose either the most similar pair out of three possible pairs or select one stimulus as the odd one out. The advantage is it is more robust and makes it easier to combine data across participants, since the choice is only 'which one' instead of 'how much' (Heikinheimo and Ukkonen, 2013; Li et al., 2021; Tamuz et al., 2011). In principle, this allows us to scale up experiments by having a larger painting selection and more participants at our disposal. However, the triplet task has one major disadvantage: the number of trials increases cubically, $n(n-1)(n-2)/6$ where n is the number of stimuli. We came up with multiple solutions to reduced the number of required trials, which will be discussed later in section 1.3.3. But first, we wanted to compare the performance between triplet and pairwise tasks. To this end, we conducted Pilot Experiment 3. It is a replication of Pilot Experiment 1, except that we replaced the pairwise rating task with a triplet comparison task. The similarity between two stimuli was calculated based on the frequency they were selected as the most similar pair. Again it led to a 2D MDS solution, as shown in Figure 1.3C. Note that the triplet result has almost identical structure as the pairwise rating result (Figure 1.3A).

ZOOM IN ON A TIMELESS OBJECT

From the three pilot experiments with flower still life paintings, we learned two major takeaways: 1) subject matter plays an important role in style judgement; we can control this by zooming in on fragments of paintings, preferably on a single object; 2) the triplet task works well; it provides robust and repeatable results while allowing us to scale up the experiment. With this knowledge, we conducted the first study on style perception. We tried to investigate the existence and description of style by isolating style from medium and subject matter. We focused on a single medium, oil painting, and a single object: an apple. In this way, we hope to exclude the possible influence of subject matter (*what*), period cues such as items tightly connected to a certain era (*when*), and occasionally signatures of the artists (*who*), thereby mainly focusing on the *how*.

Oil painting was chosen because of its long history and wide usage while the apple was selected since it is a rather *'timeless'* object. For example, an apple appeared in Adam's hand in medieval paintings, in still life paintings from Dutch 17th century, but also in modern paintings such as *The Listening Room* by René Magritte from the 20th century. We collected cut-outs of apples from a wide range of oil paintings covering diverse creation times and regions. By zooming in on cut-outs of depicted apples, we not only isolated the same object, but also removed background information such as clothing, interiors or exteriors that could indicate the period the painting was created. We first reached a perceptual style space by collecting human similarity judgements, then described the dimensions by fitting attributes rating data. This study will be thoroughly discussed in Chapter 2.

ZOOM OUT TO EXPAND MEDIA COVERAGE

If style can be defined as the way someone does something, then for artists, it also includes the medium they choose. In fact, artists decide the medium of their choice as the first step, before any content is created. In other words, medium is a stylistic choice that can affect the appearance of objects and materials. This can also be seen in the way modern content is created via generative AI. For instance, a prompt can be 'in the style of Caravaggio', or 'in the style of watercolor'. In the first study, we focused on the stylistic differences within the same medium, oil painting. In the second study, we would love to expand the media coverage to investigate the influence of medium.

Throughout history, artists experimented with various media, which evolved alongside technological advancements and cultural shifts, with each medium offering distinct perceptual characteristics. Fresco, prominent in ancient and classical periods, is characterized by its durability and matte finish, though colors tend to be muted due to the plaster base. Tempera, widely used in the Middle Ages, produces sharp lines and bright, opaque colors but lacks the smooth blending found in later media. Engraving, developed in the 14th century, enabled precise, high-contrast images with intricate detail, facilitating the reproduction of artwork. Oil painting, emerging in the 15th century, became dominant due to its vivid colors, smooth transitions, and ability to create depth and texture. Watercolor, popularized in the 16th century, is known for its translucency, lightness, and fluidity, making it ideal for capturing delicate atmospheres. Lithography, introduced in the 19th century, allowed for soft gradients and subtle textures, suitable for mass production of images. Photography, also from the 19th century, revolutionized visual representation by capturing fine details and natural light with unparalleled realism. Digital media in the late 20th century enabled precise manipulation of color, form, and texture, offering endless creative possibilities with virtual tools. Note that some media refer to the technique while others focus on the material. For example, photography is about the technique, regardless if the photo is presented on the digital screen or printed on paper; and oil painting is about the material used.

However, comparisons between different media have rarely been studied, possibly due to factors similar to those affecting the style studies: the subject matter acts as a confounding factor, making it difficult to isolate the effects of the medium. And we want to use real-world art as stimuli instead of producing our own stimulus images (Delanoy et al., 2021). We found a solution by comparing oil paintings and their engraved reproductions. Before the invention of photography, besides being a standalone form of art, engraving was used to reproduce paintings. The identical content gave us a perfect opportunity to compare material perception from two distinct media.

As can be seen from Figure 1.4, engraving is essentially line art consisting of the white of the paper and the black of the ink. It is a very different medium from oil painting where colors are used. In addition, in oil paintings shading can be achieved by a smooth gradient, while engravers can only play with line weight and patterns. From these pairs we were able to isolate areas in different materials (as illustrated by the red outlined areas in Figure 1.4), such as fabric or skin, and compare perception of these selections. This study will be thoroughly discussed in Chapter 3.

ZOOM OUT TO EXPAND CREATOR COVERAGE

While in the first two studies we dived into different appearances created by human artists, for the last study, we zoomed out to include non-human creators, that is, generative AI. Just like the artists in the last two chapters, the generative AI models in this last study work within 2D planes. Interestingly, the emerging generative AI models can be seen as a different creator than humans and/or as a new medium (technique) that

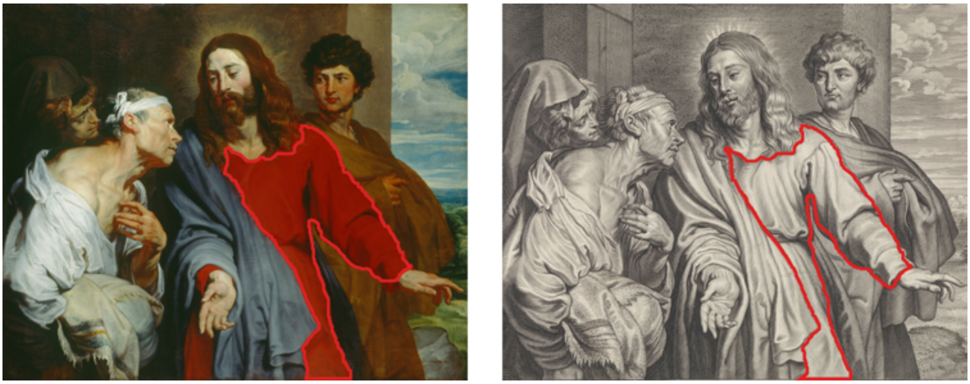


Figure 1.4: Engraving was used to reproduce oil paintings, so that the pair shares identical subject matter. The red outlines are examples of highlighted area for participants to judge. Left: Anthony van Dyck, *Christ healing the paralytic*, 1619. Downloaded from commons.wikimedia.org. Right: Pieter de Jode II, *Christ healing the paralytic*, 1641-1670. Downloaded from the online repository of National Galleries of Scotland, Edinburgh. The engraving is mirrored horizontally to match the original oil painting. Both are slightly cropped to achieve good alignment.

blends and remixes previous human creation and novelty. In Chapter 4 we asked three text-to-image AI models (DALL-E 2, Midjourney v2, Stable Diffusion v1.5) to generate images of various materials. With similarity judgements from human participants, we could construct a perceptual space for each AI model from these image sets. We compared the perceptual spaces from these AI models with each other as well as with that from a bidirectional reflectance distribution function (BRDF) material dataset, MERL (Matusik, 2003). The text prompts for the AI models were labels attached to this last material data set.

This third study consists of two experiments. In Experiment 1 we tested DALL-E 2 and Midjourney v2 in 2022. Mentioning the year is important as generative AI models developed rapidly around this time. The only constraint for the AI models is text descriptions (prompts). In Experiment 2 we tested Stable Diffusion v1.5 in combination with an add-on, ControlNet, which allowed us to have one additional graphical constraint, a depth map. With this graphical constraint, we were able to generate materials in more complex geometries. The prompts for all AI models are the text descriptions of materials (e.g., blue acrylic) from MERL dataset. These experiments will be thoroughly discussed in Chapter 4.

1.3. METHODOLOGY

1.3.1. CONTROL OVER SUBJECT MATTER

In all three studies, we isolated specific factors such as style or medium by keeping the remaining variables as constant as possible. Among these variables, keeping subject matter as constant as possible proved especially challenging when working with existing artworks. But we still found the solutions by 1) comparing the originals and their replicas (Chapter 3) and 2) keeping the objects approximately consistent (Chapter 2 and

4).

In Chapter 2, we isolated style from subject matter by cropping to a ‘timeless’ object, apple, while having a wide coverage of real paintings (instead of generated stimuli). Chapter 3 faced the similar challenge as Chapter 2. But this time, we took the control over subject matter one step further, the two versions share the same content, since the engraving version was made as the reproduction of the painting. The identical content allowed participants to focus on medium without subject matter as a confounding factor. In Chapter 4, we controlled the geometry of material blobs by either semantic constraint (i.e., a sphere) or graphical constraint (i.e., the same depth map within the same stimuli set). Figure 1.5 shows the demonstration of subject matter control.

1.3.2. Isolating variable with image manipulation

The cropping we used in the style study can be seen as image manipulation. Besides cropping, we also used other image manipulations to isolate a single variable of our interest. In Chapter 3, to investigate the influence of color and contrast, we created different versions of stimuli. By comparing the original colored version with the created grayscale version, we were able to evaluate the influence of color on material perception. Similarly, creating the histogram matched versions allowed us to delve into the influence of contrast.

1.3.3. Trial reduction

Both Chapter 2 and 4 used a perceptual scaling method to explore unknown perceptual spaces. As we mentioned earlier, we favor triplet tasks over pairwise rating in order to scale up the experiment. However, the triplet method also has the disadvantage of a large number of required trials. We applied multiple solutions to reduce the number of required trials.

Landmark MDS (LMDS) We used 48 stimuli for Chapter 2, which would have required 17,296 unique trials without trial reduction. We used LMDS to reduce trials. LMDS was originally designed to reduce computational demand when dealing with large data sets (Silva and Tenenbaum, 2002). It uses only a portion of the data to reach the MDS solution without compromising too much on accuracy. LMDS first selects a subset of stimuli as landmarks, randomly or manually. It requires the full distance matrix for the ‘landmarks’, runs classical MDS on the landmarks and reaches an MDS space with only landmarks. Then for the remaining stimuli, the non-landmarks, LMDS uses the distance between a non-landmark and all landmarks to position it in the established space, without needing the distance between non-landmarks. In our case, however, we used LMDS in a different way than its original design. Instead of collecting the full distance matrix data and then use a portion of them, we decided which portion of the data to collect beforehand. In this way, we reduced the number of required trials from 17,296 to 4,400. The 75% trial reduction is achieved by selecting 16 stimuli as landmarks, which represent various periods and origins. LMDS only requires the full distance matrix to run a classical MDS analysis on landmarks ($16 \times 15 \times 14/6 = 560$ trials). The dimensionality and the space is first defined by the landmarks. Then each remaining 32 non-landmarks is then fitted to the space with its relation to all 16 landmarks ($32 \times (16 \times 15/2) = 3840$ trials). The significant

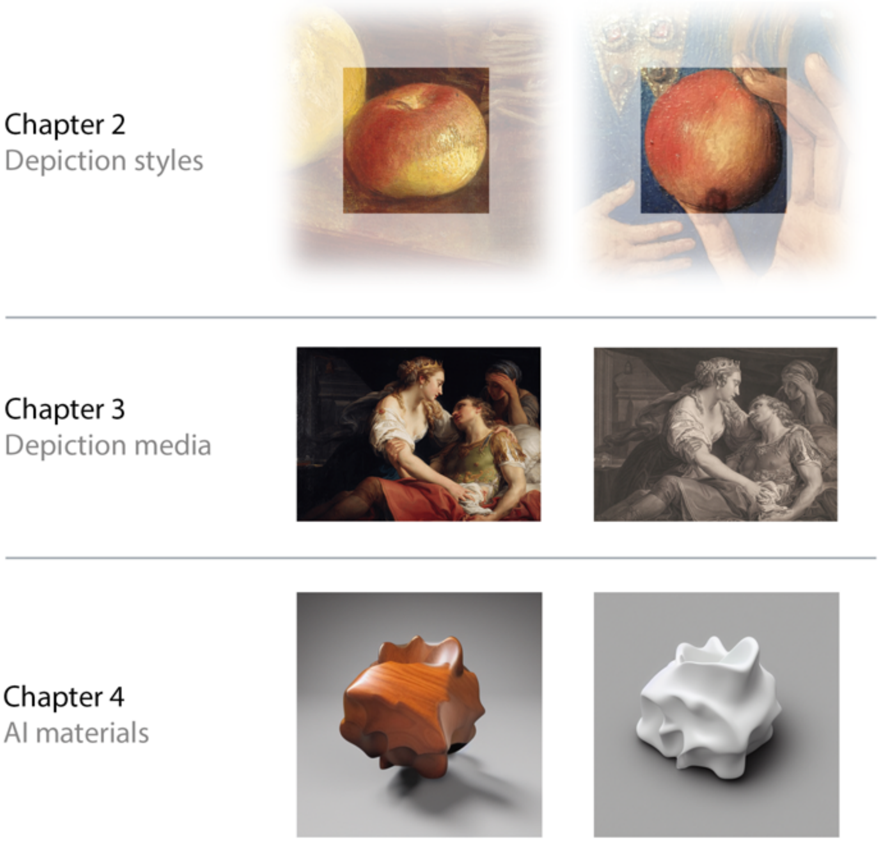


Figure 1.5: Subject matter control. Chapter 2 left: Henri Fantin-Latour, *Still Life*, 1866. Downloaded from the online repository of National Gallery of Art, Washington, D.C.. Chapter 2 right: Hans Memling, *Diptych of Maarten Nieuwenhove*, 1487. Downloaded from commons.wikimedia.org. Chapter 3 left: Pompeo Batoni, *La mort de Marc Antoine*, 1763. Downloaded from Wikipedia. Chapter 3 right: Johann Georg Wille, *La Mort de Marc Antoine*, 1778. Downloaded from the online repository of the Rijksmuseum, Amsterdam. Chapter 4: both generated by Yuguang Zhao with Stable Diffusion v1.5 and ControlNet.

reduction of trials allowed us to scale up the experiment while keeping it feasible.

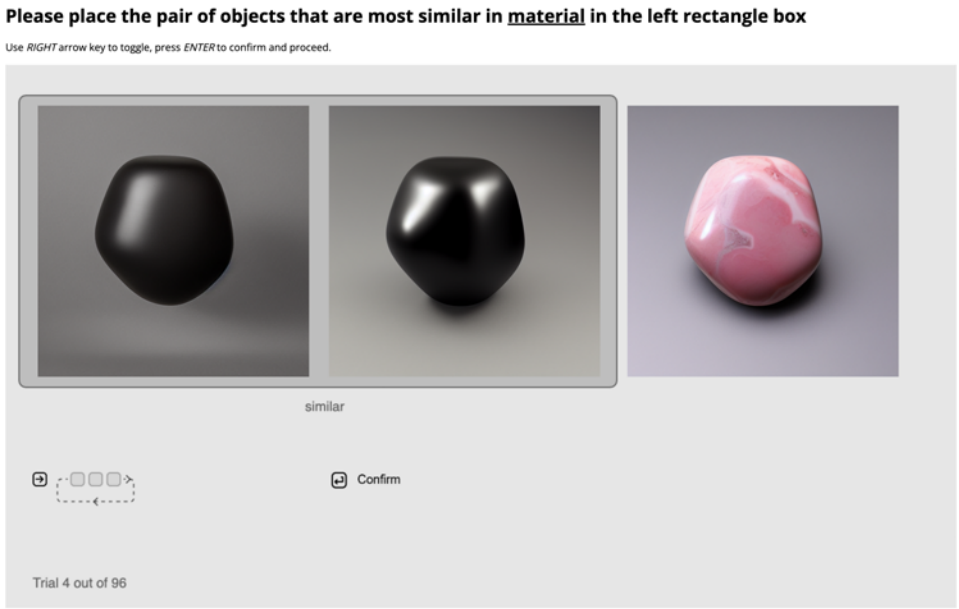
Soft Ordinal Embedding By the time we started the study of Chapter 4 where we were again interested in the unknown perceptual spaces, a new method for trial reduction was just released (Haghiri et al., 2020; Künstle et al., 2022). We were one of the early adopters for the new method, Soft Ordinal Embedding (SOE) (Künstle and von Luxburg, 2024).

Originated from machine learning, SOE takes triplet data as input and finds the perceptual space that maximizes the number of consistent triplets. The minimal numbers of required trials is only $dn\log_2 n$, where d is the estimated number of dimension(s) and n is the number of stimuli. Compared to LMDS, SOE reduced the number of required trials even further without compromising on solution accuracy. Further more, since the space is no longer determined by 'landmarks', the triplets are just a random subset of all possible triplets, which is easy to set up.

1.3.4. ENGAGING INTERFACE

In all three studies we collected human data to understand visual perception via crowdsourcing platforms (Amazon MTurk for Chapter 2 and Prolific for Chapter 3 and 4). We chose the crowdsourcing option over lab experiment so that we could scale up the studies. Besides, it was the only option during the pandemic, which began a few months after this PhD project commenced. Gathering high quality data via crowdsourcing platforms can be a challenge (Cuskley and Sulik, 2022; Keith et al., 2017; Rodd, 2024). Ensuring high data quality is crucial for the reliability and validity of our findings. To achieve this, we should first understand our participants. Collecting data online is essentially a collaboration between researchers and participants, researchers should not expect high quality data to be granted (Cuskley and Sulik, 2022). The motivations for participants to spend time on online experiments usually fall into one or more of the following categories: financial reward, altruism, knowledge seeking and entertainment (Rodd, 2024; Tinati et al., 2017). Thus, the payment is unlikely to be the only reason for participation (Göritz, 2014). Within these motivations, only financial reward and entertainment are within our control. Besides reasonable compensation, we tried to provide good user experience (UX) for our participants. The interaction designs for our experiments might not fall within the category of gamification, but at least they make the process more playful and help participants to understand the task better. Both gamification and good UX are believed to improve participants' engagement, hence improve the data quality (Carvalho et al., 2019; Rodd, 2024; Tinati et al., 2017). In addition, prevention methods such as an attention check can be seen as passive methods that detect inappropriate behaviours, and might cause long-term harm on the relation between researchers and participants (Rodd, 2024). Instead, we took a persuasive path to encourage appropriate behaviours.

As someone with experience and passion about UX and interface design, I used my programming skills to achieve the intuitive and engaging interaction designs. In both Chapter 3 and 4, we chose the triplet task as mentioned above. Figure 1.6 shows the online experiment interface of Chapter 4. The task is to select the most similar pair out of the triplet. The cyclic movement, indicated by the icon at the lower left, allows all three possible pairs to be displayed within the left rectangle for easy visual comparison. This keyboard-only operation is user-friendly, requiring only two adjacent keys. As shown by the icons below the images, the RIGHT arrow key toggles the order, while the RETURN key confirms the current selection and proceeds to the next trial.



Please place the pair of objects that are most similar in material in the left rectangle box

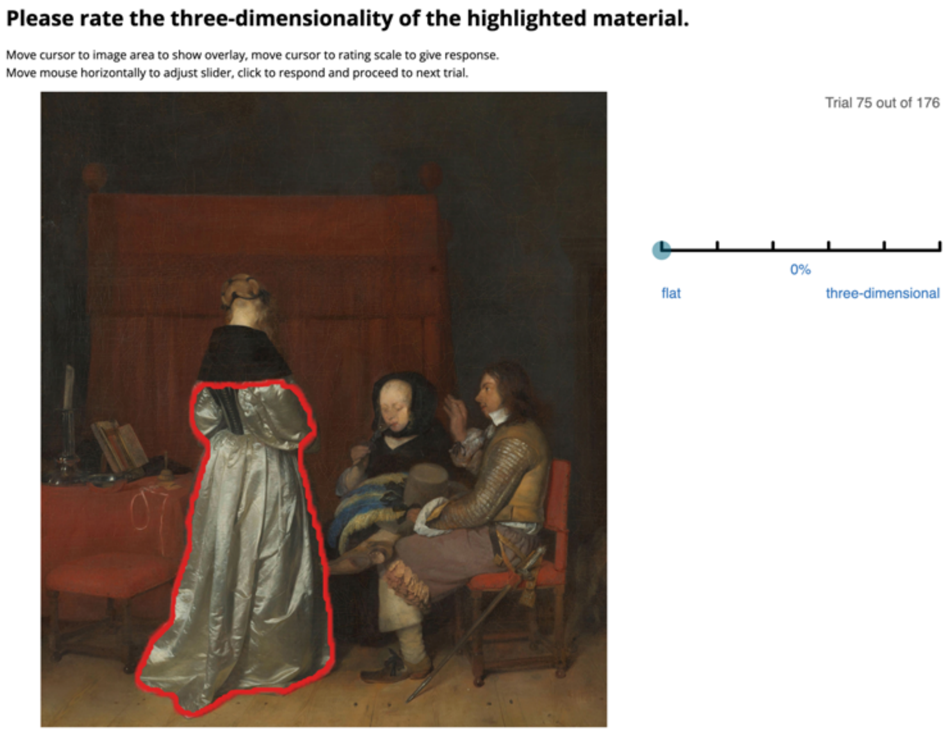
Use *RIGHT* arrow key to toggle, press *ENTER* to confirm and proceed.

Figure 1.6: Interface design for Chapter 4. As the first icon indicates, Pressing the RIGHT arrow key would toggle the order of the three images. All three possible pairs can be selected as similar in terms of material. They can press RETURN to both confirm the choice and proceed to the next trial.

In Chapter 3 where we compared material perception between two media, oil painting and engraving, I chose to have a mouse only interaction for easy rating. Figure 1.7 shows the interface design. When a participant moves the cursor to the left image area, a red outline appears to indicate the target material area for judgement. When the cursor is moved to the right rating area, its horizontal position controls the rating scale, regardless of the vertical position. Clicking confirms the rating and proceeds to the next trial. The red outline disappears when the cursor is on the right side to avoid influencing material perception with its vibrant color. For smooth and fast operation, participants can keep the cursor on the right side; the red outline will flash twice at the beginning of each trial, regardless of cursor position.

Additionally, we maintained a reasonable number of trials for each participant, ensuring the experiment duration stayed under ten minutes. This relatively short time frame helps prevent data quality degradation due to participant fatigue. The trial count at the lower left corner is always visible as an progress overview. And the instruction is always visible at the top to remind participants the task even if they did not read the instruction page carefully. To gather qualitative feedback, we included a text box at the end of each experimental session. And participants indeed provided some positive comments:

1. From study 1: “It was interesting to do. I enjoyed it.”



Please rate the three-dimensionality of the highlighted material.

Move cursor to image area to show overlay, move cursor to rating scale to give response.
Move mouse horizontally to adjust slider, click to respond and proceed to next trial.

Figure 1.7: Interface design for Chapter 3. Each time a new stimulus was shown, the red outline flashed twice to denote the area of interest. As a reminder, participants could move the cursor to the image area to show the red outline overlay. On the right side, participants moved the cursor along the rating scale to adjust the rating, and click to confirm and proceed to the next trial. Gerard ter Borch (II), *Gallant Conversation (Known as 'The Paternal Admonition')*, 1654. Downloaded from the online repository of the Rijksmuseum, Amsterdam.

2. From study 2: "The study was extremely well made and well thought. outlines were amazingly done. All in all, a really interactive and interesting study."

3. From study 3: "It was an interesting experiment and I really enjoyed doing it."

The interface designs were achieved with HTML, CSS and JavaScript. An in-depth article with animation demonstrations can be found at https://yuguang-zhao.com/design/online-experiment-UX.

Bibliography

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., & Belongie, S. (2007). Generalized non-metric multidimensional scaling. *Artificial Intelligence and Statistics*, 11–18.

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current biology*, *21*(24), R978–R983.

Berlyne, D. E., & Ogilvie, J. C. (1974). Dimensions of perception of paintings. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 181–22). Hemisphere.

Carvalho, J., Santos, A., & Paredes, H. (2019). Data quality improvement in crowdsourcing systems by enabling a positive personal user experience. *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 255–260.

Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, *434*(7031), 301–307.

Cuskley, C., & Sulik, J. (2022). The burden for high-quality online data collection lies with researchers, not recruitment platforms. *Perspectives on Psychological Science*, 17456916241242734.

Delanoy, J., Serrano, A., Masia, B., & Gutierrez, D. (2021). Perception of material appearance: A comparison between painted and rendered images. *Journal of Vision*, *21*(5), 16–16.

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.

Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., & Mazzone, M. (2018). The shape of art history in the eyes of the machine. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 2183–2191.

Ferwerda, J. A., Pellacini, F., & Greenberg, D. P. (2001). Psychophysically based model of surface gloss perception. *Human vision and electronic imaging vi*, *4299*, 291–301.

Fleming, R. W. (2017). Material perception. *Annual review of vision science*, *3*(1), 365–388.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Göritz, A. S. (2014). Determinants of the starting rate and the completion rate in online panel studies1. *Online Panel Research: Data Quality Perspective, A*, 154–170.

Graham, D. J., Friedenberg, J. D., Rockmore, D. N., & Field, D. J. (2010). Mapping the similarity space of paintings: Image statistics and visual perception. *Visual cognition*, *18*(4), 559–573.

Haghiri, S., Wichmann, F. A., & von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *Journal of vision*, *20*(9), 14–14.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Heikinheimo, H., & Ukkonen, A. (2013). The crowd-median algorithm. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *1*, 69–77.

Jacobson, J., & Werner, S. (2004). Why cast shadows are expendable: Insensitivity of human observers and the inherent ambiguity of cast shadows in pictorial art. *Perception*, *33*(11), 1369–1383.

Keith, M. G., Tay, L., & Harms, P. D. (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in psychology*, *8*, 1359.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Künstle, D.-E., & von Luxburg, U. (2024). Cblearn: Comparison-based machine learning in python. *Journal of Open Source Software*, *9*(98), 6139.

Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5–5.

Li, J., Endo, L. R., & Kashima, H. (2021). Label aggregation for crowdsourced triplet similarity comparisons. *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28*, 176–185.

Linde, L. (1975). Similarity of poetic rhythms with different amounts of semantic content-stress ratings and pairwise similarity ratings. *Scandinavian Journal of Psychology*, *16*(1), 240–246.

Mamassian, P. (2004). Impossible shadows and the shadow correspondence problem. *Perception*, *33*(11), 1279–1290.

Matusik, W. (2003). *A data-driven reflectance model* [Doctoral dissertation, Massachusetts Institute of Technology].

Mead, A. (1992). Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *41*(1), 27–39.

O'Hare, D. (1976). Individual differences in perceived similarity and preference for visual art: A multidimensional scaling analysis. *Perception & Psychophysics*, *20*(6), 445–452.

Ostrovsky, Y., Cavanagh, P., & Sinha, P. (2005). Perceiving illumination inconsistencies in scenes. *Perception*, *34*(11), 1301–1314.

Rao, A., Srihari, R. K., & Zhang, Z. (1999). Spatial color histograms for content-based image retrieval. *Proceedings 11th International Conference on Tools with Artificial Intelligence*, 183–186.

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, *134*, 104472.

Ruth, J.-E., & Kolehmainen, K. (1974). Classification of art into style periods; a factor-analytical approach. *Scandinavian Journal of Psychology*, *15*(1), 322–327.

Sablatnig, R., Kammerer, P., & Zolda, E. (1998). Hierarchical classification of paintings using face-and brush stroke models. *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, *1*, 172–174.

Sharan, L., Rosenholtz, R., & Adelson, E. (2009). Material perception: What can you see in a brief glance? *Journal of Vision*, *9*(8), 784–784.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision*, *14*(9), 12–12.

Silva, V., & Tenenbaum, J. (2002). Global versus local methods in nonlinear dimensionality reduction. *Advances in neural information processing systems*, *15*.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Tinati, R., Luczak-Roesch, M., Simperl, E., & Hall, W. (2017). An investigation of player motivations in eyewire, a gamified citizen science project. *Computers in Human Behavior*, *73*, 527–540.

Toscani, M., Guarnera, D., Guarnera, G. C., Hardeberg, J. Y., & Gegenfurtner, K. R. (2020). Three perceptual dimensions for specular and diffuse reflection. *ACM Transactions on Applied Perception (TAP)*, *17*(2), 1–26.

Wijntjes, M. W., & de Ridder, H. (2014). Shading and shadowing on canaletto's piazza san marco. *Human Vision and Electronic Imaging XIX*, *9014*, 308–313.

Wölfflin, H. (2012). *Principles of art history*. Courier Corporation.

2

Zooming in on style: exploring style perception using details of paintings

Most studies on the perception of style have used whole scenes/entire paintings; in our study we isolated a single motif (an apple) to reduce or even eliminate the influence of composition, iconography, and other contextual information. In this paper, we empirically address two fundamental questions of the existence (Experiment 1) and description (Experiment 2) of style. We chose 48 cut-outs of mostly Western European paintings (15th to 21st century) that showed apples. In Experiment 1, 415 unique participants completed online triplet similarity tasks. Multidimensional scaling (MDS) reached a non-random 3D embedding, showing that participants are able to judge stylistic differences in a systematic way. We also found a strong correlation between creation year and embedding, both a linear correlation with Dimension 2, and a rotational correlation in the first two dimensions. To interpret the embedding further, in Experiment 2 we fitted three color statistics and nine attribute ratings (glossiness, three-dimensionality, convincingness, brush coarseness, etc.) to the 3D perceptual style space. Results showed that Dimension 1 is associated with spatial attributes (Smoothness, Brushstroke coarseness) and Convincingness, Dimension 2 is related to Hue, and Dimension 3 is related to Chroma. The results suggest that texture and color are two important variables for style perception. By isolating the motifs, we could exclude higher levels of information such as composition and context. Interestingly, the results reinforce previous findings using whole scenes, suggesting that style can already be perceived in -sometimes very small- fragments of paintings.

Published as Zhao, Y., Stumpel, J., de Ridder, H., & Wijntjes, M. W. (2023). Zooming in on style: Exploring style perception using details of paintings. Journal of vision, 23(6), 2-2. doi:https://doi.org/10.1167/jov.23.6.2

2.1. INTRODUCTION

In his book Principles of Art History, Heinrich Wölfflin referred to an anecdote in which four German painters from the Romantic period tried to paint a particular scenery all 'firmly resolved not to deviate from nature by a hair's breadth' (Wölfflin, 2012). The resulting landscapes, however, differed considerably in style. Wöllflin ascribed this fact to differences in personality and vision of the artists. He also remarked that in spite of the differences we would easily see the similarities between them and recognize them as products of a particular period: the first half of the nineteenth century. For Wölfflin, such collective differences between the pictorial production of different periods were ultimately rooted in differences in artistic vision or perception. To capture the differences between sixteenth and seventeenth century painters, Wölfflin came up with five visual principles: (1) linear vs. painterly, (2) closed vs. open form, (3) planar vs. recessional, (4) multiplicity vs. unity and (5) absolute vs. relative clarity. Despite their widespread use both within and beyond art history, such as in perception research (Goude and Derefeldt, 1981; O'Hare, 1979) and computer vision (Cetinic et al., 2020; Elgammal et al., 2018), it can be seen that these principles have their limitations, and are specifically conceived to model the contrast between Renaissance and Baroque art.

To understand the matter of style we need a broader definition that is both testable and can generate novel insights. Gombrich (2009) seems to offer this broader definition:

> "Style is a distinctive, and therefore recognizable, way in which an act is performed or an artifact made."

This is clearly a general description but at the same time specifically emphasizes the role of the beholder ('recognizable'). If there are no differences to be perceived, there is no style. This fundamental aspect of style (its existence) precedes descriptions or models of style such as those of Wölfflin. In this paper, we empirically address these two fundamental questions of the existence (Experiment 1) and description (Experiment 2) of style in the context of visual perception.

STYLE MEASUREMENTS

To empirically investigate the perception of style, one ideally refrains from any explicit terminology. A disadvantage of a Wölfflinian approach is the top-down usage of terms describing style differences, instead of a bottom-up approach that does not make use of such terms. The invention of multidimensional scaling (MDS) methods (see Mead (1992) for a review) offered such an opportunity: instead of relying on explicit adjectives, attributes or descriptions the MDS approach only relies on perceived differences (or 'distances'), from which a space is constructed. This space is a low dimensional representation of the theoretical high dimensional space where each element would have its own dimension. This representation can be concisely referred to as 'embedding' and sometimes, when appropriate, as 'perceptual space'. Indeed, after substantial methodological progress was made in the 1960's (e.g. Kruskal, 1964b; Shepard, 1962), this approach became popular in style perception studies. Berlyne and Ogilvie (1974), for example, conducted a series of similarity judgements and attribute rating experiments on 52 paintings covering 14th to mid-20th (western) art. Observers were instructed "how similar or different the two pictures of each pair were" using a 7-point scale. The authors concluded that a three-dimensional (3D) space would be the most reasonable solution to explain their data, the first dimension being aligned with creation year. Interestingly, the authors had difficulty explaining the second and third dimension and only very tentatively suggested an influence of line and surface quality. Importantly, they found reasonable inter-rater reliability, meaning that observers agreed quite well on perceived style. Referring back to Gombrich's definition, the study of Berlyne and Ogilvie (1974) showed enough 'distinctiveness' between the painting styles as demonstrated by inter-rater reliability and a style space of three dimensions with reasonable stress value (< 0.2).

In addition to accessing perceptual spaces and interpreting them by means of explicit attribute ratings, similarity judgements data have also been used for classification schemes (Graham et al., 2012). Here, the similarity data can be utilized for identifying latent stylistic dimensions in an unsupervised model, or for training classification models in a supervised manner (Hughes et al., 2011).

Other studies focus more on feature statistics, such as color histogram statistics (Rao et al., 1999) or pixel information at the level of the brushstroke. Sablatnig et al. (1998), for example, used a combination of face recognition and brushstroke analysis to classify paintings into different categories. However, it can often be unclear whether the algorithms are measuring what is represented (i.e., depicted scenes) or the medium (i.e., paint on the canvas). We will come back to this issue in the General discussion.

Style descriptions

Various attempts have been made to quantify which visual features describe style. For example, Berlyne and Ogilvie (1974) asked observers in further experiments to rate the paintings on various affective, descriptive, artistic and stylistic scales. Especially interesting were the four scales of texture, lines, colors, and shapes. These scales are somewhat related to Wölfflins' principles. They were mostly significantly describing the style space. Marković and Radonjić (2008) investigated the role of implicit and explicit features in style perception. In their terminology, implicit refers mostly to subjective impressions such as aesthetic and affective judgements while explicit refers to more 'objective' features such as form, color and space. Interestingly, they found that the MDS configurations of 24 paintings could mostly be explained by explicit features. The general approach of using attribute ratings to explain stylistic differences in paintings was also used in other studies. O'Hare (1976) used a mixture of implicit (e.g. like-dislike, interesting-uninteresting, peacefulness-disturbed) and explicit (e.g. dark-bright, soft-sharp, few-many colors) features. He found significant correlations between the first MDS dimension and 'realism' and between the second MDS dimension and 'clarity' and 'symmetry'. These findings were rather robust as a follow-up study confirmed (O'Hare, 1979).

While attribute ratings have been used to explain style embeddings, they have also been used to predict style categories: Ruth and Kolehmainen (1974) performed a factor analysis on attributes in relation to existing style labels. This approach thus assumes a fixed style structure which is different from the bottom-up approach of creating style embeddings like those using MDS. An interesting different approach to looking for style features is manipulating hypothesized features of style: Gardner (1974) altered texture and color by various image manipulations. Masking impaired style recognition, making it difficult to match artworks from the same artist.

From another perspective, Wallraven et al. (2009) proposed that humans use three levels of information for style perception: high-level background information: knowledge about specific historical events, knowledge about artists and art periods in general; mid-level content information: specific objects or scenes that are depicted, type of painting or subject (landscape painting, portrait, etc.); and low-level pictorial information: technique, thickness of brush strokes, type of painting material (oil, acrylic, etc.), color composition of the scene. They conducted three experiments to perform categorizing tasks. The results showed humans definitely need high-level information (old vs. new, perspective flat vs. open, etc.) to make style categorization judgements, although mid-level information (content, realistic vs. abstract, etc.) and low-level information (brush stroke, colors, etc.) were also used by some participants. Siefkes and Arielli (2018) also suggested that high-level information is important for style perception. It was argued that humans need knowledge about culture, history or art categories to be able to perceive stylistic differences.

Computational studies

Besides behavioural research where the emphasis is on the human ability to perceive stylistic similarities of artworks, other studies have taken a computational approach. Graham et al. (2010), for example, related feature statistics to the axes of the MDS spaces reached from similarity judgements. Evidently, there has been major breakthroughs in so-called style transfer that started with Gatys et al. (2016), but this class of algorithms is not used to predict style differences and categories. Elgammal et al. (2018) used 20 style labels to train three deep convolutional neural networks (CNNs) on the WikiArt dataset. These CNNs achieved sub-spaces with fewer than 10 dimensions that explained 95% of the variance using Principle Component Analysis (PCA) (Jolliffe, 2002), with the first two dimensions cumulatively explaining between 60% and 74%. Without having creation years or artists as input training data, the 2D embeddings clearly showed a smooth temporal transition between styles, in a clock-wise U-shape structure. The angular coordinates have a Pearson correlation coefficient of 0.69 with time, again suggesting creation year can be related to the style space. Furthermore, for 1000 paintings they collected art historians' ratings of the Wölfflin principles and found correlations within the first 5 PCA dimensions. These ratings were then used by Cetinic et al. (2020) to do the reverse: train a network estimating the five principles, applying this to the original WikiArt dataset and look for patterns. They found an ascending trend of all five principles between 15th and 17th century, corresponding to style change from Renaissance to Baroque.

Our contributions

The variety of paintings used in previous studies was often rather large. For example, the selection in Berlyne and Ogilvie (1974) contained still-lifes, portraits, biblical scenes and abstract paintings. This makes it clear that style can refer to different levels (Wallraven et al., 2009), but that high-level background and mid-level content information were perhaps too dominant in their study, thus overruling potential low-level information. We hypothesize that the essence of style as defined by Gombrich will emerge more clearly when the subject matter is held constant, as in Wölfflin's anecdote.

In an attempt to limit the influence of subject matter, O'Hare (1976) conducted an experiment with twelve landscape paintings. A two-dimensional (2D) space was found where the realistic-unrealistic scale was connected to the first dimension and the clear-indefinite scale was connected to both first and second dimensions. Besides, we can observe an increase of creation year along the first dimension. Another attempt by Ruth and Kolehmainen (1974) used only paintings with similar content, 'people surrounded by nature in each painting'. Yet in both artwork selections, the variety of subject matter is still rather large: people and landscapes can both vary tremendously in comparison to having artists depict exactly the same scene. Also, other elements in the scene (e.g. means of transport or dress) can provide time-related information, which correspond to mid-level information proposed by Wallraven et al. (2009).

It is impossible to find a selection of paintings of the *exact* same subject matter, but we can isolate painting cut-outs of objects that are repeatedly depicted throughout art history. Ideally, the chosen motif does not undergo stylistic changes itself, which excludes human made objects such as clothing. The ideal motif is therefore something natural. A particular natural motif that is omnipresent throughout art history is an apple. Despite some texture and color differences, apples are relatively similar, especially concerning their shape and size. Isolating apples from their context from a wide variety of paintings and periods allows for an unprecedented control for subject matter and thus offers a unique window on the perception of style.

Secondly, attributes used to explain or create style embeddings often refer to the pictorial plane (e.g. brushstroke) and/or implicit features (e.g. aesthetic preference) usually ignoring features of pictorial representation. This may be due to the variation of subject matter, but it is undeniable that ways of depicting space and material are important aspects of style. Using square cut-outs of single objects will allow us to ask questions about object-specific properties (e.g. smoothness of depicted apple skin), regardless of the composition of the whole painting.

In most of the studies discussed above (e.g. Berlyne and Ogilvie, 1974; Elgammal et al., 2018; O'Hare, 1976), creation year could be identified in the measurements on style differences and even related to the perceived realism of the painted scenes (O'Hare, 1976). But at the same time, it could be concluded that this is confined to paintings of whole scenes only. So, as a third contribution, we looked into the question whether the time a painting was created can also be revealed in observers' style perceptions when both high-level background information and mid-level content information have been removed as much as possible.

Our fourth contribution is methodological. Many studies based on human judgements used pairwise similarity ratings. Both O'Hare (1976) and Linde (1975) have noted that pairwise similarity rating can be sensitive to individual differences: scale range can vary considerably between observers and also depends on the preceding trials. Instead, we used triplet comparison to quantify style similarities. This has various potential advantages, one of them is making it possible to scale up the experiment across various participants. This would also allow for human judgements being used in computational scenarios. The computational style studies reviewed above are all based on existing style labels (e.g. from WikiArt) and not on perceived style differences. Although the number of paintings we investigated in the present study is still relatively small compared to computational studies, a methodological advancement is needed that could use human intelligence to form a lens through which artistic style is quantified, instead of the often-used computational lens.

In the present study, we address the issues outlined above. In the first experiment, we choose 48 apples cut-outs from paintings covering 1487 to 2017, reducing variability in content matter (contributions 1 and 3). We used triplet judgements in combination with the method of Landmark MDS (De Silva and Tenenbaum, 2004) where a subset of cut-outs (the so-called landmarks) is first used to create the initial MDS embedding and then used to fit the remaining non-landmark cut-outs into this space. By doing so, we reduced number of trials dramatically (contribution 4). In the second experiment, we performed multiple linear regression on a number of attributes, including some object related features as opposed to the features about pictorial plane or implicit features (contribution 2).

2.2. Experiment 1 - Similarity triplet ranking

2.2.1. Method

Participants

The online experiments were conducted through Amazon Mechanical Turk (AMT), a crowd-sourcing website for Requesters (researchers in our case) to publish Human Intelligence Tasks (HITs) online and hire crowd-workers (participants) to perform these HITs. 415 unique participants completed Experiment 1 (98.8% were from North America, random sample).

All participants agreed with the informed consent before the actual experiment started, and received compensation via AMT. The experiment was conducted in agreement with the Declaration of Helsinki and approved by the Human Research Ethics Committee of the Delft University of Technology. All data were collected anonymously.

Stimuli

48 digital images of apple painting cut-outs were used as stimuli. All cut-outs were square cut-outs of high-resolution digital images retrieved from ‘Materials in Painting Database’ (Van Zuijlen et al., 2021) or online museum repositories. The far majority of them were oil paintings except for one or two that could have been painted in acrylic. Figure 2.1 shows an example of an original painting (on the right), and the square cut-out of an apple (on the left).

The creation years of the original paintings varied from 1487 to 2019. The selection covered artists from northern European countries (i.e., Netherlands, Germany) to southern European countries (i.e., Spain, Italy), and also paintings from France and North-America.

Most square cut-outs digital images have resolution no less than 400 by 400 pixels, and were set at 400 by 400 pixels in the online experiments. All images were embedded with an sRGB ICC color profile, so that browsers could display colors properly (Ashe, 2014).

Triplet comparison

To create a multidimensional embedding for a large set of images while distributing the judgments among many participants, we opted for a triplet comparison task over the pairwise similarity rating task. Apart from the before-mentioned advantages of this

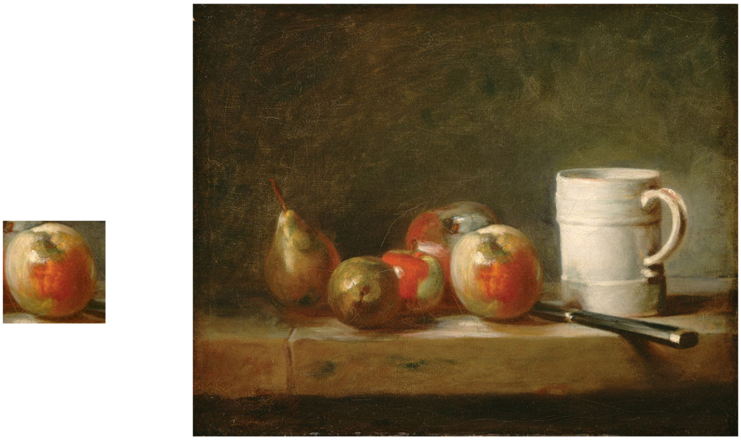


Figure 2.1: An example of a square cut-out of apple from oil painting. Jean Siméon Chardin's *Still Life with a White Mug* (1764), downloaded from the online repository of National Gallery of Art (nga.gov)

method, the disadvantage of using triplets instead of pairs is that the number of trials to create a (dis)similarity matrix increases enormously. For n stimuli, a pairwise method requires $n(n-1)/2$ trials while a triplet method requires $n(n-1)(n-2)/6$ trials. In our case, with 48 stimuli, it would be 1128 unique trials for the pairwise method versus 17,296 unique trials for the triplet method. To reduce the number of triplets to be evaluated, we used the method of Landmark MDS (LMDS).

Landmark MDS (LMDS)

The original purpose of LMDS (De Silva and Tenenbaum, 2004) was to reduce computing power, by using only a portion of the data to reach a final MDS solution without losing accuracy. In the current study, we used the method to reduce the number of trials. With LMDS, the first step is to select a subset of l stimuli as landmarks, either randomly or manually, and collect data and run a classical MDS analysis on those landmarks, i.e. all $l(l-1)(l-2)/6$ triplets are being involved. The next step is to fit the remaining data points (n non-landmarks) into the MDS space of landmarks, using distances between non-landmarks and landmarks. For the first step, the lower half of an $l \times l$ full distance matrix is required to run the MDS analysis. For the second step, only distances between non-landmarks and landmarks are required. Thus, conventional MDS would require an $(l+n) \times (l+n)$ matrix, while for LMDS only an $(l+n) \times l$ matrix is required to reach the final solution.

In the current study, 16 apple paintings were carefully chosen as landmarks, so that they represent various periods, and systematically distributed from north to south Europe (as shown in Figure 2.2, upper part with light orange background). We deliberately included two identical stimuli to verify this method. Paul Cézanne's Apples (1778-1879) was used both as landmark and as non-landmark. There were 16 landmarks (L) and 32 non-landmarks (NL). To generate the MDS space with the Ls, $16 \times 15 \times 14/6 = 560$ triplets were needed. To fit the NLs in this space, each of the 32 NLs had to be paired with all unique pairs of Ls, i.e. $32 \times (16 \times 15/2) = 3840$ triplets. In total, we presented 4400 unique triplets, compared to 17,296 triplets without LMDS, a reduction of about 75%. We split trials into 40 experimental blocks, consisting of 110 trials with an average of 10 repetitions. At least 8 unique participants completed each of 40 sub-groups (11 max, average = 10.38).

Procedure

Before the actual experiment, each participant would first read the consent form and instructions for the experiment. They could only proceed if they gave their consent by clicking 'continue' after reading the consent form. Then they were presented the following instructions:

> *STYLE: is the way things are done. People can have different driving styles, dancing styles etc. We are interested in painting styles. The aim of this experiment is to measure how humans perceive style differences. In paintings, style can show itself in various ways: the use of colors, shadows, lines, brushwork, light, shading, ordering, etc. But we preferably do not specify this exactly. In every trial, you will be shown three images of apples taken from larger paintings. You have to select the two that are most similar in style.*

Then they went through five practice trials, to familiarize our interface and operation. In each trial, three stimuli were presented side by side (as shown in Figure 2.3). Participants were asked to place the most stylistically similar two stimuli in the rectangle box on the left. They could use the Right arrow key on their keyboards to toggle the position of the three cut-outs, until the most similar pair was in the left rectangle box. They could press the Enter key to confirm their choice and go to the next trial.

Data analysis

Before data analysis, we validated the data of individual participants on the basis of two criteria. We measured how much their answer deviated from the initial random setting (criterion at >15% change). Secondly, we used a minimum medium trial time of 1 second, a threshold used in a similar study one of the authors conducted before (Van Zuijlen et al., 2020). About 20% of the participants did not meet the selection criteria, hence their data were removed for analysis, although these participants were reimbursed irrespective of this selection.

The data analysis consisted of two steps: firstly, a non-metric MDS on the landmark stimuli was performed, secondly, we fitted the other data to the LMDS configuration.

MDS on landmarks

The raw output of each participant consisted of 110 triplets. For landmark-only triplets, an output triplet [from left to right (see Figure 2.3): A, B, C] meant the participant indicated the pair of image A and image B was the most similar pair. We created the (dis)similarity matrix using a frequency based method. For a pair A-B, the similarity score was calculated as follows (this is across the results from all participants encountering the pairs A-B): s/t, where s = amount of triplets where A and B were grouped together (the first two elements were AB or BA) and t = amount of triplets containing both A and B. The corresponding dissimilarity score was $1 - s/t$.

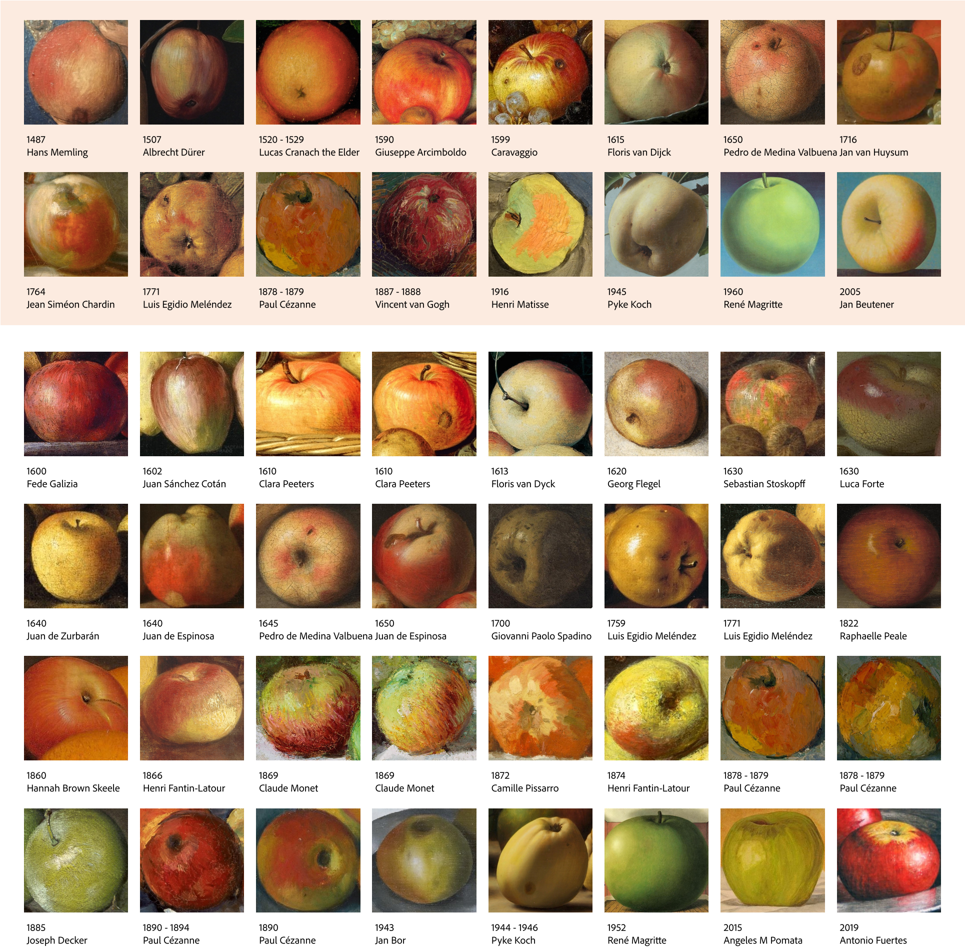
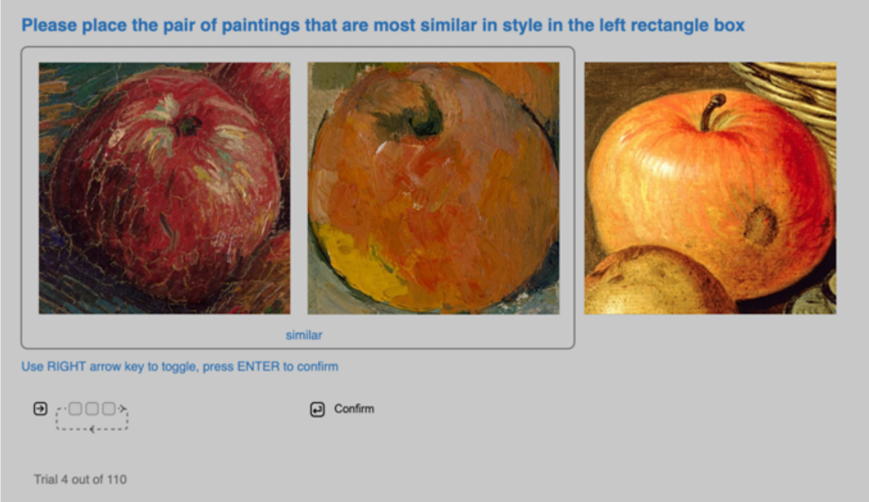


Figure 2.2: All 48 stimuli. The 16 landmarks are located in the upper part with light orange background, sorted by creation year. 32 non-landmarks are located in the lower part with white background, also sorted by creation year.



Please place the pair of paintings that are most similar in style in the left rectangle box

Use RIGHT arrow key to toggle, press ENTER to confirm

Trial 4 out of 110

Figure 2.3: Interface for Experiment 1. In each trial, participants were presented with three square cut-outs of apples. Participants could use the *RIGHT arrow key* to toggle the order of the 3 cut-outs (as the left icon indicated), until the most similar pair in their opinion were in the rectangle box frame on the left. Then they could press *ENTER* to confirm and proceed to the next trial. On the bottom-left, participants could see how many trials they still had to finish.

With a dissimilarity matrix of 16 landmarks, non-metric multidimensional scaling (NMDS) analysis was then performed with metaMDS function from vegan package (v2.5-6) in R (Oksanen et al., 2019). NMDS represents similarity data into a new configuration with the lowest possible dimensions. The best fit is achieved while the distances of landmarks are maintained as closely as possible. Compared to metric MDS, NMDS handles perceptual data better, since it arranges points to maximize rank-order correlation between real-world distance and ordination space distance (Shepard, 1962).

Non-landmarks into LMDS configuration

We fitted non-landmarks into the MDS space from the previous step using a brute-force procedure. The domain of the search extended twice the size spanned by the MDS locations and was split up in 60 evenly spaced sample points in each dimension. At each of these sampling points, fiducial triplet answers were generated on the basis of the MDS data and were compared to the participants' triplets. Simulated triplets were thus compared with real triplets from participants' answers. The cost function simply consisted of counting congruent triplets. To increase robustness, we took the average of the top 0.1 % of this congruency score (227 in the current study).

2.2.2. Results

First, we determine the dimensionality of the landmark space by calculating the stress value as defined by Kruskal (1964a). The stress for one-dimensional to six-dimensional configurations is shown in Figure 2.4A. There is no obvious "elbow" shape, a commonly used criterion to determine dimensionality. Another common criterion is to choose the dimensions where the stress value is below 0.2 (Kruskal, 1964a). The first dimension

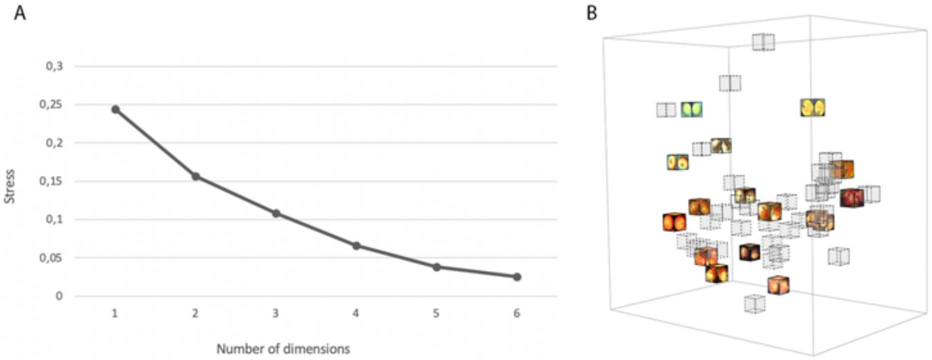


Figure 2.4: A) Scree plot of MDS configuration of landmarks. B) 3D configuration of only landmarks (gray dashed cubes are non-landmarks fitted in the space later).

that has a stress value below 0.2 is two. However, as the landmark set is only a subset of the image set, we decided to continue the analysis with three dimensions, as to not discard any potentially interesting patterns. As will be shown later, subsequent analyzes supported this choice.

Next, the 32 non-landmark samples were fitted to the MDS space (Figure 2.4B, gray cubes). The cost function is the congruency between the actual answers and the answers constructed from the (to be fitted) configuration. We followed this brute force fitting procedure for both the 2D and 3D configuration. The congruency values for both 2D and 3D solutions were well above chance level. There was a small but significant increase in congruency for the 3D embedding: ($t(31) = -5.32$, $p < 0.001$) reflecting an increase in congruency for 27 out of the 32 non-landmark points. This supported the choice for using the 3D embedding for further analysis.

Figure 2.5 shows the overall 3D embedding. As can be seen, the distribution is relatively homogeneous except for Dimension 2 where the distribution seems denser in the lower part. The two same stimuli for verification purpose locate very close to each other in the embedding, confirming the reliability of the landmark method. In addition, it seems that modern apples are on top of the space (along Dimension 2), while older apples are located lower. To further investigate this “historical dimension”, we performed a multiple linear regression for the creation year. With the set of coordinates and creation year for each stimulus as independent and dependent variable, respectively, the orientation of the vector indicates the direction that yielded the best regression, while the length indicates strength of the regression (R^2). The red arrow in figure 2.5 indicates this property vector.

Statistical analysis revealed a significant overall fit ($R^2_{adjusted} = 0.47$, $F(3{,}44) = 15.05$, $p < 0.001$). Dimension 2 received most weight distribution of the fit, and is the only significantly associated dimension with creation year. Figure 2.6A shows the positive relation between creation year and Dimension 2 ($r = 0.69$, $p < 0.001$). To further explore the temporal aspect of the embedding we plotted creation years in the first two dimensions in Figure 2.6B. This plot seems to suggest a rather clustered pattern with potentially a ro-

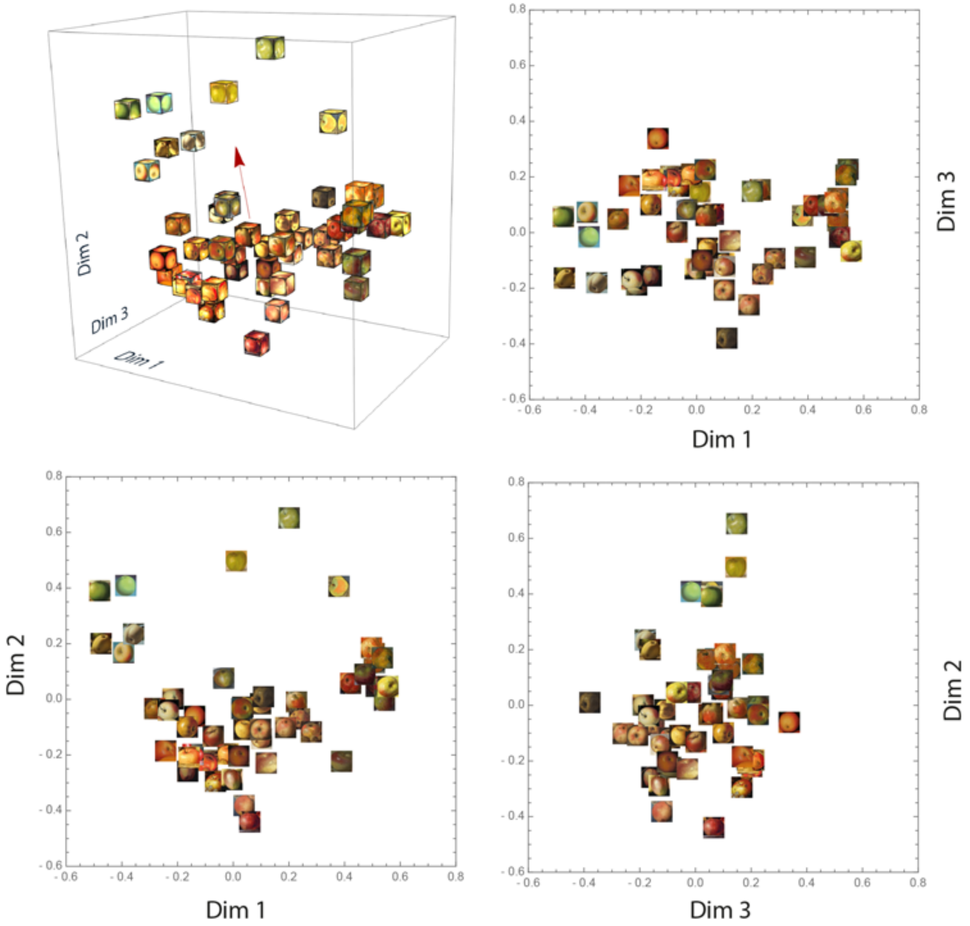


Figure 2.5: 3D space of style perception with 48 apple stimuli. 48 boxes represented 48 stimuli. Each 6 faces of a box show the same apple image, so that it is visible from any viewing angle. The red arrow represents the vector of creation year fitted in the space.

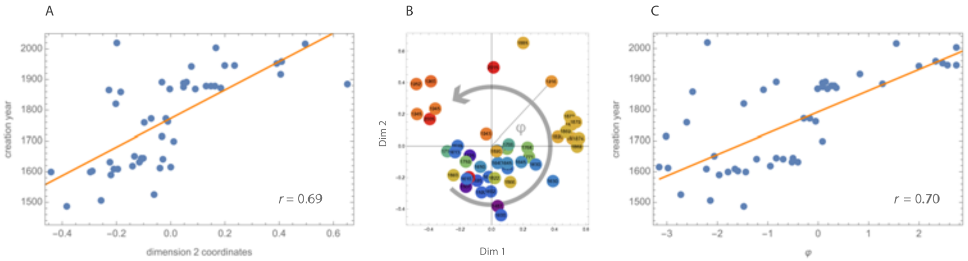


Figure 2.6: A) Correlation between creation year and Dimension 2 in MDS space. B) The first 2 dimensions with color coded creation year and indication of a potential rotational correlation. C) Correlation between creation year and rotation angle phi.

tational correlation. Coincidentally, the data is distributed such that directly calculating the angle between the data points and the positive x-axis (Dim 1) seemed to capture this trend (i.e. large negative angle and old creation year in lower left quadrant, intermediate angles and creation years in right quadrants and large positive angle with new creation year in the upper left quadrant). This was confirmed by calculating the correlation between angle ϕ and creation year ($r = 0.70$, $p < 0.001$).

What can further be tentatively observed is that apples with coarse and visible brushstroke are on the right side (along Dimension 1), while apples with fine and even invisible brushstroke are on the left side. In addition, the greenish apples seem to be at the top of the distribution (along Dimension 2) with the reddish/yellowish ones at the bottom.

2.2.3. Discussion

We found a non-random style embedding of a stimulus set where we held subject matter constant while using the landmark-MDS approach. This suggests that even when high-level background information and mid-level content information have been removed by presenting a single object (apple) only, participants can still consistently perceive style differences. Apparently, there are object properties that make these judgments possible. This will be investigated further in Experiment 2 by assessing object-related attributes like smoothness and glossiness.

The dimensionality of the MDS analysis was based on 16 landmarks. As we mentioned in Section 2.2.2, it showed relatively low stress values for dimensions higher than 2 and there was no obvious elbow shape. So, additional criteria were needed. One of these criteria came from the fit of the non-landmarks in the style space. For the majority of the non-landmarks (27 out of 32), the data fitted better in the 3D embedding, which made us decide to continue our analysis with the 3D embedding although stress levels suggested the 2D embedding to be already sufficient.

We fitted the creation year to the 3D embedding, and Dimension 2 resulted in a substantial correlation, $r = 0.69$ (Figure 2.6A). In addition, looking at the positions of the 48 cut-outs in Figure 2.5, the Dim1–Dim2 plane, a rotational pattern can be discerned as demonstrated in Figure 2.6B. This rotational component can also be associated with creation year and yielded a somewhat higher correlation, $r = 0.70$ (Figure 2.6C). Interestingly, the half hidden red circle in the third quadrant in Figure 2.6B represents a modern painting from 2019 amidst a set of much older paintings. If we consider this painting as a continuation of the modern cluster from the second quadrant, in other words, if we add 360 degrees to the same data point in Figure 2.6C (the single point in the top left corner), the rotational correlation will even increase to $r = 0.78$. This rotational pattern is particularly interesting because a similar pattern was found by Elgammal et al. (2018), even though their embedding resulted from computational methods and very different experimental parameters. They used paintings of varying subject matter analyzed by a PCA on a CNN layer resulting from training on style labels, while we reached the embedding using human similarity judgement data. As our study and Elgammal et al. (2018) are so different, our finding strengthens the possibility that a cyclical pattern is present in the history of European art during the last six centuries.

Looking at the embedding, some other observations can be made. Along Dimension 1, there appears to be a transition of brushstroke coarseness, from fine brushstroke on the left to coarse brushstroke on the right side. Brushstroke coarseness can be one of the possible features describing the embedding. As Figure 2.6B suggests, the least coarse brushstrokes belong to the modern paintings, while the coarsest ones belong to the impressionists' paintings from the 19th century. This trend in brushstroke coarseness can be one of the possible features describing the embedding and could have been used by participants as a way to differentiate styles. Another observation is a color gradient in the Dim1-Dim2 plane, from green apples on the top left, to yellow and red apples at the bottom. This gradient suggests that color could also have been used to differentiate styles. These two suggestions will be investigated in Experiment 2.

In summary, while the results clearly show a robust style space, we have yet to analyze it further. As we tentatively concluded, there appears to be a trend in Dimension 1 that relates to brushstroke coarseness, and a trend in Dimension 2 related to hue, which might imply that Dimension 3 could be associated with color saturation and/or brightness. To quantify these latent trends, we conducted a second experiment where we used both perceptual attribute ratings and color measurements.

2.3. Experiment 2 - explaining the embedding

Marković and Radonjić (2008) made a distinction between explicit and implicit features. Implicit features refer to subjective impressions (such as how pleasant a painting appears) while explicit features describe “physical properties” of the painting (such as form, color). We choose to define a number of explicit features that potentially contribute to style perception of the apple cut-outs from Experiment 1. Besides subjective rating data, we measured color statistics to account for the possible contribution of color.

2.3.1. Method

Perceptual attributes

The nine visual attributes that we used were Glossiness, Smoothness, Three-dimensionality, Convincingness, Shadow contrast, Colorfulness, Brightness, Brushstroke coarseness, and Contrast between apple and background. Glossiness and Smoothness are typical object-specific features of apples while the other features refer more to how the apple has been depicted. Some of these have been used previously by for example Marković and Radonjić (2008) who used semantic differentials: Three-dimensionality as voluminosity-flat, Convincingness as realistic–abstract, Colorfulness as multicolored–unicolored, Brightness as light–dark, Brushstroke coarseness as strong brush strokes–soft brush strokes. In addition, Convincingness (or realism in different terms) was used in several previous studies (Berlyne and Ogilvie, 1974; Chatterjee et al., 2010; O'Hare and Gordon, 1977; Ruth and Kolehmainen, 1974); Brushstroke coarseness (or clear-indefinite in different terms) was also used in several previous studies (Berlyne and Ogilvie, 1974; Chatterjee et al., 2010; Hasenfus et al., 1983; O'Hare, 1976; Skager et al., 1966). As contrast was concluded to be connected with perceived glossiness (Di Cicco et al., 2019; Marlow and Anderson, 2013), we also included Shadow contrast and Contrast between apple and background in our study.

In the online experiment, each attribute scale was defined by two contrasting concepts, listed in Table 3.2 as left and right labels at either end of the continuous rating scale. No additional information was provided about the attributes to be assessed.

Table 2.1: Keywords of rating scales for attributes rating

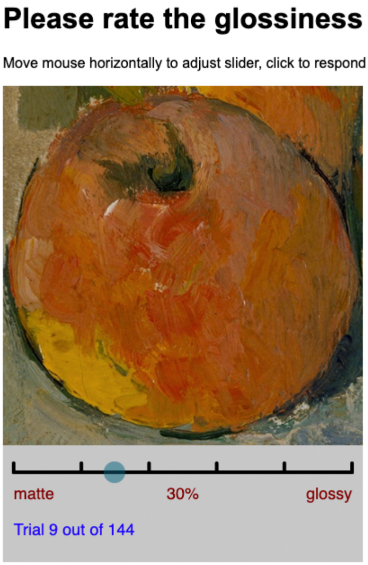
Attributes	Left label	Right label
Glossiness	matte	glossy
Smoothness	rough	smooth
Three-dimensionality	flat	three-dimensional
Convincingness	unrealistic	realistic
Shadow contrast	low	high
Colorfulness	monochrome	colorful
Brightness	dark	bright
Brushstroke coarseness	fine	coarse
Contrast between apple and background	low	high

PARTICIPANTS

224 unique participants recruited from AMT completed Experiment 2 (95.1% were from North America). Each of the nine attributes was rated by 30 unique participants, 270 responses in total. 40 participants rated more than one attribute.

STIMULI AND PROCEDURE

The same 48 stimuli as in Experiment 1 were used in Experiment 2 (as shown in Figure 2.2). Before the actual experiment started, each participant would first read the consent form and instructions for the experiment. They could only proceed if they gave their consent by clicking 'continue' after reading the consent form. Then they went through 15 practice trials, to familiarize with the interface and operation. One stimulus was displayed in each trial (as shown in Figure 2.7). Participants were asked to rate a certain attribute on a continuous scale with 6 markers and numerical feedback ranging between 0% and 100%. Each stimulus rating was repeated three times in a fully randomized set, resulting in 144 trials for each HIT.



Please rate the glossiness of the depicted apple below.

Move mouse horizontally to adjust slider, click to respond and proceed to next trial.

Trial 9 out of 144

Figure 2.7: Experiment 2 interface for Glossiness. In each trial, participants were presented with a single cutout. They could move the mouse horizontally to adjust the rating slider from matte to glossy. With a mouse click they proceeded to the next trial.

COLOR MEASUREMENTS

In addition to the subjective ratings we also computed color data from the apple images. To do so, we masked each apple image with a circular mask with a width of 75% of the image. In this way, colors almost certainly came from the apple and not from its surrounding. Colors were converted to CIELCh color space using the polar coordinates C* (chroma or relative saturation), Hue (hue angle) and L* (lightness). Chroma was defined as $\sqrt{a^{*2}+b^{*2}}$ and thus related to saturation, while Hue was defined by the hue angle, i.e. $\tan^{-1}(b/a)$, values normalized between 0 and 1.

2.3.2. RESULTS

RATING AGREEMENT

For each attribute, we first performed validity checks based on average trial time and correlation with other participants. Data from participants who spent on average less than one second per trial were omitted (but were financially compensated). This threshold was based on similar experiments one of the authors conducted before (Van Zuijlen et al., 2020) and inspection of the time distribution in the current experiment. After the exclusion of the participants that spent less than 1 second, between 19 to 23 participants remained per attribute. After initial inspection we found that a number of these participants seemed to misinterpret the polarity of the rating, i.e. had large but negative correlations with the group mean. Because the number of these cases could vary per attribute and thus result in unequal group sizes if we would use 'negative correlation' as a criterion, we decided to choose the top 15 participants.

As recommended by Martinez et al. (2020), we first performed an intra-participant reliability analysis before determining the inter-participant agreement. Figure 2.8A shows mean values and standard errors of correlations within 3 repeated measurements for each attribute.

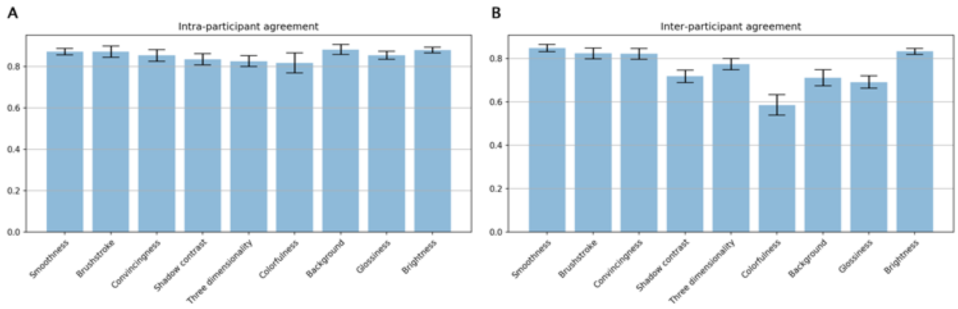


Figure 2.8: A) Mean values and standard errors of correlations within 3 repetition measurements for 15 participants of each attribute. B) Mean values and standard errors of correlation with mean for 15 participants of each attribute.

As for inter-participant agreement, we first calculated the median rating over the three repetitions. We then correlated all the individual median ratings with the group mean (excluding the individual). Figure 2.8B shows mean values and standard errors of correlation with the mean for the participants of each attribute.

We found varying degrees of inter-participant agreement which can be interpreted as perceptual ambiguities (high correlation, low ambiguity and vice versa). The inter-participant agreement varied between 0.85 and 0.59, with the highest scores for Smoothness, Brightness, Brushstroke and Convincingness. The lowest score was for Colorfulness, with the others in between. The relatively constant high intra-rater reliability correlations (all above 0.8) in Figure 2.8A suggest that differences between observers for the various attributes are truly due to inter-observer ambiguities.

Multiple linear regression of perceptual attributes

Figure 2.9 presents the results of the multiple linear regressions within the MDS embedding from Experiment 1, using the three dimensions as independent variables and the attributes as dependent variables. The orientation of the vector indicates the direction that yielded the best regression, while the length indicates strength of the regression (R^2). Table 2.2 denotes corresponding adjusted r square values, overall p-value and weights (beta coefficients) plus p-values for each dimension of the fit per attribute. The following attributes have a high overall fit within the MDS embedding: Smoothness, Brushstroke coarseness, Convincingness, Shadow contrast. The remaining attributes are moderately (Three-dimensionality, Colorfulness, Contrast between apple and background) or only weakly correlated (Glossiness, Brightness).

As Table 2.2 shows, all attributes except Colorfulness can be significantly associated with Dimension 1 from the Experiment 1 embedding, with Glossiness and Brightness only weakly associated. Dimension 2 has only weak associations with Colorfulness and

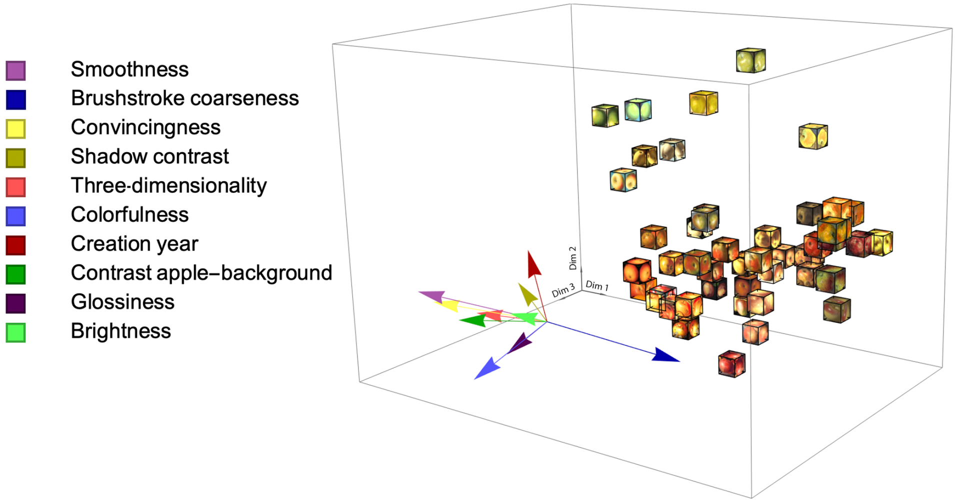


Figure 2.9: 3D MDS configuration embedded with 9 attribute vectors and creation year vector.

Glossiness. Finally, Dimension 3 has a unique high association with Colorfulness and a moderate one with Brightness.

FITTING OF COLOR DATA

Table 2.3 shows per color coordinate (Hue, Chroma, Lightness) the adjusted R^2 values, overall p-value and weights plus p-values for each dimension of the MDS embedding from Experiment 1. Hue and Chroma both have significant overall fittings, while Lightness is not significantly associated with the 3D perceptual style space. Hue is primarily associated with Dimension 2 but also has some weight on Dimension 1. The only significant weight for Chroma is on Dimension 3. We also provide a correlation matrix of the nine attributes and three color measurements in the supplementary material.

To illustrate the relation between image color coordinates Hue and Chroma and Dimensions 2 and 3 of the embedding, we plotted the Dim2-Dim3 projection of the embedding next to the Hue-Chroma plot. The result can be seen in Figure 2.10. A visual comparison between the style space and the color space underscores the strong association of Hue and Chroma with Dim2 and Dim3, respectively. It should be noted that although Hue has its primary weight on Dimension 2, the positive correlation between Hue and creation year (both highly correlated with Dimension 2) is low ($r = 0.35$), the two vectors of creation year and Hue having a substantial angle of 40.54 degrees because of the negative correlation between Hue and Dimension 1. In addition, the relation between creation year and Dim2 was explored further by calculating the partial correlation while controlling for Hue ($r = 0.66$; $p < 0.001$). The resulting correlation shows a small drop with respect to the original correlation ($r = 0.69$).

Table 2.2: Multiple linear regression of perceptual attributes.

	adjusted r square	overall p-value	dim1	p-value	dim2	p-value	dim3	p-value
Smoothness	0.88	0.000^{***}	-0.77	0.000^{***}	0.08	0.145^{NS}	0.17	0.031^{*}
Brushstroke	0.88	0.000^{***}	0.79	0.000^{***}	-0.10	0.079^{NS}	-0.12	0.167^{NS}
Convincingness	0.85	0.000^{***}	-0.64	0.000^{***}	-0.05	0.919^{NS}	0.06	0.409^{NS}
Contrast	0.70	0.000^{***}	-0.42	0.000^{***}	0.02	0.709^{NS}	-0.29	0.000^{***}
3D	0.68	0.000^{***}	-0.46	0.000^{***}	-0.07	0.248^{NS}	-0.08	0.388^{NS}
Colorfulness	0.61	0.000^{***}	-0.01	0.712^{NS}	-0.16	0.002^{**}	0.61	0.000^{***}
Background	0.50	0.000^{***}	-0.44	0.000^{***}	0.00	0.996^{NS}	0.25	0.045^{*}
Glossiness	0.26	0.001^{***}	-0.22	0.002^{**}	-0.23	0.01^{**}	0.21	0.104^{NS}
Brightness	0.20	0.005^{**}	-0.21	0.026^{*}	0.19	0.101^{NS}	0.43	0.014^{*}

Note:
Brushstroke: Brushstroke coarseness;
Contrast: Shadow contrast;
3D: Three-dimensionality;
Background: Contrast between apple and background;

* $p < 0.05$;
** $p < 0.01$;
*** $p < 0.001$;
^{NS} means not significant ($p > 0.05$).

Table 2.3: Multiple linear regression of color measurements.

	adjusted r square	overall p-value	dim1	p-value	dim2	p-value	dim3	p-value
Hue	0.55	0.000***	-0.05	0.008**	0.16	0.000***	-0.05	0.103NS
Chroma	0.55	0.000***	0.02	0.528NS	-0.01	0.802NS	0.53	0.000***
Lightness	0.04	0.186NS	-0.07	0.194NS	0.11	0.115NS	0.09	0.401NS

Note:

* $p < 0.05$;
** $p < 0.01$;
*** $p < 0.001$;
NS means not significant ($p > 0.05$).

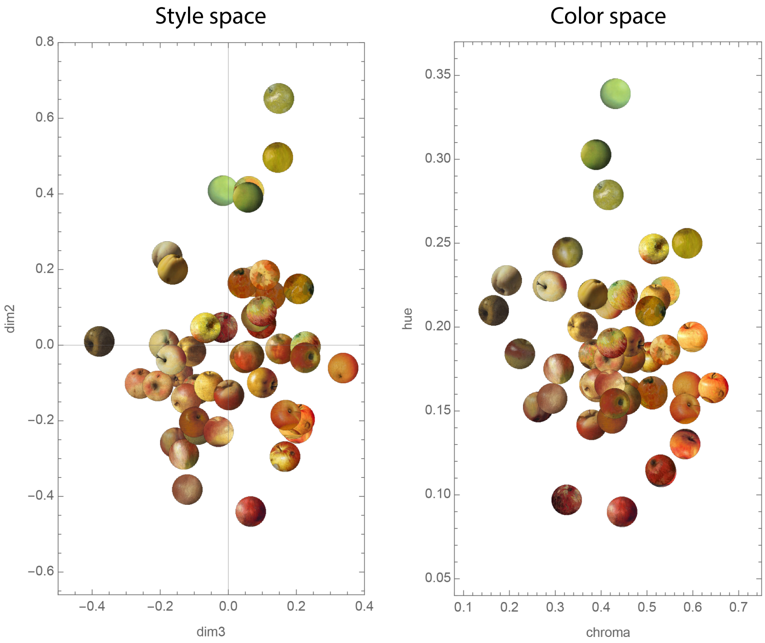


Figure 2.10: Comparison between Dim2 - Dim3 plane of the MDS embedding (on the left) and 2D plane of Hue and Chroma measurements (on the right).

2.3.3. Discussion

All visual attributes (highly to weakly) correlate significantly with the 3D embedding, as can be seen by the general adjusted R^2 values in Table 2.2. The most prominent attributes are Smoothness, Brushstroke and Convincingness and the least contributing attributes are Brightness and Glossiness. For the color measurements, Hue and Chroma have significant overall correlations with the 3D embedding. Lightness has no significant correlation with the embedding, which is in line with the low contribution of Brightness to the perceptual attributes analysis. Finally, the best fitting attributes are about spatial aspects of the paintings to which Convincingness is firmly associated, where Convincingness in its turn can be associated with realism according to O'Hare (1976).

Although all attributes (apart from Colorfulness and, to a lesser extent, Brightness) correlated significantly with Dimension 1, Brushstroke coarseness and Smoothness were the strongest ones. Similar findings were reported in previous studies (Berlyne, 1973; Gardner, 1974; Klein, 1968; O'Hare, 1976; O'Hare and Gordon, 1977; Skager et al., 1966), with similarly defined attribute names (e.g. clarity, texture). For instance, O'Hare (1976) reported in his study that the second dimension in his findings could be interpreted as clarity or clear definition of detail, from sharp outlines to diffuse and indefinite outlines. Gardner (1974) also reported that texture (brushstroke shapes, lightness gradients, etc.) makes a significant contribution to an artist's style. Elgammal et al. (2018) reported a correlation between their second dimension and Wölfflin's principle of linear vs. painterly, which is connected to clarity of outline (brushstroke).

Dimension 2 corresponded strongly with creation year as shown in Experiment 1. Interestingly, none of the attributes correlated with this dimension, except relatively weak negative correlations for Colorfulness and Glossiness. From Experiment 1 it was already visible that Hue could possibly also be related to Dimension 2. Indeed, the color statistics for Hue show a high coefficient of determination ($R^2 = 0.55$) which originates from a direction mostly in the positive Dimension 2 direction (the significant weight of 0.16 in Table 2.3) and to a lesser degree in the negative Dimension 1 direction (the significant weight of -0.05 in Table 2.3). This becomes visually clear when again looking at Figure 2.5 where a clear transition from red to green is visible in Dimension 2 and one may also see more yellow/greenish apples on the left side than on the right side in Dimension 1. Although the trend is clearly visible, the interpretation is less straightforward and we will continue this in the General Discussion.

Dimension 3 is related to Colorfulness, another suggestion that participants might have used color information for the similarity judgements in our study. In addition, fitting the results from the color measurements suggested Dimension 3 was connected to Chroma only. It should be noted that the rating scale of Colorfulness was defined by monochrome to colorful, hence it was expected that participants interpreted the term “colorfulness" as hue diversity, however, the results suggest that they interpreted the term more as saturation. In other words, semantic reasons might have caused different interpretations, which is also suggested by the lowest inter-participant correlation for Colorfulness (see Figure 2.8B, low correlation, high ambiguity).

In the attribute rating experiment, relatively high inter-participant correlations were found (Figure 2.8B), which is in line with the significant inter-subject consistency reported by Berlyne and Ogilvie (1974). Next to Brightness, the lowest agreement has been found for Shadow contrast, Contrast between apple and background and Glossiness. While this could partially be semantic, it may also be visual. Especially Glossiness is a term that is generally unambiguous, the relatively low agreement score could therefore indicate that there is not much variation in Glossiness within the 48 apples. Three-dimensionality scored higher, followed by Smoothness, Brushstroke coarseness and Brightness.

2.4. General discussion

We have measured the perception of style using a supposedly constant motif, the apple, by using square cut-outs of paintings. Gombrich (2009)'s description of style ("Style is a distinctive, and therefore recognizable, ...") was operationalized in two experiments: the first quantifying distinction by performing a landmark MDS experiment, the second describing the resulting embedding, which can be related to recognizing style. The results reveal an interesting, non-random multidimensional embedding of 48 apple depictions that are related through various visual features. The embedding is even more interesting considering only low-level information was left in the square cut-outs. Previous studies (Siefkes and Arielli, 2018; Wallraven et al., 2009) believed humans need high-level information to perceive different styles, which was removed as much as possible in our study. It suggests that low-level information might be sufficient for participants to perceive style differences.

In Experiment 1 we also found a strong correlation between creation year and our perceptual space, with both a linear fit along Dimension 2 ($r = 0.69$) and a circular fit in Dim1-Dim2 plane ($r = 0.70$). These correlations were surprising, considering all the high-level and mid-level information, in other words, all the time-related items and surroundings (e.g. clothes, house interior) that can provide information about creation time, were removed. Indeed, connections between the perceptual space and paintings' creation year has been reported in other studies, but all with the whole paintings as stimuli. Berlyne and Ogilvie (1974) found a high multiple correlation (>0.8) between their 3D perceptual space and artists' year of birth, which roughly scales with the creation year of the paintings. And Berlyne and Ogilvie (1974) interpreted the first dimension in their perceptual space as old vs. modern. Elgammal et al. (2018) also found a temporal pattern while using computational methods instead of human judgements. Their embedding was achieved by training neural networks on WikiArt style labels. Thus, similar findings from both human judgement and computer algorithm indicate a relatively robust correlation between style and time. And if we consider the circular fit, the creation year changes in a cycle of both texture and Hue.

In Experiment 2 we described the 3D perceptual style space with multiple linear regressions of nine attributes and color measurements. The first dimension of the style embedding clearly related to many of the attributes, most prominently Smoothness and Brushstroke coarseness, but also others like Convincingness, Shadow contrast, and Three-dimensionality. Except Convincingness being a higher-level attribute, all other mentioned attributes are related to spatial properties. As shown in Figure 2.9 and Table 2.2, Smoothness, Shadow contrast, Three-dimensionality and Convincingness all point in the same direction, which indicates that increasing Smoothness, Shadow Contrast and Three-dimensionality could enhance Convincingness. Similar positive correlations between Contrast, Three-dimensionality and Convincingness was reported in a previous study (Di Cicco et al., 2019). Smoothness and Brushstroke coarseness have opposite directions in the 3D embedding, implying they have an almost perfect negative correlation, which appears logical as they indeed seem semantic opposites. But it should be noted that the instructions for the Smoothness rating experiment explicitly mentioned the apple skin with the intention that Smoothness should relate to what is represented (the apple) while Brushstroke coarseness clearly relates to the medium. However, our results pointed at a transfer between these two modes, perhaps because apples painted in a rough manner cannot easily be judged as being smooth. Such phenomenon can be further tested in controlled experiments where motif and medium are systematically varied.

The remaining two dimensions are associated with color. The second dimension is associated with Hue and the third dimension with Chroma as well as the attribute Colorfulness. It seems there is some connection between Hue and creation year since they are both positively correlated to Dimension 2. Indeed, from the beginning of the Nineteenth century the production of new synthetic pigments exploded, leading to a variety of colors, unheard of in earlier centuries (Ball, 2003; Wilson-Bareau, 1991). Artists such as Rembrandt had to make do with about a dozen pigments, while Monet or Van Gogh could literally choose hundreds of different pigments. This has led to an increase of saturation of violets and greens for example. Another possibility is that it shows the history of the painted objects, in our case apples. In spite of their seemingly independence of historical developments in fashion or the development of technology, it is quite possible that European cultivated apples have a history of their own, in which there has been a gradual increase in saturated green varieties over the last century or so. However, even if we only consider the linear fit of the creation year with Dimension 2, this time dimension still cannot be fully explained by Hue change, given the creation year and Dimension 2 have a high correlation ($r = 0.69$), while the creation year and Hue have a low correlation ($r = 0.35$), and the two vectors of the creation year and Hue have a substantial angle of 40.54 degrees between them. This conclusion is convincingly supported by the fact that the partial correlation between Dimension 2 and creation year, controlling for Hue, has a value of 0.66, being close to the original correlation.

Although color measurements couldn't explain the time dimension, the contribution of color in style perception was robust and also reported in early studies. Gardner (1974), for example, concluded that both color and texture played a significant role in style detection. Interestingly, Dimension 1 in our study is mainly associated with spatial attributes, such as brushstroke coarseness, smoothness, shadow contrast and three-dimensionality, which can also be interpreted as texture. Hence, we reached the same conclusion as Gardner (1974) that texture and color are two important variables for pictorial style perception.

In this study, we showed that in judging matters of style, participants in our two experiments demonstrated high inter-subjective agreement, in line with earlier studies on the perception of style in art. We also found that participants by and large followed the historical time line when performing their matching tasks. In our case this concerned only small details of sometimes much larger paintings (our apple stimuli), thus removing such important aspects as composition, mood, or general intention of the work of art as a whole. With regard to paintings, people are apparently quite capable of looking at the 'how' of a painted subject. They show a definite sense of style.

Experiment 2 showed some of the perceptual ingredients on which this sense of style may rely, but there did not seem to be a single, one-dimensional perceptual factor explaining the results. Perhaps our sense of pictorial style is just one member of a much larger family of human sensitivities for the 'how' of something made or done by other humans: e.g. handwriting styles, dialects, speech habit, dancing styles (Hasenfus et al., 1983). In all such activities people detect various components simultaneously, like if it were Gestalts. Further research on recurring motifs in the history of art (e.g. hands, textile folds) may get us closer to discovering the various roots of this important sense of style in humans.

Bibliography

Ashe, T. (2014). *Color management & quality output: Working with color from camera to display to print*. CRC Press.

Ball, P. (2003). *Bright earth: Art and the invention of color*. University of Chicago Press.

Berlyne, D. E. (1973). Interrelations of verbal and nonverbal measures used in experimental aesthetics. *Scandinavian Journal of Psychology*, *14*(1), 177–184.

Berlyne, D. E., & Ogilvie, J. C. (1974). Dimensions of perception of paintings. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 181–22). Hemisphere.

Cetinic, E., Lipic, T., & Grgic, S. (2020). Learning the principles of art history with convolutional neural networks. *Pattern Recognition Letters*, *129*, 56–62.

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts*, *28*(2), 207–222.

De Silva, V., & Tenenbaum, J. B. (2004). *Sparse multidimensional scaling using landmark points* (tech. rep.). Stanford University.

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), 1–15.

Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., & Mazzone, M. (2018). The shape of art history in the eyes of the machine. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 2183–2191.

Gardner, H. (1974). The contribution of color and texture to the detection of painting styles. *Studies in Art Education*, *15*(3), 57–62.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Gombrich, E. H. (2009). Style. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The art of art history: A critical anthology, 2nd editon* (pp. 114029–290). Oxford University Press.

Goude, G., & Derefeldt, G. (1981). A study of Wölfflin's system for characterizing art. *Studies in Art Education*, *22*(3), 32–41.

Graham, D. J., Friedenberg, J. D., Rockmore, D. N., & Field, D. J. (2010). Mapping the similarity space of paintings: Image statistics and visual perception. *Visual cognition*, *18*(4), 559–573.

Graham, D. J., Hughes, J. M., Leder, H., & Rockmore, D. N. (2012). Statistics, vision, and the analysis of artistic style. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 115–123.

Hasenfus, N., Martindale, C., & Birnbaum, D. (1983). Psychological reality of cross-media artistic styles. *Journal of experimental psychology: Human Perception and Performance*, *9*(6), 841–863.

Hughes, J. M., Graham, D. J., Jacobsen, C. R., & Rockmore, D. N. (2011). Comparing higher-order spatial statistics and perceptual judgements in the stylometric analysis of art. *2011 19th European Signal Processing Conference*, 1244–1248.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.

Klein, S. (1968). Using points of view and multidimensional scaling analyses to describe aesthetic judgments. *Proceedings of the 76th Annual Convention of the American Psychological Association, 447*.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129.

Linde, L. (1975). Similarity of poetic rhythms with different amounts of semantic content-stress ratings and pairwise similarity ratings. *Scandinavian Journal of Psychology*, *16*(1), 240–246.

Marković, S., & Radonjić, A. (2008). Implicit and explicit features of paintings. *Spatial vision*, *21*(3-5), 229–259.

Marlow, P. J., & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of vision*, *13*(14), 2–2.

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods*, *52*, 1428–1444.

Mead, A. (1992). Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *41*(1), 27–39.

O'Hare, D. (1976). Individual differences in perceived similarity and preference for visual art: A multidimensional scaling analysis. *Perception & Psychophysics*, *20*(6), 445–452.

O'Hare, D. (1979). Multidimensional scaling representations and individual differences in concept learning of artistic style. *British Journal of Psychology*, *70*(2), 219–230.

O'Hare, D., & Gordon, I. (1977). Dimensions of the perception of art: Verbal scales and similarity judgements. *Scandinavian Journal of Psychology*, *18*(1), 66–70.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R., Simpson, G., Solymos, P., et al. (2019). Vegan: Community ecology package. r package version 2.5–6. 2019.

Rao, A., Srihari, R. K., & Zhang, Z. (1999). Spatial color histograms for content-based image retrieval. *Proceedings 11th International Conference on Tools with Artificial Intelligence*, 183–186.

Ruth, J.-E., & Kolehmainen, K. (1974). Classification of art into style periods; a factor-analytical approach. *Scandinavian Journal of Psychology*, *15*(1), 322–327.

Sablatnig, R., Kammerer, P., & Zolda, E. (1998). Hierarchical classification of paintings using face-and brush stroke models. *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, *1*, 172–174.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

Siefkes, M., & Arielli, E. (2018). The aesthetics and multimodality of style. *Experimental Research on the Edge of Theory. Bern: Lang*.

Skager, R. W., Schultz, C. B., & Klein, S. P. (1966). The multidimensional scaling of a set of artistic drawings: Perceived structure and scale correlates. *Multivariate Behavioral Research*, *1*(4), 425–436.

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), 1–17.

Wallraven, C., Fleming, R., Cunningham, D., Rigau, J., Feixas, M., & Sbert, M. (2009). Categorizing art: Comparing humans and computers. *Computers & Graphics*, *33*(4), 484–495.

Wilson-Bareau, J. (1991). Art in the making. Impressionism. London, National Gallery.

Wölfflin, H. (2012). *Principles of art history*. Courier Corporation.

3

Material Perception Across Different Media-comparing perceived attributes in oil paintings and engravings

We investigated the influence of medium on the perception of depicted objects and materials. Oil paintings and their reproductions in engravings were chosen because they are vastly distinctive media while having completely identical content.

A total of 15 pairs were collected, consisting of 88 fragments depicting different materials, including fabric, skin, wood and metal. Besides the original condition, we created three manipulations to understand the effect of color (a grayscale version) and contrast (equalized histograms towards both painting and engraving). We performed rating experiments on five attributes: three-dimensionality, glossiness, convincingness, smoothness and softness. An average of 25 participants finished each of the 20 online experimental sessions (five attributes X four conditions).

Besides clear correlations between the two media, the differences mainly show in their means (different levels of perceived attributes) and standard deviations (perceived range). In most sessions, paintings depict a wider range than engravings. In addition, it was the histogram equalization (global contrast) that made the most impact on perceived attributes, rather than color removal. This suggests that engravers compensated the lack of color by exploiting the possibilities of local contrast.

Published as Zhao, Y., Stumpel, J., de Ridder, H., & Wijntjes, M. W. (2024). Material perception across different media—comparing perceived attributes in oil paintings and engravings. I-Perception, 15(4). https://doi.org/10.1177/20416695241261140

3.1. Introduction

Around the time that in Italy linear perspective was discovered (Alberti, 1966), a material rendering innovation was taking place in Northern Europe mediated by the invention of oil paint. Although he may not have been the inventor, van Eyck was certainly the artist discovering the huge potential of oil paint for the convincing rendering of materials. The deeper colors and the slow drying, which enabled smooth transition and easy alteration, offered artists possibilities that did not exist for tempera paint (Bol, 2023). While the invention of linear perspective was related to the mathematics of projection, the material rendering revolution was related to a specific medium: that of oil paint. The perceptual influence of media is a relatively understudied topic and we made that the topic of the current chapter. However, instead of comparing oil and tempera, we choose two media that are more distant from each other: oil paintings and engravings. In the context of this research, the term 'engraving' specifically refers to monochrome engravings, and explicitly excludes any painted or color printed engravings.

The artistic handling of a medium is related to the topic of style. Within a certain medium, like oil paint, there are obviously many different styles as art history has shown. We previously found a relation between differently depicted apples and their material properties (Zhao et al., 2023). Van Zuijlen et al. (2020) took a different approach by collecting a large variety of annotated material segments from historical oil paintings. They collected material attribute ratings for 15 different material classes and compared them to a study of similar nature that used photographs (Fleming et al., 2013). Interestingly, the material 'signatures' (perceptual characterization of 10 material attribute ratings) are very similar between paintings and photographs, suggesting material perception might be independent of medium. Instead of comparing paintings and photographs, in a more controlled fashion Delanoy et al. (2021) compared realistic material computer renderings with their painting replicas by an artist. They reached the conclusion that material properties in paintings and renderings were perceived very similarly and were linked to the same image features. While these studies suggest that material perception might be independent from media, Bousseau et al. (2013) found differences between realistic renderings and painterly renderings of the same scenes. Their results showed that in painterly renderings, the range of distinguishable gloss levels reduces under increased brush size of opaque strokes, use of semitransparent strokes, or when texture of brush strokes and varnish were introduced.

The different conclusions from previous studies left the question unanswered whether medium has influence on material perception. In the current study, we wanted to answer this question using the same variety of depictions as Van Zuijlen et al. (2020) while comparing different media. At the same time, we also desired that the subject matter could be kept constant as was achieved by Bousseau et al. (2013) and Delanoy et al. (2021). The requirement of identical subject matter was difficult to meet, since artists generally compose original pictures which do not share a perfect subject matter resemblance.

We found a solution by comparing paintings and their reproductions in print media, particularly engravings. Engraving has been a form of art on its own, but also as a method to reproduce paintings from the seventeenth century onwards, before various printing techniques that made direct use of photographic images, from rotogravure, to off set and beyond. The identical pictorial content in oil paintings and their engraved reproductions provides a perfect opportunity to compare the portrayal of materials such as fabric and skin across the two media without the confounding factor of different subject matter. Maintaining constant subject matter would be much more difficult, if not impossible, if we were to compare oil and tempera paint. Furthermore, the two media are drastically different, which makes it a critical case study for the influence of media on perception.

Seemingly originating from goldsmithing, engraving emerged in the late fifteenth century in Germany and Italy. As an intaglio process, engravings are created with a burin, a wedge-shaped metal tool, to carve into the base plate usually made of copper. The plate, consisting of grooves created by burin, could hold ink. Ink would then transfer onto a damp sheet of paper under high pressure to complete a print. The early German master Martin Schongauer raised engraving from a minor craft to a major art form with compelling works, followed by Albrecht Dürer, and many other masters (Thompson, 2000). The process of engraving differs from etching. In etching, the metal plate is covered by a layer of wax or soft varnish. The artist can draw effortlessly by removing parts of this layer with a needle, upon which a chemical process with acid creates the grooves. However, in engraving the grooves are made directly by the handling of the burin which requires great skill and craftsmanship, based on years of training.

Engraving is a challenging medium not only because of the difficulty in craftsmanship, but also because it is a medium restricted by monochromatic lines and dots. Oil paintings have colored fluid brush strokes and could easily achieve smooth color transitions and color contrast. Engravings, on the other hand, are categorically different, with only ‘black’ and ‘white’ (color of the ink and the paper). Luminance contrast is achieved by the distribution of lines. Within these boundary conditions, engravers were still able to create form, texture, shading and highlights. Engravers had their own idiosyncratic approach to create engraving lines, some preferred to use lines that followed the contours, some preferred cross hatching to create shading and three-dimensional (3D) volume (Thompson, 2000).

To quantify the perceptual differences between paintings and engravings, we focused on measuring five perceptual attributes of various depicted objects. We investigated the depiction of materials by letting observers rate the smoothness, glossiness and softness. Furthermore, we let observers rate three-dimensionality to assess the depiction of shape. In addition to investigating the formal elements of material and shape, we were also interested in the overall quality of the depictions of objects. Therefore, we asked observers to rate the ‘convincingness’. We will shortly elaborate on these five attributes.

Since many old masters in both painting and engraving pursued realistic and convincing depiction, we compared convincingness of these two media. As an overall judgement, convincingness (or realism in different terms) has been widely studied in the field of visual perception (Berlyne and Ogilvie, 1974; Chatterjee et al., 2010; Di Cicco, 2022; Di Cicco et al., 2018; O’Hare and Gordon, 1977) and is often considered an important perceptual measurement. It should be noted that convincingness seems to play a role both in historical pictorial revolutions such as the invention of linear perspective and oil paint as discussed above, but also in contemporary pictorial revolutions. The recent success of AI mediated synthetic image algorithms such as Midjourney is largely attributable to their impressive convincingness (Göring et al., 2023).

Gloss is the most widely studied attribute that is important for material perception

(Marlow and Anderson, 2013; Pellacini et al., 2000), including real and photographed objects (van Assen et al., 2016; Zhang et al., 2019), computer rendered images (Wendt et al., 2008), and also for paintings (Bousseau et al., 2013; Delanoy et al., 2021; Di Cicco et al., 2019). Previous studies concluded that gloss perception is mostly determined by contrast, sharpness and coverage of the highlights (Di Cicco et al., 2019; Marlow et al., 2012). Contrast, which is manifested distinctively in oil paintings and engravings, plays a pivotal role as one of the key features and predictors of gloss perception (Di Cicco et al., 2019).

Smoothness plays an important role in perceived realism (Rademacher et al., 2001). Sometimes it has been measured as its opposite, roughness (Delanoy et al., 2021; Di Cicco et al., 2021; Zhang et al., 2019). What is furthermore interesting about smoothness is that it can refer to the smoothness of the depicted object (the motif) but also to the depiction (the medium). This could theoretically also be the case for gloss, but the glossiness of the medium (e.g. caused by the varnish) is often made invisible by the way of visual documentation: a glossy reflection in a photo copy of a painting is rather undesirable. However, the roughness of brushstrokes or hatching is difficult to ignore. Interestingly, Zhao et al. (2023) found a potential transfer of smoothness between medium (smooth brushstroke) and motif (depicted apples) in a study on style perception. In the current study, we were interested whether the visible engraving lines in the medium (see Figure 3.6) may influence the perceived smoothness of the depicted materials.

The third material attribute that we decided to investigate is softness. It is particularly related to materials such as fabric and skin, which make up the larger part of our stimulus set. Previously, it has been found that softness is not correlated to roughness in a study on depicted fabric perception (Di Cicco et al., 2021). Furthermore, softness could be seen as a more mechanical property as opposed to the optical property of gloss. Hence, softness complements the other two material attributes rather well.

A related attribute, though not a material attribute but rather a shape attribute, is three-dimensionality. There is a strong perceptual connection between gloss and three dimensional shape (Fleming et al., 2004; Norman et al., 2004; Todd and Mingolla, 1983). Contrast is also used as an effective depth cue for 3D shape perception (O'Shea et al., 1994). Since engraving has different approaches than oil painting to achieve 3D rendering, we will investigate the performance of the medium in expressing three-dimensionality.

There are a number of a priori differences between paintings and engravings that could lead to perceptual differences. Color and contrast are the most prominent differences. Being denied access to colors, engravers likely compensated by deploying all available efforts towards the luminance channel. While we empirically investigated the 'original' (albeit digitized) pictures, we additionally included image manipulations to better understand the respective roles of color and contrast.

The first image manipulation served to understand the role of color and consisted of taking gray scale versions of both stimuli. This was established by converting the colors into luminance values. To understand the role of luminance contrast we equalized the respective luminance histograms. However, because the luminance histogram of an engraving theoretically consists of two single peaks at the white of the paper and the black of the ink, we first blurred the engraving such that hatchings became smooth gradients. To counterbalance the blurring manipulations on the engravings, we applied the same procedure on the paintings. In sum, we added two manipulation conditions to the original condition: gray scale and equalized luminance histogram. As the latter condition can be applied both from the engraving to the painting and vice versa, this condition consisted of two versions. Thus, a total of 4 conditions (original, grayscale and two histogram equalizations) were measured in the following experiment.

3.2. METHOD

3.2.1. STIMULI

We collected 15 pairs of digital copies of color oil paintings and their engraving reproductions. Identical content gave us the opportunity to take medium as a controlled variable and minimize the influence of content, or 'subject matter'. Most oil paintings are portraits or scenes of daily life to ensure the diversity of materials. Both original oil paintings and their engraving reproductions covered a wide range of creation years. The creation year of the original oil paintings varied from 16th to 18th century, while the creation year of engravings ranged from 17th to 19th century. Figure 3.1 shows an overview of all stimuli.

Before further processing, we first endeavoured to align all the pairs and crop them into the same framing. Since engravings are not photo copies, their framing and aspect ratio can differ slightly from the original oil paintings. Better aligned content can further reduce the influence of subject matter. A few images were mirrored for the alignment. Besides, some engravings have text below the figures, which is different from oil paintings. Removing text reduced the possibility for participants to easily infer the media. All the aligned and cropped high resolution images were then rescaled in Adobe Photoshop (Adobe Inc., 2021). Since they have different aspect ratios, we set the longer edge to be 1500 pixels.

Then we created two manipulations to understand the effect of color and contrast. Firstly, we removed chromatic information by creating a grayscale version. The conversion was performed in Mathematica (Wolfram Research Inc., 2020), the formula from sRGB to grayscale is $Grayscale = 0.299R + 0.587G + 0.114B$. Secondly, we removed the difference in global luminance contrast by equalizing histograms (towards both painting and engraving, hence 2 versions). The histogram matching was performed with the 'HistogramTransform' function in Mathematica (Wolfram Research Inc., 2020). Before the histogram equalization we removed high frequency information by blurring the images. This was necessary as a 'sharp' engraving is essentially a bitmap: black when there is a line, white in background, without intermediate grayscale values which only emerge when viewed from a distance, i.e. blurred. The Gaussian blur radius for each stimulus was determined per picture individually so that engraving lines became invisible. It should be noted that the purpose of the blurring was to facilitate histogram matching, and not a purpose on its own, hence we only have two image manipulation conditions: grayscale and histogram equalization. That we end with 4 experimental conditions (including the 'original' condition) is due to histogram equalized images had two versions with either the painting or the engraving functioning as source.

After the manipulations, all stimulus images were converted in Photoshop (Adobe Inc., 2021) to PNG format, and embedded with an sRGB ICC color profile, for browsers

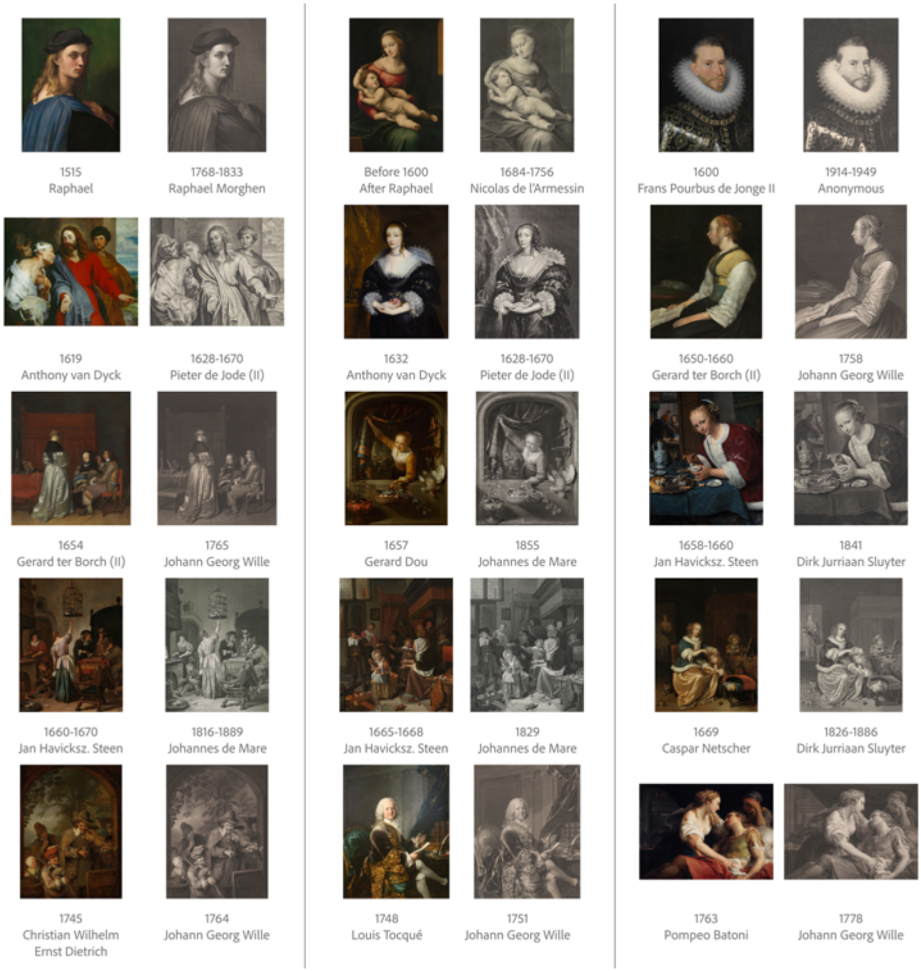


Figure 3.1: An overview of 15 pairs of stimuli, sorted by creation year of oil paintings. In some cases, where there is no precise creation year information available, we presented the estimated range of creation year, or the lifespan of the artist.

Table 3.1: Number of selections for each material category

	3D/Gloss/Convincingness/Smoothness	Softness
Fabric	54	54
Skin	18	18
Lace	4	4
Fur	2	2
Metal	3	na
Wood	6	na
Ceramic	1	na

Table 3.2: Keywords of rating scales for attributes rating

Attributes	Left label	Right label
Three-dimensionality	flat	three-dimensional
Glossiness	matte	glossy
Smoothness	rough	fine
Softness	hard	soft
Convincingness	unrealistic	realistic

to display colors properly (Ashe, 2014).

Lastly, from each picture pair we selected multiple objects, including fabric, skin, lace, wood, metal and ceramic, marked with a red outline in the experiment interface (see Figure 3.3). In total, we selected 88 objects from these 15 pairs. Table 3.1 shows numbers of selections in detail. A preview of all 88 selections can be found in the supplementary material.

3.2.2. EXPERIMENTAL DESIGN

The study consisted of 20 online experimental sessions. In each session, a unique group of participants judged one of five attributes for the two media (oil paintings and engravings) in one of four conditions: original (ori), grayscale (bw), histogram of painting matched to that of engraving (hmp), histogram of engraving matched to that of painting (hme) (see Figure 3.2). Per attribute, this resulted in a two by four mixed design, with medium as a within-subject and condition (manipulations) as a between-subject variable. The five attributes to be judged were: three-dimensionality, glossiness, smoothness, softness and convincingness. Each attribute scale was defined by two contrasting terms, listed in Table 3.2 as left and right labels at either end of the continuous rating scale. No additional information was provided about the attributes to be assessed.

All attributes have 88 material selections in total except softness that has 78, since metal, wood or ceramic are not relevant for softness. As a result, three-dimensionality, glossiness, smoothness and convincingness had (88 times two) 176 trials and softness had (78 times two) 156 trials for each session.

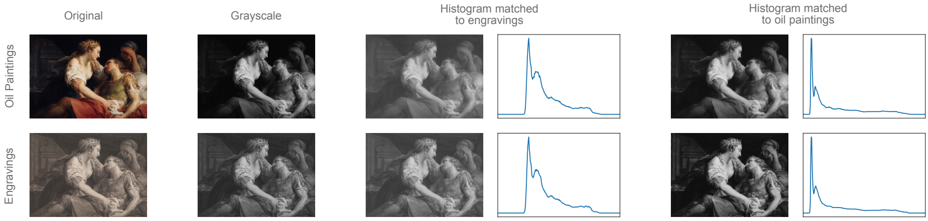

Original
Grayscale
Histogram matched to engravings
Histogram matched to oil paintings
Oil Paintings
Engravings

Figure 3.2: Four conditions: original, grayscale, histogram of painting matched to that of engraving (hmp), histogram of engraving matched to that of painting (hme). For hmp and hme conditions, we first applied the same Gaussian blur to both engravings and paintings before histogram matching so that the engravings have smooth histograms and no visible engraving lines. After histogram matching, they have the same overall luminance distribution. Note that blurred oil paintings usually have higher contrast than blurred engravings. The oil painting: Pompeo Batoni, *La mort de Marc Antoine*, 1763. Downloaded from Wikipedia. The engraving: Johann Georg Wille, *La Mort de Marc Antoine*, 1778. Downloaded from the online repository of the Rijksmuseum, Amsterdam. Both images were cropped to the same framing.

3.2.3. Participants

600 unique participants were recruited for our experiment, 30 participants for each experimental session. However, we lost some responses due to server issues which resulted in an average of 25 participants for each session. All participants were recruited from Prolific (www.prolific.co) from all available countries. The experiment was conducted in agreement with the Declaration of Helsinki and approved by the Human Research Ethics Committee of the Delft University of Technology. All data were collected anonymously.

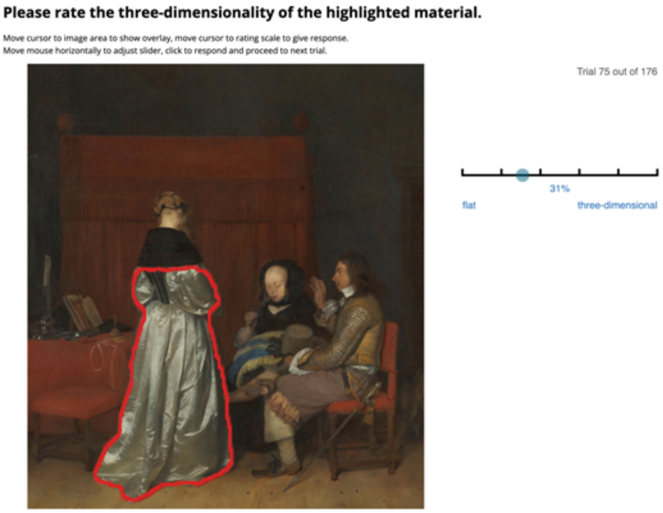
3.2.4. Procedure

Each participant would first read instructions and the consent form before the actual experiment. Then they would perform ten practice trials to get familiar with both the interface and the variety of stimuli. Their task was to rate one of the five attributes regarding the selection marked by a red outline (see Figure 3.3). Each participant just rated one attribute (e.g. softness) in one condition (e.g. original), in two different media (engraving and painting). The order of trials was randomized across participants.

The interface was designed to minimize the influence of the red outlines: they would first flash twice when the trial started. Then the participant could receive a reminder by moving the cursor to the image. When participants moved the cursor to the right side, the red outlines disappeared and the cursor controlled the rating scale automatically. They could click to rate and proceed to the next trial. Clicking on the image was disabled to avoid accidental ratings.

3.2.5. Data analysis

We first performed validity checks for the raw data. We excluded participants who spent less than 1 second on average for each trial. This threshold was based on previous experience (Van Zuijlen et al., 2020) in our group. It is very likely that too short answering time means clicking without paying attention, which can result in noisy data. After filtering, each session had on average of 24.4 participants. Then we performed z-score normalization on the rating data per participant, so that we can later combine data from different



Please rate the three-dimensionality of the highlighted material.

Move cursor to image area to show overlay, move cursor to rating scale to give response.
Move mouse horizontally to adjust slider, click to respond and proceed to next trial.

Trial 75 out of 176

Figure 3.3: Experiment interface of original condition regarding three-dimensionality. Each time a new stimulus was shown, the red outline flashed twice to denote the area of interest. As a reminder, participants could move the cursor to the image area to show the red outline overlay. On the right side, participants moved the cursor along the rating scale to adjust the rating, and click to confirm and proceed to the next trial. Gerard ter Borch (II), *Gallant Conversation (Known as 'The Paternal Admonition')*, 1654. Downloaded from the online repository of the Rijksmuseum, Amsterdam. Cropped to the same framing as the engraving reproduction.

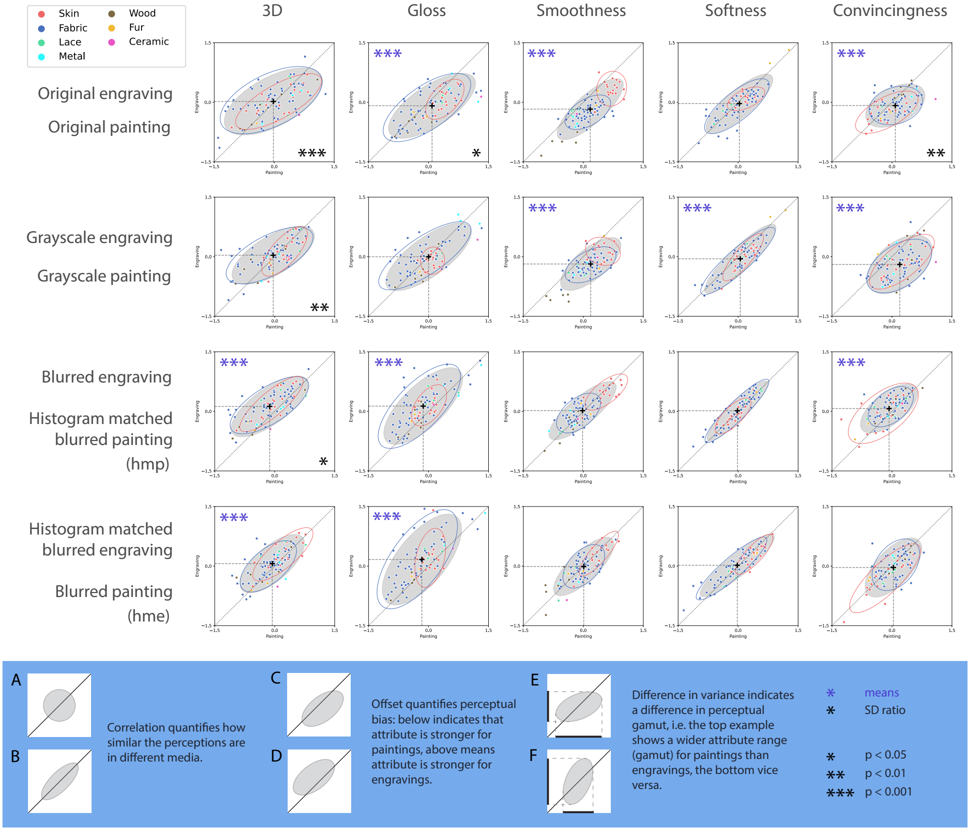
participants with different internal scales, and reduce noise. For further analysis, we always used the mean score across all the participants for each material selection.

3.3. RESULTS

The overall results are summarized in Figure 3.4. Each subplot presents the results of one experimental session with each data point denoting the mean ratings (z-score) of a given material selection. The x-coordinates denote painting ratings, the y-coordinates denote engraving ratings. These scatter plots allow for various qualitative inferences that can be made by the eye, but do require some prior intuitions that we will try to provide before discussing the data in more detail. A visual explanation is also given at the bottom of Figure 3.4.

The scatter data is summarized by covariance ellipses. The gray ellipse denotes all data, the red and blue ellipse denote the subsets of skin and fabric, respectively. The position of the ellipse with respect to the diagonal denotes a perceptual bias: A point above the diagonal line implies that engravings were rated higher than oil paintings and vice versa. An example where this is robustly present is the smoothness data in the original condition: almost all data points are clearly below the diagonal indicating that participants judged materials in paintings to be smoother than in engravings.

A second characteristic, besides position denoting the perceptual bias, is the correlation itself. For example, it can easily be seen that the correlation between painting and engraving is higher for softness than for convincingness. High correlations suggest that



Correlation quantifies how similar the perceptions are in different media.

Offset quantifies perceptual bias: below indicates that attribute is stronger for paintings, above means attribute is stronger for engravings.

Difference in variance indicates a difference in perceptual gamut, i.e. the top example shows a wider attribute range (gamut) for paintings than engravings, the bottom vice versa.

Figure 3.4: Results overview. Each subplot is an experimental session. Row one to row four represent original, grayscale and two histogram matched conditions, respectively. Each data point in these scatter plots represents the mean rating of engravings as a function of the mean rating of oil paintings of the same material selection. Each data point is color coded with respect to material category, as indicated in the legend on the top left corner. The ellipses are confidence ellipses from bivariate normal distributions kept constant at 1.96 standard deviation. The gray ellipses are based on all data, the blue and red ellipses denote fabric and skin, the two largest material categories. The legend on the bottom with the blue background illustrates a few possible scenarios. Purple asterisks on the top left corner in a given subplot indicates that the mean ratings for oil paintings and engravings were significantly different for that session. Black asterisks on the bottom right corner in a given subplot indicates that the standard deviations for oil paintings and engravings were significantly different from each other for that session.

the individual ratings judgements are preserved across medium change: something soft in a painting is also perceived soft in the engraving, which is less so for convincingness. Therefore, this correlation points to a perceptual constancy with respect to medium.

A third quality of the ellipses is the slope, which indicates whether the range of judgements is different for the two media. If the slope is smaller than one, the perceptual range in the paintings is larger than that for the engravings, which can for example be observed for three-dimensionality judgements in the original (top) condition.

In the following section we will statistically verify the qualitative observations that we just made. The section after that is devoted to differences between the materials skin and fabric.

3.3.1. OIL PAINTINGS VERSUS ENGRAVINGS

Looking at the overall data in Figure 3.4, it can be seen that the major axis of the fitted ellipses always points in the positive direction. This is in line with the finding that all correlations are positive and significant (p < 0.001) ranging from 0.45 to 0.90 (with a mean of 0.71). The correlation coefficients are shown in Table 3.3.

The ratio of the standard deviation of engravings and that of oil painting varies between 0.69 and 1.14 (with a mean of 0.71). The ratio is smaller than 1 for 17 out of 20 ratios, suggesting that in most cases, the standard deviation for the oil paintings is larger than that for engravings. Levene's test shows that only 5 ratios are significant with ratios varying between 0.69 and 0.79: original 3D (p < 0.001), original gloss (p < 0.05), original convincingness (p < 0.01), grayscale 3D (p < 0.01) and hmp 3D (p < 0.05). In the above significant cases, oil paintings have a broader range of perceived attributes than engravings. These sessions are marked with black asterisks on the bottom right corner of the corresponding plots in Figure 3.4. Note that there is a tendency for this ratio to increase towards one from the first row (original condition) to the last row (hmp). This is particularly visible in the 3D column.

The means of the oil painting and engraving ratings determine the centroid of the ellipses. In Figure 3.4, the black plus signs indicate the position of the centroids of gray ellipses (all data). The corresponding values can be found in Table 3.3. To test for significance, we performed 20 paired t-tests for unequal variances. To compensate the increased chance of Type I error from multiple t-tests, we applied Bonferroni correction, and set the critical α value at 0.05/20 = 0.0025. For the original condition (the first row in Figure 3.4), there was no significant difference between paintings and engravings for three-dimensionality and softness. However, oil paintings were rated significantly higher for glossiness, smoothness and convincingness (all with $p < 0.001$).

In Figure 3.5, the mean ratings are shown for all conditions, which essentially presents the streamlined information of Figure 3.4, with less distraction from other statistical properties. On the y-axis only the engraving ratings are shown as the painting ratings are the opposite due to the z-transformation. It can thus be viewed as a relative difference. Even more in this representation, it can be seen that gloss, smoothness and convincingness are all judged significantly higher in paintings than engravings.

After removing the colors resulting in the grayscale condition (second row in figure 3.4, gray data in Figure 3.5), there was no significant change from the original condition for three out of five attributes: three-dimensionality, smoothness and softness. For

Table 3.3: Bivariate probability density functions statistics

	3D			Gloss			Smoothness			Softness			Convincingness		
	mean	SD	corr	mean	SD	corr	mean	SD	corr	mean	SD	corr	mean	SD	corr
ori_P	-0.02	0.61	0.59	0.09	0.54	0.69	0.17	0.41	0.79	0.03	0.42	0.81	0.08	0.37	0.45
ori_E	0.02	0.42		-0.09	0.41		-0.17	0.39		-0.03	0.37		-0.08	0.27	
bw_P	-0.04	0.52	0.66	0.00	0.54	0.73	0.18	0.38	0.70	0.05	0.45	0.89	0.20	0.41	0.49
bw_E	0.04	0.37		0.00	0.43		-0.18	0.33		-0.05	0.44		-0.20	0.37	
hmp_P	-0.13	0.48	0.72	-0.13	0.49	0.71	-0.02	0.42	0.81	-0.02	0.38	0.90	-0.07	0.36	0.64
hmp_E	0.13	0.38		0.13	0.48		0.02	0.37		0.02	0.38		0.07	0.31	
hme_P	-0.06	0.40	0.69	-0.17	0.50	0.66	0.01	0.38	0.76	-0.02	0.47	0.90	0.04	0.36	0.63
hme_E	0.06	0.37		0.17	0.57		-0.01	0.37		0.02	0.40		-0.04	0.37	

All correlations have p-value lower than 0.001.
P indicates oil painting, E indicates engravings.
The significance of the t-tests is indicated by the purple stars in Figure 3.4.

glossiness, engravings were rated significantly higher after removing colors; for convincingness, engravings were rated significantly lower, both are marked with light gray asterisk signs in Figure 3.5.

When we applied blurring and luminance histogram matching, the differences between paintings and engravings changed rather substantially. Three-dimensionality was larger for engravings than paintings, while in the original and grayscale versions there was no significant difference between the two media. Glossiness was also larger for engravings than paintings while the reverse was true for the original condition. The differences in smoothness vanished, which also holds for softness although in the original condition there already was no difference. Lastly, the convincingness was significantly higher for engravings than paintings in one condition (hmp), and non significant in the other histogram matched condition (hme), while in the original and grayscale condition the paintings were judged as more convincing.

3.3.2. COMPARISON BETWEEN MATERIAL CATEGORIES

To investigate possible differences between materials, we compared the results for fabric and skin. These materials were best represented with 54 and 18 elements, respectively. We mainly focus on qualitative observations about the bivariate normal distributions, denoted in figure 3.4 by the red and blue confidence ellipses for skin and fabric, respectively. Looking at the red (skin) and blue (fabric) ellipsoids, we observe various configurations: overlapping (some position and size), enclosing (one smaller and withing area of other) or complementary (inhibiting different areas). Note that these three possibilities also hold for a uni-dimensional representation of the data, i.e. on one of the axes. The first row of figure 3.4 shows the data for the 'original' condition and illustrates the three qualitative configurations well: three-dimensionality and (to a lesser extent) convincingness show overlapping data, glossiness and softness show encapsulating configurations and smoothness shows a complementary configuration with the ellipse for skin systematically above that for fabric. The interpretation is relatively straightforward and will be presented in the discussion section.

3.4. DISCUSSION

As Figure 3.4 and Table 3.3 show, all conditions and attributes show positive correlations, indicating that oil painting and engraving media elicit similar perceptions for these five attributes. In other words, engravers did an excellent job to replicate the oil paintings and provoke similar perceptions for the five attributes we tested, although engraving is a challenging medium with only monochromatic lines and dots. This finding is in line with the conclusions from Delanoy et al. (2021) and Van Zuijlen et al. (2020) that different media provoked similar material perception. This 'perceptual constancy over medium' for both our study and Van Zuijlen et al. (2020) could be partially driven by semantic information, as Fleming et al. (2013) has shown: relatively similar perceptual spaces were found for mere material classes defined by their word as by their photographic representations. Yet, the role of semantics vanishes when trying to explain the variance within material categories, such as fabric or skin.

The differences between oil paintings and engravings are mainly found in their means

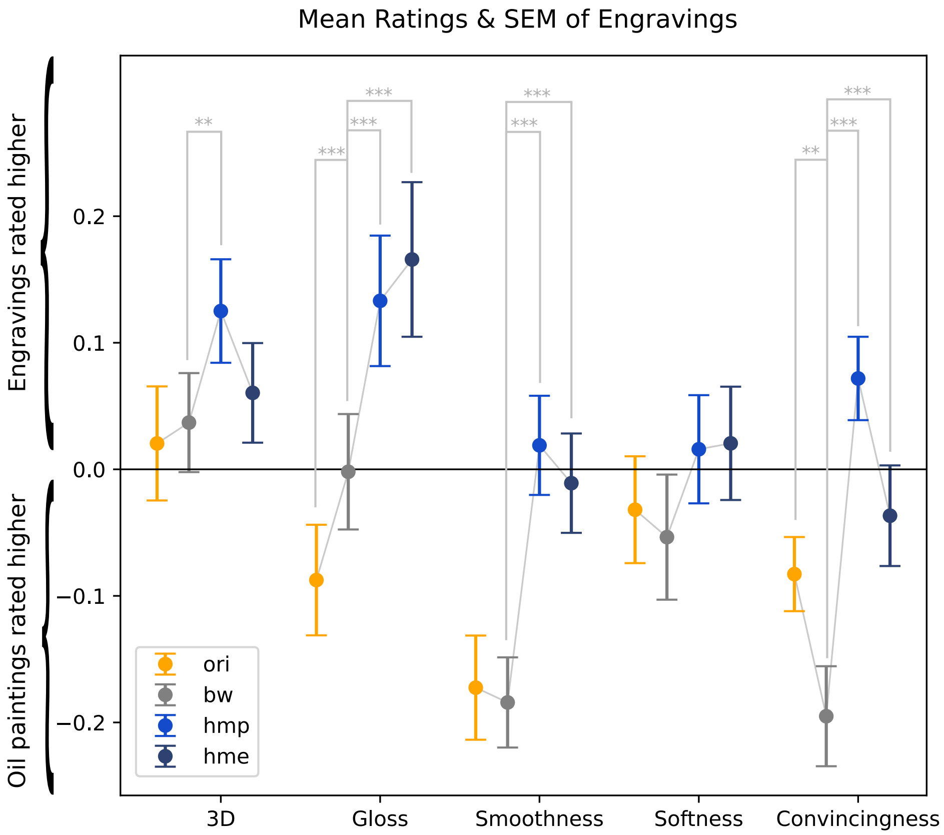


Figure 3.5: Mean ratings and standard error of means of engravings. It can be seen as a streamlined visualization of Figure 3.4, showing the overall trend for the engravings. Positive values indicate engravings were rated higher, negative values indicate oil paintings were rated higher. Since we used z-score data, the means of oil paintings always equal to the negative means of engravings, as shown in Table 3.3, hence we only plotted engravings. The lines for oil paintings and engravings would be symmetrical about the x-axis. The light gray asterisk signs indicate significance of differences between conditions: ** $p < 0.01$; *** $p < 0.001$. For clarity, only the significance between original and grayscale is indicated, as well as the significance between grayscale and the two histogram matched ones (hmp and hme).

and standard deviations. Different means indicate different levels of perceived attributes. Different standard deviations indicate different perceived range of certain attributes. In the original condition the oil paintings always show a broader range of perceived attributes, regardless of the significance of variance differences. To be precise, in almost all sessions (17 out of 20) oil paintings have broader perceptual gamut than engravings. Bousseau et al. (2013) found that the range of perceived gloss in painterly renderings is narrower than that in realistic renderings. Our current study shows that the perceived ranges of three-dimensionality, gloss and convincingness in engravings are significantly narrower than those in oil paint. For the original pictures, three out of five attributes showed a smaller range for engravings, but after removing chromatic information (color), only three-dimensionality showed this difference between painting and engraving, and differences vanish completely in one of the two histogram matched conditions. It should furthermore be noted that, although not significant, for Gloss the perceptual range of engravings seems to trump that of paintings in the case of histogram matching.

3.4.1. Comparisons in the original condition

The first row in Figure 3.4 shows the comparison between oil paintings and engravings for the original condition. For glossiness, smoothness and convincingness, representations in oil paintings were rated significantly higher, meaning materials in oil paintings were perceived as glossier, smoother and more convincing. The difference in convincingness is to be expected: the combination of colorlessness and the visibility of hatching lines likely lack the convincingness found in oil paintings. Less expected is that convincingness showed a larger perceptual range in paintings. This finding is less straightforward to explain than the larger perceptual range for three-dimensionality and gloss (discussed in more detail in the next paragraphs). As gloss and three-dimensionality vary in reality, it makes sense to depict these variations and the painting medium apparently affords depiction of a larger variety of the pictorial attributes than engraving. However, convincingness is not an attribute of a pictorial object but rather an overall quality of the depiction itself. Convincingness does not vary in reality, as reality itself is an ultimate aim achieved through convincingness. There does not seem a need or motivation for a larger convincingness range in paintings than in engravings. Therefore, this range difference may reflect that differences in style may be larger within paintings than engravings, which would be an interesting observation. This would imply that in copying a painting into an engraving, idiosyncratic style elements are lost and depictions converge towards a more homogeneous ‘engraving style’. It seems feasible to investigate this conjecture empirically, although it is beyond the scope of the current study.

The difference in mean ratings for smoothness is rather large. One possible explanation is that in the original condition, the brushstrokes in oil paintings were fine and not very visible, while engravings have visible engraving lines (see an example in Figure 3.6). As discussed in the introduction, we previously found an interaction between the smoothness of the medium (visible brushstrokes) and pictorial smoothness (of the motif) in a study on apple depictions (Zhao et al., 2023). Although we specifically instructed the participants to rate the smoothness of the depicted material, it could be a similar case of observers unable to discount for the smoothness of the medium while judging

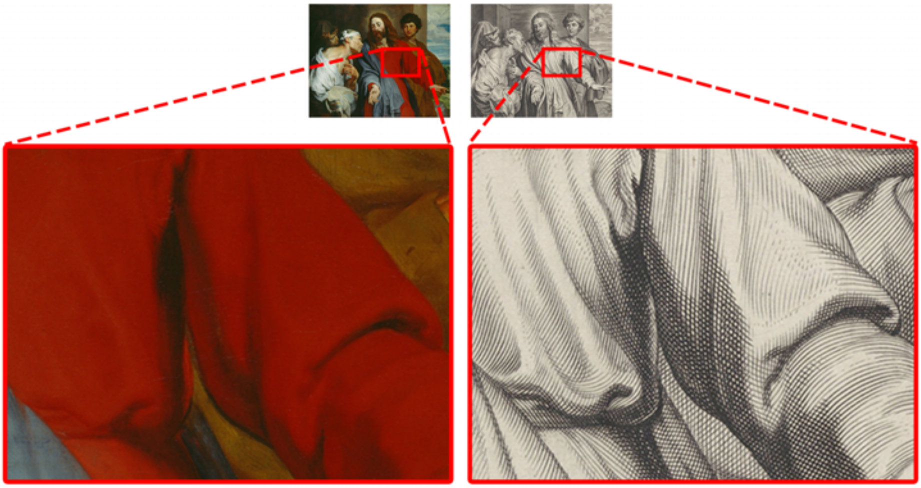


Figure 3.6: A zoomed in look at details. The engraving on the right shows visible engraving lines. Oil painting on the left: Anthony van Dyck, *Christ healing the paralytic*, 1619. Engraving on the right: Pieter de Jode (II), *Christ healing the paralytic*, 1628 - 1670.

the smoothness of the motif.

The difference in perceived gloss is more challenging to explain. Indeed, the painter possesses more control over the gloss parameters, especially being able to vary the amount of blur at the edge of highlights. That would not explain an overall higher gloss ratings for paintings, but it could contribute to the larger perceptual range as found by comparing the variances (indicated by the black asterisk in Figure 3.4). If we observe the material specific categories (blue ellipse for fabric, red ellipse for skin), we observe that skin dominates the gloss bias. Apparently, painted skin appears more glossy than engraved skin. A look at the skin fragments in Figure 3.7 may suggest a possible explanation. While both engraving and painting make use of tonal differences to articulate shape and material, it seems easier to disentangle the specular reflections from the shading patterns in paintings than in the engravings.

As for three-dimensionality and softness, we did not find differences in mean ratings between paintings and engravings. However, we did find a larger perceptual range for three-dimensionality in paintings. To understand the three-dimensionality range difference we show some stimuli that seem responsible for this effect in Figure 3.8. The stimuli that elicited low three-dimensionality ratings for paintings in comparison to engravings (left rectangle in the figure) all seem to show objects that were painted without contrast, rather homogeneous without much shading detail in comparison to their engraved counterparts. The stimuli in the right rectangle should show the opposite effect, i.e. very three-dimensional in paintings and less so for engravings. Indeed, the paintings show well articulated shading patterns, especially in comparison to the paintings with low three-dimensionality. However, the engravings for this second group of stimuli look quite similar to the paintings; they also show shading articulation. If anything,

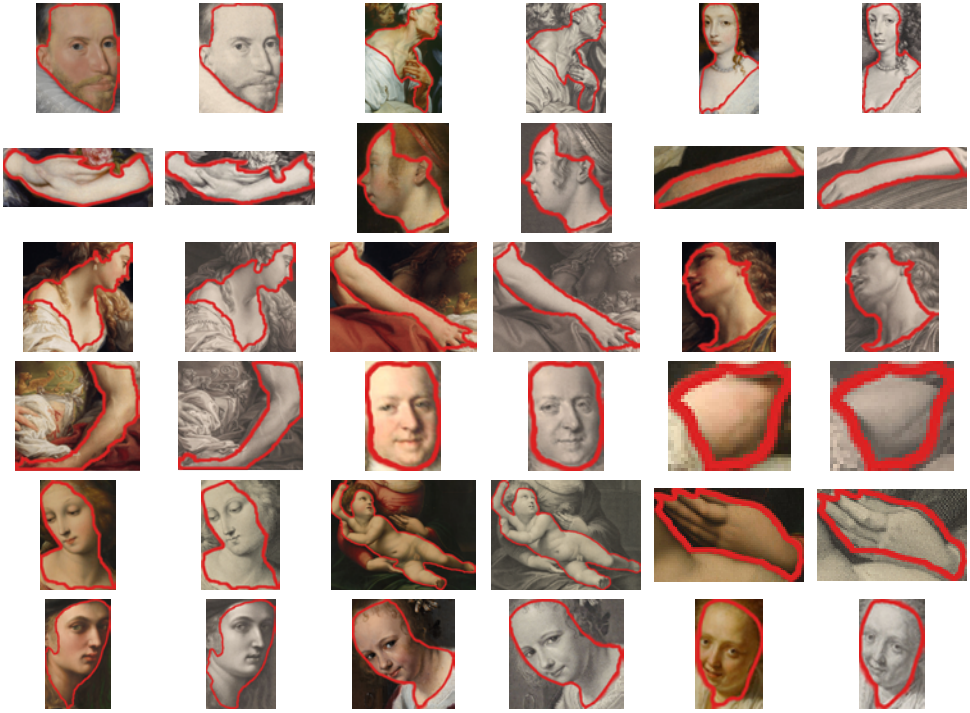


Figure 3.7: Examples of skin fragments from paintings (on the left) and engravings (on the right).

the paintings seem to include both shading and (cast) shadowing while engravings seem mostly involved with shading patterns. In sum, when looking at individual stimuli we can indeed see a relatively large range in three-dimensionality for paintings and a much shallower range (more similar) for engravings.

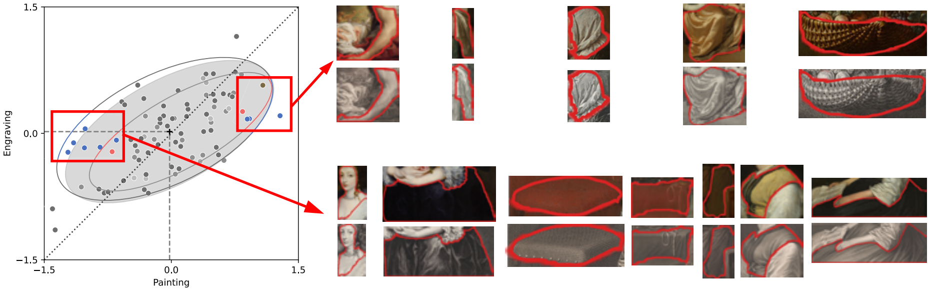


Figure 3.8: Some examples from the three-dimensionality ratings in the 'original' condition that illustrate a potential cause for the difference in perceptual range between paintings and engravings. From the left rectangle: some objects in paintings with low three-dimensionality ratings hardly show tonal contrast while their engraved equivalents do. From the right rectangle: some objects that show similar level of shading and detail.

3.4.2. Effect of color removal & histogram equalization

We manipulated the images to reduce the two most prominent differences between paintings and engravings: chromatic information and the luminance histogram. The color manipulation was performed for the obvious reason that engravings lack color information. The rationale behind the luminance histogram equalization was to reduce the difference in global luminance statistics (for the whole image) such as mean luminance, contrast (as quantified by the variance) and skewness.

By only removing colors, the evoked perceptions of the two media did not change much compared to the original condition (the second and first rows in Figure 3.4, or the gray and orange points in Figure 3.5). This suggests that color did not affect perception much. It was the blurring and histogram equalization that had a more substantial overall impact. We will now discuss the results in more detail.

A somewhat surprising result is that the difference between paintings and engravings in convincingness was enhanced instead of mitigated when removing chromatic information. Many facets can underlie the perception of convincingness. In the computer science literature, the closest equivalent to 'convincingness' is 'realism' and Rademacher et al. (2001) found that shadow sharpness and surface texture visibility contribute significantly towards the perception of realism, both in photos and renderings. Although there is interesting literature comparing realism across various art styles, Hagen (1986) mainly focuses on the depiction of pictorial space and various types of perspective. An extension towards computer rendering (Ferwerda, 2003) offers three varieties of realism: physical, photo(metric) and functional realism. While broadening the scope towards other formal elements than pictorial space, the categorisation seems too coarse to offer an explanation for our finding. One plausible speculation could be that in the original condition, the styles are so far apart that each is judged on its own merit but as differences become smaller, the two media are more directly compared by the observers. Again, this is mere speculation in need of further empirical evidence. What is certain is that when we removed differences in luminance histograms, convincingness differences vanished and for half of the data even reversed: when histograms were matched to the painting the engravings were judged to be more convincing. Initially, this manipulation aimed at histogram matching to equalize the luminance characteristics, such as mean, variance (i.e., contrast), and skewness. However, a side effect was the necessity to blur the engravings in order to compute a continuous histogram. In hindsight, the blurring alone would have merited an independent manipulation as in the case of convincingness the effect may well have depended on the visibility of engraved lines.

For three-dimensionality, the color removal did not cause much difference: the larger perceptual range persists for paintings and the mean three-dimensionality ratings are again not significantly different between paintings and engravings. However, when applying the luminance histogram equalization, we see that the perceptual range difference vanishes for half of the data (the histograms matched to the engravings). This could potentially be due to contrast equalization. As we showed in Figure 3.8, this seemed a potential difference between paintings and engravings. Moreover, we found a significant difference in mean three-dimensionality ratings. Given that chromatic information and the (global) luminance distributions are similar between the paintings and engravings, these rather robust findings are likely due to local contrast, i.e. the detailed shading contrast on certain objects seems to be stronger in engravings than paintings.

A similar shift in mean ratings was found for gloss perception. While in the original condition glossiness ratings were higher in paintings than engravings, removing color caused this difference to vanish and luminance histogram equalization even reversed the effect: engravings are perceived to be more glossy. In the original condition we conjectured that the bias towards paintings could be attributed to skin, as illustrated in Figure 3.7. The removal of color did not seem to change much about the position of the red ellipse (denoting the skin samples) with respect to the diagonal although the position itself shifted downwards. Yet, the vanishing of the mean gloss difference in the grayscale condition seems to be due to fabrics samples (engravings show higher gloss) counterbalancing the skin samples (paintings show higher gloss). With the removal of luminance histogram differences the engravings robustly received higher ratings. We believe that this bias is also due to local contrast, as shown in Figure 3.9. The effect seems similar to the three-dimensionality data, although the underlying mechanism differs: for gloss the contrast between highlight and background is an important cue (Marlow and Anderson, 2013) while for three-dimensionality contrast in general likely plays a role. While a change in contrast can theoretically be attributed to either a change in light direction or to a change in depth (Belhumeur et al., 1999), it has been shown that participants often attribute it to shape: Ho et al. (2006) tested surface roughness with a rather coarse texture stimulus and found that increasing contrast by lowering the light direction was attributed to the roughness, i.e. depth variation as the texture was rather coarse.

For smoothness, in both original and grayscale conditions, oil paintings received higher ratings. After blurring and histogram matching, the performance of these two media became very similar. One possible explanation is that in both original and grayscale

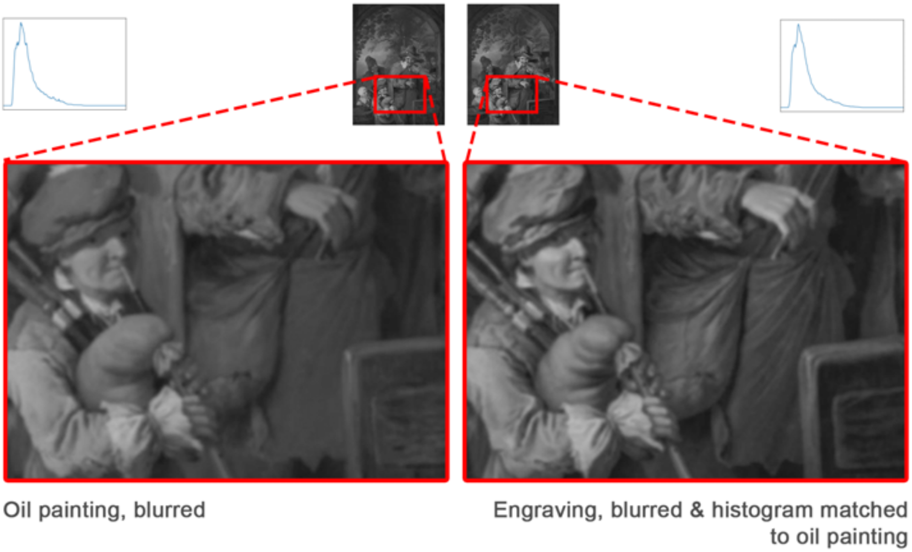


Figure 3.9: A zoomed in look at details on blurred oil painting and blurred engraving with histogram matched to oil painting. The engraving on the right shows higher local contrast than oil paintings, although they share the same histogram. Oil painting on the left: Christian Wilhelm Ernst Dietrich, *The Wandering Musicians*, 1745, from The National Gallery, London. Engraving on the right: Johann Georg Wille, *The Wandering Musicians*, 1764, from Rijksmuseum, Amsterdam.

conditions, the visible lines led to less perceived smoothness for engravings. After blurring, the engraving lines became invisible, resulting in similar smoothness ratings between these two media. As mentioned earlier, in a previous study about style perception (Zhao et al., 2023), we found a potential transfer between the smoothness of depicted apple skin and brushstroke coarseness of the medium. Although we cannot dissociate whether smoothness perception similarity relies on blurring or histogram equalization, we hypothesis that it is indeed due to the vanishing engraving lines. This would imply that again we found a transfer of smoothness/roughness from medium to depicted objects/materials.

Softness was the only attribute in the original condition that neither showed a significant difference between the means nor the variances of paintings and engravings. This changed when we removed color information: objects were perceived softer in paintings than engravings. It is tempting to believe that smoothness and softness are correlated and that the solution of the softness bias towards paintings finds it origin in the smoothness discussion from the previous paragraph. Yet it can be readily inferred that smoothness and softness are judged differently by observers: for smoothness the skin and fabric samples are clearly segregated while for softness there is much overlap. This leaves us with the open question of why painted objects are perceived softer than engraved objects when color is removed. The second manipulation (histogram equalization) let the softness bias disappear again, which could either mean that global contrast or hatching visibility contributed to the bias we found in the achromatic condition. What is furthermore interesting to note is that for softness the correlations were all rather high: in the original condition about 0.8 and in all manipulated conditions about 0.9, as can be read in table 3.3 and also observed in Figure 3.4. These values are all substantially higher than for the other attributes. This implies that the softness of materials is the most medium-invariant attribute.

3.4.3. Comparison between material categories

As we showed in Section 3.3.2, the two material categories of skin and fabric have different configurations. For three-dimensionality and convincingness, they have an overlapping configuration, indicating similar perceptual ranges. For gloss and softness, they show an enclosing configuration. Skin has a lower glossiness and softness range than fabric. A possible explanation is that fabric is a more diverse material category than skin. It can vary from matte cotton to glossy satin, or from heavy stiff damask to soft silk. Skin, on the other hand is much more consistent. For smoothness, skin has overall higher values than fabric. The possible explanation is that skin is in general smooth, while fabric is in general less smooth than skin, and can vary in terms of smoothness.

Additionally, for each attribute the configurations of these two material categories demonstrate similar trends across the manipulations. This suggests fabric and skin have similarly been affected by the color and luminance manipulations.

3.4.4. Conclusion

We investigated the perceptual influence of media by measuring judgements about materials, shape and the pictorial quality (convincingness). We choose to compare engravings and paintings as they are both famous art media and because of the engraved copies of paintings we could study a similar pictorial scene differently depicted in the respective media. Furthermore, paintings and engravings span an important historical style axis as defined by Heinrich Wölfling who in his “Principles of Art History" (Wölfflin, 2012) defined the first dimension of style and form that between ‘linear’ and ‘painterly’.

Our overarching interest is how engravers handled the limited boundary conditions of their medium. How to cope with the lack of color and the binary nature of tonal variations? Indeed, when directly compared to paintings, engravings lack convincingness. But this difference vanishes when the boundary conditions are equalized for the media. Moreover, gloss and three-dimensionality judgements are higher for engravings than for paintings in the equalized conditions, and for softness and smoothness perceptual differences vanish. We have hypothesized that engravers show a stronger local articulation of the shading details, which likely compensated, or was meant as compensation, for the lack of color and smooth transitions afforded by oil paint. A more detailed study on what types of pictorial ingredients engravers use to convey material properties would be highly desirable. Our study has generated a number of other interesting follow up questions. First, we found more evidence for the interaction between medium and motif, in our case for smoothness perception. Second, as three-dimensionality relies on both shading and shadowing, the clear visibility of these two is necessary for an optimal three-dimensionality percept. For engravings, however, the discernibility between shading and shadowing seems to be limited. Thirdly, a difference in the depiction of skin became apparent where there again seemed to be dissociation difficulties for engravings, this time between shading and highlight. Fourthly, although this may be more art-historically interesting: what is the role of paint degradation when comparing engravings and paintings, particularly the local shading patterns. It seemed that some parts of the paintings were rather dully shaded while their engraved counterparts were highly articulated. Was this the engraving compensating as just discussed, or was the original painting equally articulated? A future study could investigate whether some of our paintings did in fact degrade over time, although this may require some technical art history effort.

In conclusion, engravings can render materials and shapes well and elicit similar perceptions as oil paintings. Nevertheless, there were some differences in performance for portraying certain attributes, as well as differences in perceptual range, which has resulted in interesting new research leads. In addition, we showed the role of color and luminance distribution via manipulations of color removal, blurring and histogram equalization. The manipulations close the gap between them. In some case, engravings even show advantages over oil paintings.

Bibliography

Adobe Inc. (2021). Adobe photoshop. https://www.adobe.com/products/photoshop.html

Alberti, L. B. (1966). *On painting: Revised edition* (Vol. 175). Yale University Press.

Ashe, T. (2014). *Color management & quality output: Working with color from camera to display to print*. CRC Press.

Belhumeur, P. N., Kriegman, D. J., & Yuille, A. L. (1999). The bas-relief ambiguity. *International journal of computer vision*, *35*(1), 33–44.

Berlyne, D. E., & Ogilvie, J. C. (1974). Dimensions of perception of paintings. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 181–22). Hemisphere.

Bol, M. (2023). The varnish and the glaze: Painting splendor with oil, 1100–1500. In *The varnish and the glaze*. University of Chicago Press.

Bousseau, A., O'shea, J. P., Durand, F., Ramamoorthi, R., & Agrawala, M. (2013). Gloss perception in painterly and cartoon rendering. *ACM Transactions on Graphics (TOG)*, *32*(2), 1–13.

Chatterjee, A., Widick, P., Sternschein, R., Smith, W. B., & Bromberger, B. (2010). The assessment of art attributes. *Empirical Studies of the Arts*, *28*(2), 207–222.

Delanoy, J., Serrano, A., Masia, B., & Gutierrez, D. (2021). Perception of material appearance: A comparison between painted and rendered images. *Journal of Vision*, *21*(5), 16–16.

Di Cicco, F. (2022). The legacy of willem beurs–bridging the gap between art and material perception. *Art & Perception*, *10*(2), 111–136.

Di Cicco, F., van Zuijlen, M. J., Wijntjes, M. W., & Pont, S. C. (2021). Soft like velvet and shiny like satin: Perceptual material signatures of fabrics depicted in 17th century paintings. *Journal of vision*, *21*(5), 10–10.

Di Cicco, F., Wiersma, L., Wijntjes, M., Dik, J., Stumpel, J., & Pont, S. (2018). A digital tool to understand the pictorial procedures of 17th century realism. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

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), 1–15.

Ferwerda, J. A. (2003). Three varieties of realism in computer graphics. *Human vision and electronic imaging viii*, *5007*, 290–297.

Fleming, R. W., Torralba, A., & Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of vision*, *4*(9), 10–10.

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of vision*, *13*(8), 9–9.

Göring, S., Rao, R. R. R., Merten, R., & Raake, A. (2023). Analysis of appeal for realistic ai-generated photos. *IEEE Access*.

Hagen, M. A. (1986). *Varieties of realism: Geometries of representational art*. CUP Archive.

Ho, Y.-X., Landy, M. S., & Maloney, L. T. (2006). How direction of illumination affects visually perceived surface roughness. *Journal of vision*, *6*(5), 8–8.

Marlow, P. J., & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of vision*, *13*(14), 2–2.

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), 1909–1913.

Norman, J. F., Todd, J. T., & Orban, G. A. (2004). Perception of three-dimensional shape from specular highlights, deformations of shading, and other types of visual information. *Psychological Science*, *15*(8), 565–570.

O'Hare, D., & Gordon, I. (1977). Dimensions of the perception of art: Verbal scales and similarity judgements. *Scandinavian Journal of Psychology*, *18*(1), 66–70.

O'Shea, R. P., Blackburn, S. G., & Ono, H. (1994). Contrast as a depth cue. *Vision research*, *34*(12), 1595–1604.

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.

Rademacher, P., Lengyel, J., Cutrell, E., & Whitted, T. (2001). Measuring the perception of visual realism in images. *Rendering Techniques 2001: Proceedings of the Eurographics Workshop in London, United Kingdom, June 25–27, 2001 12*, 235–247.

Thompson, W. (2000). The printed image in the west: Engraving. *Heilbrunn Timeline of Art History. New York: The Metropolitan Museum of Art*.

Todd, J. T., & Mingolla, E. (1983). Perception of surface curvature and direction of illumination from patterns of shading. *Journal of Experimental Psychology: Human perception and performance*, *9*(4), 583.

Van Zuijlen, M. J., Pont, S. C., & Wijntjes, M. W. (2020). Painterly depiction of material properties. *Journal of vision*, *20*(7), 1–17.

van Assen, J. J. R., Wijntjes, M. W., & Pont, S. C. (2016). Highlight shapes and perception of gloss for real and photographed objects. *Journal of Vision*, *16*(6), 6–6.

Wendt, G., Faul, F., & Mausfeld, R. (2008). Highlight disparity contributes to the authenticity and strength of perceived glossiness. *Journal of Vision*, *8*(1), 14–14.

Wölfflin, H. (2012). *Principles of art history*. Courier Corporation.

Wolfram Research Inc. (2020, December 16). Mathematica, Version 13.0. https://www.wolfram.com/mathematica

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), 11–11.

Zhao, Y., Stumpel, J., de Ridder, H., & Wijntjes, M. W. (2023). Zooming in on style: Exploring style perception using details of paintings. *Journal of vision*, *23*(6), 2–2.

4

PROMPTS AND APPEARANCES: COMPARING PHYSICALLY BASED RENDERINGS AND GENERATIVE AI IMAGES THROUGH MATERIAL PERCEPTION

Generative Artificial Intelligence (AI) models unlock new ways to create images, which are distinct from physically based rendering engines creating 2D images from 3D environments. In two experiments, we chose human material perception to compare the perceptual similarity embeddings of three generative AI models with that of a computer-generated BRDF stimulus set.

In Experiment 1, we used the text descriptions of 32 materials (e.g., blue acrylic) from MERL, a BRDF dataset, as prompts for DALL-E 2 and Midjouryney v2, two text-to-image models, to generate 32 images of spheres with comparable materials. We collected human similarity judgements for each data set and then constructed perceptual spaces for all three sets via Soft Ordinal Embedding. Both AI models resulted in a 2D space while the MERL set was confined to 1D, probably due to lack of surface texture. The perceptual spaces were found to be unrelated, suggesting that the AI models generated unique and different images of materials from identical text prompts.

In Experiment 2, the open-source text-to-image AI model Stable Diffusion v1.5 was combined with ControlNet allowing the additional constraints of depth maps. We kept the same 32 material descriptions from MERL and generated three sets using three different shapes as depth maps. The three perceptual spaces from Experiment 2 are all 2D and

Submitted

exhibit high similarity, indicating a robust and non-random structure. They also show a similar structure as the MERL embedding and perceptual spaces from other material studies using real-world photos, computer renderings and depictions.

4.1. Introduction

We are surrounded by a large variety of materials signaling various properties, for example, physical properties such as hardness, roughness or viscosity (Fleming, 2017). While the 'natural' environment already contains a large variety of materials, the contemporary (build) environment also includes an increasing number of manufactured materials. This apparent material complexity is an interesting topic for the study of visual perception as humans likely reduce this complexity by grouping materials into categories, enabling them to estimate properties such as mentioned above over a large variety of materials (Fleming, 2017; Schmidt, 2019). Moreover, we encounter an ever-increasing number of *images* of materials, such as a photo of a glass building facade on a phone screen, a computer-rendered rock in a game on a laptop, or a painting depicting ocean water.

There are various approaches to understand human material perception. Most studies use images of materials instead of the actual physical objects. Having control over physical characterizations of materials has been the norm in material perception studies over the past decades, either using photos in combination with physical measurements or using physics-informed computer renderings. Images may have different properties, characters or styles, both among themselves, and compared to material perception in actual environments. Therefore, it may be important to probe possible differences between different varieties of generated images of materials. Over the past decades, computer-generated imagery (CGI) has become a dominant technique in the movie and gaming industry and rendering innovations have been developed in tandem with insights from perception research (Khan et al., 2006; Thompson et al., 2011; Vangorp, 2009). When studying specific material properties, the complex relationship between visual cues and material perception often requires restricting the study to a single material category, or sometimes even maintaining a constant object shape. For example, to understand gloss, Wills et al. (2009) used computer rendered bunnies with different bidirectional reflectance distribution functions (BRDFs), while Ferwerda et al. (2001) used rendered spheres as stimuli. These studies used computer renderings as they afford precise control over the various distal (or 'world') parameters that define materials, such as reflectance characteristics.

At the same time, it appears possible to investigate cue-perception relations in uncontrolled stimuli as shown in the glossiness study by Di Cicco et al. (2019). They investigated painterly practice and defined cue intensities by measuring various image properties such as contrast and blur. Using art images instead of renderings has the added value that a certain pictorial approach of the maker is automatically incorporated which may reveal additional insights into mechanisms of material perception. Furthermore, paintings are made on surfaces such as canvas, not in a physics rendering engine. As a result, painters are not limited by the rigidity of physically based rendering algorithms (Cavanagh, 2005).

As anyone living in the time of our study must have noticed, a new medium has become available: generative AI. Although synthetic textures have already existed for several decades (Efros & Leung, 1999; Heeger & Bergen, 1995; Portilla & Simoncelli, 2000), deep neural networks revolutionized the production of images with the invention of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Instead of being trained to only recognize and classify depicted objects such as in AlexNet (Krizhevsky et al., 2012), this new type of network made it possible to expand the kinds of classifications, for example, classifying whether an image is a photo or not. This architecture resulted in rather photorealistic images albeit with a certain uncanniness. Interestingly, this type of 'bug' is often regarded by artists such as Mario Klingeman and Helene Sarin as a positive 'feature' to help creating aesthetically pleasing images (Hertzmann, 2020; Wang et al., 2020). Combining text and images became the next significant innovation, for example by using the CLIP model (Radford et al., 2021). This resulted in various generative image synthesizers based on so-called 'prompts', text describing what (and how) the image should depict. Around 2022, various platforms started their online services of text-based image generation (e.g. DALL-E and Midjourney) or released their model (Stable Diffusion).

Generative AI pictures conceptually resemble paintings as the generation takes place in the picture plane, i.e. the RGB matrix and canvas, respectively. In contrast, for computer rendering and photography there is always a 3D source of which the image is the projection. Generative AI depictions do not originate from distal scene properties and, consequently, cannot be directly linked to physical parameters. However, it is likely that there are latent space correlates for various visual phenomena. (Goetschalckx et al., 2019; Liao et al., 2023). The absence of a direct link to physical parameters also means that generative AI depictions are not limited by the laws of physics, just like paintings, drawings, etc. This is important as the human visual system is also not bound to the laws of physics but rather uses its own 'alternative physics' (Cavanagh, 2005) to model the outside world.

In this study we want to explore the use of generative AI depictions for material perception. We were particularly interested in the perceptual dimensions that generative AI depictions span. Quantifying perceptual spaces can be used to explore core dimensions in material perception (Schmidt et al., 2022) but also in style perception (Zhao et al., 2023) and in many other fields where an a priori structure is lacking. The traditional method to create a perceptual space is Multidimensional Scaling (MDS) (Mead, 1992) where observers are asked to rate the difference between each pair of stimuli and the resulting scores are transformed into distances in a multidimensional space. To avoid individual scaling differences, various other methods have been developed, for example, the ones that make use of triplets where the observer is asked to select the two most similar stimuli per trial. This ordinal information can then be processed with, for example, (landmark) MDS (De Silva & Tenenbaum, 2004; Zhao et al., 2023) or specific neural networks (Hebart et al., 2020). These two methods address the challenge of the quickly increasing number of possible triplets—scaling cubically with the number of stimuli—by applying various techniques to reduce the required minimum number of triplets. A relatively new and promising method that seems to require the least data for generating robust perceptual space reconstructions is Soft Ordinal Embedding (SOE). Originated from machine learning, the goal of SOE is to find an embedding (perceptual space) that maximizes the number of consistent triplets (Haghiri et al., 2020; Künstle & von Luxburg, 2024; Künstle et al., 2022; Terada & Luxburg, 2014).

We wanted to compare the new type of text-to-image generated visual stimulus with that from an already established technique of stimulus generation. We choose computer rendering and specifically choose the stimuli by Lagunas et al. (2019) who in turn used the data driven BRDFs of the Mitsubishi Electric Research Laboratories (MERL) dataset (Matusik, 2003) to generate a large dataset. Lagunas et al. (2019) first collected human similarity judgements for a subset of the stimuli and then used these data to train a deep learning model that can measure (predict) the appearance similarity between different materials. One of the shapes they used was a sphere, being one of the very few geometric shapes that is unambiguously captured by text (which is the input for our stimulus generation). A cube would also be possible, but objects of tessellated structure can fail to evoke correct reflectance properties (Vangorp et al., 2007). In our second experiment, however, we explored an alternative technique for generating similar shapes using ControlNet. (Zhang et al., 2023).

4.2. Experiment 1 - Influence of generative AI model

Experiment 1, conducted in 2022, investigated two popular text-to-image generative AI models, DALL-E 2 (Ramesh et al., 2022) and Midjourney v2 (https://www.midjourney.com/). We compared images generated from these two models with a computer graphics rendering dataset by Lagunas et al. (2019) who rendered spheres with various BRDFs from the MERL dataset (Matusik, 2003) under various light probes (Debevec, 2008). For the generative AI models, the only constraint of the output images is the text description (in contrast to the image constraints we used in Experiment 2).

4.2.1. Method

Stimuli

We used three sets of images, each containing 32 comparable materials, an overview is shown in Figure 4.1. The first set, MERL, is a BRDF dataset based on real-world measurements. We chose to include images with different environment maps to ensure diversity and anticipate on a variety of 'lighting' settings in the generative AI stimuli. Six environment maps were used, Uffizi, Grace, Pisa, Ennis, Glacier and Doge (Debevec, 2008), for 10 images, 7 images, 7 images, 4 images, 3 images, and a single image, respectively. Each material comes with a text description (e.g., 'blue acrylic') as specified by the BRDF name in the original MERL dataset. The other two sets were generated with generative AI models, DALL-E 2 and Midjourney v2. Each set contains 32 comparable materials, as we used the BRDF names as prompts to generate the images. To control the shape, we added the word 'sphere' in the text prompt. Examples of prompts are 'a blue acrylic sphere' and 'a chrome sphere'. Note that participants only saw the images of materials, not the text description.

Procedure

Since we are interested in the perceptual spaces from these three image sets, we chose a similarity judgement task. We conducted three online experimental sessions, one for the MERL, one for the DALL-E 2 and one for the Midjourney v2 image sets. Each session

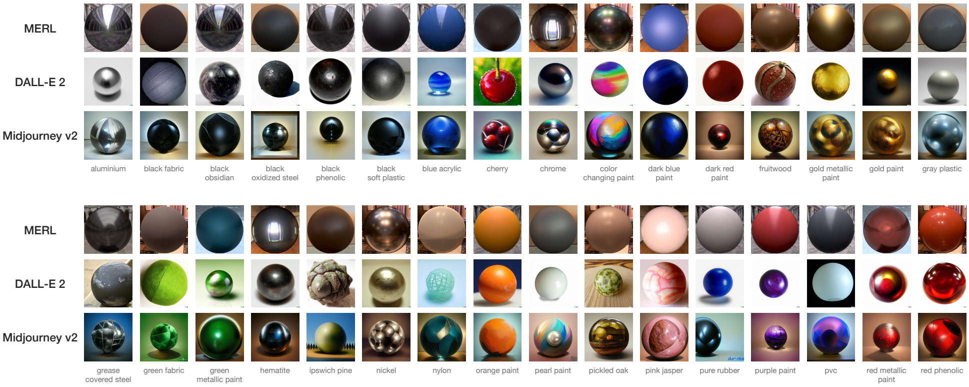


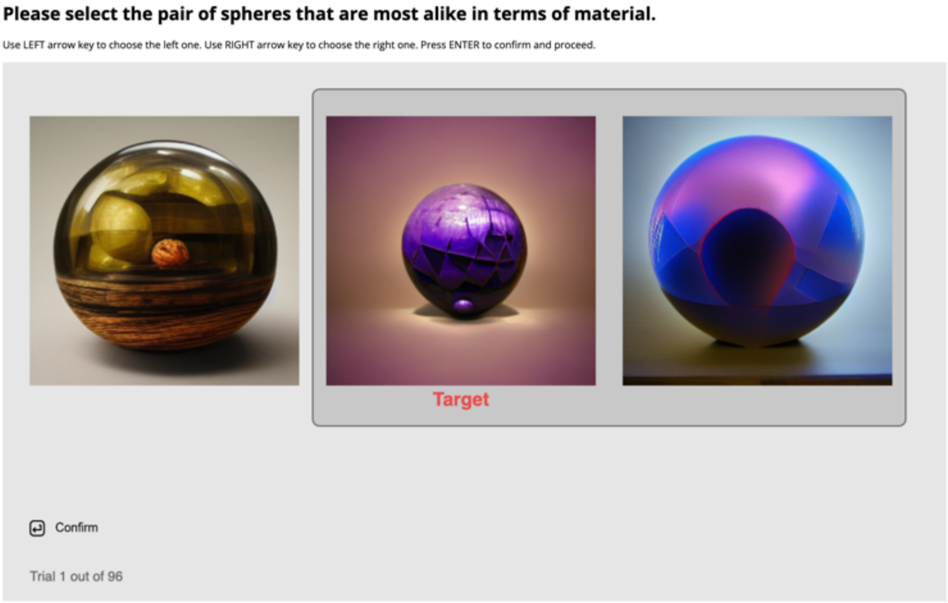
Figure 4.1: An overview of all stimuli used in Experiment 1. The word(s) below each group of images are the text descriptions from the MERL dataset. We used the same descriptions as prompts for the AI models but added the word 'sphere'. (i.e. 'aluminium' becomes 'an aluminium sphere')

contained 96 trials after 15 practice trials for participants to get familiar with the concept and operation. In each trial, participants were presented with a triplet of images, both the selection and order were randomized. The center stimulus was set as target, the task was to select either the left one or the right one as the one most alike the center target[1] in terms of material. Figure 4.2 shows the experiment interface. Participants could use the left and right arrow keys to indicate their choice, then use the 'return' key to both confirm and proceed to the next trial. One benefit of the triplet judgement task (over similarity rating) is the ability to scale up the experiment by combining data across multiple participants, without the issue of different internal scales (Linde, 1975; O'Hare, 1976). This advantage makes triplet method better suited for crowd-sourcing studies (Heikinheimo & Ukkonen, 2013; Li et al., 2021; Tamuz et al., 2011).

Participants

150 unique participants were recruited for Experiment 1, 50 participants for each session. A server issue caused some data loss. Eventually, we recorded 45 (for MERL), 34 (for DALL-E 2) and 39 participants (for Midjourney v2) for three sessions. All participants received compensation regardless of their data being recorded. All participants were recruited from Prolific (www.prolific.com). The following prescreen criteria were used: 1) approval rate 95% - 100%, 2) number of previous submissions 100 - 1000, 3) highest education level completed higher than high school, 4) fluent in English, 5) from the USA or UK, 6) exclude participants from our previous studies. Note that criteria 3 to 5 were used to make sure participants could understand the instructions properly without language barrier. The experiment was conducted in agreement with the Declaration of Helsinki and approved by the Human Research Ethics Committee of the Delft University of Technology. All data were collected anonymously.

[1] We initially were under the impression that the embedding algorithm would only be able to process comparisons with the middle target, instead of choosing the more intuitive 'odd one out'. Later we learned that Soft Ordinal Embedding *was* able to process this type of data, which we then used in Experiment 2



Please select the pair of spheres that are most alike in terms of material.

Use LEFT arrow key to choose the left one. Use RIGHT arrow key to choose the right one. Press ENTER to confirm and proceed.

Trial 1 out of 96

Figure 4.2: Experiment interface of Experiment 1, showing images from Midjourney v2. Participants were shown three images of different materials, where both the selection and order were randomized. Their task was to select either the left image or the right image that is most alike the center one (target). Participants could use the LEFT and RIGHT arrow keys to select their choice by sliding the window, then press RETURN to confirm and proceed to the next trial.

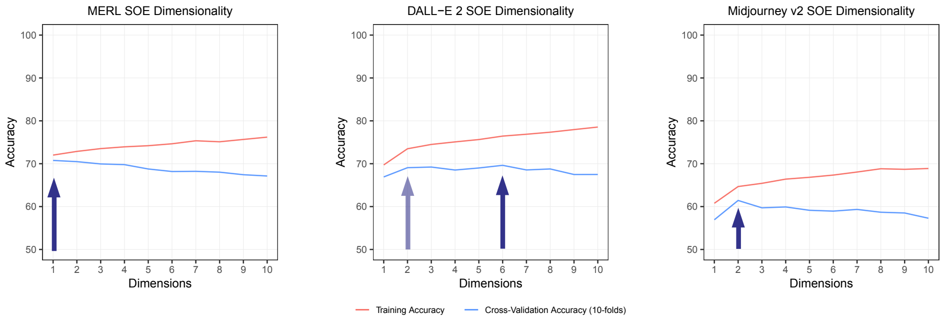


Figure 4.3: The dimensionality of MERL, DALL-E 2 and Midjourney v2 embeddings. The blue cross-validation curves show the overall performance of the fit. The peak of the cross-validation accuracy curve, indicated by the blue arrows, stands for its optimal dimensionality. Error bars are not included because the variance between folds is very small.

DATA ANALYSIS

We used Soft Ordinal Embedding (Haghiri et al., 2020; Künstle et al., 2022) to convert the triplet data into perceptual embeddings. This method does not necessitate the data of all possible triplets, but rather requires only $2dn\log_2 n$, where d is the estimated number of dimension(s) and n is the number of stimuli. Based on previous research, we expected at least a 2D solution. To anticipate a possible 5D or 6D solution (1920 triplets), we aimed for 50 participants (50 x 96 trials = 4800 triplets). Having more triplets also improves the accuracy of the results.

After getting the embeddings from the Soft Ordinal Embedding algorithm (Künstle & von Luxburg, 2024), we first conducted Procrustes analysis for embeddings with the same dimensionality, where scaling, rotation, translation and reflection were applied separately to the DALL-E 2 and Midjourney v2 embeddings so that both embeddings were optimally aligned with the MERL embedding. Then we calculated the canonical correlations between the three embeddings, quantifying the similarities between them. Besides the overall correlation coefficient and its significance, the calculation also yields weights on the dimensions, indicating the importance of each dimension, thus helping us to interpret the canonical correlations.

4.2.2. RESULTS

First, we determined the dimensionality by looking at the cross-validation accuracies as shown in Figure 4.3. The blue curves denoting the 10-fold cross-validation accuracy measured the percentage of triplet data that can be correctly predicted by the embedding. To this end, the data were split multiple times into training and validation data to exhaustively utilize the entire dataset for both training and validation in a systematic manner. A higher value indicated a better fit. The peak of a cross-validation curve indicated its optimal dimensionality. In theory, this method would prevent both underfitting and overfitting, since decreasing or increasing the number of dimensions will not provide any accuracy gain. Our data suggested a 1D solution for MERL materials, a 2D or 6D solution for DALL-E 2 materials and a 2D solution for Midjourney v2 materials. For

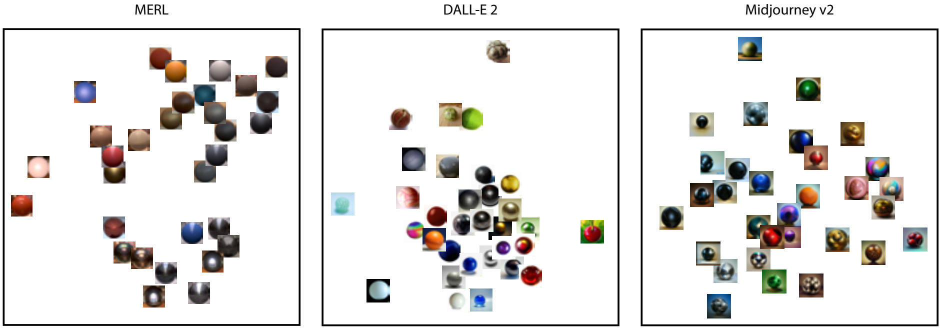


Figure 4.4: The 2D embeddings of MERL, DALL-E 2 and Midjourney v2. Both DALL-E 2 and Midjourney v2 embeddings are aligned with MERL after Procrustes analysis. The Y-axis of the MERL embedding is positively associated with the 1D solution for MERL.

direct comparison, we plotted the 2D embeddings for all three spaces. Figure 4.4 shows the three 2D embeddings after Procrustes analysis with the DALL-E 2 and Midjourney v2 embeddings aligned with MERL embedding. From observation, they show low similarity even after the Procrustes alignment. Since the MERL materials embedding is in fact one dimensional, we fitted the 1D solution into the 2D MERL space by means of multiple linear regression. This resulted in a positive correlation with the Y-axis of the 2D space ($r = 0.99$; $p < 0.001$).

Secondly, we applied canonical correlation analysis to quantify possible similarities between these embeddings. Table 4.1 shows the results of this analysis. The canonical weights indicated how much each dimension contributes to the overall correlation. The three correlation coefficients appeared to be relatively low, suggesting hardly any relation between these three embeddings. This underscores the visual inspection of Figure 4.4. The only case that yielded a significant correlation was between the MERL and Midjourney v2 embeddings, with a relatively low correlation coefficient ($r = 0.550$, $p < 0.05$). As for the weights, the y-axis (0.691) from the MERL embedding contributed slightly more than the x-axis (-0.584), where the x-axis (0.907) from the Midjourney v2 embedding contributed much more than the y-axis (0.437) to the overall correlation. This may be attributed to the corresponding presence of metallic, glossy stimuli in the lower half of the MERL embedding and the lower left quarter of the Midjourney v2 embedding.

4.2.3. Discussion

Looking at the dimensionality plot from Figure 4.3, MERL and the two AI models yielded different dimensionalities. Both DALL-E 2 and Midjourney v2 have higher dimensions than MERL, i.e. need more dimensions to explain the perceptual differences between the stimuli. A possible explanation is the difference in surface texture: being a BRDF material dataset, MERL has no surface texture, where both DALL-E 2 and Midjourney v2 have texture on the surface. The additional information of surface texture increases stimulus complexity which may cause an increase in dimensionality.

Both the different number of dimensions and low correlations as shown in Table 4.1

Table 4.1: Canonical correlation results for Experiment 1

space1-space2	correlation coefficient	p-value	canonical weights space1-X	 space1-Y	 space2-X	 space2-Y
MERL-Dalle	0.283	0.616	-0.071	0.968	-0.513	0.898
MERL-Midj	0.550	0.017	-0.584	0.691	0.907	0.437
Dalle-Midj	0.073	0.997	-0.996	-0.005	-0.705	-0.722

The canonical weights indicate how much each dimension contribute to the overall correlation.

suggest low similarity among these perceptual spaces. Besides difference in surface texture, another explanation for this low similarity might be related to semantics. The MERL dataset comes from measurements of real-world materials, with text descriptions. For AI generated images, however, the text prompts are the starting point. The output images are the interpretation of the described material by the AI models. In some cases, the interpretation from the AI models is rather literal. For example, the three datasets have very different appearances of the material 'cherry' as shown in Figure 4.1. Both DALL-E 2 and Midjourney v2 depicted the object cherry instead of the cherry wood material. Similar for Ipswich pine. In MERL, it stands for a type of wood, but in Midjourney v2 this text prompt is interpreted, very creatively, as a sphere in front of a pine forest. At the same time, other materials show rather consistent appearances among the three datasets, for instance black soft plastic and orange paint. Generative models are also known to be more strongly triggered by specific words that were frequently represented in the training set. As a result, visual differences in the images created by these models may simply be due to a higher familiarity with certain words, which can vary across MERL keywords.

Another interesting observation that can be made from Figure 4.3 is overall differences in training and cross-validation accuracies, indicating how coherent triplet judgements contribute to the embedding (Künstle et al., 2022). The Midjourney v2 embedding yielded a substantially lower training and cross-validation accuracy than the other two embeddings, indicating a higher noise level. In the context of the current study, noise suggests lapses, imprecision or disagreement between participants. One possible explanation is the Midjourney v2 created unique materials that are different from existing material datasets. The various unique and interesting patterns within the sphere shape might introduce ambiguity, which leads to a higher noise level. As shown in Figure 4.1, compared with MERL and DALL-E 2, Midjourney v2 has more textures or patterns within the material spheres, as well as more diverse backgrounds. The overall visual style can be described as fantasy-like. Note that DALL-E 2, the other generative AI, produced less ambiguity than Midjourney v2. One possible reason is the visual style of DALL-E 2 is truer to life. This diversity between Midjourney v2 and DALL-E 2 has also been observed by Göring et al. (2023) who concluded that the former has a more artistic style and the latter a more realistic one.

In summary, in Experiment 1, we used two AI models to generate material images according to text description from a classic BRDF material dataset, MERL. The resulting materials from AI are unique and different, not so comparable with those from MERL. Probably the AI models have quite some freedom in the interpretation and generation. Later in early 2023, we learned a tool that can provide more control over image generation. This allowed us the explore more complex shapes than the spheres from Experiment 1, as we did in Experiment 2.

4.3. Experiment 2 - Influence of shape

The generative AI text-to-image model Stable Diffusion (Rombach et al., 2022) is regularly being used by artists to generate images. Unfortunately, this model faces the same limitation as DALL-E 2 and Midjourney v2, where the text prompt is the only means of controlling the spatial composition of the output images, which can be insufficient. However, in late February 2023, a new add-on for Stable Diffusion was released: ControlNet (Zhang et al., 2023). This add-on provides various ways of precisely controlling the spatial condition of the output images, such as desired human posture, depth map, Canny edge, etc.

As stated above, before the introduction of ControlNet, one major limitation for AI image generation was that the stimulus shape could be controlled through text prompts only. For example, shapes more complex than a sphere are difficult to describe using only text. However, in addition to text, ControlNet can achieve control over the spatial conditions of the output as we show later in Figure 4.5 and Figure 4.6. This option afforded us to further explore AI generated materials using more complex shapes than the spheres we used in Experiment 1. Moreover, it provided the means to investigate the potential influence of shape on material perception in the domain of generative AI.

4.3.1. Method

Stimuli

Compared to Experiment 1 where the image output from AI models is controlled by text prompts only, in Experiment 2 image generation by combining Stable Diffusion v1.5 with ControlNet has one more constraint: the precise control of the shape by using depth maps. Figure 4.5 presents three examples of different depth maps combined with a single text prompt. ControlNet is a neural network architecture that adds spatial conditioning by locking parameters (layers) within a large text-to-image diffusion model such as Stable Diffusion. The weight of the control can be adjusted to vary the degree in relaxation of locking parameters (Zhang et al., 2023). It should be noted that there is no ground truth 3D shape, ControlNet merely converges to solutions that visually resemble the specific 3D shape used as input. We used the same 32 materials from Experiment 1, with depth maps we generated to control the shape of the material blobs. The weight for ControlNet 1.0 was set to one (i.e., halfway) for all images. For Experiment 2, we changed the text prompts from ‘sphere’ to ‘object’ (e.g., ‘1 gray plastic object’). Figure 4.6 shows an overview of all 32 x 3 stimuli of Experiment 2. We choose three different shapes, two ‘globular’ shapes of high (Shape 1) and low (Shape 2) complexity, and a topologically different, more regular shape of a torus (Shape 3).

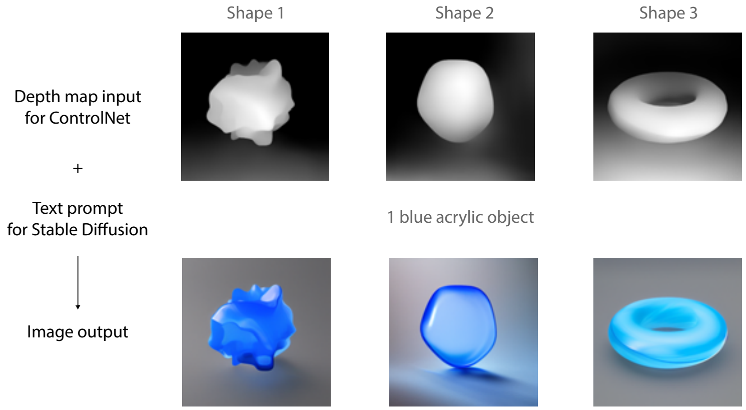

Shape 1
Shape 2
Shape 3
Depth map input for ControlNet
+
Text prompt for Stable Diffusion
1 blue acrylic object
Image output

Figure 4.5: Image generation for Experiment 2, using ControlNet with depth maps and Stable Diffusion v1.5 with text prompts. The upper row presents the depth maps of three different shapes, input for ControlNet. The lower row are the final images.

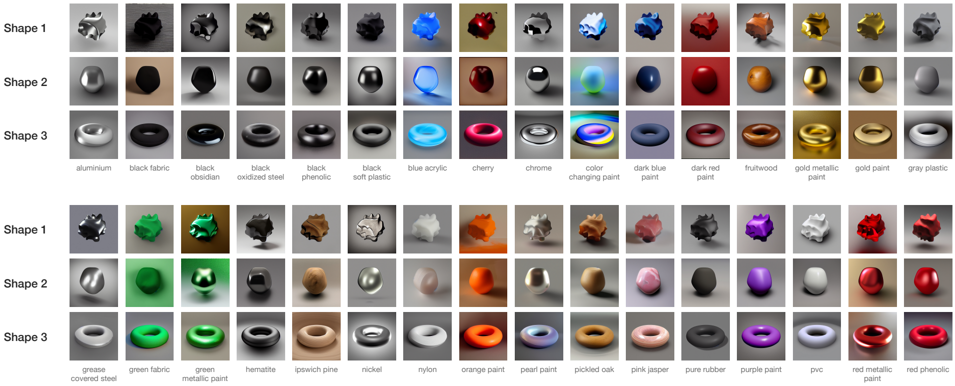
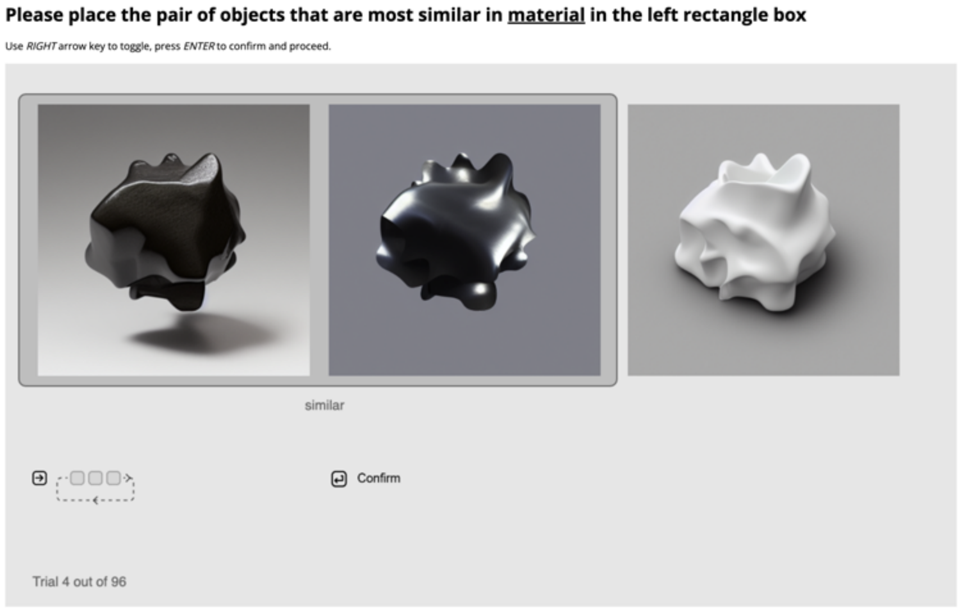


Figure 4.6: An overview of all 32 x 3 stimuli used in three separate sessions in Experiment 2.



Please place the pair of objects that are most similar in material in the left rectangle box

Use *RIGHT* arrow key to toggle, press *ENTER* to confirm and proceed.

Trial 4 out of 96

Figure 4.7: Interface for online Experiment 2. As the icon on the left-hand side indicates, Pressing the RIGHT arrow key would toggle the order of the three images. All three possible pairs can be selected as being similar in terms of material. The participants can press RETURN to both confirm their choice and proceed to the next trial.

PROCEDURE

We used the same approach as in Experiment 1: in three sessions, one per unique shape, we instructed participants to judge the similarity in materials where each session consisted of 96 trials. The only difference is that, instead of having one fixed target image, participants were now free to choose any pair from the triplet. See Figure 4.7 for the new interface. We changed the task to reduce noise in the data. We noticed from Experiment 1 that a fixed target sometimes makes the choice more difficult when the fixed target is the odd one. Without a target, participants could freely choose from all three possible pairs. In all other respects the data analysis was the same as in Experiment 1.

PARTICIPANTS

60 unique participants were recruited for Experiment 2, 20 participants for each session. We recruited all participants from Prolific with the same prescreen criteria as in Experiment 1. The number of recruited participants was less than in Experiment 1 since we anticipated that the new task would produce slightly less noise. The same server issues caused some data loss. Eventually, we recorded 18, 18 and 13 participants for shape 1 to 3. Yet, the number of triplets we got was well beyond the minimal requirement.

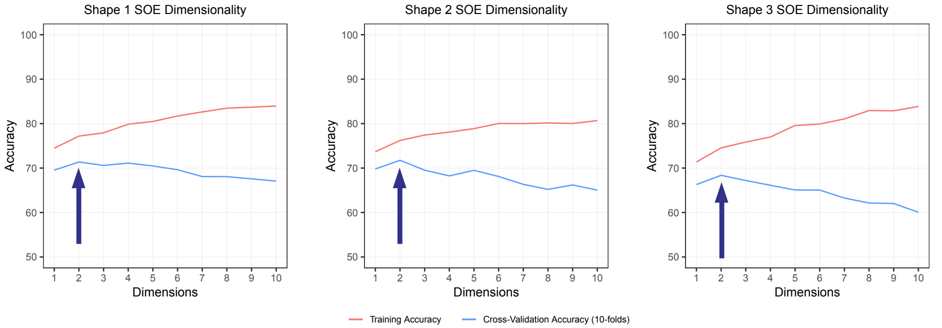


Figure 4.8: The dimensionality of embeddings of shape 1 to 3. The peaks of the blue cross-validation curves indicate the optimal dimensionality. All three embeddings yielded a 2D solution.

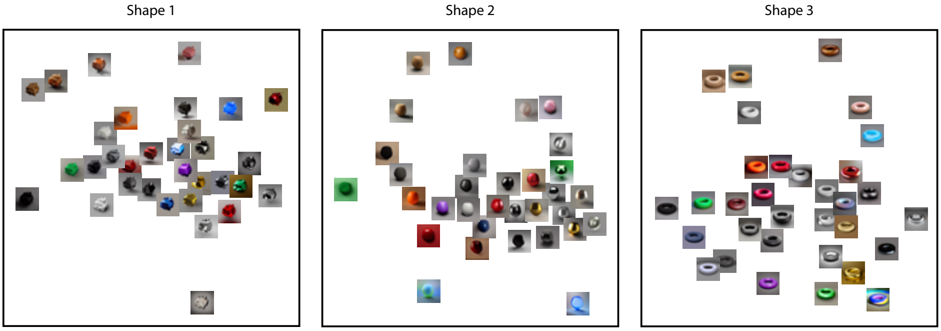


Figure 4.9: The 2D embeddings of shape 1 to 3 after Procrustes analysis, the embeddings for shape 2 and 3 being aligned with that of shape 1.

4.3.2. Results

Figure 4.8 shows the dimensionality plots for Experiment 2. All three embeddings suggest a 2D solution. Note that the peaks for cross-validation accuracy were all around 70% and more pronounced than in Experiment 1, indicating a reduction in data noise (Künstle et al., 2022).

Figure 4.9 shows the three 2D embeddings after Procrustes analysis, the embeddings for shapes 2 and 3 being aligned with that of shape 1. Similarities can already be noticed by observation. To strengthen this observed similarity, we classified the 32 text descriptions from the MERL dataset into the material categories from (Fleming et al., 2013). In doing so, about the same clustering can be seen in the three embeddings: three wood materials are positioned in the top left corner and glossy metallic materials in the bottom right side, while matte fabric-like materials show up mainly on the left side.

We also found statistical support for the observed similarity from canonical correlation analysis results, as denoted by Table 4.2. Each pair of the three embeddings shows high correlations ($r = 0.815, 0.836, 0.764$) which all are significant ($p < 0.001$).

To compare all perceptual embeddings representing the various generative models

Table 4.2: Canonical correlation results for Experiment 2

space1-space2	correlation coefficient	p-value	canonical weights			
			space1-X	space1-Y	space2-X	space2-Y
Shape1-Shape2	0.815	0.000	-0.411	0.815	-0.482	0.793
Shape1-Shape3	0.836	0.000	-0.547	0.712	-0.621	0.750
Shape2-Shape3	0.764	0.000	0.572	-0.722	0.548	-0.807

p-value 0.000 means p<0.001.
The canonical weights indicate how much each dimension contributes to the overall correlation.

and the BRDF stimulus set, we combined the results from our two experiments as follows. First, we correlated the 2D embeddings of MERL, DALL-E 2 and Midjourney v2 from Experiment 1 with the three embeddings from Experiment 2. This led to nine extra correlations, where correlations between MERL and the three shape embeddings were all significant (shape 1: $r = 0.567$, $p = 0.019$; shape 2: $r = 0.619$, $p = 0.004$; shape 3: $r = 0.727$, $p < 0.001$). Similarly, all correlations for DALL-E 2 were significant (shape 1: $r = 0.701$, $p < 0.001$; shape 2: $r = 0.781$, $p < 0.001$; shape 3: $r = 0.567$, $p = 0.024$). In contradiction, all correlations for Midjourney v2 were not significant (shape 1: $r = 0.311$, $p = 0.245$; shape 2: $r = 0.136$, $p = 0.969$; shape 3: $r = 0.204$, $p = 0.807$). Second, we combined these nine correlations with the values from Table 4.1 and 4.2 into one correlation matrix on which we performed an MDS analysis using the correlations as similarity measures. The outcome of the MDS analysis can be found in Figure 4.10 with the smallest/largest distance between two embeddings representing highest/lowest correlation.

4.3.3. Discussion

All three embeddings from Experiment 2 yielded a clear 2D solution with relatively high correlations among each other. The canonical weights as shown in Table 4.2, imply that both dimensions contribute about equally to the correlations. Both consistent dimensionality and high similarity among the three embeddings suggest that the results are non-random and reasonably robust. This also suggests that semisystematic variations in the combination of predefined geometry with uncontrolled illumination have only minor influence on material appearance and perception. Olkkonen and Brainard (2011) investigated the joint effects of illumination and object geometry on material perception and found strong interactions between them. Since the illumination in the current study was not controlled due to the nature of generative AI models, the variation in illumination might have interfered with the influence of object geometry. In addition, Vangorp et al. (2007) considered the blob shape (with a gently changing smooth surface) to be one of the best choices for material discrimination. All three shapes used in the current study have relatively smooth surfaces, but none of them has a flat surface, which could also explain the limited influence of object geometry.

Recently, Göring et al. (2023) evaluated the perceived realism and image appeal of 135 AI-generated images created by several text-to-image models, including DALL-E 2,

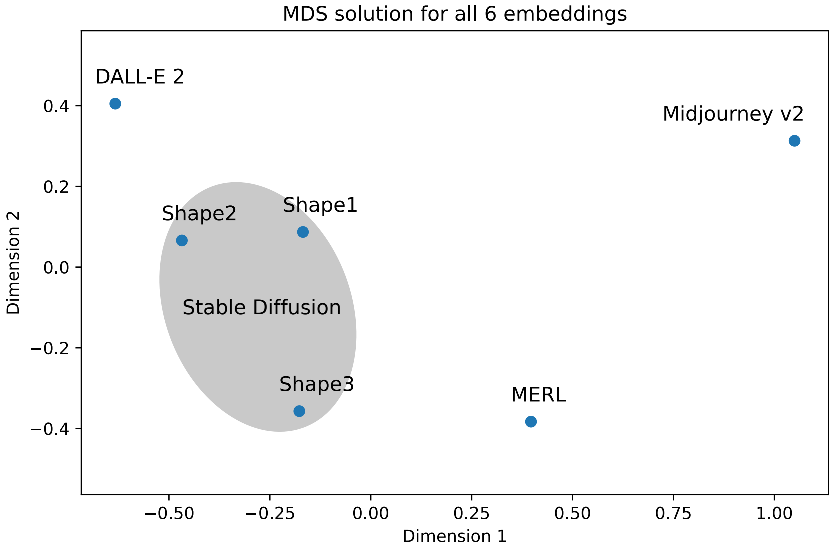


Figure 4.10: The MDS solution for all six embeddings. The canonical correlation coefficients were used to build the distance matrix.

Midjourney and Stable Diffusion. They found that DALL-E 2 and Midjourney were assessed to generate the most and least realistic images, respectively, with Stable Diffusion in between. The Midjourney images were described as "...more artistic similar to a painting...". With respect to image appeal, the images from both DALL-E 2 and Midjourney were judged to be about the same but more appealing than the ones created by Stable Diffusion. The MDS solution as shown in Figure 4.10 suggests that MidJourney v2 cannot be associated with any of the other embeddings, that MERL and DALL-E 2 come up with different solutions, and that shape 3 seems closest to MERL and shapes 1 and 2 to DALL-E 2. Interestingly, the order of the three AI models on the first dimension appears to be compatible with the realism evaluation from Göring et al. (2023) study with DALL-E 2 and Midjourney being the most and least realistic and both Stable Diffusion and MERL in between. Looking at the second dimension, the order seems consistent with the appeal judgements from Göring et al. (2023) study in that both DALL-E 2 and Midjourney v2 have the highest values and Stable Diffusion v1.5 (shape 3) the lowest, together with MERL.

4.4. General Discussion

In two experiments, we explored human material perception using generative AI stimuli and compared the perceptual embeddings between three different generative AI models (DALL-E 2, Midjourney v2, Stable Diffusion v1.5) and one computer rendered BRDF stimulus set (MERL). Unlike the computer rendered material stimuli, the generative AI stimuli are not accompanied by their distal characterization (i.e. reflectance parameters and illumination information) and solely rely on the so-called prompt. This limits the range of psychophysical paradigms available to quantify the perceptual appearances of these artificial objects. We chose to explore perceptual embeddings using Soft Ordinal Embedding (Künstle et al., 2022) and found that in all generative AI experiments, the embedding turned out to be 2-dimensional. In contrast, the embedding of the BRDF dataset from which the prompts for the AI models were taken, was 1-dimensional. The difference in dimensions may be attributed to varying levels of stimulus complexity. While no distal parameters were involved, the generative AI images appeared to contain rich texture information. Some even displayed translucency, a feature absent in the BRDF rendered stimuli due to their exclusion of subsurface scattering.

One difference between Experiment 1 and 2 is the numbers of constraints for generative AI models. In Experiment 1, text description from MERL dataset was the only constraint for the generative AI models, while in Experiment 2, we also introduced a depth map as the second visual constraint in addition to the semantic one. The text descriptions for materials could cause semantic ambiguity for AI models. As we mentioned in the discussion of Experiment 1, Both DALL-E 2 and Midjourney v2 could have their own interpretations of the descriptions. Some interpretations were literal, not necessarily correct or wrong. For example, cherry can be both interpreted as wood or fruit. In addition, AI can generate images beyond reality, for example, a single cherry with stem from DALL-E 2 and a bunch of cherries pressed within one sphere from Midjourney v2.

Although the 2-dimensional optimum occurred robustly for every generative AI stimulus set (except the possible 6D solution for DALL-E 2), independent of shape (Experiment 2) or generative model (Experiment 1 and 2), the dimensionality is relatively low in comparison to other material perception studies based on triplet data. For example, Filip et al. (2024) studied the perceptual dimensions of wood, which is only one category in our experiment, and found the optimal number of dimensions to be between five and nine. Next to using a different algorithm, i.e., the VICE model by Muttenthaler et al. (2022), they also computed the optimal embedding dimensionality by means of Künstle et al. (2022)'s Soft Ordinal Embedding method and found an optimum at six dimensions. A related study that used a wide variety of photos from the STUFF dataset (Schmidt et al., 2022)) revealed that 36 dimensions were needed to describe similarities between material photos. Hebart et al. (2020) using 'THINGS' dataset instead of 'STUFF' dataset, came up with 49 dimensions. Different dimensionalities may arise from focusing on either a single material category or a diverse range of materials. Each approach can lead to distinct criteria, resulting in unique perceptual spaces with varying dimensions.

While the dimensionalities vary widely between these studies and ours, it is interesting to see that the accuracies are rather similar. Accuracy means the percentage of raw triplet data being the same to the triplet data predictions that arise from a model or directly from the embedding itself. The lowest accuracy was found in a study with the largest diversity in pictures, i.e. the THINGS database Hebart et al. (2020) reporting approximately 65% accuracy. Although this appears low, it is high when compared to their upper limit of approximately 67%, which was computed on the bases of repeated trials by different observers. The STUFF database (Schmidt et al., 2022) yielded an accuracy of 71.86% with an upper limit of 73.84% while the study on wood (Filip et al., 2024) resulted in an accuracy of approximately 76% with an upper limit of 82%. This accuracy is comparable to the 68-72% cross-validation accuracy we found in Experiment 2. Note that we did not measure repeated trials and can therefore not compute an upper limit. Experiment 1 yielded a somewhat different picture with 71% cross-validation accuracy for MERL, 70% for DALL-E 2 and 61% for Midjourney v2. Mind that all these accuracies were calculated at the optimal number of dimensions (i.e., two in all our cases except MERL). In summary, while it is challenging to make direct comparisons between studies due to differences in dimensionality and methodology, the perceptual embeddings from our synthetic generative AI stimuli have accuracy levels comparable to those of previous studies using photos. Lastly, it should be noted that Lagunas et al. (2019) also applied a triplet similarity task to their stimulus. Yet they did not explore the perceptual embedding as a (potentially) interpretable global space.

We are not the first using AI generated stimuli for research into material perception, being aware of studies using prompt-based diffusion models. These studies used a variational auto-encoder (VAE) on the perception of glossiness (Storrs et al., 2021) and a Generative Adversarial Network (GAN) on the perception of translucency (Liao et al., 2023). In both studies, the architectures were specifically trained on predefined image datasets (albeit unsupervised) and with different research scopes from ours. Storrs et al. (2021) found that VAEs clustered glossy and matte objects in a manner similar to humans and proposed that the unsupervised model (as opposed to a supervised model) could well predict human gloss perception. Liao et al. (2023) found that distinct layers in their generative model corresponded to different perceptual attributes, where the middle layers corresponded to translucency while higher layers corresponded to body color. Finding paths in latent space that correspond to the intensity of material attributes brings generative AI images closer to traditional CGI in which, for example reflectance parameters can be manipulated. Altering material appearance via latent space was also explored by Delanoy et al. (2022) using GANs and by Sharma et al. (2024) using a diffusion model (Stable Diffusion v1.5). While Delanoy et al. (2022) used the same MERL dataset (Lagunas et al., 2019) as we used, in their study the images were the starting point to generate novel stimuli, while in our study BRDF labels were the starting point. Hence, our study complements other studies using generative AI for material perception as we used text-based images and explored their perceptual embeddings.

AI generated imagery forms a new 'medium' to explore material perception. Using this new medium we find that our results correspond with studies using real-world photos (Fleming et al., 2013), CGI renderings (Zhang et al., 2019) and paintings (Van Zuijlen et al., 2020). This is illustrated in Figure 4.11 where Figure 4.11A presents the perceptual material space as found by Fleming et al. (2013), and Figure 4.11B summarizes the three embeddings from Experiment 2 in the form of the five centroids of five material categories. The perceptual space on the left has wood and stone at the top, and from left to right, fabric, plastic and metal in the middle. The space from Experiment 2 on the right side has a similar structure, only the position of plastic seems to be shifted upwards. Both Zhang et al. (2019) and Van Zuijlen et al. (2020) found similarly structured perceptual spaces as Fleming et al. (2013). Yet it should be noted that we did not have a complete and evenly distributed material category set, as our stimuli were confined to the names of the MERL dataset.

Finally, the word medium can be used to describe not only what the image is made

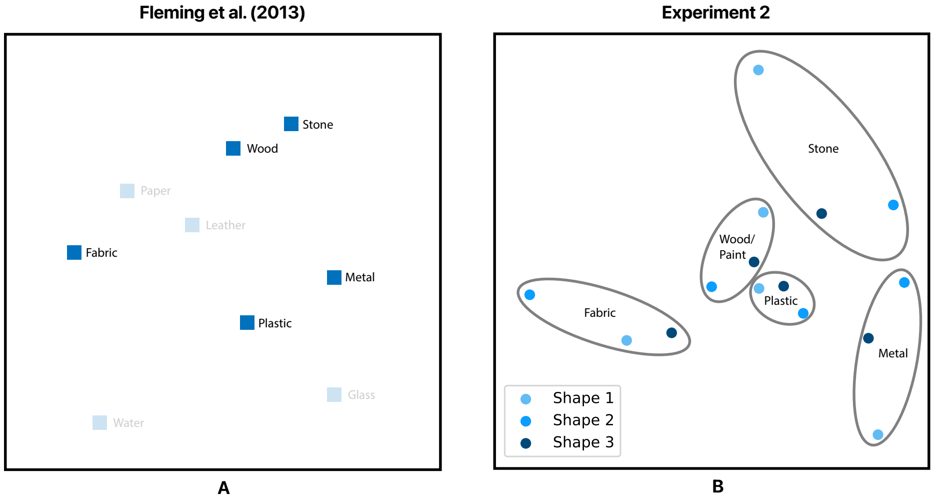


Figure 4.11: A) Perceptual material space adapted from Fleming et al. (2013). B) Perceptual space from Experiment 2 with the centroids of five material categories for three shapes.

of (e.g., oil paint on canvas, pixels on a screen, drawing on paper) but also the technique with which the image has been made (e.g. painting, animating, sketching, photographing). The few studies that directly compared different media yield a mixed picture. Delanoy et al. (2021) compared paintings and renderings and found comparable material perceptions across these two media, while Zhao et al. (2024) compared paintings with engravings and did find differences that could be attributed mainly to contrast. In the latter study, the lack of color seemed to have been (over)compensated through additional local contrast applied by engravers. Generative AI is clearly not a traditional medium and does not depend as much on the interaction between artist and material in the same way as rendering, painting and engraving. Although it appears a fundamentally distinct medium, it does have similarities. The role of the artist or creator is conceptually similar, but the practice of handling the brush is transitioned into handling the prompt. Another similarity with other media is that it has limitations, and that these limitations may lead to specific creativity. One of the limitations of generative AI is its stochastic nature: the relation between prompt and image is rather undeterministic and it is impossible to control every pixel of the image. So the unexpected findings might invite serendipity, probably more so than the traditional act of image making.

To summarize, we evaluated the visual output of AI Generative models using identical material prompts taken from MERL, a BRDF dataset. To this end, we compared their perceptual spaces derived from triplet similarity judgments. In the first experiment, the perceptual spaces of DALL-E 2 and Midjourney v2 turned out to be unrelated, suggesting that these models have different styles in realistically visualizing materials, in line with earlier observations on perceived realism and appeal of AI Generative models (Göring et al., 2023). So, like painters choosing the medium (oil paint, pencil, charcoal, etc.) to visualize materials, it seems wise to do the same when selecting the most appropriate

AI Generative model. The results of the second experiment indicating minor influence of shape on material representation suggest that this choice does not depend critically on the object's shape. In this experiment, the shape was controlled by combining the open-source text-to-image AI model Stable Diffusion v1.5 with ControlNet allowing the additional constraints of depth maps. The resulting perceptual space showed not only a similar structure as the MERL embedding but was also like perceptual spaces from other material studies using real-world photos, computer renderings and depictions. So, Generative AI models have unlocked new methods to generate images. Our comparative study has made clear that they may indeed provide a rich and valuable source for the production of visual stimuli in order to study material perception.

Bibliography

Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, *434*(7031), 301–307.

De Silva, V., & Tenenbaum, J. B. (2004). *Sparse multidimensional scaling using landmark points* (tech. rep.). Stanford University.

Debevec, P. (2008). Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Acm siggraph 2008 classes* (pp. 1–10).

Delanoy, J., Lagunas, M., Condor, J., Gutierrez, D., & Masia, B. (2022). A generative framework for image-based editing of material appearance using perceptual attributes. *Computer Graphics Forum*, *41*(1), 453–464.

Delanoy, J., Serrano, A., Masia, B., & Gutierrez, D. (2021). Perception of material appearance: A comparison between painted and rendered images. *Journal of Vision*, *21*(5), 16–16.

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), 1–15.

Efros, A. A., & Leung, T. K. (1999). Texture synthesis by non-parametric sampling. *Proceedings of the seventh IEEE international conference on computer vision*, *2*, 1033–1038.

Ferwerda, J. A., Pellacini, F., & Greenberg, D. P. (2001). Psychophysically based model of surface gloss perception. *Human vision and electronic imaging vi*, *4299*, 291–301.

Filip, J., Lukavskỳ, J., Děchtěrenko, F., Schmidt, F., & Fleming, R. W. (2024). Perceptual dimensions of wood materials. *Journal of Vision*, *24*(5), 12–12.

Fleming, R. W. (2017). Material perception. *Annual review of vision science*, *3*(1), 365–388.

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of vision*, *13*(8), 9–9.

Goetschalckx, L., Andonian, A., Oliva, A., & Isola, P. (2019). Ganalyze: Toward visual definitions of cognitive image properties. *Proceedings of the ieee/cvf international conference on computer vision*, 5744–5753.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Göring, S., Rao, R. R. R., Merten, R., & Raake, A. (2023). Analysis of appeal for realistic ai-generated photos. *IEEE Access*.

Haghiri, S., Wichmann, F. A., & von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *Journal of vision*, *20*(9), 14–14.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, *4*(11), 1173–1185.

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 229–238.

Heikinheimo, H., & Ukkonen, A. (2013). The crowd-median algorithm. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *1*, 69–77.

Hertzmann, A. (2020). Visual indeterminacy in gan art. In *Acm siggraph 2020 art gallery* (pp. 424–428).

Khan, E. A., Reinhard, E., Fleming, R. W., & Bülthoff, H. H. (2006). Image-based material editing. *ACM Transactions on Graphics (TOG)*, *25*(3), 654–663.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Künstle, D.-E., & von Luxburg, U. (2024). Cblearn: Comparison-based machine learning in python. *Journal of Open Source Software*, *9*(98), 6139.

Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5–5.

Lagunas, M., Malpica, S., Serrano, A., Garces, E., Gutierrez, D., & Masia, B. (2019). A similarity measure for material appearance. *arXiv preprint arXiv:1905.01562*.

Li, J., Endo, L. R., & Kashima, H. (2021). Label aggregation for crowdsourced triplet similarity comparisons. *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28*, 176–185.

Liao, C., Sawayama, M., & Xiao, B. (2023). Unsupervised learning reveals interpretable latent representations for translucency perception. *PLOS Computational Biology*, *19*(2), e1010878.

Linde, L. (1975). Similarity of poetic rhythms with different amounts of semantic content-stress ratings and pairwise similarity ratings. *Scandinavian Journal of Psychology*, *16*(1), 240–246.

Matusik, W. (2003). *A data-driven reflectance model* [Doctoral dissertation, Massachusetts Institute of Technology].

Mead, A. (1992). Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *41*(1), 27–39.

Muttenthaler, L., Zheng, C. Y., McClure, P., Vandermeulen, R. A., Hebart, M. N., & Pereira, F. (2022). Vice: Variational interpretable concept embeddings. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 33661–33675, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/da1a97b53eec1c763c6d06835538fe3e-Paper-Conference.pdf

O'Hare, D. (1976). Individual differences in perceived similarity and preference for visual art: A multidimensional scaling analysis. *Perception & Psychophysics*, *20*(6), 445–452.

Olkkonen, M., & Brainard, D. H. (2011). Joint effects of illumination geometry and object shape in the perception of surface reflectance. *i-Perception*, *2*(9), 1014–1034.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*, 49–70.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schmidt, F. (2019). The art of shaping materials. *Art & Perception*, *8*(3-4), 407–433.

Schmidt, F., Hebart, M. N., Fleming, R. W., et al. (2022). Core dimensions of human material perception. *PsyArXiv, doi:10.31234/osf.io/jz8ks*.

Sharma, P., Jampani, V., Li, Y., Jia, X., Lagun, D., Durand, F., Freeman, B., & Matthews, M. (2024). Alchemist: Parametric control of material properties with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24130–24141.

Storrs, K. R., Anderson, B. L., & Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, *5*(10), 1402–1417.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Terada, Y., & Luxburg, U. (2014). Local ordinal embedding. *International Conference on Machine Learning*, 847–855.

Thompson, W., Fleming, R., Creem-Regehr, S., & Stefanucci, J. K. (2011). *Visual perception from a computer graphics perspective*. CRC press.

Van Zuijlen, M. J., Pont, S. C., & Wijntjes, M. W. (2020). Painterly depiction of material properties. *Journal of vision*, *20*(7), 1–17.

Vangorp, P. (2009). *Human visual perception of materials in realistic computer graphics* [Doctoral dissertation, Citeseer].

Vangorp, P., Laurijssen, J., & Dutré, P. (2007). The influence of shape on the perception of material reflectance. In *Acm siggraph 2007 papers* (77–es).

Wang, X., Bylinskii, Z., Hertzmann, A., & Pepperell, R. (2020). Toward quantifying ambiguities in artistic images. *ACM Transactions on Applied Perception (TAP)*, *17*(4), 1–10.

Wills, J., Agarwal, S., Kriegman, D., & Belongie, S. (2009). Toward a perceptual space for gloss. *ACM Transactions on graphics (TOG)*, *28*(4), 1–15.

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), 11–11.

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhao, Y., Stumpel, J., de Ridder, H., & Wijntjes, M. W. (2023). Zooming in on style: Exploring style perception using details of paintings. *Journal of vision*, *23*(6), 2–2.

Zhao, Y., Stumpel, J., de Ridder, H., & Wijntjes, M. W. (2024). Material perception across different media—comparing perceived attributes in oil paintings and engravings. *i-Perception*, *15*(4), 1–21.

5

Conclusion

The thesis investigated human visual perception through the lens of art. No matter how the art is created, on canvas, or copper plate and damp paper, or from the black box of AI models, they are illusions from which we perceive various materials. From another aspect, the thesis studied *appearance* from different angles. Chapter 2 explored the appearance of the same object under the paint brushes of various artists throughout history. We tried our best to isolate style from subject matter by zooming in onto the same single motif from fragments of paintings. Chapter 3 zoomed out to examine appearance differences between two distinct media. We controlled subject matter by comparing oil paintings and their engraved reproductions. Beyond human creation abilities, Chapter 4 analyzed the potential of an emerging tool, generative AI. Both human artists and generative AI models brought us interesting insights on visual perception.

5.1. Implications of the thesis

In Chapter 2 we measured and described the perception of depiction style, using fragments of paintings. One interesting finding I would like to highlight is the circular temporal pattern we find from the style perceptual space. It is intriguing that the time dimension emerges, even though we only used fragments of the paintings and only asked laymen for similarity judgements. The cut-outs removed all the time-related cues, such as fabric or room interior. And we assume the majority of the participants we recruited from AMT have no art background. A similar relation between creation time and art perception has been reported before. Berlyne and Ogilvie (1974) reported a linear correlation between their three-dimensional perceptual space of paintings and the artist's year of birth. Note that they used full paintings to construct the space, which might contain more time-related cues. Elgammal et al. (2018) found a similar circular temporal pattern using computational method instead of human judgement. Moreover, the attributes that formed the basis of their embedding were the formal elements described by Wöllflin (Wölfflin, 2012). Similar findings from a distinct method can indicate the robustness of the phenomenon. It is possible that the circular pattern of time is in line with some universal patterns, similar as fashion, even as history, that everything evolves in a ‘spiral’ pattern, repetitive but not the same (Dalio, 2021).

By the time we conducted the style study, we were aware of style transfer neural networks, but it was two years before the generative AI rise in 2022. Interestingly, in the last AI study we found that different AI models have their own visual styles. And most images from generative AI have the general ‘AI look’, at least by the time I write the current chapter in 2024. To some extent, AI is like a black box, as we usually give input and receive output without knowing all details. In some cases we do know the architecture and training data but we are ignorant of the inner workings. In other cases, like contemporary services like DALL-E and Midjourney, even the training data and architecture are unknown because they are proprietary. Similarly, we can also argue that an artist is a ‘perceptual black box’. Without knowing the internal mechanism, we only see depictions as the final output. Depictions can be the results of information extraction and interpretation of the world we live in. Artists encoded the information in the form of depictions, and viewers decode the information with possibly the same visual perception system. In most cases, we do not even know the input of the depictions. But we can still perceive the rendered materials. For example, it is unknown whether the girl with a pearl earring (painted by Vermeer) ever existed, or whether the fruits from van Dyck are imaginary. On the contrary, some engraved replicas are the rare cases where we still have access to the original paintings they refer to (if the original paintings survived history). Arguably, both human artists and generative AI models generated appearances based on their own understanding of materials, or more generally, the world around us.

Both the engraving and the AI studies suggest that the perceptual space of material is medium independent (as shown in Figure 4.11 from Chapter 4 on page 88). The comparison between engravings and oil paintings did reveal different characters of the two media. Yet all 20 experimental sessions yield significant positive correlations between the two distinct media, regardless of the manipulations. With the manipulations, we investigated the effects of color and contrast in a controlled fashion. Still, a piece of fabric depicted in colorful oil paint and black engraving lines are perceived as similarly glossy. It suggests that material perception is rather universal, even with different appearances from distinct media. Generative AI, being a new method of image making, produced images that lead to similar perceptual space of materials as reported by other studies, regardless of the stimuli being photographs (Fleming et al., 2013; Zhang et al., 2019), CGI renderings (Zhang et al., 2019), paintings (Van Zuijlen et al., 2020) or AI generated images. It suggests that the way of image making might have very minor influence on material perception, and people have the ability to see through the appearance and somehow capture the essence of materials.

Interestingly, words of materials also point to the same material space (Fleming et al., 2013), indicating human’s interpretation of materials can be tightly connected to semantics. And semantic input is the starting point for text-to-image generative AI models. It is an attempt to connect the two modalities of language and image. We notice the inequality between these two modalities. When describing some image features, language has its limitation, for example, for spatial information. When generating images using text-to-image models, complex geometries can get difficult to describe. If we try to describe using only language, it leaves room for different interpretations and imaginations.

In other words, it introduces ambiguity. Another example, when announcing the exact location or entrance of a conference, text description is often paired with a map or photograph. Liao et al. (2024) has also reported the limitation of verbal description of materials that the verbal descriptions are unable to convey the visual nuances of material appearances, although they could capture material qualities on the coarse level. Moreover, Muttenthaler et al. (2024) suggest that human perceptual judgement data can be used to improve the representation of visual-semantic models.

5.2. LIMITATIONS AND FUTURE WORKS

Of course no research is perfect, in this section we would like to discuss the limitations of our work and some thoughts on future work. In the style study, we concentrated on a single medium, oil painting, and a single motif, the apple. While this focused approach revealed insights on style perception, a broader exploration covering various media and motifs might bring additional insights into the field. Besides, to explain the perceptual space of style, we fitted a limited number of attributes, most of which were subjective perceptual judgments. Future studies might benefit from exploring the relationship between these subjective judgments and objective image statistics, potentially enhancing our understanding of style perception.

Similar suggestions can also be applied to the media comparison study. Collecting pairs of oil painting and their engraved reproductions was no easy task. However, we would love to expand the selection coverage. When examining the stimulus images, we noticed different engravers have their personal styles. A wider coverage could further rule out the potential influence of personal styles. As for the phenomenon that engravings usually have higher local contrast and more details than oil paintings, it is difficult to argue whether engravers compensate for the lack of color with contrast and details, or if oil paintings have lost contrast and details due to degradation of the paint chemicals. A multidisciplinary approach involving chemical analysis and X-ray imaging could shed light on this issue by revealing the material composition and any degradation processes affecting the artworks. Additionally, insights from art history could provide context regarding the techniques and intentions of the artists. These combined perspectives suggest promising directions for future studies to explore and better understand the differences between media.

The next limitation is related to the difference between viewing real artworks in museum or gallery environments and viewing digital images of art on computer monitors. First, the original art have various sizes. And when viewing the original art in a gallery or a museum, a viewer can freely change their viewing distance. In the online experiment setup, on the other hand, all the images of art have the same fixed size on the screen, and the viewing distance remains approximately constant. However, we are not particularly concerned about this because the visual angles remain rather constant. Carbon (2017) reported a strong correlation ($r^2 = 0.929$) between canvas size and viewing distance by observing visitors in a museum environment. The second difference between viewing art in museums and via screens is viewing time. The same study by Carbon (2017) reported a much longer viewing time (average 33.9s) in the museum than that in experiment context (often between 1 and 3s). The significant time difference might be attributed to the purpose of the viewers. The visitors in the museum might have more interest in art so that they spend more time to appreciate the artworks, while participants mainly aim for completing the perceptual tasks. The third difference specifically pertains to oil paintings. Created by brushstrokes and layered paint application, the original oil paintings have the microscopic three-dimensional surface structure, which is absent in digital images. The difference in surface texture might affect the perception of color and gloss (Elkhuizen et al., 2019). While the differences on viewing distance, viewing time and surface structure might be less relevant for style perception and material perception, future research could investigate further.

Another related limitation concerns the crowd sourcing method we used, as they have advantages and disadvantages. On the one hand, it allowed us to scale up the study, involving much more participants compared to lab experiments. We were able to easily recruit participants across different countries, with our own prescreen criteria. Furthermore, it allowed us to keep gathering perceptual data during Covid time. On the other hand, online studies also have their limitations. One significant drawback of conducting online experiments for visual perception is the lack of control over participants' viewing conditions. Variables such as screen size, pixel density, screen brightness, and color calibration can vary widely between devices, potentially affecting how visual stimuli are perceived. However, some researchers suggest that these factors have a minimal influence on visual perception in the context of online experiments (Hoßfeld et al., 2020). Since each participant views all stimuli under the same viewing conditions, any inconsistencies tend to cancel out within the individual across multiple images. This consistency allows for reliable comparisons within participants, even if absolute measures of perception might vary between different users. Researchers should be mindful of these limitations and consider them when designing and interpreting online vision studies.

Nonetheless, we made effort to provide participants intuitive and playful interactions for the online experiments and received positive feedback from them. To sum up, we investigated visual perception through the lens of art, revealed insights to visual perception of different appearances, further approved the value of art in scientific research.

Bibliography

Berlyne, D. E., & Ogilvie, J. C. (1974). Dimensions of perception of paintings. In D. E. Berlyne (Ed.), *Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation* (pp. 181–22). Hemisphere.

Carbon, C.-C. (2017). Art perception in the museum: How we spend time and space in art exhibitions. *i-Perception*, *8*(1), 2041669517694184.

Dalio, R. (2021). *Principles for dealing with the changing world order: Why nations succeed or fail*. Simon; Schuster.

Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., & Mazzone, M. (2018). The shape of art history in the eyes of the machine. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 2183–2191.

Elkhuizen, W., Essers, T., Song, Y., Geraedts, J., Weijkamp, C., Dik, J., & Pont, S. (2019). Gloss, color, and topography scanning for reproducing a painting's appearance using 3d printing. *Journal on Computing and Cultural Heritage (JOCCH)*, *12*(4), 1–22.

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of vision*, *13*(8), 9–9.

Hoßfeld, T., Wunderer, S., Beyer, A., Hall, A., Schwind, A., Gassner, C., Guillemin, F., Wamser, F., Wascinski, K., Hirth, M., et al. (2020). White paper on crowdsourced network and qoe measurements–definitions, use cases and challenges. *arXiv preprint arXiv:2006.16896*.

Liao, C., Sawayama, M., & Xiao, B. (2024). Probing the link between vision and language in material perception using psychophysics and unsupervised learning. *bioRxiv*.

Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R. A., Hermann, K., Lampinen, A., & Kornblith, S. (2024). Improving neural network representations using human similarity judgments. *Advances in Neural Information Processing Systems*, *36*.

Van Zuijlen, M. J., Pont, S. C., & Wijntjes, M. W. (2020). Painterly depiction of material properties. *Journal of vision*, *20*(7), 1–17.

Wölfflin, H. (2012). *Principles of art history*. Courier Corporation.

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), 11–11.

Curriculum Vitæ

Yuguang Zhao

1991	Born in Taiyuan, China.

Education

2011–2015	BSc in Communication Engineering Tongji University, Shanghai, China
2015–2017	MSc in Human-Technology Interaction Eindhoven University of Technology, Eindhoven, The Netherlands
2019–2025	PhD in Visual Perception Delft University of Technology, Delft, The Netherlands *Thesis:* Appearance rendering by painters, engravers and generative AIs - Material perception and depiction across different styles and media *Promotors:* Dr. M.W.A. Wijntjes, Prof. dr. H. de Ridder

Awards

2024	Early Career Best Paper Prize (Chapter 3), i-Perception (Sage)
2015	ALSP Gold Scholarship, TU/e
2014	Second Prize of 8th Student Innovation Training Program
2014	Tongji Scholarship of Social Practice
2013	Third Prize of Creative Design of China Package
2013	Tongji Scholarship of Social Practice

List of Publications

Papers

4. **Zhao, Y**., Stumpel, J., de Ridder, H., Van Assen, J. J. R., Wijntjes, M. W. (2024). Prompts and appearances: Comparing physically based renderings and generative AI images through material perception. *Submitted*.

3. **Zhao, Y**., Stumpel, J., de Ridder, H., Wijntjes, M. W. (2024). Material perception across different media—comparing perceived attributes in oil paintings and engravings. *i-Perception., 2024* 15(4), 20416695241261140.

2. **Zhao, Y**., Stumpel, J., de Ridder, H., Wijntjes, M. W. (2023). Zooming in on style: Exploring style perception using details of paintings. *Journal of vision., 2023* 23(6), 2-2.

1. Di Cicco, F., **Zhao, Y**., Wijntjes, M. W., Pont, S. C., Schifferstein, H. N. (2021). A juicy orange makes for a tastier juice: The neglected role of visual material perception in packaging design. *Food Quality and Preference, 2021* 88, 104086.

Abstracts

7. Wijntjes, M. (2024)., **Zhao, Y**. Material perception with GPT-Vision. *European Conference on Visual Perception (ECVP), Aberdeen, 2024.*

6. **Zhao, Y**., de Ridder, H., Wijntjes, M. (2023). The appearance of depictions. Vision Sciences Society (VSS). *Journal of Vision,* 20(11), 1741-1741.

5. **Zhao, Y**., de Ridder, H., Stumpel, J., Wijntjes, M. (2023). Perceiving style at different levels of information. Vision Sciences Society (VSS). *Journal of Vision,* 23(9), 5388-5388.

4. Wijntjes, M. W. A., **Zhao, Y**., van Assen, J. J. R. (2022). The similarity space of fictional materials. *European Conference on Visual Perception (ECVP), Nijmegen, 2022.*

3. **Zhao, Y**., de Ridder, H., Stumpel, J., Wijntjes, M. (2022). Material perception across different media, comparing perceived glossiness and softness on paintings and engravings. *Vision Sciences Society (VSS). Journal of Vision,* 22(14), 4298-4298.

2. **Zhao, Y**., de Ridder, H., Stumpel, J., Wijntjes, M. (2022). A sense of style; comparing style perception between local and global. *European Conference on Visual Perception (ECVP), Nijmegen, 2022. PERCEPTION (Vol. 51, pp. 170-170).*

1. **Zhao, Y**., de Ridder, H., Stumpel, J., Wijntjes, M. (2022). A sense of style; stylistic features as perceived by non-experts. *European Conference on Visual Perception (ECVP), online, 2021. PERCEPTION (Vol. 50, No. 1_ SUPPL, pp. 220-220).*