Aligning non-exact copies of artworks with their original

Tristan Quin
Supervisor: Ruben Wiersma
EEMCS, Delft University of Technology, The Netherlands

23-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

This research investigates the efficacy and reliability of geometric matching for the specific case of aligning non-exact copies of artistic works with the original from which they were derived. The purpose of which is to provide a foundation for comparison in any further analysis conducted by conservators and art historians. An overview of the image alignment field as a whole is provided as well as a discussion of the particulars of aligning artworks. A variety of demonstrative results indicate the strengths and weakness of the technique in this domain and are accompanied by a discussion of the headline conclusions.

Figure 1: Example of a semantically similar artwork pair. Left: *Girl with a Pearl Earring* by Johannes Vermeer (c. 1665, Mauritshuis). Right: Artificially generated interpretation of *Girl with a Pearl Earring* using the DALL·E 2 system taken from openai.com/dall-e-2.

1 Introduction

Art conservators of the modern-day make use of a plethora of non-invasive scanning techniques to capture and better analyse pieces of artwork [1]. There is great value in finding an accurate spacial correspondence between an original piece of artwork and any physical or digital copies that have been derived from it. It can reveal hidden alterations and improve our historical view of an artwork. For instance, using copies of an artwork from different stages in its composition, details or lack thereof that are not easily visible in the final artwork can be revealed to construct a more complete picture of the artist's creative process and intentions throughout the composition.

Previous work has explored the modern selection of image alignment techniques for general cases [2] and furthermore for specific cases like medical imaging [3]. Alignment techniques motivated by the need for alignment of real-world classes of objects [4, 5] often focus on the specific case of registering a pair of images with large intraclass differences. On the other end of the spectrum, applications like medical imaging narrow down the scope of the task to image pairs with much greater correspondence. However, little explicit interest has been given to the case of aligning image pairs that are semantically very similar but exhibit feature-level differences. An example of such an image pair is illustrated in Figure 1. Circling back to the research topic, the alignment of non-exact copies of artwork is an application that falls into this lesser-explored category. It is therefore a field worthy of exploration.

This research aims to explore the usefulness and applicability of geometric matching and transformation using a pretrained feature extraction convolutional neural network (CNN) as a means for aligning originals and copies of artwork. The outcome of such an exploration takes the form of both qualitative and quantitative metrics accompanied by some discussion and interpretation.

2 Related Work

To the best of our knowledge, there exists no research specifically investigating alignment in the painting domain. However, the general task of image alignment has long garnered the attention of researchers across disciplines within computer science [6]. Subsequently, there is no shortage of research both proposing new techniques and improvements to existing techniques for registering images. To register two images means to transform and overlay them in such a way that their features are roughly in agreement. Alignment is a component of registration. The earliest proposals of registration techniques made use of hand-crafted descriptors that perform well on "correctable" sets of differences [7]. "Correctable" loosely means that the transformation necessary to register the image pair is not difficult to model. This occurs under the presumption that features exhibit identical pixel values between images like in the case of stereo alignment. The image registration field has since evolved to encompass many classes of alignment scenarios. Liu *et al.* nicely summarise the modern variety of differences considered for alignment scenarios [8]. These include pixel values, background clutter, perspective and occlusion, intraclass variation, and multiple subject occurrences. The category of paintings and their copies can exhibit any of the types of differences to some degree depending on the specifics of the image pair being considered. In reviewing the literature, an important takeaway is that it would be irresponsible to say that any single technique will always perform best for a given scenario in any domain. With that in mind, it should be said that this paper is purely investigative and does not intend to present geometric matching as the best approach for aligning artworks but rather as a candidate that ticks many of the boxes for this domain.

3 Background

Rocco et al. provide a comprehensive summary of modern techniques [5]. Their consensus supports the intuitive trend that techniques have moved from utilising hand-crafted descriptors to employ trainable components at different stages in the alignment process, making recent techniques far more capable of registering semantically related images. The driving force behind this move is the limited ability of hand-crafted descriptors (e.g. SIFT [9]) to cope with differences exceeding those of moderate viewpoint changes of the same scene/object [10].

The terminology that is used throughout the proceeding

text is expanded below:

- **Source Image** The image that will undergo alignment.
- **Target Image** The image that the alignment process aims to mimic i.e., the form that the aligned image should take on.
- **Resulting/Warped Image** The image that has undergone alignment.
- **Transform** The mathematical operation necessary to perform the alignment.

Rocco *et al.* demonstrate that their implementation of an end-to-end trainable geometric CNN, called *WeakAlign*, set the standard for the state of the art at the time of publication [5]. No better performing techniques were discovered in conducting a review of available later published materials at the time of writing this paper. *WeakAlign* is a network architecture and training procedure for semantic image alignment and is a suitable candidate technique for this task as it is proven to perform well on semantically related image pairs [5]. A limited overview of the transformation component, excluding the training component, of the pipeline is detailed below:

Siamese feature extraction CNN Features are extracted from the source and target images respectively to produce collections of tensors representing d-dimensional features.

Pairwise feature matching Match scores are generated by exhaustive pairwise similarity computations. Scores are normalised in a similar fashion to that of the second nearest neighbour test.

Geometric transformation estimation Parameters of a K-dimensional transform are estimated through a regressive transformation CNN. $K = 6$ & $K = 18$ for the case of affine and thin plate spline transforms respectively.

A far more detailed description of the entire pipeline is of course available in the original paper [5]. The geometric transformation component is all that is considered for this paper as training will not be applicable. With that in mind, it is appropriate to defer to an earlier paper from the same authors that only implements the geometric matching CNN component, *CNNGeo* [4]. The relevance of *CNNGeo* to this paper is that it is a working and high performing implementation [4,5] of the technique being investigated. Subsequently, it is used to conduct all of the proceeding experiments.

Contribution. The results of this research should deliver a review of geometric matching networks as a mechanism to cope with the types of variations that often exist between originals and copies of artworks. This would be a valuable contribution, in the eyes of the author, as scholarly research into this specific application of image alignment is thus far non existent.

4 Methodology and Experiment Setup

The term ”non-exact copies” is divided into two separate categories according to how they were created:

Recreations. Examples that attempt to mimic the original as closely as possible. These could take the form of an artwork composed by the original artist or composed by another artist in imitation.

Reinventions. Examples that maintain a general similarity with the original but include intentional alterations or differences in the overall composition.

4.1 Manual keypoint annotation

A custom manual annotation tool was made for the purpose of selecting keypoints and storing them in the correct format to be interpreted by *CNNGeo*. An example showing the 11 keypoints used for *Girl with a Pearl Earring* is depicted in Figure 2.

Figure 2: Left: annotated target image. Right: annotated source image. *Girl with a Pearl Earring*.

4.2 Foreground segmentation

Foreground segmentation is achieved by first roughly removing the background on the image and then making finer adjustments with a brush tool to better define the edges between the background and foreground. An example of a foreground segmented image pair is illustrated in Figure 3.

Figure 3: Left: source image. Right: foreground mask. *Girl with a Pearl Earring*.

4.3 Visual comparisons

Comparisons of source, target, and aligned images are illustrated visually with the use of interlacing. A checkerboard pattern mask is applied between the images being compared to produce an image in which no pair of directly adjacent tiles come from the same image.

4.4 Experimental setup

To restate the goal: providing insight into the effectiveness of geometric matching and transformation for the purpose of aligning originals and non-exact copies. The experiments will require the *CNNGeo* implementation, data and different means by which to make evaluations. These components are detailed below.

CNNGeo implementation

The implementation is mostly unaltered for use in these experiments. However, a handful of dependencies in the code base have become outdated since the original publication. This necessitated changes to bring the code to a functional state. These changes do not alter the overall function of the code and would not make for a particularly interesting inclusion in the body of this paper.

Data

There is no single dataset over which all experiments are conducted. With that being said, the examples have been retrieved from the following sources:

Mauritshuis collection. The provided archive contains high-resolution captures of a variety of art pieces in their collection.

DALL·E 2 Girl with a Pearl Earring Samples. This is a collection of ten artificially generated images that take inspiration from the original painting [11]. All ten images are included in Appendix A and numbered accordingly.

Pre-trained networks

There are five pre-trained networks provided in *WeakAlign* that are all trained using the training set from the PF-PASCAL dataset [12]. Note that the implementation being used for experiments is *CNNGeo* but the networks were trained using the full training procedure in *WeakAlign*. Network architectures and geometric models are combined to achieve the following variations:

Resnet101 [13] Available in modes affine, thin-plate spline (TPS), and two-stage (affine followed by TPS).

VGG-16 [14] Available in modes affine and thin-plate spline. The VGG-16 architecture is not compatible with multiple dimension stages: 6 (affine) vs. 18 (TPS) parameters.

Evaluation Metrics

The quality of an alignment result can be measured in both qualitative and quantitative manners. To ensure both are accounted for, the following evaluation metrics are chosen:

Percentage of Correct Keypoints (PCK). This metric is included as a quantitative measurement. It is the count of keypoints which exhibit a transfer error under some threshold compared to an expected value derived from a manually annotated ground truth set. The procedure for computing the threshold closely follows that found in [15] in which the keypoint must not deviate more than $\alpha \times h$ vertically and $\alpha \times w$ horizontally to be classified as correct (h, w are the height and width of the source image respectively and α is a relative factor)

Distance from Ground Truth (DIST). An additional quantitative measurement. This is the average deviation from the expected position of each keypoint after transformation. This is measured as a percentage of the image dimensions.

Intersection Over Union (IOU). Although also measured quantitatively, this often-used metric in image processing should directly correspond with how a human would perceive alignment i.e., the amount of overlap between the foreground and background regions in the target and aligned images.

5 Ethical Considerations and Reproducibility

The entire premise of this research relies on making alterations to pieces of artwork in some way or another. While many of the examples that are used for experimentation are in the public domain, it is nevertheless important to identify and respect the wishes of an artist or their kin when using their work in any capacity. In the eyes of the author, the relatively minor alterations to pieces of artwork in this research are both fair and appropriate in attempting to contribute to the knowledge base in this research sphere. No attempt to tarnish any artist's image or misappropriate their oeuvre is intended.

This research has been conducted using only publicly available resources with the exception of some of the images provided privately in the Mauritshuis collection. The original implementations for *CNNGeo* and *WeakAlign* are both available through their respective citations. However, as mentioned before, alterations were necessary to bring both into a working state. A fork has been created with the alterations[1]. *WeakAlign* encompasses the functionality of *CNNGeo* so will suffice for both. Working with either will require a good amount of familiarisation with the code, especially for using custom data and custom evaluation. Custom scripts were used for annotation, interlacing, *etc*. These are available on GitLab[2].

6 Experiment Results

In this section selective results across different data samples are presented and discussed in some detail.

6.1 Recreation of eye: *Girl with a Pearl Earring*

The first experiment intends to qualitatively gauge the effectiveness of the technique when acting on a recreation. Specifically, the region surrounding the right eye (left side from the viewer's perspective) in Figure 4. The sample taken from the original image is compared to a recreated version provided by *Mauritshuis*. Note that for this recreation only the eye was composed. The original minor misalignment in areas like the curvature of the face and position of the specular highlight of the retina is made evident on the left in Figure 5 through interlacing the two images. The aligned image, obtained by warping the source image, is then shown similarly on the right in Figure 5.

Speaking in purely qualitative terms, the result does appear to be well aligned in regions like the edges of the sclera (white

[1] github.com/tquintudelft-source/weakalign

[2] gitlab.ewi.tudelft.nl/cgv/cse3000/pixel-perfect-paintings/aligning-non-exact-copies

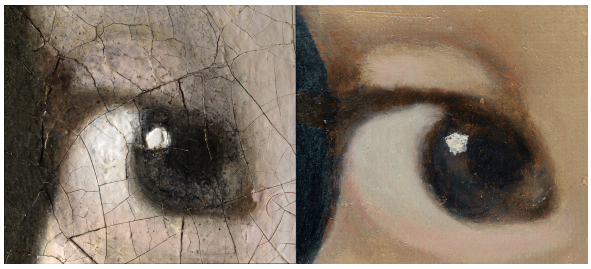

Figure 4: Left: target image sample (original painting). Right: Source image sample (reconstruction). Using full images. *Girl with a Pearl Earring*.

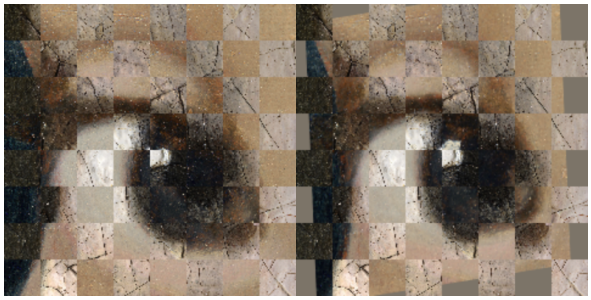

Figure 5: Left: target image interlaced with the source image. Right: target image interlaced with the aligned image. *Girl with a Pearl Earring*.

of the eye) and specular highlight when compared with the original. As an initial indicator, this is quite promising.

6.2 Reinvention: *Girl with a Pearl Earring*

Figure 6 illustrates a handful of comparisons between the original full image of *The Girl with a Pearl Earring* and the images sourced from the DALL·E 2 samples. The average PCK, DIST, and IoU metrics across the DALL·E 2 images are provided in Table 1. As one would expect, the PCK decreases with α regardless of the transformation or network being used (smaller α results in less tolerance for keypoint deviation). The average PCK and DIST metrics across the DALL·E images are provided in Table 1. As one would expect, the PCK decreases with α regardless of the transformation or network being used (smaller α results in less tolerance for keypoint deviation).

Taking a closer look at some specific images, Table 2

		Affine	TPS	Affine + TPS
VGG	PCK@0.1	78.18	82.73	No model
	PCK@0.05	60.91	56.36	No model
	DIST	5.93	6.07	No model
	IoU	74.57	78.85	No model
Resnet	PCK@0.1	82.73	82.73	80.91
	PCK@0.05	57.27	63.64	62.73
	DIST	6.53	5.65	5.82
	IoU	69.85	75.39	78.74

Table 1: Evaluation metric scores using the DALL·E 2 samples of *Girl with a Pearl Earring*

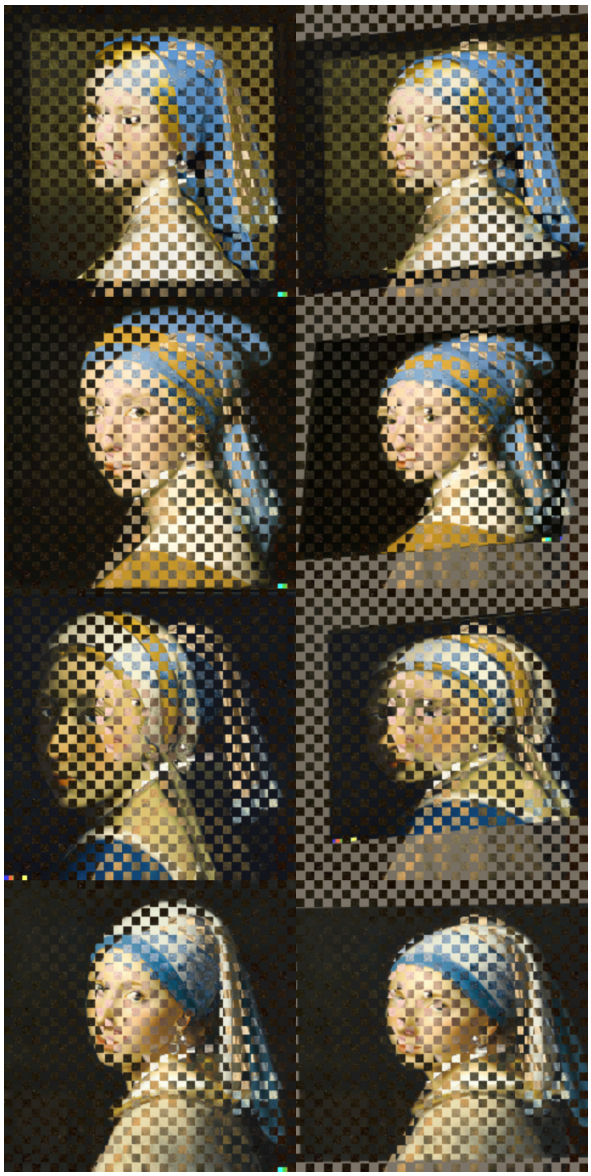

Figure 6: Left column: target image interlaced with source images from DALL· 2 [11]. Right column: target image interlaced with aligned images. *Girl with a Pearl Earring*.

	IoU S-T	PCK@0.1	PCK @ 0.05
Image 2	68.09	72.73	72.73
Image 7	82.31	90.91	63.64
Image 10	36.75	54.55	18.18

Table 2: Comparison of PCK scores across a selection of images from the DALL·E 2 samples of *Girl with a Pearl Earring* using VGG and affine mode.

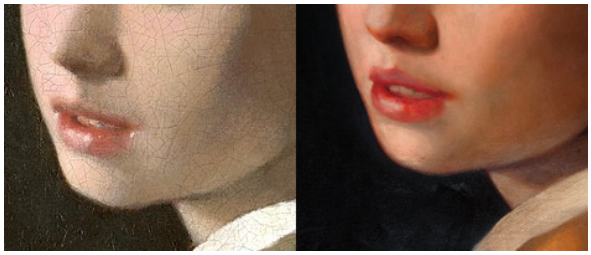

Figure 7: Left: target image sample. Right: source image sample. Images patches taken from identical co-ordinates in the respective full images. *Girl with a Pearl Earring*.

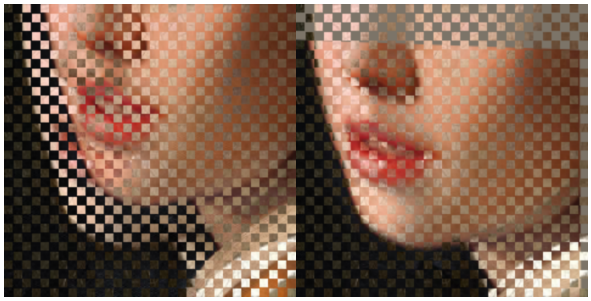

Figure 8: Left: target image sample. Right: source image sample. Images patches taken from identical co-ordinates in the respective full images. *Girl with a Pearl Earring*.

shows the individual PCK and source-target IOU scores for images 2, 7 & 10 (see Appendix A for numbering explanation). Images 2 & 7 are examples in which the foregrounds of the source and target images already overlap to a fair degree. The IoU between the source and target image, seen in Table 2, reflect this considered in the context of the entire set averages in Table 1. However, image 9 is massively misaligned, which is evidenced by the interlaced target-source image in the third row of Figure 6 and the correspondingly low initial IoU score. Although an admittedly small comparison, it would suggest that is quite difficult to approximate a transform for severe cases.

Figure 7 illustrates the case of comparing the original and a reinvention of *The Girl with a Pearl Earring*. The image patches are sampled from identical coordinates in the target and source images respectively (this is done by anchoring the corners of the images and then selecting a sampling region). With that in mind, the spacial mismatch is evident. However, the visual similarity is quite apparent. The results illustrated in Figure 8 would suggest that this is the easiest class of misalignment to correct as the alignment result in the second interlaced image near perfect. The quantitative metrics for this particular pair are as follows:

- PCK@0.05: 81.82 %
- DIST: 6.13 %

These numbers are very much in line with the results obtained in Table 1. This suggests that the qualitative evaluation does not always agree with the quantitative measurement as one might expect the PCK score to be closer to 100% given the visual quality of the alignment. This will of course vary across different image pairs, but is nonetheless worth bearing in mind when interpreting results.

6.3 Small variation: *Portrait of Maria Mancini*

It is necessary to also experiment with image pairs that exhibit substantial feature similarity (exceeding that of just semantic correspondence). In theory, the alignment should only slightly alter the source image. Figure 9 illustrates such a pair of images.

Figure 9: Left: target image, *Portrait of Maria Mancini* by Jacob-Ferdinand Voet (c. 1670, Rjiksmuseum). Right: source image, Copy of *Portrait of Maria Mancini* taken from auktionsverket.com/sv.

Figures 10 & 11 show the results using the different network variations. The purpose of this experiment is to investigate how well the technique can perform on a feature-rich image pair with very minor differences.

This is arguably a scenario in which this technique fails noticeably. The results using the VGG-16 model exhibit visible misalignment artifacts, indicating that the geometric transformation has been poorly estimated. While the Resnet101 model results appear passable at first, a closer inspection into the face reveals the extent of the misalignment. This result would indicate that the pre-trained network being used "expects", speaking abstractly, some minimal degree of misalignment to perform well. This is evidenced when comparing this result with the results of the previous sections where smaller misalignment errors are overlooked in focusing on the correction of more substantial misalignment.

Figure 10: Aligned results using Resnet101. Left: affine. Right: TPS. *Portrait of Maria Mancini*.

Figure 11: Aligned results using VGG-16. Left: affine. Right: TPS. *Portrait of Maria Mancini*.

6.4 Alignment on an image patch basis: *Girl with a Pearl Earring*

An alternative approach is to segment the source and target image into smaller patches beforehand in an effort the simplify the necessary warping for alignment for each patch. Figure 12 shows the results of doing such with some samples from *Girl with a Pearl Earring*. The average IoU of the three aligned patches with the original image is 96.47 % (Note that the grey boundary regions in the aligned images were excluded for this calculation) indicating promise for this approach. The viability of this approach is discussed in Section 8.1.

Figure 12: Aligned image patches overalyed onto the target image *Girl with a Pearl Earring*.

7 Limitations

Drop-in networks are often pre-trained using large real-world data sets, e.g. the ResNet and VGG networks in *WeakAlign* [5], which are not always applicable or available when dealing with artwork. Training augmentation using synthetic data is a possible remedy. This is expanded on further in section 8.3. In addition, manual annotation is necessary for training and evaluation with native artwork data. A small sample of images has been annotated to produce the evaluation metric results in this case, but more concrete results will require the annotation of a much larger data set. This would of course be a time-consuming task.

8 Discussion

In this section, different facets of the overall alignment process and notable findings are discussed.

8.1 Global alignment vs. local alignment

Comparing pieces of artwork is often done to uncover specific and non-obvious changes between versions or copies. For such an analysis, it may be beneficial to only focus on aligning smaller regions between images if the interest only pertains to a single feature or facet of the artwork as demonstrated in the results of Section 6.4. This may be a way to circumvent any clutter and obtuse features that are not of interest in the original image. It is not a far extension to realise that as the patch size decreases, any deformation is easier to model with a simpler transform.

8.2 Scene complexity

Admittedly, this paper has only focused on artwork where the content of the scene being depicted is limited to that of a single subject with a well-defined set of differences between the image pairs. A piece like Rembrandt's *Nachtwacht*, depicted in Figure 13 along with a copy, incorporates many more of the classes of differences and would undoubtedly pose a more substantial challenge in using this technique when trying to align the entire painting. This is a challenge in all domains of image alignment and knowledge of how to circumvent is still very much developing, as Rocco *et al.* mention in their closing remarks [5].

Figure 13: Left: *Nachtwacht* by Rembrandt (1642, Rjiksmuseum). Right: copy of *Nachtwacht* by Gerrit Lundens (c. 1642 Rjiksmuseum).

8.3 Synthetic data augmentation

As mentioned previously, there is likely not enough publicly, or privately for that matter, available data required to natively train the geometric CNN model in the *WeakAlign* pipeline. Synthetic data augmentation has emerged as a technique for partially overcoming such a hurdle as machine learning finds further applications across scientific fields. Applications in the medical imaging field have found substantial classification accuracy [16], potentially showing promise for this application as well.

9 Conclusions and Future Work

Returning to the initial research topic: the applicability of geometric matching for alignment of non-exact copies of artwork. Qualitative results indicate that models trained using real-world data from outside of the domain of artwork generalise well for this application. This suggests that a geometric matching network is certainly a suitable candidate for this admittedly niche application. While the qualitative results are easy to interpret, the quantitative results lack a proper basis for comparison as only geometric matching as implemented has been explored. Further research into this application using other techniques will be necessary to contextualise these results. Improvements are of course possible if limitations bought about by the lack of training data can be overcome. This research has provided an analysis that encourages further experimentation. An interesting avenue for further research is to explore the ability of synthetically generated data to mimic subtle differences between originals and copies. This would open up the possibility of creating natively trained networks for semantic artwork alignment.

References

[1] A. Vandivere, "The technical (re-)examination of Vermeer's girl with a pearl earring," Heritage Science, vol. 8, no. 1, 2020.

[2] B. Zitová and J. Flusser, "Image registration methods: A survey," Image and Vision Computing, vol. 21, no. 11, pp. 977–1000, 2003.

[3] J. B. A. Maintz and M. A. Viergever, "A survey of Medical Image Registration," Medical Image Analysis, vol. 2, no. 1, pp. 1–36, 1998.

[4] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[5] I. Rocco, R. Arandjelovic, and J. Sivic, "End-to-end weakly-supervised semantic alignment," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[6] D. Mistry and A. Bannerjee, "Review: Image Registration", International Journal of Graphics and Image Processing (ISSN 2249 – 5452), 2012.

[7] L. G. Brown, "A survey of Image Registration Techniques," ACM Computing Surveys, vol. 24, no. 4, pp. 325–376, 1992.

[8] Ce Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 978–994, 2011.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.

[10] B. Ham, M. Cho, C. Schmid and J. Ponce, "Proposal Flow," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[11] OpenAI, "Dall·E: Creating images from text," OpenAI, 21-Jun-2021. [Online]. Available: https://openai.com/blog/dall-e/. [Accessed: 08-Jun-2022].

[12] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow: Semantic correspondences from object proposals," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 7, pp. 1711–1725, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[14] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR, vol. abs/1409.1556, 2014.

[15] K. Han, R. S. Rezende, B. Ham, K.-Y. K. Wong, M. Cho, C. Schmid, and J. Ponce, "SCNet: Learning semantic correspondence," 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[16] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing, vol. 321, pp. 321–331, 2018.

A DALL·E 2 Samples

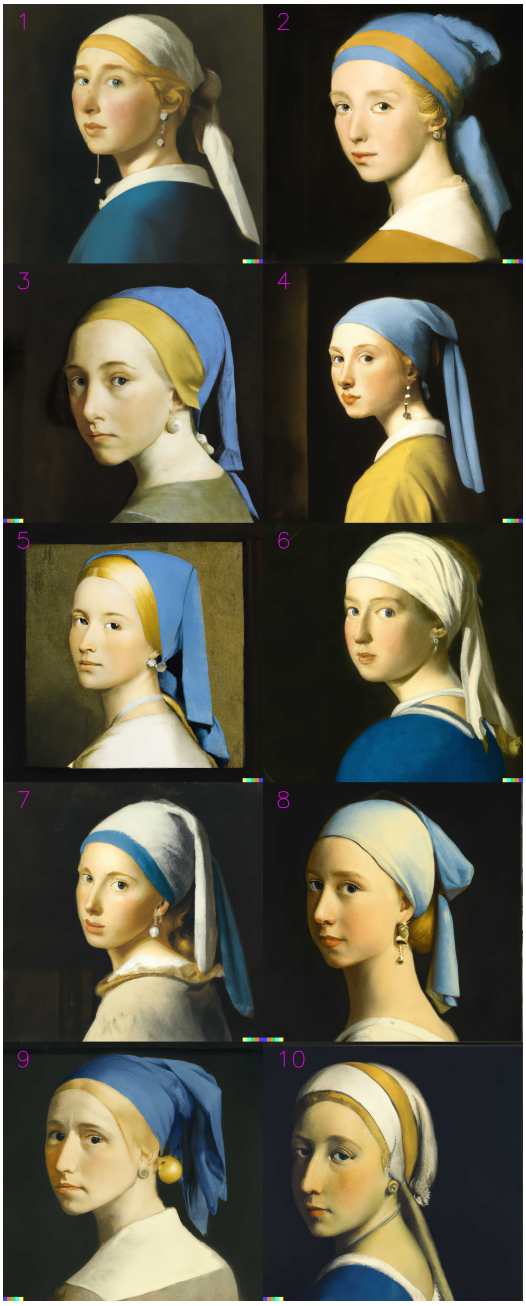

Figure 14: Numbered image samples from DALL·E 2.