Reel to Real: Unleashing Artistic Synergies in Film and Painting through Style Matching

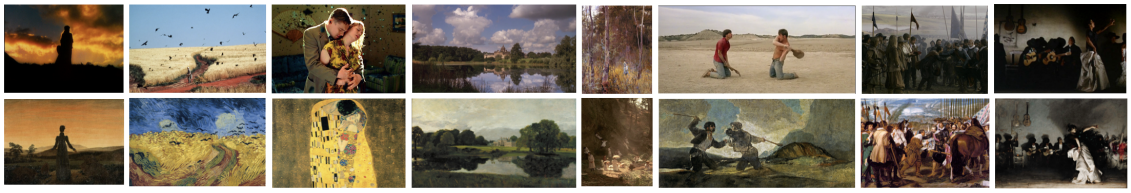


Figure 1: Style matching results of movie frame and painting. Given one film frame, our model can accurately identify the painting with the highest visual and semantic similarity.

Abstract

The rich tapestry of art history and filmography offers a unique vista into the human creative expression, with both mediums often drawing upon similar stylistic and thematic reservoirs. This paper introduces a novel deep learning framework that bridges the gap between the static canvas of paintings and the dynamic frames of films, facilitating the discovery of stylistic parallels and influences. By integrating advanced image processing techniques, our methodology delves into the realms of visual and semantic similarity, extracting and juxtaposing elements such as color distribution, depth of field, and shadow patterns. We harness the power of a Tripartite Contrastive Network, employing triplet loss functions to train on a meticulously curated dataset of historical artworks, enabling the model to discern and match correlated visual narratives. The results showcase the model's capacity to not only identify matching frames and paintings with high accuracy but also to provide insightful observations into the shared language of art and cinematography. The implications of our research extend into a broader understanding of visual affinity, establishing a new paradigm in cross-disciplinary stylistic analysis.

1 Introduction

> *"There are a thousand ways to point a camera, but really only one."*
> — Ernst Lubitsch

We recognize the work of artists through their unique style, such as color choices or brush strokes. The same goes for the work of a cinematographer or a director through their use of light or lens selection. This project deploys a deep learning approach to match a given film frame to a similar painting, thus identifying resonances between the evocative imageries of art and the mystical beauty of motion pictures.

Art Selfie and GPT4 Our task has proven to be far from trivial. In 2018, Google introduced a project called Art Selfie which was supposed to match your selfie with a portrait [11]. However, this achieved slightly underwhelming results as seen in Figure 2.

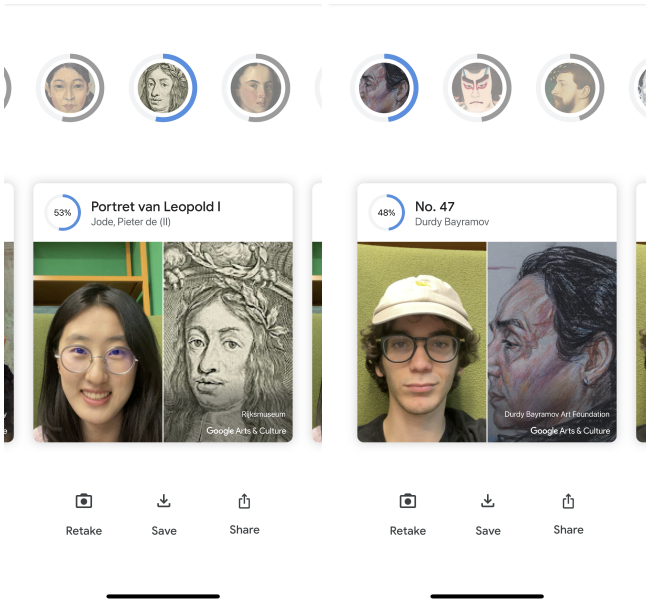


Figure 2: Google Art Selfie unsuccessfully attempts to match our selfies with portraits.

Similarly, AI tools like GPT4, struggle to recognize a particular painting given an image from a different domain. To understand this intricacy we need to delve into the definition of visual and semantic similarity.

Cues for similarity identification Traditional image similarity models rely on visual categorization, meaning that, as seen in Figure 3, two images with common visual characteristics could be deemed as similar even if they are semantically unrelated [14]. In essence, semantic similarity is about what the images represent, while visual similarity is about how the images look. Therefore, unlike current alternatives a suitable model for this task should leverage both semantic and visual cues. Our work will introduce a novel approach to achieve this milestone thus leading to a better understanding of cross-domain style matching.

Figure 3: These images have similar visual cues but are semantically different.

2 Related Work

Generative models for style transfer Techniques for artistic style transfer and analysis have evolved in recent years. Gatys et al. introduced an approach using neural representations to separate and recombine the content and style of arbitrary images [7]. Furthermore, the concept of stable diffusion for artistic style was proposed by Zhang et al [17]. Nonetheless, only a few studies have explored the relationship between art and other visual media through semantic parallelism [4]. These mostly focused on the use of computational models to relate paintings and photographs in cross-domain style transfer, however, advanced cross-domain style-matching (i.e. film frames to paintings), introduces additional challenges which to our knowledge have not been addressed in literature.

Visual similarity feature extraction The extraction of features to evaluate image similarity has been a critical aspect of image analysis, subsequent developments have focused on enhancing the granularity and accuracy of feature extraction. Techniques such as Siamese networks and triplet loss functions have been employed to refine the process of distinguishing between similar and dissimilar images [15]. These methods involve training networks on pairs or triplets of images to learn a feature space where distances directly correspond to the level of visual similarity. Moreover, attention mechanisms have been integrated into CNN architectures to enable models to focus on the most relevant aspects of an image for similarity assessment [6]. In recent years, the field has witnessed a surge in applying deep learning for extracting semantic features, which goes beyond mere pixel-level analysis. This involves interpreting images in a context that aligns more closely with human perception [13]. Such approaches have proven particularly effective in applications where the visual similarity transcends straightforward color and shape matching, requiring an understanding of the image's content and context.

Triplet in image classification The utilization of triplet loss in image classification represents a significant stride in the domain of metric learning and feature discrimination.

This approach, central to fine-grained image categorization, revolves around the principle of understanding and quantifying the relative similarities and differences within image datasets. Introduced by Schroff et al. in their seminal work on FaceNet in 2015 [15], the triplet loss function has been instrumental in training models to differentiate between complex and nuanced image features. At its core, the triplet loss framework operates by considering triplets of images at a time – an anchor, a positive example (similar to the anchor), and a negative example (dissimilar from the anchor). The objective is to train the neural network in such a way that the distance between the anchor and the positive example is minimized, while the distance between the anchor and the negative example is maximized in the learned feature space. This results in a powerful discriminative learning process that is especially effective for tasks where subtle differences between categories or styles are significant.

One of the key challenges in implementing triplet loss effectively is the selection of appropriate triplets. The process of mining hard positives and hard negatives – examples that are just on the boundary of being classified incorrectly – has been a focal area of research, as it significantly impacts the efficiency and accuracy of the learning process [10]. Moreover, recent advancements have sought to integrate triplet loss with other deep learning approaches to enhance performance. The incorporation of convolutional neural networks (CNNs) for feature extraction, combined with triplet loss for classification, has shown promising results in achieving higher accuracy and robustness in image classification tasks [5].

Twin networks and Contrastive Loss function The concept of twin networks, often paired with the contrastive loss function, has emerged as a pivotal element in deep learning, especially for tasks involving image comparison and verification. This methodology has been integral in advancing models' abilities to discern and learn from the similarities and dissimilarities between images [13]. Twin networks, also known as Siamese networks, consist of two parallel neural network branches with identical architecture and shared weights. These networks are designed to process two separate inputs concurrently and learn to differentiate between them. This architecture is particularly suited for comparison tasks, as it enables the model to encode images into a feature space where distances between them reflect their similarity or dissimilarity [3]. The role of the contrastive loss function in training twin networks is crucial. It functions by minimizing the distance between similar (or "positive") pairs of images while maximizing the distance between dissimilar (or "negative") pairs [8]. This loss function effectively compels the network to concentrate on the most distinguishing features between image pairs, leading to a more discriminative embedding space.

In conclusion, twin networks combined with the contrastive loss function offer a sophisticated approach to learning image similarities and differences. Their continuous development and application significantly contribute to advancements in areas requiring nuanced and precise image comparison capabilities. Despite these advancements, challenges remain, especially in handling diverse and complex image datasets. Issues such as the variability in lighting, orientation, and scale continue to pose difficulties in achieving consistent feature extraction.

3 Methodology

3.1 Disentangling visual and semantic similarity

As we seek to unleash artistic synergies across film and painting, the uniqueness of our approach lies in our model's interpretation of similarity. We aim to define this style similarity on the basis of both semantic affinity and visual resemblance. Visual resemblance can be determined by the degree to which the two images are related in terms of their high-level features (similar objects or scenes) without accounting for variations in color, viewpoint, or other visual attributes. On the other hand, we will consider semantic affinity as those low-level features that do not necessarily imply that the images represent the same concept but provide a strong insight into their visual style such as lighting, color distribution, and shadow patterns. Bearing this in mind, we designed our scoring function for our problem in the following manner:

$$0.5S_a + 0.5V_r$$

Where $V_r(y_1, y_2)$ (visual resemblance between two paintings y_1 and y_2) is defined as:

$$V_r(y_1, y_2) = \frac{1}{1 + d(\Gamma(y_1), \Gamma(y_2))}$$

And: S_a (semantic affinity) can be computed the following way:

$w_1(AvgDoF) + w_2(MaxDoF) + w_3(MinDoF) + w_4(DC) + w_5(Sa) + w_6(SP) + w_7(Ba) + w_8(Co)$

Note that: $d(\Gamma(y_1), \Gamma(y_2))$ is the Euclidean distance between the feature vectors of the two paintings and V_R is calculated using: AvgDoF (Average Depth of Field), MaxDoF (Maximum Depth of Field), MinDoF (Minimum Depth of Field), DC (Dominant Colors), Sa (Saturation), SP (Shadow Patterns), Ba (Balance), Co (Contrast).

As seen in the formula above, a pivotal step for us to define semantic similarity is through the calculation of Depth of Field, which we achieved by using MiDaS v3.1 a deep learning framework renowned for its proficiency in monocular depth estimation [1]. The model provided us with a heatmap output for each image which encapsulates the estimated depth information. An example of these images can be found in Figures 4. Alongside this visual representation, we also extracted numerical depth values; specifically minimum, maximum, and average depth which will be later used in our similarity definition.

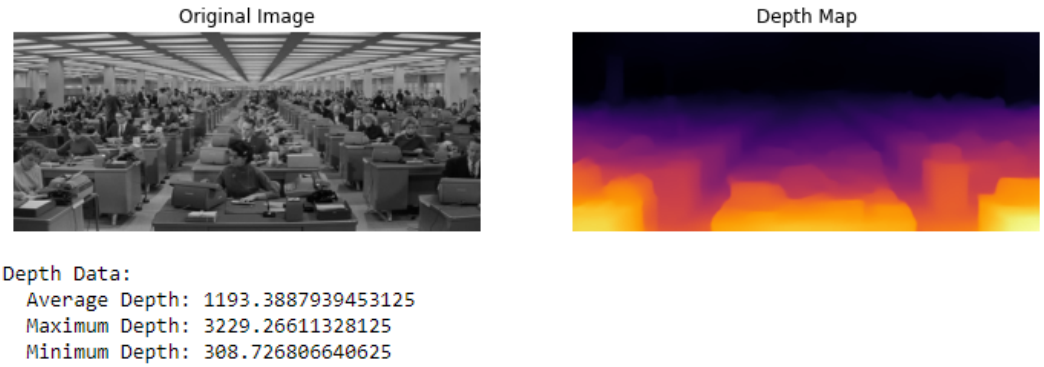


Figure 4: DoF in film ("The Apartment", Billy Wilder)

In our quest to enhance the definition of semantic similarity, we also consider the evaluation of color, a crucial attribute of artistic imagery. We have methodically extracted the RGB values of the three predominant colors in each image, assigning these as supplementary labels. Along the same line, we opted for including saturation, This metric pertains to the vividness or mutedness of colors within an image, effectively gauging their intensity and purity. High saturation is characterized by colors that are striking and vibrant, while low saturation corresponds to tones that are more restrained and subtle.

Essentially, this parameter measures how significantly a color diverges from a neutral gray, providing insight into the emotional and aesthetic tone of the imagery. Additionally, our focus extends to the analysis of shadow patterns, where we assess the relative darkness and positional aspects of shadows, encapsulating both their intensity and spatial distribution within the image.

Completing our suite of semantic similarity metrics are the concepts of balance and contrast. Balance in our analysis refers to the spatial distribution of elements within an image, examining whether these elements are organized symmetrically or asymmetrically. Contrast, on the other hand, is concerned with the luminance disparities within objects from the same visual field. The computation of the aforementioned array of metrics is a pivotal part of our research, enabling us to perform a more pertinent analysis of the semantic stylistic characteristics of a given image, potentially allowing us to achieve a better understanding of cross-domain style matching.

3.2 Tripartite Contrastive Network

Enabling latent space representations of the input images is made possible through our encoder model architecture built upon ResNet [9] as the backbone, followed by an MLP consisting of dense, batch normalization, and activation layers (See figure 5).

Each image in the triplet will be passed to the encoder, and the three latent space representations generated will be used to calculate the triplet loss function:

$$L(A, P, N) = \max(d(f(A), f(P)) - d(f(A), f(N)) + \alpha, 0) \tag{1}$$

where f denotes the latent space representation, $d(\cdot, \cdot)$ denotes the distance (our customized distance function as detailed in Section 3.1), and α is the margin added to maximize the distance between negative pairs. This approach adheres to the fundamental principle of the Siamese neural network [12], characterized by the utilization of twin sub-networks that work in tandem, process image pairs independently, and have identical weights.

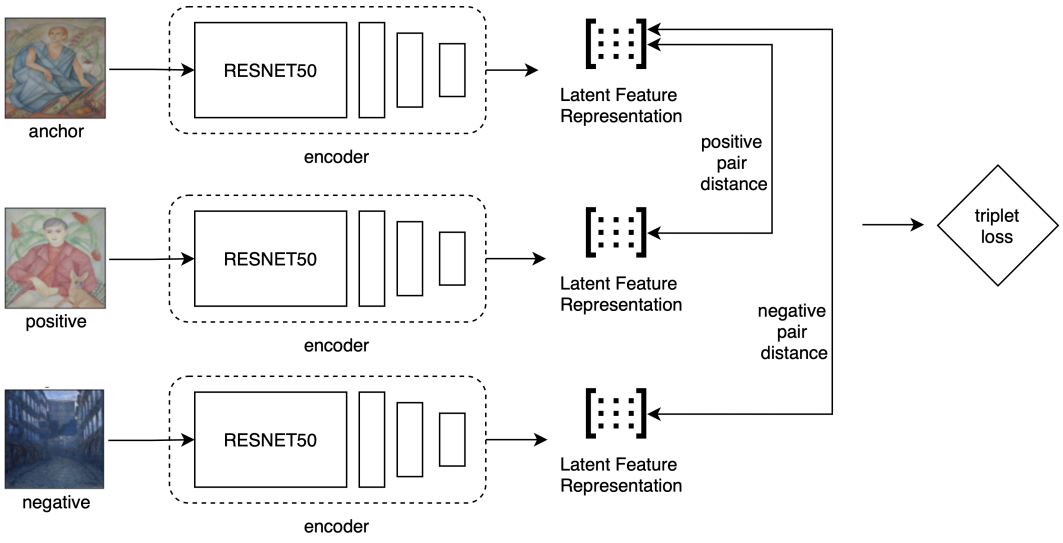


Figure 5: Model Architecture

3.3 Constrained triplet formation

Our neural network ought to proficiently identify similarities and dissimilarities in images. This requires a training dataset consisting of similar and dissimilar pairs. We established a baseline by employing a naive approach for the generation of triplets within the WikiArt dataset. Each triplet is composed of an anchor image, a similar image (positive pair), and a dissimilar image (negative pair) relative to the anchor. While positive pairs are formed from artworks of the same artists and movements, negative pairs are constituted by different artists and movements. Nonetheless, this method may not account for variations within the same artist or movement and overlooks stylistic

overlaps across different artists or movements, leading to positive pairs lacking resemblance or negative pairs being insufficiently contrastive. It is imperative to consider additional attributes, including depth of field, light distribution, and predominant color schemes, to generate triplets that offer greater significance.

Consequently, we have developed an advanced algorithm (see Algorithm 1) designed to integrate these metrics more comprehensively. The algorithm employs a heuristic methodology to identify the first image meeting all predefined constraints relative to the anchor image, thereby selecting images as either the positive or negative counterpart within the matching process. This yields a more adequate set of triplets potentially increasing the performance of our model for the given task as you can see in Figure 6.

Figure 6: Triplets formed with our optimized approach

In conclusion, our methodology shows how we navigated beyond conventional methods to develop a robust algorithm that meticulously forms triplets by integrating a comprehensive set of artistic attributes. This algorithm not only enhances the precision of our model but also enriches the training dataset with meaningful and contrasting image pairs. As we transition into the experiments section, it is with confidence that our methodological foundations are competent enough to unveil the subtle nuances that define artistic synergy across domains.

Algorithm 1 Pseudocode for metric integration

$available \leftarrow \{(\text{movement}_1, \text{artist}_1) : \{\text{img}_1, \text{img}_2, \ldots, \text{img}_n\}, \ldots, (\text{movement}_n, \text{artist}_n) : \{\text{img}_1, \text{img}_2, \ldots, \text{img}_n\}\}$
$threshold\_min \leftarrow \{\text{DC, AveDoF, MaxDoF, MinDoF, SP, Sa, Ba, Co}\}$
$threshold\_max \leftarrow \{\text{DC, AveDoF, MaxDoF, MinDoF, SP, Sa, Ba, Co}\}$
$triplets \leftarrow \{\}$
while len($available$) > 3 **do**
    $anchor \leftarrow$ pop img from $available$
    $metricsAnchor \leftarrow$ compute_metrics($anchor$)
    $movement \leftarrow$ get_movement(img)
    $artist \leftarrow$ get_artist(img)
    $positiveCandidates \leftarrow$ get_positive_candidates($movement, artist$)
    $negativeCandidates \leftarrow$ get_negative_candidates($movement, artist$)
    $triplet \leftarrow \{\}$
    append $anchor$ to $triplet$
    for all img in $positiveCandidates$ **do**
        $candidate \leftarrow$ pop img from $positiveCandidates$
        $metricsCandidate \leftarrow$ compute_metrics($candidate$)
        for $metricA$ in $metricsAnchor$ **do**
            $diff \leftarrow$ abs($metricA - metricC$ in $metricsCandidate$)
            if $diff >$ corresponding threshold in $threshold\_max$ **then**
                continue
            end if
            append $candidate$ to $triplet$
            pop $candidate$ from $available$
            break
        end for
    end for
    for all img in $negativeCandidates$ **do**
        $candidate \leftarrow$ pop img from $negativeCandidates$
        $metricsCandidate \leftarrow$ compute_metrics($candidate$)
        for $metricA$ in $metricsAnchor$ **do**
            $diff \leftarrow$ abs($metricA - metricC$ in $metricsCandidate$)
            if $diff <$ corresponding threshold in $threshold\_min$ **then**
                continue
            end if
            append $candidate$ to $triplet$
            pop $candidate$ from $available$
            break
        end for
    end for
    if len($triplet$) $== 3$ **then**
        append $triplet$ to $triplets$
    else
        push matched $positiveCandidate$ and/or $negativeCandidate$ back to $available$
    end if
end while

4 Experiments

4.1 Wikiart dataset

A fitted dataset is of paramount significance to model training thus, we would like to use the WiKiArt General dataset [16] as our training dataset, it contains approximately 25.000 artworks from 15 art movements. The data is labeled by movement, author, and painting category ("Portraits", "Landscape" and "Other"). Albeit large and diverse, the dataset is observed to be unstructured. In light of our primary objective to match a given film frame to a similar painting, it is essential to undertake a data-cleaning process. This will ensure that only relevant images are incorporated into the dataset for utilization. Therefore, we have removed sketches and photographic representations of sculptures, electing to preserve solely paintings. We also strategically omitted contemporary artworks from periods after the invention of cinema as such works are less likely to have influenced the cinematic visual language [2]. Consequently, we have successfully distilled the dataset to a collection of 21,300 paintings, which are now available for training and testing purposes.

4.2 Fine-Tuning of Threshold Parameters for Triplet Formation

The training of the Tripartite Contrastive Network was conducted in a progressive manner using triplets that were optimized for the task. In order to extract triplets that are semantically informative and conducive to effective encoding, we fine-tuned the threshold parameters that govern the formation of these constrained triplets. Empirical analysis indicates that the threshold values presented in Table 1 yield the most substantial triplets, thereby facilitating fast convergence and optimal test accuracy.

Table 1: Thresholds for Positive and Negative Pairs

	DC	AveDoF	MaxDoF	MinDoF	Sa	Ba	SP	Co
Max for positive pairs	201	1719	3236	539	72	5	72	43
Min for negative pairs	430	2597	4116	1527	118	21	408	61

4.3 Relative Importance of Metrics

The perceptibility of certain features in an image varies. It is seemingly easy to identify differences in color themes between a painting and a movie frame. In contrast, more intricate and subtle elements like shadow patterns might not be immediately apparent. To address this, we evaluated the relative importance of all defined metrics, assigning appropriate weights to each. We selected 100 paintings and 100 movie frames at random, calculated the metrics for each, and manually examined whether the weighted sum of these metrics effectively reflects their semantic affinity. While expensive and time-consuming, this manual process proved crucial for calibrating our performance evaluation, especially considering the challenge of quantifying the somewhat subjective and qualitative task of assessing the similarity between film frame and painting.

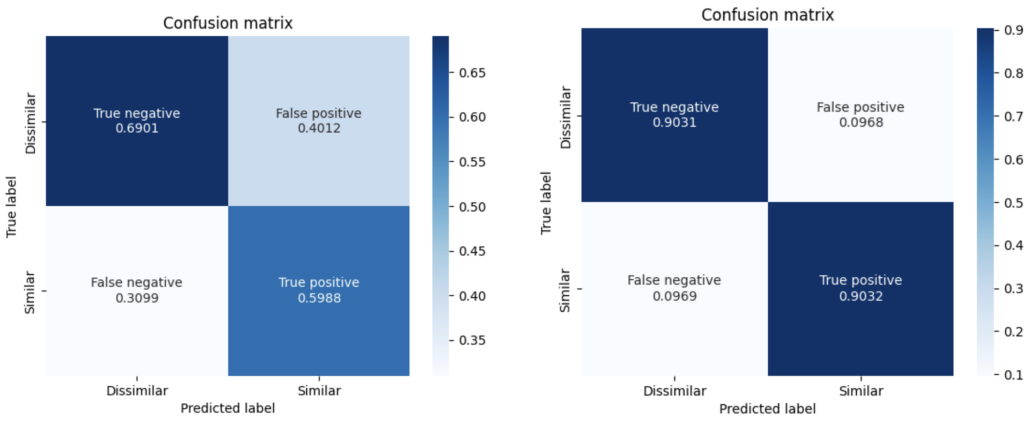
Table 2: Weight Percentages for Semantic Metrics

	DC	AvgDoF	MaxDoF	MinDoF	Sa	Ba	SP	Co
Weight (%)	18	11	14	14	13	10	10	10

4.4 Performance

Similar to the training phase, the model's performance evaluation involves testing with triplets of images: an anchor, a positive image, and a negative image. The testing procedure is reduced to a classification task, wherein the model must discern the image that is visually similar to the anchor from the dissimilar one. The confusion matrix derived from the classification results demonstrates the efficacy of our model in encoding features and its proficiency in discerning similarities and dissimilarities as you can see in Figure 7.

Additionally, from a randomly selected subset of 100 image pairs in our test set, we computed the difference in the weighted sum of normalized metrics for each pair, as well as the normalized



(a) Naive approach. (b) Optimized approach.

Figure 7: Performance before and after optimization.

Euclidean distance between the pairs in their latent space representations. We followed equation 2 for normalization.

$$x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{2}$$

A notable strong positive correlation is evident between these two quantitative approaches as see in Figure 8. This positive correlation supports the effectiveness of our tripartite contrastive network in capturing image features and performing cross-domain stylistic matching, evidenced by not only achieving optimal loss and high classification accuracy but also by its ability to discern subtle semantic details such as depth of field, dominant colors and other metrics defined in this project.

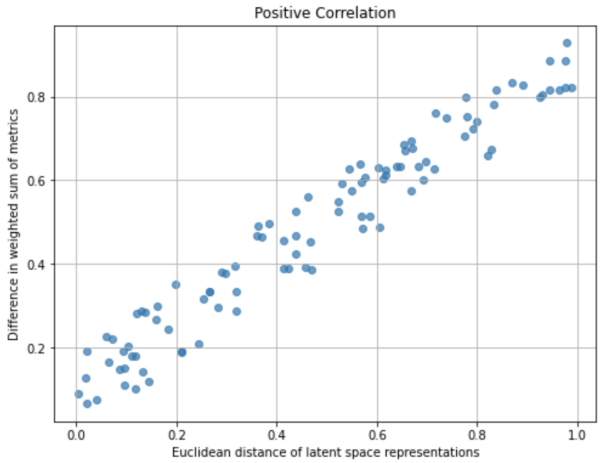


Figure 8: Positive correlation between the difference in the weighted sum of metrics and Euclidean distance of latent space representations.

The experimental phase of this research project was marked by meticulous dataset curation and robust model training, culminating in a fine-tuned algorithm adept at discerning artistic nuances. The resulting performance of the model was not only quantifiable in terms of accuracy and loss metrics but also qualitatively apparent through the enhanced ability to match cinematic frames to their stylistically similar painting counterparts. The successful implementation of the triplet network, reinforced by empirical validation of the weighted semantic metrics, paves the way for innovative applications in cross-domain style recognition as we will discuss in the next section.

5 Conclusion

In this work, we ventured into the compelling domain of cross-disciplinary artistic analysis, pioneering a deep learning approach to discern stylistic affinities between film and painting. We have successfully completed the given task of matching film frames to similar paintings whilst achieving an understanding of the intricacies of style, encompassing new approaches for defining visual and semantic similarity in image analysis.

In conclusion, our journey through the realms of film and painting has not only been a technical achievement but also an ode to the enduring spirit of creativity. It is a celebration of the unseen threads that connect different forms of human expression, the hidden nuances in the work of the old masters that make us enjoy the evocative imageries of art and the eternal beauty of motion pictures.

References

[1] R. Birl, D. Wofk, and M. Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

[2] I. L. Blom. Quo vadis? from painting to cinema and everything in between. In *La decima musa: il cinema e le altre arti: atti del VI Convegno DOMITOR, VII Convegno internazionale di studi sul cinema: Udine, Gemona del Friuli, 21-25 marzo 2000*, pages 281–296. Forum, 2000.

[3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[4] N. Cohn. A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, 146:304–323, 2016.

[5] Y. Ding, Z. Ma, S. Wen, J. Xie, D. Chang, Z. Si, M. Wu, and H. Ling. Ap-cnn: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Transactions on Image Processing*, 30:2826–2836, 2021.

[6] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017.

[7] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[8] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] J. Herman and W. Usher. Salib: An open-source python library for sensitivity analysis. *Journal of Open Source Software*, 2(9):97, 2017.

[11] G. Inc. Is your portrait in a museum? Google Arts & Culture Blog, 2018.

[12] G. Koch, R. Zemel, R. Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[13] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[14] O. Risser-Maroiux, A. Marzouki, H. Djeghim, C. Kurtz, and N. Loménie. Learning an adaptation function to assess image visual similarities. *arXiv preprint*, 2021.

[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[16] WikiArt. Wikiart dataset general. WikiArt Official Website, 2021.

[17] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023.