An Approach to Style Image Selection in Neural Style Transfer

Andrea Nappi
Department of Computer Science
University of Twente
Enschede, Netherlands
a.nappi@student.utwente.nl

***Abstract*—Neural style transfer (NST) is a branch of computer vision that generates an output combining the style of one image with the content of another. Since its introduction in 2016, NST has evolved significantly, with improvements largely focused on model optimization and new architectures.**

This paper takes a different approach by attempting to address a key practical challenge: selecting an appropriate style image. It is investigated an optimal style identifier that automatically selects the most suitable style image from a predefined gallery. Users can specify a general style category (e.g., "Picasso" or "cosmic"), and the identifier will determine the best image to achieve their desired result. By simplifying the style selection process, this approach relieves users from the burden of choice and possibly improves the output quality.

***Index Terms*—Neural style transfer, style selection, optimal style**

I. Introduction

In 2016, Gatys et al. [1] introduced Neural Style Transfer (NST), a method that uses Convolutional Neural Networks (CNNs) to blend the content of one image with the style of another. Their algorithm optimizes a starting image to match the content features of a photo and the style statistics of an artwork, enabling stylized image creation without training on specific datasets. This approach revolutionized image processing by enabling flexible style application, paving the way for numerous successful industrial applications such as Prisma, Ostagram, and Deep Forger [2], and finding utility in ares such as scene rendering for the film industry [3]. Building on the foundational work of Gatys et al. [1], researchers have made significant progress by focusing primarily on optimizing deep learning models, reducing computational costs, and addressing limitations in resolution and scalability [4], with state-of-the-art techniques leveraging advanced generative models [5], [6], enabling more flexible and precise style transfer results. However, practical challenges, such as selecting an appropriate style image, which would also simplify the user experience, have received comparatively little attention. This gap in the literature presents an opportunity to improve the accessibility and usability of NST.

This paper investigates the feasibility of an optimal style identifier to streamline the style selection process. The proposed identifier determines, based on the NST architecture, the content image, and the desired style intensity, the most suitable style image from a predefined gallery. This approach aims to empower the NST technology, simplify the process for users, and has the potential to improve the overall quality of stylized outputs by ensuring better alignment between content and style.

While this work focuses on a single architecture, VGG19, it opens new avenues for research into user-centric tools in the field of neural style transfer, complementing ongoing efforts to advance model performance.

II. Methodolgy

A. Framework

This paper aims to build on top of the work of Xiang Zhou, 2024 [7], therefore, it will try to replicate as closely as possible its framework, which mostly resembles the one presented in [1], to ensure consistency of the fundamental findings.

The employed architecture is VGG19 [8], a deep convolutional neural network featuring 19 layers (16 convolutional ones and 3 fully connected ones), that uses small 3×3 convolution filters and a consistent architecture. VGG19 performs well in NST tasks due to its depth and ability to capture both low-level textures and high-level semantic content, with its pre-trained layers providing a robust feature space for separating and recombining content and style effectively [1]. The image loading function constrains images to 512x512 pixels: this size allows the optimization to be computationally manageable while still producing high-quality images [1]. The adopted optimizer is Adam [9], which is often used in deep learning due to its adaptive learning rates, efficiency, and robustness to noisy gradients and complex architectures; and the learning rate is 0.02. The (total) loss function (3) is the same as in [1], it is composed of a content loss part (1), measuring how different the output's content is from the content reference, and of a style loss part (2), measuring how different the output's style is from the style reference.

$$L_{\text{content}}(p, x, l) = \frac{1}{2} \sum_{ij} \left(F_{ij}^{l}(x) - P_{ij}^{l} \right)^2 \tag{1}$$

Equation (1) calculates the difference between the feature representations of the generated image (x) and the content image (p) at a specific layer l in the neural network. Here,

$F_{ij}^l(x)$ and P_{ij}^l represent the activations of the generated and content images, respectively, at the i-th feature map and j-th spatial position. The term $\left(F_{ij}^l(x) - P_{ij}^l\right)^2$ measures the squared error, and the summation over i and j aggregates this error across all feature maps and spatial positions. The factor $\frac{1}{2}$ is included for convenience in gradient computation. This loss ensures that the generated image retains the structural content of the target content image.

$$L_{\text{style}}(a, x) = \sum_l w_l E_l(a, x) \tag{2}$$

Equation (2) calculates the difference in style between the generated image (x) and the style image (a) across multiple layers of the neural network. Each layer l contributes with a weighted term, $w_l E_l(a, x)$, to the total style loss, where $E_l(a, x)$ is the per-layer style loss defined as:

$$E_l(a, x) = \frac{1}{4N_l^2 M_l^2} \sum_{ij} \left(G_{ij}^l(x) - A_{ij}^l\right)^2.$$

The per-layer style loss measures the difference between the Gram matrices (matrices able to summarize the texture and style of an image) of the style and generated images at layer l. The constants N_l and M_l are the number of feature maps and the number of spatial positions, respectively, at layer l. The term $\frac{1}{4N_l^2 M_l^2}$ normalizes the loss to account for variations in the size of the feature maps.

$$L_{\text{total}} = \alpha L_{\text{content}} + \beta L_{\text{style}} \tag{3}$$

The total loss is the weighted sum of (1) and (2). The weights α and β are hyperparameters that control the trade-off between the content loss and the style loss in (3); a higher α relative to β prioritizes preserving the structure of the content image, while a higher β relative to α emphasizes transferring the artistic style. Because it is the ratio that matters, it is common practice to set the content weight α to 1 and adjust the style weight β to control the trade-off, and variations in the trade-off will be referred to as style weight adjustments in this paper.

The model will start iterating from a copy of the content image, following [1] [7], which is the approach that yields the best results [10], with each NST operation being optimized through 1000 iterations, ensuring consistency with [7].

B. Proposed Approach

Assessing the quality and aesthetic appeal of an NST product has been a persistent challenge since the development of the technique [11]. Artworks are subjective, and different people will likely have different preferences about the style intensity and the content preservation. A criterion for style image selection should take into account such user preference, and this is indeed the starting constraint of this paper's investigation. Style intensity is determined by adjusting the style weight, therefore, the first thing to do is establish the user's style weight preference.

One of the findings of Zhou's work [7] is that for a given framework (regardless of content and style images), there exists an optimal starting total loss range, above which the optimizer struggles to efficiently reduce the loss, likely leading to poor convergence, with the risk of generating low-quality (with inadequate style and/or content) images with artifacts and with random noise. Conversely, below this range, the optimizer fails to produce meaningful style transfer results, as the optimization process lacks sufficient space to progress, resulting in shallow outputs. For this specific framework, Zhou's work [7] suggests an optimal initial loss of around 10^8. Given an optimal starting loss range and the content and style images, it is possible to infer the style weight that practically yields that optimal starting loss.

This paper investigates whether, because of the uniqueness of each pair of content and style images, they will all require tailored style weights to reach the optimal loss, which means that some of them will be a better match for the user's desired style intensity. If this is the case, it will be possible to define a suitability score, according to user preference, for each style image, and therefore select the most suitable style among a gallery of potential style images.

III. Execution

To accurately determine the user's preferred style intensity, the proposed solution involves creating a comprehensive library of images. Each image is generated by applying the style image to a content image using a wide range of style weights. The greater the variety of weights represented in the library, the more precisely user preferences can be understood. For example, as shown in Fig. 1, the library illustrates how varying style weights influence the final output, enabling users to select their ideal balance between content preservation and stylistic intensity. While style images may naturally exhibit some slight variability in their intrinsic intensity, this framework assumes that the observed differences are primarily due to the applied style weight, ensuring a consistent and effective approach to capturing user preferences.

Once the style weight preference ($\overline{\beta}$) has been assessed, the next step is to infer the optimal style weight (5) for each of the pictures of an existing gallery (e.g. "Paintings from Picasso" or "Cosmic pictures in a folder").

$$\hat{\beta}_{ij} = \frac{\hat{L}_{total}}{L_{\text{style}_{ij}}} \tag{4}$$

$\hat{\beta}_{ij}$ is the optimal style weight to apply the style of the image "j" to the content image "i", and is computed as the ratio between the optimal initial total loss $\hat{L}_{total}$ and the initial style loss $L_{\text{style}_{ij}}$. A key factor here is that the image from which the model starts iterating is a replica of the content image, which means that in the initial total loss, the content loss contribution is null, therefore the only weight influencing the initial total loss is the style weight. Had this not been the case, there would have been a situation where any pair of

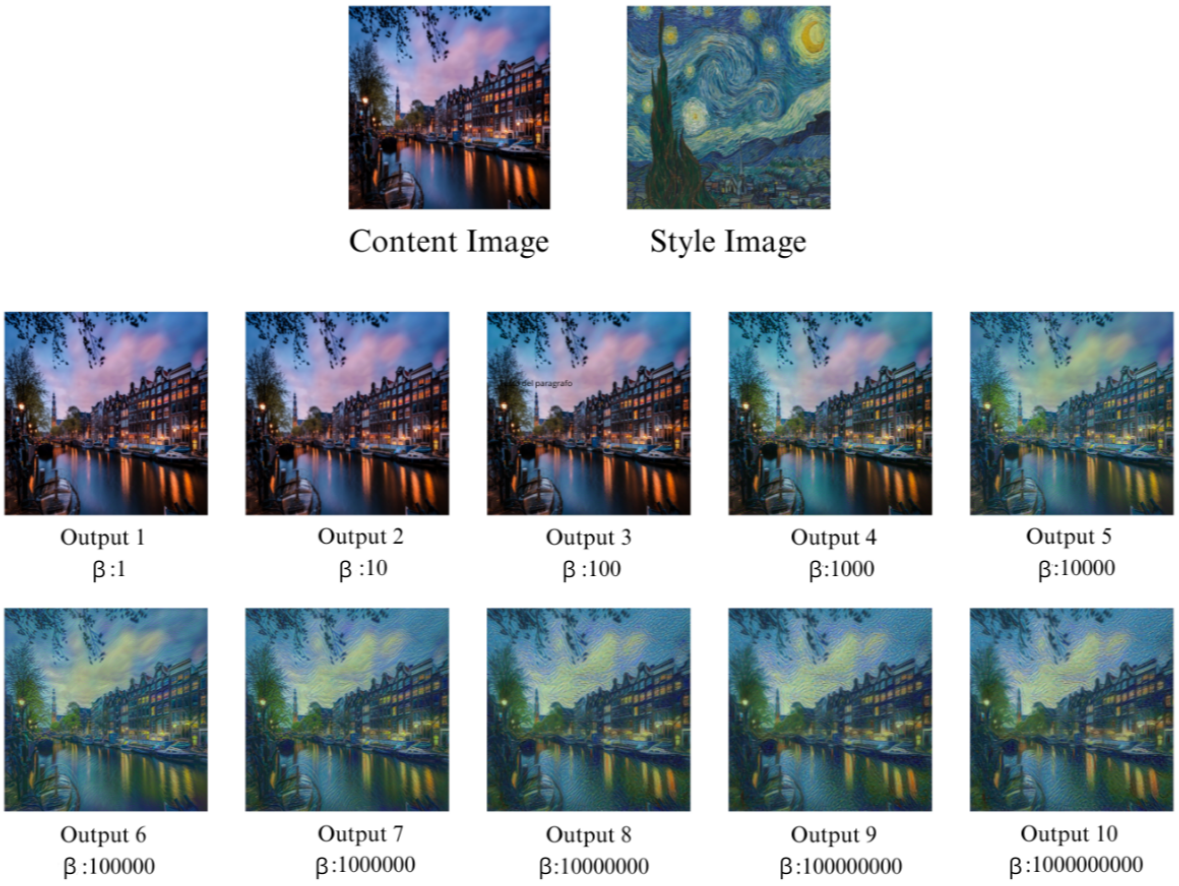

Fig. 1: Library for Style Preference Selection

images could have reached the initial total loss through non-unique style weights, adjusting the ratio through the content weight. Now, it is possible to define the suitability score (6) for each image of a gallery, given a content image.

$$\hat{\beta}_{ij} = \frac{\hat{L}_{total}}{L_{\text{style}_{ij}}} \tag{5}$$

$\hat{\beta}_{ij}$ is the optimal style weight to apply the style of the image "j" to the content image "i", and is computed as the ratio between the optimal initial total loss $\hat{L}_{total}$ and the initial style loss $L_{\text{style}_{ij}}$. A key factor here is that the image from which the model starts iterating is a replica of the content image, which means that in the initial total loss the content loss contribution is null, therefore the only weight influencing the initial total loss is the style weight. Had this not been the case, there would have been a situation where any pain of images can reach the initial total loss through multiple style weights, adjusting the ratio through the content weight. Now, it is possible to define the suitability score (6) for each image of a gallery, given a content image.

$$S_{\text{ij}} = \frac{1}{|\hat{\beta}_{ij} - \overline{\beta}| + \mu} \tag{6}$$

S_{ij} is the suitability score of style image "j" for the content image "i", and is computed as the inverse of the absolute difference between the desired style intensity ($\overline{\beta}$) and the optimal style weight ($\hat{\beta}_{ij}$). The addition of μ is needed to avoid the very unlikely scenario of an undefined score, in case of $\overline{\beta} = \hat{\beta}_{ij}$.

Given content images 1, 2, 3, 4, 5, in Fig. 2 (chosen to showcase a variety of styles and subjects), the Picasso gallery (Fig. 3), and the Cosmic gallery (Fig. 4); Tables I and II present the suitability scores, and indicate which style image would be the chosen one (marked in green), given the content image and a style preference (using the described framework). The content images are accordingly edited "Picasso style" and "Cosmic style", and the results are showcased in Fig. 5 and Fig. 6.

TABLE I: Suitability Score of the Picasso Gallery images for the 5 contents, $\overline{\beta}$:10^10

	PG 1	PG 2	PG 3	PG 4	PG 5	PG 6	PG 7	PG 8	PG 9	PG 10
CI 1	2.02e-10	4.60e-10	1.56e-11	2.86e-11	2.90e-11	6.68e-11	5.51e-11	1.45e-11	1.10e-11	2.54e-11
CI 2	1.19e-10	1.30e-10	1.23e-10	1.25e-10	1.23e-10	1.22e-10	1.26e-10	1.20e-10	1.24e-10	1.27e-10
CI 3	1.74e-10	1.95e-9	5.55e-11	4.40e-11	5.80e-11	1.13e-10	5.78e-11	1.22e-10	5.17e-11	4.60e-11
CI 4	2.04e-10	1.24e-8	1.55e-10	1.19e-10	1.93e-10	9.27e-10	2.10e-10	3.81e-10	1.36e-10	5.66e-11
CI 5	1.82e-10	2.84e-10	2.67e-10	3.33e-10	2.55e-10	2.37e-10	2.95e-10	2.07e-10	2.45e-10	2.67e-10

TABLE II: Suitability Score of the Cosmic Gallery images for the 5 contents, $\overline{\beta}$:10^10

	CG 1	CG 2	CG 3	CG 4	CG 5
CI 1	1.34e-10	3.96e-10	5.69e-11	2.73e-11	3.96e-11
CI 2	1.14e-10	1.21e-10	1.21e-10	1.23e-10	1.22e-10
CI 3	1.32e-10	3.27e-10	2.50e-10	7.58e-11	1.18e-10
CI 4	1.31e-10	3.59e-10	4.51e-9	1.28e-10	1.71e-10
CI 5	1.30e-10	2.15e-10	2.10e-10	2.99e-10	2.49e-10

Content Image 1 Content Image 2 Content Image 3 Content Image 4 Content Image 5

Fig. 2: Content Images

PG1 Woman with Watch | PG2 Still life with Lamp | PG3 Portrait of dora Maar | PG4 Weeping Woman | PG5 Self Portrait

PG6 Portrait of a Young Girl | PG7 Gilr before a Mirror | PG8 The old Guitarist | PG9 Le Domoiselles d'Avignon | PG10 Guernica

Fig. 3: Picasso Gallery

CG1 CG2 CG3 CG4 CG5

Fig. 4: Cosmic Gallery

Fig. 5: Picasso style application to the Content Images

IV. CONCLUSION

A. Evaluation and Observations

While the proposed approach is technically viable, its practical implementation reveals important limitations. The suitability-based distinction among style images proves too weak to produce a meaningful impact, as the suitability scores consistently exhibit low magnitudes. Although standardizing these scores helps from the interpretability point of view, it fails to effectively differentiate between varying levels of style intensity. This limitation likely arises from the significant disparity in scale between the optimal total initial loss and the style loss computed between paired images. As a result, the system struggles to handle style intensities outside the impactful range of 10^9 to 10^{11}. Specifically, when style intensities fall outside this range, the system has very little effectiveness, often favoring a single dominant style image in its gallery, thereby limiting the influence of user preference.

About the results in Table 5, which highlight PG 2 as the optimal style for most content, they appear to be coincidental rather than systematic, since this outcome is generated from a style weight preference that falls within the impactful range.

The approach is sensitive to even minor pixel differences, as this can significantly alter results. For transparency and reproducibility, all related materials and details are provided in Appendix A.

B. Exploration of alternatives

Addressing the issue of style preferences falling outside the impactful range presents an opportunity for improvement. A potential solution involves constraining the assessment of preferences exclusively to styles within the impactful range or, more specifically, within the bounds defined by the lowest and highest optimal style weights of images in a gallery. This narrower range could ensure more reliable results and better alignment with the system's capabilities.

Future research should also explore the potential efficiency gains of the proposed approach under alternative frameworks.

Fig. 6: Cosmic style application to the Content Images

These could include experimenting with different architectures, optimizers, learning rates, and iteration schemes. Frameworks with simpler overall characteristics might be likely to exhibit lower optimal initial losses. However, it is crucial to maintain a balance; overly simple frameworks risk failing to deliver satisfactory results for Neural Style Transfer (NST). Addressing this trade-off requires careful investigation to ensure the system remains both effective and efficient.

Additionally, pursuing this line of research might provide valuable insights into the interplay between user style preferences and framework selection. Such findings could guide recommendations for optimal frameworks based on specific user preferences, further enhancing the adaptability and usability of the approach.

Appendix

All the images used and the code are available at: https://github.com/ANDREAaNAPPI/An-Approach-to-Style-Image-Selection-in-Neural-Style-Transfer

References

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423. [Online]. Available: https://api.semanticscholar.org/CorpusID:206593710

[2] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural Style Transfer: A Review," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 11, pp. 3365–3385, 2020, doi: 10.1109/TVCG.2019.2921336.

[3] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative Adversarial Networks for Photo Cartoonization," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9465–9474, doi: 10.1109/CVPR.2018.00986.

[4] Q. Cai, M. Ma, C. Wang, and H. Li, "Image neural style transfer: A review," *Computers and Electrical Engineering*, vol. 108, p. 108723, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790623001477. doi: https://doi.org/10.1016/j.compeleceng.2023.108723.

[5] X. Gao and Y. Zhang, "SRAGAN: Saliency Regularized and Attended Generative Adversarial Network for Chinese Ink-Wash Painting Style Transfer," *Pattern Recognition*, vol. 162, p. 111344, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320325000044. doi: https://doi.org/10.1016/j.patcog.2025.111344.

[6] J. Chung, S. Hyun, and J.-P. Heo, "Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 8795–8805.

[7] X. Zhou, "Neural Style Transfer with Automatic Style Weight Searching," *Applied and Computational Engineering*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268480021.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14124313.

[9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106.

[10] Y. Nikulin and R. Novak, "Exploring the Neural Algorithm of Artistic Style," *ArXiv*, vol. abs/1602.07188, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:8800936.

[11] M. Wright and B. Ommer, "ArtFID: Quantitative Evaluation of Neural Style Transfer," in *Pattern Recognition*, B. Andres, F. Bernard, D. Cremers, S. Frintrop, B. Goldlücke, and I. Ihrke, Eds. Cham: Springer International Publishing, 2022, pp. 560–576.