Style transfer of computer-generated orthophoto landscape images into a realistic look

Nejc Krajšek[1] and Ciril Bohak[1]

Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113,
1000 Ljubljana, Slovenia
{nk7629@student.uni-lj.si, ciril.bohak@fri.uni-lj.si}
https://lgm.fri.uni-lj.si

Abstract. In this paper, we present a novel application of Neural Style Transfer (NST) for converting procedurally generated orthophoto landscape images into highly realistic representations. Traditionally used to apply artistic styles to images, we adapt NST to transfer the photorealistic qualities of real aerial orthophoto images to synthetic terrain images. This enables the creation of realistic visuals from generated landscapes, addressing the common issues of stylization, abstraction, and inaccuracy in synthetic imagery. Our approach involves upgrading and modifying existing NST techniques and their comparison. The evaluation demonstrates that our methods produce more convincing and realistic results than general generative models. These findings highlight the potential of NST in enhancing the realism of computer-generated landscapes, with possible applications in urban planning, environmental simulations, video games, and the film industry.

Keywords: neural style transfer, orthophoto images, terrain generation, texturing, 3D models, virtual worlds

1 Introduction

In recent years, deep neural network techniques have transformed the field of computer vision and machine learning, enabling the development of innovative approaches to generate and transform visual content in ways that go beyond previous limitations. One such use is image style transfer. It has mostly been used to transfer photos into stylized images in the style of a specific artist or to stylize images and or photos in a sketch-like style. However, there are other use cases arising such as one presented in this paper.

We present an approach for transferring the style of the generated landscape to a realistic appearance. This makes it possible to create images that reflect realistic surface properties from synthetically generated landscapes, which are often stylized, abstract, or simply too inaccurate. This has important applications in urban planning, environmental change simulations, video games, and the film industry where a high degree of visual fidelity is required.

The motivation for this research comes from the need to effectively bridge the gap between large data sources such as aerial imagery and the ability to generate realistic visualizations of landscapes with different characteristics for different applications. To address such needs, we can use a procedure called style transfer, which has recently been addressed by methods for Neural Style Transfer (NST) [7]. NST is a technique that uses the power of Convolutional Neural Network (CNN) to synthesize images that combine the content of one source image with the stylistic characteristics of another. Approaches mainly focus on transferring artistic styles between images, while style transfer between generated landscapes and realistic visualizations is an unexplored area.

This work investigates the potential of NST in solving the presented problem. By using NST in the field of generating orthophoto images, we want to obtain very realistic computer-generated images. These images can be used as a cheaper and faster alternative to real orthophoto images and can help reduce the space consumption of approaches that need large amounts of such images for learning or optimization. The newly developed approach based on Gatys*et al.* [7] was compared with photorealistic style transfer using wavelet transforms based on Yoo*et al.* [19].

The main contributions of this work are:

- NST, built on top of work by Gatys *et al.* [7], with adapted weights transferring photorealistic style to computer-generated landscapes.
- Evaluation of the developed approach and its comparison with photorealistic style transfer with wavelet transformations and an approach on computer-generated images of the terrain and on real topographic maps.

The presented research introduces a new way of bridging the problem of large data sources by using images of the earth's surface, which allows us to generate a large amount of data on the fly and transform it into a realistic look by transferring the style. This also allows us to prepare a collection of realistic data that meets certain criteria much faster than we would achieve by searching large collections.

2 Related Work

Neural Style Transfer (NST) refers to a class of algorithms that manipulate images or videos to leverage deep neural networks for image transformation, marking a significant advancement in non-photorealistic rendering. The concept of NST was initially explored by Efros and Freeman [6] and Drori *et al.* [4] with texture synthesis algorithms using image patches. These methods utilized pairs of input-output images (*e.g.,* a photograph and a stylized image) to learn a transformation applicable to new input images. While these methods produced satisfactory results and had short learning times, they struggled when aligned input-output pairs were unavailable. This limitation was partially addressed by Efros and Freeman [6] through quilting techniques, though the results were still constrained in handling more complex structures.

The formal introduction of NST as a deep learning technique was pioneered by Gatys *et al.* [7]. Their approach employed the *VGG19* network architecture presented by Simonyan and Zisserman [17], traditionally used for image classification, repurposed to extract important image features for style transfer. NST consists of two main components: content representation, achieved by extracting feature maps from the deeper layers of the network, and style representation, quantified using the Gram matrix. The final stylized image is obtained by minimizing a combined loss function that balances content and style losses, fine-tuned with user-defined parameters. Subsequent advancements have included image segmentation techniques to apply style transfer to individual objects within an image, as presented by Luan *et al.* [15]. Huang *et al.* [9] introduced arbitrary style transfer using adaptive instance normalization (AdaIN), which aligns the mean and variance of content features with those of style features. Mechrez *et al.* [16] proposed contextual loss for image transformation with non-aligned data, addressing the challenges of preserving spatial context in style transfer. Li *et al.* [14] explored the use of deep photo style transfer to maintain photorealism in stylized images. Yoo *et al.* [19] introduce wavelet corrected transfer based on whitening and coloring transforms WCT^2 that allows features to preserve their structural information and statistical properties.

While the original NST approaches were not real-time, significant progress has been made to accelerate the process. Johnson *et al.* [11] introduced a perceptual loss function combined with advanced CNN architectures, enabling real-time style transfer. This innovation led to the development of real-time camera filters, as demonstrated by Dumoulin *et al.* [5]. Chen and Schmidt [2] developed Fast Neural Style Transfer through instance normalization and residual learning, achieving real-time performance with high-quality results. Gatys *et al.* [8] also extended their initial work to video, introducing temporally coherent neural style transfer for smooth style application in video sequences.

The introduction of Generative Adversarial Network (gan)s further revolutionized style transfer by eliminating the need for paired images. Notable approaches include CycleGAN by Zhu *et al.* [21] and DiscoGAN by Kim *et al.* [12]. CycleGAN ensures coherent style transfer between different domains without paired examples, while DiscoGAN introduces reconstruction loss to maintain semantic relationships in the transferred styles.

Recent works in style transfer cover different scenarios, such as text conditioning presented by Kwon and Ye [13] and a recent survey by Jin *et al.* [10], transformers-based architecture for style transfer presented by Deng *et al.* [3] and unbiased image style transfer via reversible neural flows presented by An *et al.* [1].

3 Background

Neural Style Transfer (NST) utilizes CNNs to blend the content of one image with the style of another. A CNN is a deep learning model trained to recognize patterns in images through a hierarchy of layers. Lower layers detect basic features like edges and textures, while deeper layers identify complex structures such as objects and scenes. The NST process begins by passing a content image through a pre-trained CNN to extract its content features from deeper layers. A style image is similarly processed to extract style features from multiple layers, represented by the Gram matrix, which captures the correlations between different feature maps. NST aims to generate a new image that maintains the content of the input image and the style of the reference image by optimizing a loss function comprising two components:

1. *Content Loss*: measures the similarity between the content features of the generated image and the input content image, typically using the Mean Square Error (MSE).
2. *Style Loss* measures the similarity of the style features between the generated image and the style image, using the mean squared error of their Gram matrices across multiple layers.

The total loss is a weighted sum of the content and style losses, allowing control over the balance between content and style. Gradient descent is used to minimize this loss function. This iterative optimization algorithm adjusts the pixel values of the generated image to reduce the loss, gradually transforming the image to match the desired style while preserving the content. Through this process, NST produces an image that artistically merges the structural content of the input image with the stylistic elements of the reference image, showcasing the powerful capabilities of deep learning in creative applications.

4 Method

In this work, we examined two distinct approaches to style transfer, applying them to orthophoto images of computer-generated terrain and real topological maps.

For terrain generation, we utilized publicly available applications to rapidly and efficiently create diverse landscapes. Primarily, we employed World Machine Basic[1], which proved highly effective for generating terrains with intricate features such as vegetation, snow, water, and roads. For urban areas requiring greater accuracy, we used Esri ArcGIS City Engine[2].

In selecting textures, we decided on less realistic textures to evaluate the performance of different style transfer methods on simpler image data. This approach allowed us to comprehensively analyze and understand the efficacy of style transfer algorithms within terrain generation.

4.1 Classic NST

We first present our implementation of style transfer, which ranks as the simplest of the pair. Observing how implementing the classical NST method addressed

[1] World Machine: https://www.world-machine.com/

[2] Esri ArcGIS City Engine: https://www.esri.com/en-us/arcgis/products/arcgis-cityengine/

the challenge was insightful. Image preprocessing was crucial for the successful implementation of NST. This required resizing the images to uniform dimensions and normalizing the pixel values. Additionally, it was necessary to ensure that all images were in the RGB color space. These preprocessing steps were performed using the transforms module from the torchvision[3] library.

For the model architecture, we selected the *VGG19* network [17] (see Figure 1), known for its simplicity and efficiency. Although originally developed for object recognition, this architecture is highly suitable for NST due to its ability to capture various feature levels within images.

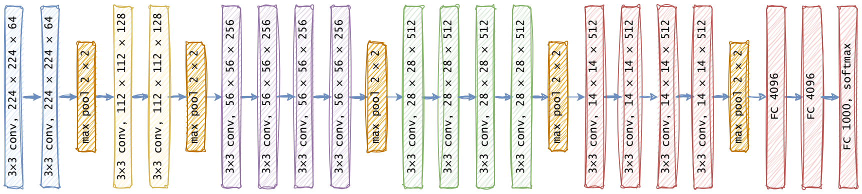

Fig. 1. *VGG19* architecture.

Determining the appropriate weights is a critical step in NST implementation, as they significantly influence the final output. Selecting suitable weights is time-consuming and requires extensive testing of multiple variations. This experimental process involves adjusting both the overall style and content weights and the individual layer weights to achieve a desired balance between content retention and style representation.

Generally, assigning more weight to deeper layers results in the generated image exhibiting more pronounced and complex stylistic characteristics of the reference image, potentially at the expense of content features from the original image. Conversely, focusing on the early layers highlights simpler stylistic features such as textures, edges, and basic shapes, preserving the content characteristics of the original image and producing a more recognizable representation.

To experimentally determine the best weights, we started with a balanced set of initial weights and iteratively adjusted them based on visual feedback. With our goal in mind, to create orthophoto landscape pictures that look convincingly realistic, we added specific constraints and objectives to our weight adjustment process.

For instance, if the generated image lacked fine details, we increased the weights in early layers to enhance textural details such as grass and water. This was crucial because these fine textures contribute significantly to the realism of landscape images. However, if the image appeared too noisy or overly detailed, reducing the early layer weights and increasing the deeper layer weights helped.

[3] Torchvision: https://pytorch.org/vision/stable/index.html

This adjustment ensured that broader stylistic features like lighting, color gradients, and overall atmosphere were well represented, contributing to a more cohesive and realistic look.

Moreover, we carefully balanced the content weight with the deeper layer weights to maintain clear and accurate structural details like the layout of landforms, roads, and buildings. Increasing the content weight helped preserve these crucial elements. In contrast, strategic adjustments to deeper layer weights ensured the style did not overpower the content, keeping the overall scene coherent and lifelike.

This iterative process involved running the NST algorithm with different weight configurations, visually evaluating the results, and making incremental adjustments until the desired balance was achieved. By carefully fine-tuning the weights, we emphasized textures and fine details where necessary while ensuring that the larger stylistic elements contributed to a lifelike appearance without overwhelming the content. Additionally, it was crucial to minimize artifacts to maintain the overall realism and quality of the images. Artifacts can detract from the visual coherence and authenticity of the landscape, so each adjustment aimed to enhance the style while keeping the image clean and free of distracting anomalies.

We used the originally proposed loss functions: MSE was used for content loss, the sum of MSEs of Grahm matrices for selected layers was used for style loss, and a weighted sum of content loss and style loss was used for total loss function.

The optimization process is iterative, performed over a specified number of steps. Each iteration incrementally refines the target image toward the desired artistic result. If the number of steps was too low, the model did not have enough iterations to fully integrate the style, resulting in incomplete or unsatisfactory images. On the other hand, an excessively high number of steps would lead to diminishing returns, where further iterations did not significantly improve the image quality and only increased computational costs. Control of the learning rate is crucial, as it dictates the step size in each optimization iteration. In our implementation, we employed the Adam optimizer with a learning rate of 0.01, which proved suitable for our purposes.

4.2 Style transfer using Wavelet Corrected Transfer

Style transfer using Wavelet Corrected Transfer (WCT^2) (see Figure 2) upgrades the basic style transfer algorithm and relies on whitening and coloring transformations. This advanced version allows efficient style transfer without the need for additional post-processing procedures such as smoothing and filtering, which helps reduce the algorithm's running time.

The wavelet transforms used in the algorithm are based on Haar transforms, which are used to group and ungroup information. These transforms include four cores (LLT, LHT, HLT, HHT) where low-pass and high-pass filters capture different aspects of the image, such as smooth surfaces, textures, and vertical, horizontal, and diagonal edges. An important property of Haar transforms is that

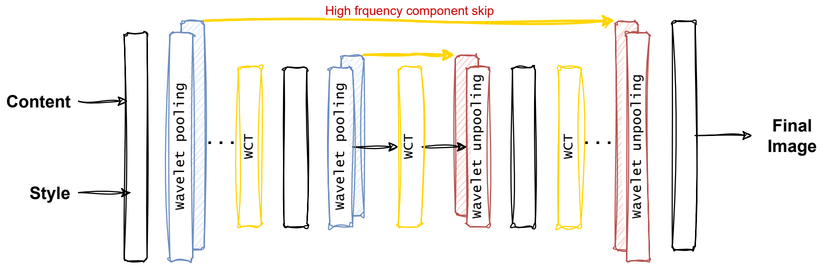

Fig. 2. WCT^2 architecture

the original signal can be accurately reconstructed, allowing image stylization with minimal loss of information.

The architecture of the WCT^2 model replaces all concatenation and deconvolution layers with wavelet transforms, while the basic architecture of the *VGG19* network remains unchanged. Layers 1 $(3 \times 3\ conv, 224 \times 224 \times 64)$ to 9 $(3 \times 3\ conv, 28 \times 28 \times 512)$ serve as the encoder and represent the image style, while the high-frequency components are directly transferred to the decoder. This allows only low-frequency components to be transmitted, leading to minimal information loss and preserving the original image's quality.

A special feature of the WCT^2 algorithm is the use of progressive stylization, which enables repeated use of the WCT^2 stylization. This simpler approach uses only one encoder-decoder pair, allowing for faster implementation.

However, with this method, artifacts may appear in the output image due to the amplification of errors during repeated execution of the stylization. Like the basic NST, adjusting the weights to our needs was necessary here.

5 Results

We evaluated the presented approaches on orthophoto data. We first present qualitative results and then quantitative evaluation. All the approaches were tested on NVIDIA GTX 1070 GPU with 8GB VRAM, Ryzen 3600 processor, and 16GB RAM.

5.1 Qualitative evaluation

In Figure 4, we show a comparison of content and style input images and images generated using NST and WCT^2. Content images are generated using Word Machine Basic from an orthophoto view. Style images were acquired using Google Earth[4]. The images contain diverse terrain and surface structures for content and style images to illustrate the models' adeptness in different conditions.

[4] https://earth.google.com

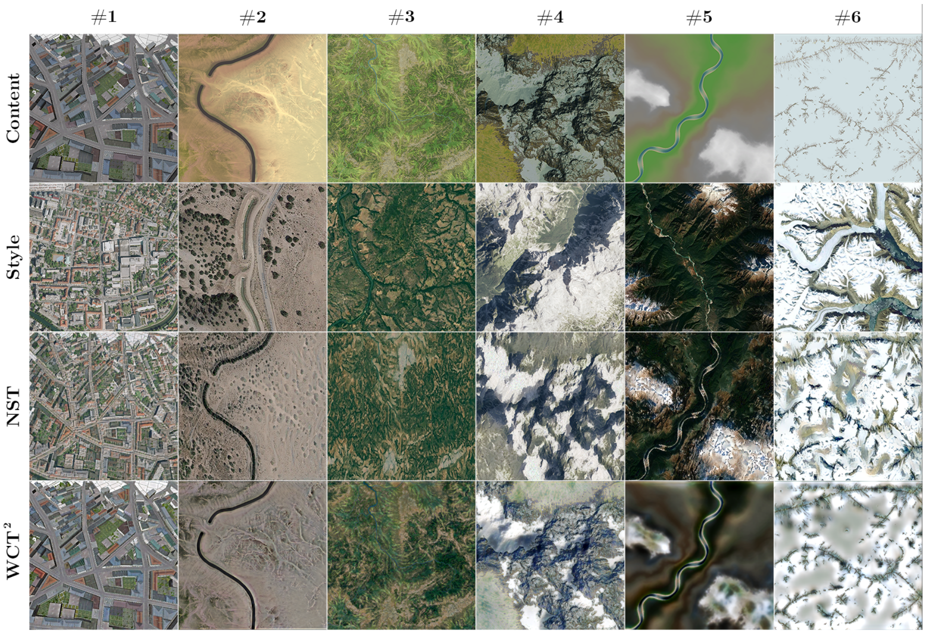

Fig. 3. Qualitative comparison of results from NST and WCT^2 and the input content and style images.

One can see that the results of both methods are quite similar. In most cases, NST images contain more high-frequency details from the style image, which reflects how we manually tuned weights to work best with orthophoto images. WCT^2 results preserve more structural details from the content images.

5.2 Quantitative evaluation

To make a quantitative evaluation meaningful, we generated images using map images as content and orthophoto images as style images. The pairs of images contained the same regions and were aligned.

For quantitative evaluation, we calculated image similarity metrics Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR) [18], and Learned Perceptual Image Patch Similarity (LPIPS) [20]. The results are presented in Table 1.

6 Discussion

The qualitative results show that both NST and WCT^2 approaches have their strengths and limitations. One key observation is that NST images tend to incorporate more high-frequency details from the input style images. This can be attributed to the manual tuning of weights to optimize performance with

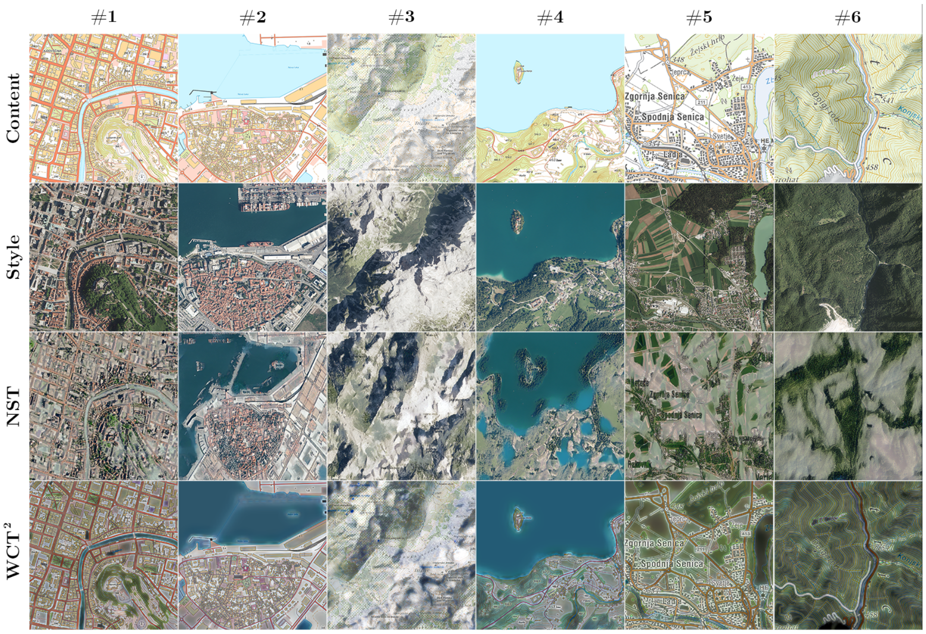

Fig. 4. Quantitative comparison of results from NST and WCT^2 where the content image is map and style image an aligned orthophoto image.

orthophoto images. This characteristic is particularly evident in Figure 3 for NST examples #2 and #3, where the generated images successfully capture textural details such as vegetation and water bodies. In contrast, WCT^2 tends better to preserve structural details from the input content images, maintaining a more coherent layout of landforms and man-made structures, as seen in examples #4 and #5.

Despite these strengths, both approaches exhibit challenges in reconstructing high-frequency details in certain scenarios. For instance, NST example #4 and WCT^2 examples #5 and #6 fail to retain sufficient detail, resulting in a less realistic appearance. This suggests that while NST is adept at capturing detailed textures, it may struggle with maintaining the overall structure, whereas WCT^2, although preserving structural integrity, might lose finer details due to its emphasis on low-frequency components during wavelet transformations.

Quantitatively, the evaluation metrics reveal that, on average, WCT^2 achieves better similarity metrics values for all metrics compared to NST, indicating a closer resemblance to the original orthophoto images (the content image). The results show that the methods' performance is on pair and that a more extensive study would be needed to draw more meaningful conclusions.

We can see from the resulting images that both methods have problems coping with the text in the content images. This could be addressed with additional preprocessing steps for text removal in content images using image inpainting.

Table 1. Quantitative comparison between the style transferred images of maps to orthophoto and the original orthophoto image using image similarity metrics SSIM, PSNR, and LPIPS.

	Neural Style Transfer (NST)			Wavelet Corrected Transfer (WCT^2)		
Example no.	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
1	0.087	11.383	0.434	0.115	11.801	0.494
2	0.260	11.626	0.533	0.354	12.899	0.495
3	0.115	9.013	0.589	0.123	9.212	0.701
4	0.441	14.474	0.588	0.606	12.274	0.387
5	0.200	11.846	0.678	0.250	12.118	0.560
6	0.172	11.317	0.720	0.232	14.612	0.722
mean	0.213	11.610	0.590	0.280	12.153	0.560

As demonstrated in this research, the adaptation and evaluation of the methods for orthophoto images expand the applicability of style transfer approaches to more realistic and practical domains such as urban planning and environmental simulations. Their application to synthetic terrain data showcases a novel use case with significant potential for improving procedural landscape generation.

7 Conclusion

In conclusion, this research demonstrates the feasibility and effectiveness of using NST and WCT^2 for transforming computer-generated orthophoto landscape images into realistic representations. The results underscore the potential of these techniques in various applications requiring high visual fidelity and open up new possibilities for future advancements in the field of realistic image synthesis.

References

1. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 862–871 (2021). https://doi.org/10.48550/arXiv.2103.16877
2. Chen, T.Q., Schmidt, M.: Fast patch-based style transfer of arbitrary style. arXiv preprint arXiv:1612.04337 (2016). https://doi.org/10.48550/arXiv.1612.04337
3. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11326–11336 (2022). https://doi.org/10.48550/arXiv.2105.14576
4. Drori, I., Cohen-Or, D., Yeshurun, H.: Example-based style synthesis. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 2, pp. II–143. IEEE (2003). https://doi.org/10.1109/CVPR.2003.1211464
5. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016). https://doi.org/10.48550/arXiv.1610.07629

6. Efros, A.A., Freeman, W.T.: Image Quilting for Texture Synthesis and Transfer. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. p. 341–346. Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/383259.383296
7. Gatys, L., Ecker, A., Bethge, M.: A Neural Algorithm of Artistic Style. Journal of Vision **16**(12), 326–326 (2016). https://doi.org/10.1167/16.12.326
8. Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling perceptual factors in neural style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3985–3993 (2017). https://doi.org/10.1109/CVPR.2017.397
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017). https://doi.org/10.1109/ICCV.2017.167
10. Jin, D., Jin, Z., Hu, Z., Vechtomova, O., Mihalcea, R.: Deep learning for text style transfer: A survey. Computational Linguistics **48**(1), 155–205 (2022). https://doi.org/10.1162/coli_a_00426
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016). https://doi.org/10.1007/978-3-319-46475-6_43
12. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International conference on machine learning. pp. 1857–1865. PMLR (2017). https://doi.org/10.48550/arXiv.1703.05192
13. Kwon, G., Ye, J.C.: CLIPstyler: Image style transfer with a single text condition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18062–18071 (2022). https://doi.org/10.1109/CVPR52688.2022.01753
14. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. Advances in neural information processing systems **30** (2017). https://doi.org/10.48550/arXiv.1705.08086
15. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4990–4998 (2017). https://doi.org/10.1109/CVPR.2017.740
16. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European conference on computer vision (ECCV). pp. 768–783 (2018). https://doi.org/10.48550/arXiv.1803.02077
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014). https://doi.org/10.48550/arXiv.1409.1556
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
19. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9036–9045 (2019). https://doi.org/10.48550/arXiv.1903.09760
20. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068

21. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2242–2251 (2017). https://doi.org/10.1109/ICCV.2017.244