Journal of Intelligent Systems and Internet of Things
Vol. 16, No. 02, PP. 117-122, 2025
DOI: https://doi.org/10.54216/JISIoT.160209

Research on Image Generation Style Transfer and Reconstruction Loss Reduction Based on Deep Learning Framework

Wei Zou[1], Mohd Alif Ikrami Bin Mutti[1,*]
[1]University of Science Malaysia, Penang, 11700, Malaysia
Email: Wei_Zou1@outlook.com; alifikrami@usm.my

Abstract

Nixi black pottery has a unique place in Chinese black pottery art. In this article, we have developed a style transfer model based on deep learning, which automatically transforms Nixi black pottery into images of other styles. This is of great value for the dissemination of this art. In this paper, we propose a method called DualTrans that utilizes a pure Transformer architecture to enable context-aware image processing, effectively addressing the issue of low receptive field. Additionally, we introduce a Location Information Encoding Module (LIM) and a Style Transfer Control Module (STCM) to tackle the problem of long-range dependencies while ensuring that the generated target image remains structurally and stylistically consistent throughout the style transfer process, without being influenced by the content and style images. During the mapping process, the LIM encodes the original image block information and concatenates it with the projected image block information. To alter the final produced style of the picture, the STCM leverages a set of learnable style-controllable factors. Extensive trials have shown that DualTrans exceeds previous approaches in terms of stability.

Received: December 10, 2024 Revised: February 08, 2025 Accepted: March 05, 2025

Keywords: Image Style Transfer; Transformer; Construction Loss; Art Style Transfer

1. Introduction

Deep convolutional neural networks (DNN) have been shown to be extremely successful in the field of visual style transfer. For example, generative adversarial networks (GANs) have demonstrated great effectiveness in a variety of tasks. [1] Unfortunately, GANs are notorious for their training instability, and significant efforts have been made to stabilize GAN training by introducing various regularization terms [3,4], better loss functions [2,5], and training techniques. However, convolutional operators have limited receptive fields, which means that neural networks struggle to capture long-range dependencies unless they have enough layers. This approach is inefficient and can result in the loss of fine details and feature resolution, in addition to making the optimization process more challenging. The Transformer architecture is a potential solution to address these challenges. Its capacity to represent structural features and extract similar structural information across multiple blocks allows it to efficiently capture long-term interdependence. Moreover, self-attention allows the Transformer to learn global information about the input [6], providing a comprehensive understanding at each Block. Because of this, the Transformer exhibits strong feature representation capabilities, which helps avoid the loss of fine details during feature extraction while preserving the overall structure intact.

Building upon the insights mentioned above, this paper proposes a fully non-convolutional dual-layer Transformer model for image style transfer. The specific network architecture is illustrated in Figure 1, aiming to enhance the information overlooked by CNNs during the style transfer process. Additionally, to have more control over the style of the produced images and incorporate positional information, we introduce the Style Transfer Control

Module (STCM) and the Location Information Module (LIM) into the model. In particular, the STCM considers style information to be modulators rather than statistical data, and it provides a set of learnable transfer control factors to alter the style of the generated pictures. The LIM further integrates individual information from the current image, allowing the network to perceive the differences between different images. In summary, our main contributions can be summarized as follows:

We propose a dual-layer Transformer model for image style transfer, which overcomes limitations of convolutional operators in capturing long-range dependencies and preserves fine details and overall structure during the stylization process.

We introduce the Style Transfer Control Module (STCM) and the Location Information Module (LIM) to enhance style controllability and incorporate positional information. The STCM utilizes learnable transfer control factors to alert generated image's style, while LIM integrates individual information to capture the differences between different images.
We undertake thorough experiments to show that our suggested method is successful and stable, exceeding existing approaches in terms of stylization quality, style controllability, and visual variety.

2. Related works

2.1 Statistics-based parametric method

Even though Gatys [8] were able to generate produce aesthetically attractive styled images, their iterative optimization process resulted in slow inference speed. To address this issue, Ulyanov [13] proposed an instance normalization model to significantly enhance the quality of image stylization as a replacement for batch normalization.

2.2 Image-Iteration-based style transfer

Deepdream [14] was the first to use convolutional neural networks to transmit style in 2015, displaying significant transfer effects. Image-iteration-based style transfer frequently gives superior outcomes in practical situations. This is since the convolutional neural network (CNN) model utilized is often pre-trained, eliminating the requirement to build a new model. However, because a picture must be created iteratively, there are issues with low efficiency and model generalization.

Difference of previous work: The Dual Trans model we propose is based on a pure Transformer architecture, effectively addressing the issue of potential detail loss during feature extraction by CNNs while preserving the structural integrity of the generated images. Additionally, we have designed two effective modules, the Location Information Module (LIM) and the Style Transfer Control Module (STCM), which not only capture long-term dependencies but also regulate the structure and style of the generated images.

The LIM module incorporates positional information into the model, allowing it to better understand the spatial relationships within the image and capture fine-grained details. This helps in maintaining the overall structure of the generated images.

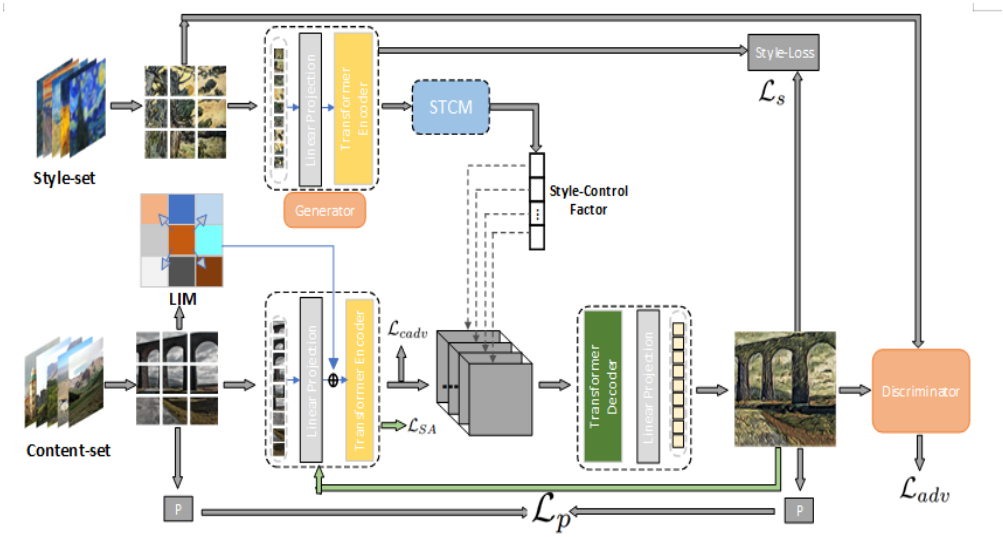

Figure 1. The structure of Dual Trans is presented, wherein the adversarial network utilizes TransGan.

3. DualTrans for image style-transfer

Firstly, Style transfer's purpose is to produce stylised graphics at any resolution. Significant differences in position encoding result in an exponential increase in positional deviations and poor picture resolution output quality. Given that, Transformers may capture long-range relationships; this research proposes a dual-network topology, as seen in Figure 1. The model is made up of three parts: a content Transformer encoder, a style controller, and a position decoder. The position encoder encodes picture long-range information in the content domain. On the other hand, the style controller utilizes TransGan to encode the desired style from the target domain and then integrates it with the BottleNeck in the main network. For the sake of clarity, let X and Y represent the content set and style set, respectively. As mentioned above, our objective is to learn the overall artistic style from Y, Specific artistic styles from y∈ Y may then be applied to any content image x∈ X to generate new images.

Style Adversarial Learning: We employ Transgene [15], currently the best-performing model, to match the output photos with the style images from Y. Generative adversarial network makes up of a generator G and a style discriminator $\mathcal{D}_s$, which compete against each other. The specific equation is as follows:

$$\mathcal{L}_{adv} = \mathop{\mathbb{E}}_{y\sim Y}[\log(\mathcal{D}_s(y))] + \mathop{\mathbb{E}}_{x\sim X}[\log(1-\mathcal{D}_s(D(E(x),\tau)))] \quad (1)$$

Where τ is the control factor that controls the style of the generated images. Similar to previous works, we utilize the final loss function of TransGan as a fixed artistic style loss. Inspired by [6], we formulate the artistic style loss as follows:

$$\mathcal{L}_s := \sum_{i=1}^{n} \|\mu(\phi_i(D(E(x),\tau))) - \mu(\phi_i(y))\|_2 + \sum_{i=1}^{n} \|\sigma(\phi_i(D(E(x),\tau))) - \sigma(\phi_i(y))\|_2 \quad (2)$$

Where μ and σ are the channel-wise mean and standard deviation. ϕ_i represents the specific block in TransGan used to calculate the perceptual style loss.

Style Transfer Control Module: previously [6], image synthesis was led by computing statistical transformations from pictures to content. In this work, we design a style transfer control module that enable automatically learn to generate parameters (styles) τ that contain style features from y:

$$\tau = SCB(\phi_{\text{Block}}(y)) \quad (3)$$

Where the output from the Block is used to control each generated image $E(x) \in \mathbb{R}^{B\times H\times W\times C}$ .This simple statistical style transfer has two advantages: (1) it eliminates the need for complex statistical computations, making the process more efficient, and (2) the style control factors can adaptively learn based on different style instances, allowing for more flexible and personalized style transfer.

Location Information Module: When using Transformer-based methods, to directly capture the structural information of the image, position encoding should be included in the input sequence. The attention factor between the i-th image block and the j-th image block can be represented as follows:

$$A_{ij} = \left(e_j \times W_k + p_j \times W_k\right) \times \left(e_i \times W_q + p_i \times W_q\right) \quad (4)$$

Where W_q. W_k are parameter matrices used for querying and computing. In the 2D case, the patch-to-patch relationship between patches at pixel positions $(x_i, y_i)$and$\left(x_j, y_j\right)$ can be represented as follows:

$$p(x_i,\ y_i)^T p\left(x_j, y_j\right) = \sum_{k-0}^{\frac{d}{4}-1}\left[cos\left(w_k \times x_j - w_j\right) + cos\left(w_k \times y_j - w_k \times x_j\right)\right] \quad (5)$$

Therefore, in this paper, we propose a position encoder that is scale and resolution invariant, and relevant to content semantics, making it more suitable for style transfer. $P_{LI}(x, y)$encoded information is as follows:

$$P_{LI}(x,y) = \sum_{k=0}^{S}\sum_{l=0}^{S}\left(a_{kl}F_{pos}\left(AvgP_{n\times n}(\varepsilon)\right)(x_k, y_l)\right) \quad (6)$$

The $AvgP_{n\times n}$denotes the average pooling, and F_{pos} represents a 1x1 convolution operation used as a learnable location information function. In this work, n= 18, a_{kl} is the interpolation weight, while S represents the number of neighboring slices.

Structure Reconstruction Evaluation: As mentioned above, during the style transfer procedure, we must also retain the original image's content. To meet this requirement, we assume that the image x is compared with the final result $D(E(x),\tau)$ to compute a forced reconstruction loss as follows:

$$\mathcal{L}_p := \mathop{\mathbb{E}}_{x\sim X}[\| P(D(E(x),\tau)) - P(x) \|_2^2] \quad (7)$$

However, the final output $D(E(P_{LI}.x)$ has a weak resistance to noise, making it easy for both the input and output to converge towards zero during gradient updates. Although $\mathcal{L}_{SA}$ decreases, this kind of update is meaningless. Therefore, we introduce another feature discriminator $\mathcal{D}_f$ with the following specific loss formulation:

$$\mathcal{L}_{\text{cadv}} = \mathop{\mathbb{E}}_{x\sim X}\left[\log\left(\mathcal{D}_f(E(x))\right) + \log\left(1 - \mathcal{D}_f(E(D(E(x),\tau)))\right)\right] \quad (8)$$

$\mathcal{L}_{\text{cadv}}$ attempts to reduce the disparity between the style input and the produced output distributions. Therefore, it enforces E to jump out of local minima during gradient updates.

The overall loss function of DualTrans is obtained by combining all the aforementioned loss functions. As a result, the final loss function of DualTrans is as follows:

$$\mathcal{L}_{\text{all}} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_s\mathcal{L}_s + \lambda_p\mathcal{L}_p + \lambda_{SA}\mathcal{L}_{SA} + \lambda_{cadv}\mathcal{L}_{cadv} \quad (9)$$

Where the above-mentioned hyperactive parameters are demonstrated in the following experiments.

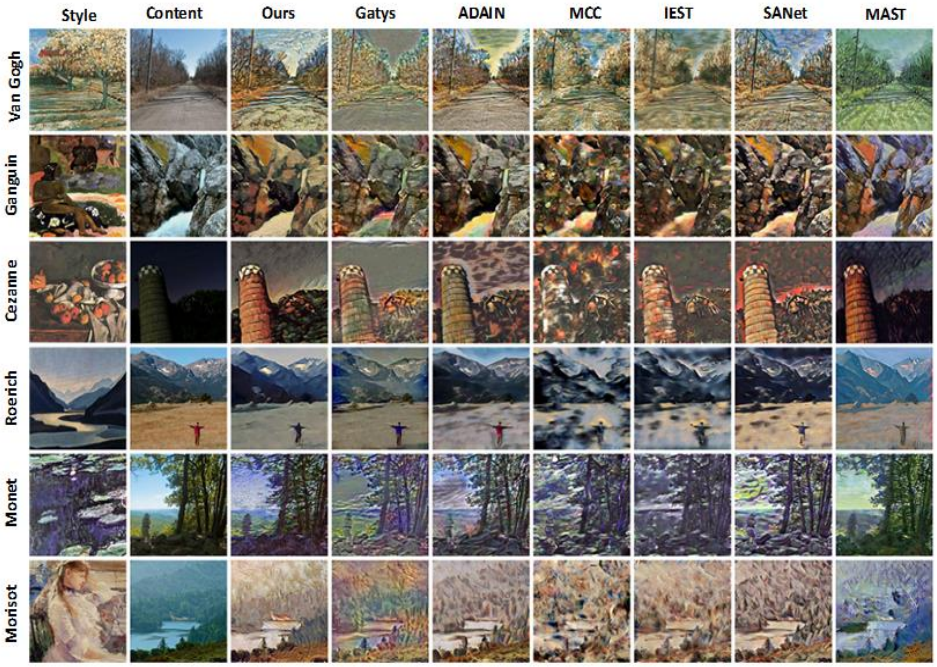

Figure 2. The performance of various methods.

4. Experiment and analyse

Datasets: In this work, we used two training ways to assess the performance of our suggested method :(1) The content dataset was MS-COCO [16], while the style dataset was Wiki-Art [17].

Baseline: We used AdaIN [9], SANet [10], AST [7], MCC [11], IEST [12] and Gatys [8] as baseline methods for comparison in our study. All the baselines were trained using their respective default configurations. These methods were chosen because they are SOTA approaches to style transfer and provide a strong foundation for evaluating the success of our proposed method.

Quantitative Analysis: To validate the efficiency of our suggested technique, we compared the findings to the above-mentioned baseline methodologies. Gatys, AdaIN, and SANet, for example, transfer styles from a single

style picture, resulting in aesthetically restricted image variants. MCC lacks in preserving the main content, leading to style confusion. IEST and MAST show deficiencies in expressing the content structure, tending to apply repetitive styles to stylized images. In contrast, our proposed DualTrans can simultaneously learn the style while preserving the fixed image structure, demonstrating good performance. As shown in Figure 2, DualTrans earns the best ratings and beats other approaches significantly, suggesting its supremacy in style transfer performance.

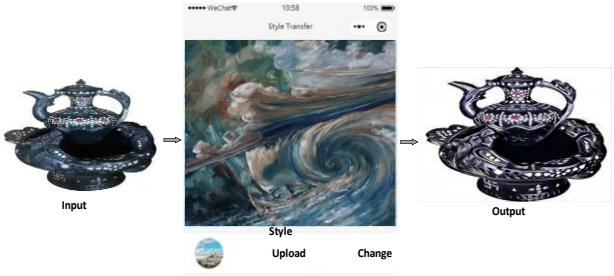

Figure 3. The application of WeChat mini program is demonstrated, where users can generate the final output by inputting content and style images, the style transformation effect of Nixi black pottery is also very good, which will effectively greatly enhance our dissemination of this culture.

Further Applications: In practical engineering applications, we believe that this technology can provide excellent APIs, such as style transfer for WeChat mini programs. The process can be outlined as follows:

(1) User uploads an image through the WeChat mini program.
(2) The image is transmitted to the server to be processed for style transfer.
(3) The processed image is displayed in the WeChat mini program.

The development process can be summarized as follows:

(1) To upload the picture and parameters, use the WeChat API call wx.uploadFile.
(2) Upon receiving the data, the backend server processes the image, performs the style transfer, and generates the resulting image.
(3) The generated result is saved in the WeChat development database for future searches before being returned to the client for display.

This workflow, as depicted in Figure 3, allows users to easily select images, perform style transfer, and view the generated results within the WeChat mini-program environment.

5. Conclusion

In this research, we offer DualTrans, a new style transfer technology that approaches the challenge of artistic style transfer from a novel angle. DualTrans' fundamental concept is to study both the general visual structure and distinct artistic approaches. We introduce a position information-encoding module during data processing, which effectively captures the image structure and enhances the model's understanding of it. Additionally, we incorporate a style transfer control module that further enables self-learning and facilitates style transfer according to the desired style images. This approach has the potential to improve the efficacy and stability of picture style transfer in future applications. The combination of learning picture structure and managing style transfer brings up new creative possibilities in the realm of artistic image alteration.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, et al., “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, et al., “Improved training of Wasserstein GANs,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[3] K. Kurach, M. Lučić, X. Zhai, et al., “A large-scale study on regularization and normalization in GANs,” in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 3581–3590.

[4] B. Zhou, A. Lapedriza, J. Xiao, et al., “Learning deep features for scene recognition using Places database,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[5] X. Mao, Q. Li, H. Xie, et al., “Least squares generative adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.

[6] Y. Deng, F. Tang, W. Dong, et al., “Arbitrary video style transfer via multi-channel correlation,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1210–1217.

[7] A. Sanakoyeu, D. Kotovenko, S. Lang, et al., “A style-aware content loss for real-time style transfer,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 698–714.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.

[9] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.

[10] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv: 1511.06434*, 2015.

[11] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Comput. Vis. – ECCV 2016*, vol. 9906, Springer, 2016, pp. 694–711.

[12] C. Li and M. Wand, “Precomputed real-time texture synthesis with Markovian generative adversarial networks,” in *Comput. Vis. – ECCV 2016*, vol. 9907, Springer, 2016, pp. 702–716.

[13] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6924–6932.

[14] A. Mordvintsev, C. Olah, and M. Tyka, “Inceptionism: Going deeper into neural networks,” 2015.

[15] Y. Jiang, S. Chang, and Z. Wang, “TransGAN: Two pure transformers can make one strong GAN, and that can scale up,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 14745–14758, 2021.

[16] T.-Y. Lin, M. Maire, S. Belongie, et al., “Microsoft COCO: Common objects in context,” in *Comput. Vis. – ECCV 2014*, vol. 8693, Springer, 2014, pp. 740–755.

[17] F. Phillips and B. Mackintosh, “Wiki Art Gallery, Inc.: A case for critical thinking,” *Issues Account. Educ.*, vol. 26, no. 3, pp. 593–608, 2011.