Research for the Development of Image Style Migration

Yitao Lin[a]
School of Petroleum, China University of Petroleum-Beijing at Karamay, Karamay, Xinjiang Province, 834099, China

Keywords: AdaAttN Model, CAST Model, StyTr2 Model, StyleID Model, StyleShot Model.

Abstract: As a cutting-edge image processing technology, the influence of image style migration technology has spanned multiple industries such as art, design, and advertising, demonstrating its strong creativity and application potential. With the emergence of image datasets and the proposal of various deep learning model networks, computer vision technology has entered a phase of rapid development. To provide a comprehensive and detailed overview of this rapidly developing research area, this paper introduces the methodology and architectural design of the current mainstream models in the field of image style migration. Through systematic review and analysis, the article elaborates on various types of models, including but not limited to those based on Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), Transformers, etc., as well as their variants and hybrid architectures. To comprehensively and objectively evaluate the performance of these models in complex style migration tasks, this paper introduces the Contrastive Language–Image Pre-training (CLIP) score and ArtFID score as key evaluation metrics. By combining the CLIP score and ArtFID score, this paper not only realizes the quantitative evaluation of the performance of various style migration models but also reveals their potential advantages and limitations in handling complex style migration tasks. In addition, the paper also discusses the significance of these evaluation results for model selection and optimization, as well as the possible directions of future research in terms of improving model performance and expanding application scenarios.

1 INTRODUCTION

Traditional image style migration methods mainly rely on basic techniques such as filters and color mapping, which make it difficult to cope with the representation of complex artistic styles. However, with the rise of deep learning technology, neural style migration has become an important breakthrough in solving this challenge. Gatys for the first time used deep convolutional networks to extract content and style features of images and achieved the fusion of content and style through optimization strategies (Gatys et al., 2016). This approach marked the entry of image-style migration into the deep learning era and greatly stimulated the research enthusiasm of the academic community in this field.

Compared to the neural style migration method of Gatys which mainly relies on global feature matching, the Adaptive Attention Normalization (AdaAttN) method introduces an adaptive attention mechanism that can adaptively adjust to different regions based on the content and features of the image (Gatys et al., 2016; Liu et al., 2021). Zhang proposed the Contrastive Arbitrary Style Transfer (CAST) model, an approach that completely abandons Gatys' traditional reliance on second-order statistics and instead focuses on learning and characterizing style styles directly from the high-dimensional feature space of images (Zhang et al., 2022). With the success of Transformer in Natural Language Processing (NLP), Transformer-based architectures have been used for a variety of visual tasks. Deng's proposed Style Transfer Transformer (StyTr2) model is based on the Transformer architecture, which, through the mechanism of self-attention, can naturally capture long-range dependencies and global contextual information in an image, which enables StyTr2 to better preserve and fuse the global semantic structure of an image during the style migration process and can effectively address the limitations of Gatys in this regard (Deng et al., 2021). Style Injection in

[a] https://orcid.org/ 0009-0001-8020-1018

Lin and Y.
Research for the Development of Image Style Migration.
DOI: 10.5220/0013511700004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 154-158
ISBN: 978-989-758-754-2

Diffusion (StyleID) is a style migration method based on a diffusion model. A technique for expanding the application of untrained style migration to large-scale pre-trained DMs using a diffusion model (Chung, Hyun & Heo, 2023). The diffusion process of StyleID is gradual, and thus style injection is incremental, allowing the model to fine-tune the image in multiple steps. This gradual diffusion makes the style migration more natural and smooth, avoiding the problems of style over-application or localized imbalance in traditional methods.

In addition, some approaches tend to decouple style features in the CLIP feature space, resulting in unstable style transfer performance (Liu et al., 2023; Ngweta et al., 2023). To solve the above problem, Gao developed a generalized style migration technique called StyleShot, which can capture any open-domain styles without the need for style adjustment during testing (Gao et al., 2024). This model is the first style migration method based on Stable Diffusion, which introduces a style recognition encoder and a content fusion encoder (Rombach et al., 2022). StyleShot not only captures and parses a unique artistic style from any reference image but also seamlessly applies this style to the content of the target image to generate a style that matches the specified style while maintaining the integrity of the content (Rombach et al., 2022). The high-quality stylized artwork produced is consistent with the specified style and retains the content's integrity. In particular, StyleShot eliminates the need for complex debugging or tweaking of each new style, increasing the flexibility and efficiency of style migration.

These models mentioned above represent the frontiers of current style migration techniques. Their appearance makes style migration evolve from the initial Convolutional Neural Network (CNN)-based approach to multi-style, arbitrary style, and real-time style migration, and further promotes the development of this field by introducing advanced techniques such as attention mechanism and diffusion model.

In this paper, the paper focuses on the diversified model architectures in the field of image style migration, and this research will systematically elaborate on the different innovations of each model, and then reveal their respective advantages and limitations through comparative analysis. On this basis, this paper also looks forward to the future development trend and potential research direction of image style migration technology.

2 INTRODUCTION TO THE METHODOLOGY

2.1 AdaAttN

AdaAttN is an attention and normalization module proposed to adaptively perform point-by-point attention normalization on images (Liu et al., 2021). To achieve accurate style migration, the core idea revolves around generating spatial attention scores through the learning of shallow and deep features of content and style images. The model integrates multiple AdaAttN modules at different layers of the Visual Geometry Group (VGG) network (ReLU3_1, ReLU4_1, and ReLU5_1) to fully utilize the features at different depths through a multi-level strategy. A style feature point is used to compute the perpoint weighting statistic by analyzing the attention-weighted output of all style features. Ultimately, the model normalizes the content features so that their local feature statistics match the computed per-point weighted style feature statistics, resulting in a more natural and fine-grained integration between style and content.

2.2 CAST

The CAST framework can compare arbitrary style transfers, unlike traditional neural deep learning, which implements learning styles directly from image features rather than through second-order statistics (Zhang et al., 2022). The CAST framework consists of three key components, namely, a multilayer style projector (MSP) for encoding the style code, a domain augmentation module for efficiently learning the distribution of styles, and a generative network for image style migration. Among them, the MSP module is based on the VGG-19 architecture, which integrates the style feature extractor with the multilayer projector, surpassing the limitations of single-layer or multilayer feature fusion in traditional methods.

2.3 StyTr2

To overcome the limitations of CNNs in global information extraction and maintenance, traditional neural-style migration methods are often limited by the problem of biased content representation. To address this challenge, StyTr2 is a Transformer-based framework that is incorporated into the long-range dependency of the input image into the image style migration process. StyTr2 overcomes the traditional

Transformer by introducing two customized Transformer encoders focusing on domain-specific sequence generation of content and style, respectively (Deng et al., 2021). limitations (Deng et al., 2021). The content sequences are stylized finely based on the style sequences by a multilayer Transformer decoder that is fed these sequences. The model identifies the drawbacks of current positional coding methods and suggests Content-Aware Positional Encoding (CAPE) that is scalable and better for image-style migration tasks.

2.4 StyleID

Chung proposed a novel approach to art style migration that relies on a pre-trained large-scale diffusion model and does not require any additional optimization (Chung, Hyun & Heo, 2023). In particular, the model alters the characteristics of the self-attention layer to resemble the cross-attention mechanism's operation. During the generation process, the key and value of the content image are replaced with the key and value of the style image, especially in the second half of the decoder associated with the local texture. In addition, to mitigate the problem of raw content interruptions, the model introduces query retention and attentional temperature scaling. By maintaining the query of the content image in a self-attentive mechanism, the backpropagation process can preserve the spatial structure of the original content. Attention temperature scaling aims to preserve the integrity of the content structure by dealing with the ambiguous self-attentive graph triggered by key substitutions. Finally, initial potential AdaIN corrects for color discordance by adjusting the statistics of initial noise in the diffusion model (Huang & Belongie, 2017).

2.5 StyleShot

The StyleShot model proposed by Gao is capable of capturing the style of any open domain and requires no additional style adjustments during testing (Gao et al., 2024). The core components of the model include a style-aware encoder and a structured style dataset called StyleGallery. The style-aware encoder is specifically designed for style learning and is capable of extracting richer and more expressive styles from reference images. To handle advanced styles (e.g., 3D, planar, etc.), the encoder employs a Mixture-of-Expert (MoE) structure to lightly extract multi-level image patch embeddings. In addition, StyleShot uses a content fusion encoder to further enhance image-driven style-based migration, and StyleGallery is a style-balanced dataset that provides a diverse style reference for models.

3 CONTRASTIVE ANALYSIS

3.1 CLIP Score and ArtFID Score Assessment

For model comparison and evaluation, I introduced the CLIP score and the ArtFID score (Hessel et al., 2021; Wright, & Ommer, 2022). The CLIP score captures whether the generated images successfully capture complex stylistic elements while maintaining content integrity. In addition compared to other metrics, CLIP has the highest correlation with human judgment. So it first used the CLIP score for the evaluation of each model. In addition, it used the ArtFID score as a comprehensive evaluation criterion for the models. The difference between this score and the CLIP score is that ArtFID mainly evaluates the content fidelity and style fidelity of images after image style migration. ArtFID is calculated as $(ArtFID = (1 + LPIPS) * (1 + FID))$. LPIPS assesses the content realism between the stylized image and its counterpart, and FID assesses the style realism between the stylized image and its counterpart. Meanwhile, it further introduces the Content Feature Structure Distance (CFSD) as a complement to provide a purer view of content fidelity evaluation.

For the CLIP score, it used StyleBench as a benchmark for style assessment (Gao et al., 2024). For the ArtFID score, it followed Chung's experiment for preparation (Chung, Hyun & Heo, 2023).

Table 1: Quantitative comparison of clip scores on image alignment with the SOTA image-driven style migration method (Gao et al., 2024).

CLIP	AdaAttN	CAST	StrTR-2	StyleID	StyleShot
Image ↑	0.569	0.575	0.586	0.604	0.660

Table 2: Comparison between different models of correlation values and CFSD values for ArtFID (Chung, Hyun & Heo, 2023)

Metric	AdaAttN	CAST	StrTR-2	StyleID
ArtFID ↓	30.350	34.685	30.720	28.801
FID ↓	18.658	20.395	18.890	18.131
LPIPS ↓	0.5439	0.6212	0.5445	0.5055
CFSD ↓	0.2862	0.2918	0.3011	0.2281

As can be seen from the table above:

StyleShot: CLIP scores the highest, and it is clear that this model achieves the best results in terms of correlation between image-text pairs.

StyleID: CLIP score comes second and ArtFID score has the lowest score among all the metrics in Table 2, indicating that among these models, the StyleID model achieves the best balance between content fidelity and style fidelity. It excels in the technique of generating images from text.

AdaAttN: performs better overall in the ArtFID score metrics, approaching StyleID in most of the metrics, and also has very good content and style fidelity metrics. However, the CLIP score scores the lowest, reflecting the model's technical shortcomings in text image generation.

StrTR-2: The ArtFID score is moderate, but is slightly weaker than AdaAttN due to the higher CFSD, which is influenced by the image style information, and the CLIP score is better than the ArtFID score, which suggests that the StrTR-2 model still has a place in the correlation between image-text pairs.

CAST: The poor performance in CLIP score and ArtFID score, and the highest score in ArtFID score composite metrics indicate a weaker performance in combined content and style retention, and poorer fidelity of the model's content and style. and poor ability to migrate text images.

3.2 Advantages and disadvantages of the models

Table 3: Advantages, disadvantages, and application prospects of StyleShot, StyleID, StrTR-2, CAST, and AdaAttN models. (Xin & Li, 2023)

Models	Advantages	Disadvantages	Applications
StyleShot (Gao et al., 2024)	Can directly capture styles from any open domain, simplifying the application process and improving user experience and operational efficiency. The model shows better generalization ability when dealing with diverse styles	For simple or flat styles, the model's representation may be relatively overly complex, with overfitting or unnecessary complexity issues.	StyleShot has a much wider range of applications and can be adapted to a wider variety of artistic styles. Examples include virtual reality and game design, movie post-production, advertising, and branding.
StyleID (Chung, Hyun & Heo, 2023)	This approach saves time and computational resources by eliminating the training or tuning steps. Reduced diffusion modeling is time-consuming in terms of style migration methods. The query retention mechanism and attention temperature scaling method further enhance the balance between style and content.	The generation process of the diffusion model often requires multiple steps for gradual denoising, which can lead to longer inference time and larger computational overhead.	Since no optimization is required, the model is ideally suited for on-the-fly artistic style transformation applications, such as artists generating multiple stylized images in a short period. Not only that, the model is also suitable for inter-frame style migration in video.
StrTR-2 (Deng et al., 2021; Strudel et al., 2021)	StyTr2 models two customized Transformer encoders to better capture long-distance dependencies in images. The Transformer-based framework overcomes the limitations of CNN in global information extraction and maintenance.	When processing high-resolution images, the computational and storage requirements of the attention matrix increase rapidly with high computational complexity.	Vision Transformer (ViT) can be used as a basis to make it extend to semantic segmentation application areas.
CAST (Zhang et al., 2022)	The CAST framework learns styles directly from image features. This approach captures the details and diversity of styles more comprehensively, avoiding the problems of style inconsistency and localized distortion that can result from traditional approaches.	framework's MSP module is based on the VGG-19 architecture, which means that it relies to some extent on the properties of that architecture.	Educational institutions can use the CAST model to create stylized teaching materials, such as presenting pictures of historical events in different artistic styles to attract students' attention

			and enhance memorization.
AdaAttN (Xin & Li, 2023; Liu et al., 2021; Huang & Belongie, 2017)	Key features can be captured efficiently, improving the retention of content semantic features. and localized details are more prominent compared to AdaIN.	Prioritizing the retention of semantic features may result in compromising stylistic elements, leading to a potential breakdown of stylistic coherence.	Generating high quality stylized images can be effectively improved by applying AdaAttN module. And it can be applied to edge detection algorithms with excellent performance.

4 CONCLUSIONS

As an important part of the computer vision field, image-style migration technology has made significant progress and demonstrated powerful creativity in several application scenarios. This paper comprehensively analyzes the design ideas, methodological innovations, and development stages of each model by systematically reviewing the current mainstream style migration models, including AdaAttN model, CAST model, StyTr2 model, StyleID model, and StyleShot model. Through comparison, it can be found that the models from only supporting a single art style image generation, to the current stage can realize any style of image generation. To compare the models, this paper introduces CLIP scores and ArtFID scores as the key evaluation indexes, enabling quantitative analysis of the model performance in complex style migration tasks. The advantages and limitations of the different models in handling diverse styles, as well as content and style fidelity, are revealed. Among them, StyleShot and StyleID models are strong contenders in the field. Both can generate superior images of arbitrary styles. By exploring these models in depth, this paper not only clarifies the challenges of the current technology but also points out the different application scenarios that can be carried out in the direction of future research. Finally, the hope is that this research will contribute to the advancement of computer vision and lead to more pertinent discussions.

REFERENCES

Chung, J., Hyun, S. & Heo, J., 2023. Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. ArXiv, abs/2312.09008.

Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L. & Xu, C., 2021. StyTr2: Image Style Transfer with Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11316-11326.

Gao, J., Liu, Y., Sun, Y., Tang, Y., Zeng, Y., Chen, K. & Zhao, C., 2024. StyleShot: A Snapshot on Any Style. ArXiv, abs/2407.01414.

Gatys, L.A., Bethge, M., Hertzmann, A. & Shechtman, E., 2016. Preserving Color in Neural Artistic Style Transfer. ArXiv, abs/1606.05897.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R. & Choi, Y., 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. ArXiv, abs/2104.08718.

Huang, X. & Belongie, S.J., 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1510-1519.

Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q. & Ding, E., 2021. AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6629-6638.

Liu, G., Xia, M., Zhang, Y., Chen, H., Xing, J., Wang, X., Yang, Y. & Shan, Y., 2023. StyleCrafter: Enhancing Stylized Text-to-Video Generation with Style Adapter. ArXiv, abs/2312.00330.

Ngweta, L., Maity, S., Gittens, A., Sun, Y. & Yurochkin, M., 2023. Simple Disentanglement of Style and Content in Visual Representations. ArXiv, abs/2302.09795.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B., 2021. High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674-10685.

Strudel, R., Garcia Pinel, R., Laptev, I. & Schmid, C., 2021. Segmenter: Transformer for Semantic Segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7242-7252.

Wright, M. & Ommer, B., 2022. ArtFID: Quantitative Evaluation of Neural Style Transfer. Proceedings of the German Conference on Pattern Recognition (GCPR).

Xin, H.T. & Li, L., 2023. Arbitrary Style Transfer with Fused Convolutional Block Attention Modules. IEEE Access, 11, pp. 44977-44988.

Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T. & Xu, C., 2022. Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning. ACM SIGGRAPH 2022 Conference Proceedings.