IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 May 2023, accepted 1 June 2023, date of publication 6 June 2023, date of current version 12 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283260

RESEARCH ARTICLE

Image and Video Style Transfer Based on Transformer

SUN FENGXUE[1], SUN YANGUO[2], LAN ZHENPING[1], WANG YANQI[1], ZHANG NIANCHAO[1], WANG YURU[1], AND LI PING[1]

[1]Electronic Information Department, Dalian Polytechnic University, Dalian 116034, China
[2]Information Center, The Second Hospital of Dalian Medical University, Dalian 116023, China

Corresponding authors: Sun Yanguo (563083592@qq.com) and Lan Zhenping (8091811@qq.com)

This work was supported in part by the Office of Liaoning Provincial Department of Education under Grant LJKZ0515 and Grant LJKFZ20220206.

ABSTRACT The essence of image style transfer is to generate images that both maintain in the original content image and present the effect with artistic features under the guidance of style images. Deep learning's quick rise has resulted in even more achievements in image style transfer, an otherwise popular study area. Nevertheless, due to the limitations of Convolutional Neural Networks (CNN), extracting and retaining the input images is problematic. Therefore, image style transfer based on traditional CNN is biased in the representation of content images. To address the above problems, this paper proposes STLTSF (Style Transfer based on Transformer), a transformer-based method that may achieve image style transfer based on the long-range dependencies of the input images. Compared with traditional visual transformers, STLTSF has two different transformer encoders, one for generating domain-specific content and the other for generating styles. We may convert the encoder to a decoding method that can be styled based on content sequences by using numerous layers of transformers. The suggested STLTSF approach outperforms traditional CNN-based methods in both qualitative and quantitative experiments.

INDEX TERMS Encoder, neural network, transformer, style transfer, STLTSF.

I. INTRODUCTION

Image style transfer is a difficult subject in both image processing and art. It is the process of transforming one image (called a content image) into the style of another image (called a style image). Deep Convolutional Neural Networks (DNN) have proven to be quite effective in the field of visual style transfer. DNNs can encode the content information of low-level images and extract the style information of images from high levels. The limitation of the convolutional operation field is that deeper convolution is necessary to capture long-distance dependencies during convolution. The image details will be significantly damaged, the resolution will be drastically lowered, and the quality will become exceedingly poor as the network depth increases.

In recent years, there are various prominent style transfer algorithms based on DNN [1], such as AdaIN (Adaptive Instance Normalization) [2], WCT (Whitening and Coloring Transformation) [3], MST (Multimodal Style Transfer) [4] and SEMST (Multimodal Style Transfer with Emphasis on Structure) [5]. Nevertheless, these algorithms adjust the content image using only one style or one technique, which tends to blur the foreground and background simultaneously.

It is well known that Transformer has been used with great success in the field of NPL (Natural Processing Language) [6] in recent years, and we can also see that Transformer-based architectures are starting to be applied in a variety of computer vision tasks. In this paper, we aim to eliminate the biased representation issue of CNN-based style transfer methods and propose a novel image Style Transfer Transformer framework called STLTSF. Different from the original transformer, we design two transformer-based encoders in our STLTSF framework to obtain domain-specific information. Following the encoders, the transformer decoder is used to progressively generate the output sequences of image patches. In summary, our main contributions include:

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

FIGURE 1. Effect of image style transfer.

- A transformer-based style transfer framework called STLTSF, to generate stylization results with well-preserved structures and details of the input content image.
- A content-aware positional encoding scheme that is scale- invariant and suitable for style transfer tasks.
- Comprehensive experiments showing that STLTSF outper-forms baseline methods and achieves outstanding results.

II. RELATED WORKS

A. NEURAL STYLE TRANSFER

The purpose of generic style transfer algorithms is to transfer arbitrary visual styles to content pictures. Early transfer algorithms involved color transfer and texture synthesis. Nonparametric sampling, texture matching, and edit propagation were among the color transfer and texture synthesis. With the development of deep neural network technology, it was applied to art style transfer and achieved better results. Gatys et al. [7] suggested an art transfer algorithm based on Convolutional Neural Networks. They discovered that the network's higher levels could convey style while the lower layers could describe content. The algorithm uses neural networks to separate and reorganize the content and style of arbitrary images. However, the iterative optimization process of this algorithm is slow. To solve the speed problem, Huang and Belongie [8] proposed an Adaptive Instance Normalization (AdaIN) layer based on the VGG-19 network, which aims to overcome the speed problem by aligning the mean and variance of content features with the mean and variance of style features for quick transfer.

Style transfer was defined by Li et al. as an image reconstruction approach mixed with feature change, specifically Whitening and Coloring Transformation (WCT). Zhang et al. [9] proposed a Multimodal Style Transfer (MST) method that explicitly analyzes matching local content aspects and semantic patterns in sub-style components. Inspired by MST, Chen [10] suggested a Structurally Emphasized Multimodal Style Transfer (SEMST) that captures structural information and automatically finds the ideal number of clusters, inspired by MST. The survey includes more neural-type transfer mechanisms.

B. IMAGE TO IMAGE TRANSLATION

Image-To-Image Translation [11] is the process of learning the mapping from an input image to an output image, for example, from an edge to an actual object.pix2pix [12] employs a conditional GAN-based network, which requires pairs of data for training. Unfortunately, in many applications, collecting paired data is problematic. As a result, approaches needing unpaired data have been investigated. Liu and Tuzel proposed Coupled GAN (CoGAN) [13], which learns the joint distribution of two domains by weight sharing. Later, Liu [14] extended CoGAN to unsupervised image-to-image translation. Recently, Zhu et al. [15] proposed the Cyclic Consistent Adversarial Network (CycleGAN), which can perform many visual and graphical tasks well.

Based on previous work, this paper introduces STLTSF, a unified image style transfer and video style transfer framework that allows transfer models to be well extended to new styles and contents. Unlike the original Transformer structure,this paper designs two transformer-based encoders in the STLTSF framework to obtain domain-specific information. The Transformer decoder is used to generate the output sequence of image patches step by step after the encoders. In addition to the natural language processing-based location coding method, this paper introduces a Content-Aware Position Encoding (CAPE)that learns location coding based on image semantic attributes and dynamically scales the location to fit changing image sizes.

The framework can also be used to video style transfer. The smoothness of the video is boosted when the framework is involved in video style transfer by adding temporal and spatial smoothing loss.

III. STLTSF CONVERTER-BASED APPROACH

A. IMAGE STYLE TRANSFER

The Transformer has made tremendous advances in the field of Natural Language Processing (NLP) in recent years and may be seen in a variety of vision tasks. First, the structure attributes of the Transformer-relational modeling allows it to extract similar structural information in different layers; second, the Transformer's ability to learn the global information of the input [16] under the guidance of the self-attentive mechanism [17] allows it to gain an overall understanding of the input at each layer. As a result, Transformer has a strong ability to represent features, which allows it to avoid the problem of detail loss that can occur when extracting features while keeping its created structure intact.

The image style transfer based on CNN has significant bias in content representation, so this paper offers a new image style transfer algorithm, namely STLTSF algorithm. In contrast to the previous ones, the STLTSF framework includes two transformer-based encoders [18] that are designed to acquire information on certain domains. The Transformer decoder is employed after the encoders to create progressively. Two considerations are presented for the suggested nonlinear programming location coding method.

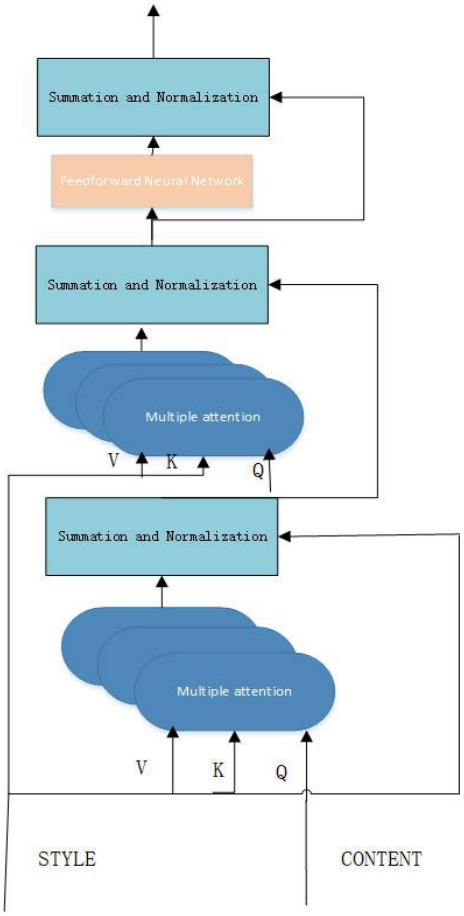

FIGURE 2. Coding layer of transformer.

Image sequence tags, unlike logically ordered sentences, are associated with semantic information about the image content.

First, the goal of the style transfer is to generate stylized images of any resolution. Significant changes in position encoding will come from exponential growth in image resolution, resulting in huge position deviations and poor output quality. Then as the Transformer may capture long-term dependencies, this paper develops a network structure as depicted in Figure. The model consists of three parts: the content Transformer encoder, the style Transformer encoder, and the Transformer decoder. The content Transformer encoder encodes the images' long-range information in the content domain. The style Transformer encoder, on the other hand, encodes information in the style domain [19]. The Transformer decoder is used to style the content features, transforming them into stylized visuals.

1) CONTENT-AWARE LOCATION CODING

When using the transformer-based model, the location encoding (PE) should be included in the input sequence to obtain the structural information. The attention score based on the i-th patch as well as the j-th patch is calculated as:

$$A_{i,j} = (\varepsilon_j \times W_k + \rho_j \times W_k) \times (\varepsilon_i \times W_q + \rho_i \times W_q) \quad (1)$$

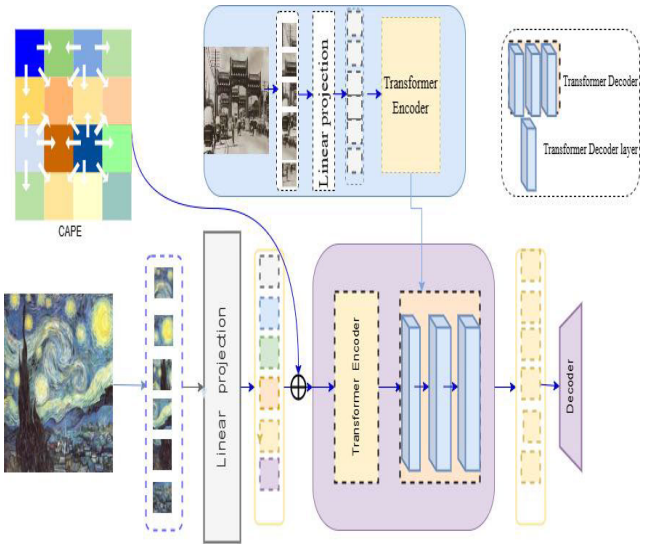

FIGURE 3. Network structure.

where W_q, W_k are the parameter matrices used to query and calculate, in the 2D case, the patch at pixel (x_i, y_i) corresponds to the position at pixel (x_j, y_j) as

$$\rho\,(x_i, y_i)^T\ \rho\,(x_j, y_j) = \sum_{k-0}^{\frac{d}{4}-1} \begin{bmatrix} \cos\left(w_k \times x_j - w_k \times y_j\right) \\ + \cos\left(w_k \times y_j - w_k \times x_j\right) \end{bmatrix} \quad (2)$$

Therefore, this paper proposes Content-Aware Positional Encoding (CAPE), namely location encoding with content-awareness [20], which is characterized by scale invariance, is related to content semantics, and is more suitable for style transfer. Its formula in $P_{CA}\,(x, y)$ is.

$$P_{CA}(x, y) = \sum_{k=0}^{s} \sum_{l=0}^{s} \left(a_{kl} F_{pos}\left(AvgPool_{n\times n}(\varepsilon)\right)(x_k, y_l)\right) \quad (3)$$

where $Avgpool_{n\times n}$ is the average pooling function and F_{pos} is the convolution operation of 1×1 used as the learnable position encoding function. In this experiment, n is set to 18, a_{kl} is the interpolation weight, and s is the number of neighboring face slices.

2) TYPE CONVERSION TRANSFER

a: TRANSFORMER ENCODER

Sequential visual representations are learned by using a transformer-based structure to capture the long-distance dependencies of image blocks. Unlike other vision tasks, the input to the style transformation task in this paper comes from two different domains corresponding to natural images and art paintings, respectively. Therefore, STLTSF has two transform encoders to encode domain-specific [21] features for transforming the sequences into domains in the next stage.

b: TRANSFORMER DECODER

The Transformer decoder is used to translate the encoded content sequence Y_s in a regressive manner based on the

TABLE 1. Time statistics of the three different methods.

resolution	Gatys et al.	AdaIN	SANet	MCC	IEST	STLTSF
256*256	14.17	0.018	0.015	0.013	0.065	0.116
*512*512*	46.75	0.065	0.019	0.015	0.092	0.661

encoding style sequence Y_c. Unlike in natural language processing tasks, all sequence facets are used as a single input to predict the output. As shown in Figure 2(a), each Transformer decoder layer contains two MSA layers and one FFN. The input to the Transformer decoder consists of the encoded content sequences,

$$\widehat{Y}_c = \{Y_{c1} + P_{C.A1}, Y_{c2} + P_{C.A2}, \cdots, Y_{cL} + P_{C.AL}\} \quad (4)$$

and the style sequences,

$$Y_s = \{Y_{s1}, Y_{s2}, \cdots, Y_{sL}\} \quad (5)$$

c: CNN DECODER

The output sequence of the transformer is in the shape of $X = (HW \times C)/64$. A three-layer CNN decoder is used to optimize the transformer decoder output instead of upsampling the output sequence to construct the final result. For each layer, the scale is extended by employing a series of operations, including 3×3 $Conv + \mathrm{Re}LU + 2XUpsample$ Finally, the final result with a resolution of $H \times W \times 3$ is obtained.

B. VIDEO STYLE TRANSFER

When the above methods are directly applied to video style transfer, the output is not very satisfactory in terms of smoothness. So this paper increases the smoothness of the video by adding smoothing loss on timing and smoothing loss on image space.

The smoothing loss on the image space is calculated by computing the average gradient of the generate image on the horizontal and vertical axes as a loss function

$$L_{tv} = Aver(grad(I_{cs})_h^2) + Aver(grad(I_{cs})_w^2) \quad (6)$$

Consistency on timing is mainly implemented in two ways, one type of association from the model perspective introducing voxels and the other type of restriction from the loss function perspective. In video style transfer, the more common method on the temporal loss function is to calculate the optical flow of the input image before and after frames, which uses the output image of frame t-1 and the optical flow to predict the image of frame t, and then calculates the difference with the output result of frame t. However, two issues were discovered when testing the loss function on homemade video datasets in this paper: first, the training time increases more, and second, the difference between images of adjacent frames after prediction using optical flow is sometimes larger than the difference calculated directly. For these reasons, this paper chooses the idea proposed in the paper [22] to use the difference of feature maps between adjacent frames directly as the loss in time series.

$$L_t = Aver \left\| \phi(I_{cs}^t) - \phi(I_{cs}^{t-1}) \right\|^2 \quad (7)$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASET

For the image style transfer process, this paper employs the MS-COCO S-COCO dataset as the content dataset and WikiArt as the style dataset.

And in video transfer, in order to improve the effect of the model on video transfer, this paper makes a special content image dataset using video sequences. Sixty-three short videos, all of which are below 30s in length, are randomly downloaded from video.net and unframed, of which 60 short video sequences are used as the training set with about 27000 images and 3 as the test set.

B. EVALUATION INDICATORS

The method of this paper is compared with AdaIN, SANet [23], MAST [24], MCC [25], IEST [26] and Gatys et al.

1) TIME COMPARISON

The speed with which images are generated is a critical evaluation factor in the field of image style transfer. The model was trained on two NVIDIA Tesla P100 GPUs and two NVIDIA GeForce RTX3090 GPUs for one day. Multiple methods' generation times at two resolutions were compared.

2) QUALITY COMPARISON

PSNR, SSIM, IS, and FID are common evaluation metrics. Other methods, such as pixel-by-pixel difference-squared summation to calculate L1 and L2 distances and manual scoring, are available (recruiting volunteers and picking actual image samples). PSNR [27] (peak signal-to-noise ratio) is the peak signal-to-noise ratio that is used to evaluate and measure the image's distortion and noise. The lower the distortion and the more realistic the generated image, the higher the PSNR value. The higher the initial score, the greater the diversity and quality of the generated image. Compared with the recent chatGPT [28] and GPTB [29], STLTSF leverages a transformer-based network, which has better feature representation to capture long-range dependencies of input image features and to avoid missing of content and style details. Therefore, our results can achieve well-preserved content structures and desirable style patterns.

3) ANALYSIS OF CAPE

When calculating PE, we should take the semantic information of content images into account. To compare the proposed CAPE with sinusoidal PE which is not semantics-aware, we show two cases where the input content image has repetitive patterns or is simply collaged by repeating one image four times.

Moreover, handling input resolution different from the training examples is generally challenging for a

TABLE 2. Comparison results on MS-COCO.

MS-COCO	PSNR.	IS
Gatys et al.	6.767	3.43
AdaIN	21.463	7.43
SANet	20.686	12.14
MCC	25.653	8.96
IEST	23.788	10.19
STLTSF	26.881	13.75

TABLE 3. Comparison results on S-COCO.

S-COCO	PSNR.	IS
Gatys et al.	7.834	5.32
AdaIN	25.085	25.28
SANet	24.017	13.04
MCC	26.262	18.95
IEST	24.859	23.76
STLTSF	27.866	37.43

learning-based method. To this end, an ideal PE for vision tasks should be scale-invariant, but a drastic change of image resolution leads to a significant difference in traditional PE. Different from learnable PE where the encoding is conditioned on the whole dataset, our CAPE dynamically encodes different input and thus can easily generalize to various resolutions.

C. EXPERIMENTAL RESULTS

In training, all images were randomly cropped to a fixed resolution of 256 × 256, while any image resolution was supported in testing. The Adam optimizer [30] was used and the learning rate was set to 0.0005 using a warm-up adjustment strategy. set the batch size to 8 and trained the network with 160,000 iterations.

First, 2000 images were randomly selected from MS-COCO and S-COCO datasets. Compare the image generation effect of STLTSF with other style transfer on different datasets. PSNR and IS were selected as evaluation metrics, and the test results obtained on the two datasets are shown in Table 2 and Table 3. It can be seen that the PSNR and IS values of the method in this paper are higher than other similar evaluation results, which can indicate that STLTSF generates more realistic images and better image diversity and image quality.

A comparison of image processing effects is also made. Because CNN-based image style transfer includes local style transfer, the approach presented in this paper is compared to AdaIN, SANet, MCC, and IEST as a whole (the effect of Gatys et al. is extremely different from the other methods, thus it is not compared here). As illustrated in Figure 3, the first method produces a more complex and realistic style transfer effect that accurately simulates the nuances of the style image. When compared to this method, STLTSF is specifically built on Transformer's network, which has excellent feature expression ability and can capture the long-standing relationships of the input image. The most notable benefit is that it can prevent the loss of style and content details. As a result, the style transfer achieved is of greater quality.

AdaIN, on the other hand, produces insufficient style patterns due to the alignment simplification of means and variances. Cracking abnormalities appear in the programmed images, affecting overall transmission quality. The SANet approach leads to delivering extremely fuzzy photos. The MCC employs a self-attentive transformation formula, but the lack of nonlinear operations restricts the network output's maximum value, resulting in overflow issues around object borders. IEST has a higher visual quality than the other approaches. However, the style of the generated results may differ from the style references in the input. As a result, in terms of image generation results, the method described in this study outperforms the other methods.

D. VIDEO STYLE TRANSFER

The experiments for video transfer were performed by fine-tuning the pre-trained model in the text, adding the loss functions L_{tv} and L_t, and using the dataset in 3.1 and the WikiArt dataset for training. The weights of the loss functions are $\lambda t = 2$ and $\lambda tv = 20$ respectively. The extraction method of the training data was also adjusted so that the randomly selected image would return its previous frame at the same time. The training time on NVIDIA Titan X is about 30 hours for 100000 iterations.

1) ADD LOSS FUNCTION

Due to the purpose of the original text being the style transfer of images, when directly applied to video style transfer, the output results are not very satisfactory in terms of smoothness. So we increase the smoothness of the video by adding smoothing losses in temporal and image space.

The smoothing loss in image space comes from [31]. The average gradient of the generated image on the horizontal and vertical axes is calculated as the loss function.

$$L_{tv} = Aver(grad(I_{cs})_h^2) + Aver(grad(I_{cs})_w^2) \tag{8}$$

By looking up the literature, we found that the consistency in time series is generally realized in two ways, one is to introduce voxel correlation from the perspective of model, and the other is to limit it from the perspective of loss function. In video style migration, the common method of timing loss function is to calculate the optical flow of the input image before and after the frame. It uses the output image of t-1 frame and optical flow to predict the image of t frame, and then calculates the difference with the output result of t frame.

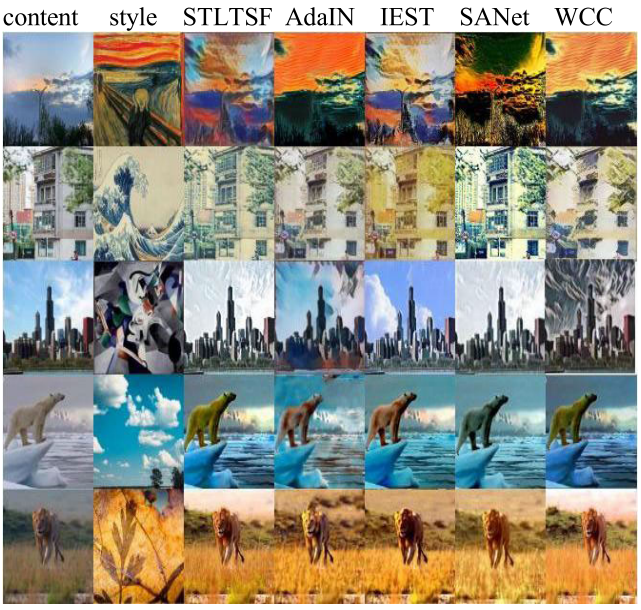

FIGURE 4. Comparison of image style transfer effects.

However, when we tested the loss function on the self-made video dataset, we found two problems: one is that the training time increased more, and the other is that the difference between adjacent frames' images predicted by optical flow is sometimes greater than the difference calculated directly. Based on the above reasons, we chose the approach proposed in paper [30], which directly uses the difference in feature maps between adjacent frames as the temporal loss. However, I personally feel that although this loss can ensure the smoothness of the video, it may also suppress any changes that should occur between frames.

$$L_t = Aver \left\| \phi(I_{cs}^t) - \phi(I_{cs}^{t-1}) \right\|^2 \tag{9}$$

To demonstrate the superiority of this method, this paper also compares this video transfer method with SCTNET, AdaIN, SANet and other methods.

As shown above, in SCTNET photos, the texture details of the land are blurred and show unwanted colors; AdaIN has non-negligible artifacts in almost all scenes and is relatively oversmoothed; and SANet is difficult to distinguish between the background or the active people when processing and migrating without difference. And the STLTSF results in this paper achieve effective preservation of realism and true details, maintaining the stylized effect while further eliminating distortion.

E. ANALYSIS OF STLTSF

STLTSF leverages a transformer-based network, which has better feature representation to capture long-range dependencies of input image features and to avoid missing of content and style details. Therefore, our results can achieve well-preserved content structures and desirable style patterns. Our method can significantly alleviate the content leak issue. Therefore, our model captures precise content representation

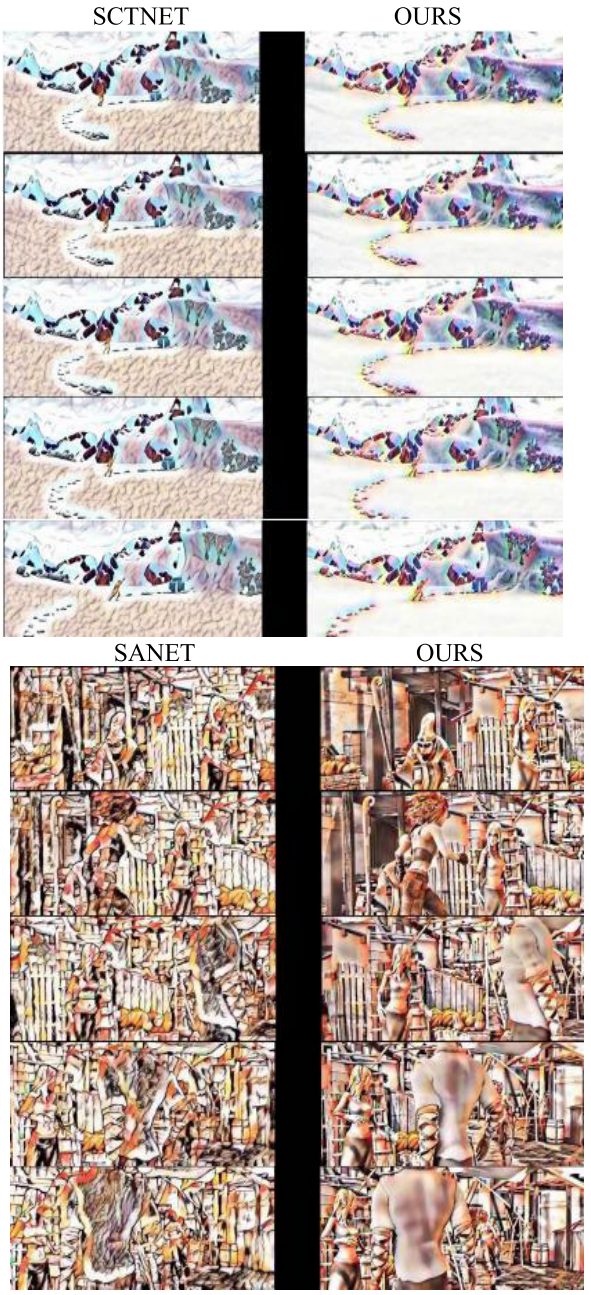

FIGURE 5. Comparison of video style transfer effect.

leading to superior style transfer results while effectively alleviating the content leak issue. At present, the test-time speed of our method is not as fast as some CNN-based approaches. Incorporating some priors from CNNs to speed up the computation would be an interesting future approach.

V. APPLET DASIGN AND DEVELOPMENT

As a photo processing product, providing a positive user experience is critical. WeChat applets are handy and have a large user base, thanks to the growing popularity of WeChat. As a result, the WeChat applet was chosen for this project's user interaction interface. After uploading the images selected by users through WeChat applet and using

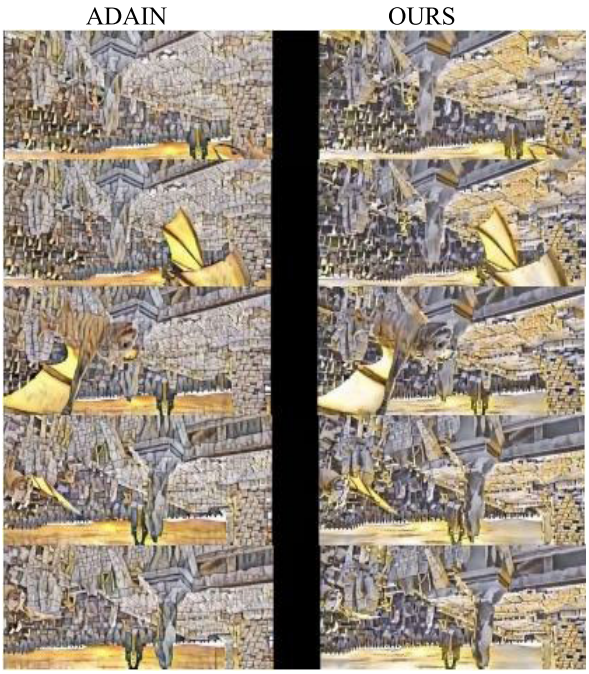

FIGURE 5. *(Continued.)* **Comparison of video style transfer effect.**

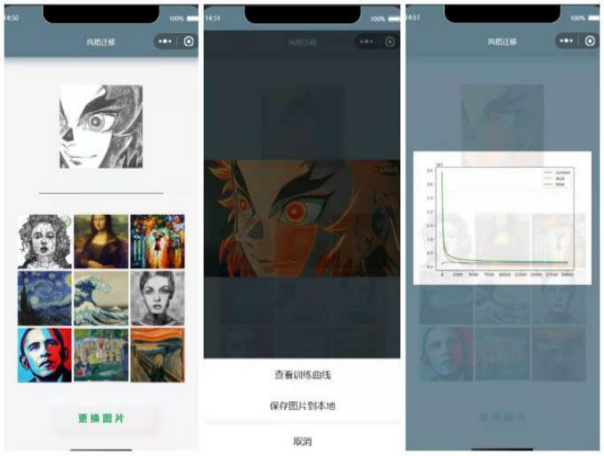

FIGURE 6. **The effect of small program implementation.**

the server for style transfer, the processed images are then returned to WeChat applet for display.

Image style transfer process, for the main page of the applet, users transfer the relevant style by uploading the corresponding photos. The specific process is as follows.

Enter the style transfer-select corresponding style-select photos and parameters-generate photos-complete the style transfer. WeChat small program side style transfer, using the WeChat API function wx.uploadFile(), to upload images and parameters; after the images are generated, then the generated results are stored in the WeChat development database for the history of the query.

VI. CONCLUSION

This paper introduces STLTSF, a Transformer-based image and video style transfer framework. For capturing domain-specific remote information, STLTSF comprises a content converter encoder and a style converter encoder. Transformer decoders are designed to convert content sequences into reference style sequences. A content-aware positional encoding scheme that is semantic-aware and suitable for scale-invariant visual generation tasks is also proposed. STLTSF alleviates the content leakage problem of CNN-based models and provides new insights into the difficult style transformation problem as a first baseline for style transformation utilizing visual converters. Due to the increase in system complexity, the training time of the system is relatively slow. Currently, the method's test rate is slower than that of some deep learning-based methods. This system cannot be applied to tasks with high training time requirements. However, it outperforms other methods in terms of image processing results. Furthermore, we apply the model to video style transfer and am astounded by the improved processing outcomes.

REFERENCES

[1] S. Bruckner and M. E. Gröller, "Style transfer functions for illustrative volume rendering," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 715–724, Sep. 2007.

[2] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[3] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1–7.

[4] E. Risser, P. Wilmot, and C. Barnes, "Stable and controllable neural texture synthesis and style transfer using histogram losses," 2017, *arXiv:1701.08893*.

[5] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[6] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 386–396.

[7] Z. Wang, L. Zhao, H. Chen, L. Qiu, Q. Mo, S. Lin, W. Xing, and D. Lu, "Diversified arbitrary style transfer via deep feature perturbation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7786–7795.

[8] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, "A closed-form solution to photorealistic image stylization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 453–468.

[9] T. Lin, Z. Ma, F. Li, D. He, X. Li, E. Ding, N. Wang, J. Li, and X. Gao, "Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5137–5146.

[10] J. An, T. Li, H. Huang, L. Shen, X. Wang, Y. Tang, J. Ma, W. Liu, and J. Luo, "Real-time universal style transfer on high-resolution images via zero-channel pruning," 2020, *arXiv:2006.09029*.

[11] M. Lu, H. Zhao, A. Yao, Y. Chen, F. Xu, and L. Zhang, "A closed-form solution to universal style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5951–5960.

[12] J. An, H. Xiong, J. Huan, and J. Luo, "Ultrafast photorealistic style transfer via neural architecture search," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 10443–10450.

[13] H. Wang, Y. Li, Y. Wang, H. Hu, and M. Yang, "Collaborative distillation for ultra-resolution universal style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1857–1866.

[14] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5873–5881.

[15] Y. Deng, F. Tang, W. Dong, W. Sun, F. Huang, and C. Xu, "Arbitrary style transfer via multi-adaptation network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2719–2727.

[16] Y. Deng, F. Tang, W. Dong, H. Huang, C. Ma, and C. Xu, "Arbitrary video style transfer via multi-channel correlation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 1210–1217.
[17] Y. Yao, J. Ren, X. Xie, W. Liu, Y. Liu, and J. Wang, "Attention-aware multi-stroke style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1467–1475.
[18] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 2001, pp. 341–346.
[19] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, "AdaAttN: Revisit attention mechanism in arbitrary neural style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6629–6638.
[20] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," 2021, *arXiv:2102.07074*.
[21] J. An, S. Huang, Y. Song, D. Dou, W. Liu, and J. Luo, "ArtFlow: Unbiased image style transfer via reversible neural flows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 862–871.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosu khin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 1–12.
[23] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," 2021, *arXiv:2105.07581*.
[24] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" 2021, *arXiv:2108.08810*.
[25] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 694–711.
[26] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 702–716.
[27] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "StyleBank: An explicit representation for neural image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2770–2779.
[28] X. Li, S. Liu, J. Kautz, and M. Yang, "Learning linear transformations for fast image and video style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3804–3812.
[29] S. Liu and T. Zhu, "Structure-guided arbitrary style transfer for artistic image and video," *IEEE Trans. Multimedia*, vol. 24, pp. 1299–1312, 2022.
[30] Q. C. Tian and M. Schmidt, "Fast patch-based style transfer of arbitrary style," 2016, *arXiv:1612.04337*.
[31] X. Wu and J. Chen, "Preserving global and local temporal consistency for arbitrary video style transfer," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1–9.

SUN FENGXUE was born in Changchun, Jilin, China, in November 1997. She is currently pursuing the master's degree with Dalian Polytechnic University, Dalian, China. Her research interest includes AI image processing technology.

SUN YANGUO received the M.S. degree in mobile communication from the Lanzhou University of Technology, Gansu, China. Since 2004, he has been the Director of the Information Center, The Second Affiliated Hospital of Dalian Medical University, and he has been engaged in hospital information construction for over ten years. His research interest includes medical image processing. He has been a member of the Remote Medical Informatization Professional Committee of the China Health Information and Health Big Data Society, a member of the Health Card Application Management Professional Committee of the China Health Information and Health Big Data Society, the Liao Executive Director of the Ning Provincial Health Information Society, an Executive Committee Member of the Informatization Professional Committee of the Liaoning Provincial Hospital Association, and a member of the Medical Information Branch of the Liaoning Medical Association.

LAN ZHENPING received the M.S. degree in mobile communication from the Lanzhou University of Technology, Gansu, China. Since 2004, she has been an Assistant Professor with the Communication Engineering Department, School of Information Science and Engineering, Dalian Polytechnic University, Dalian, China. She has published more than 30 journals. She has presided and participated in more than 20 vertical and horizontal projects, with a total of more than two million, one invention patent, and two provincial third prizes for scientific and technological achievements. Her research interests include deep learning and mobile communication.

WANG YANQI was born in Chaoyang, Liaoning, China, in May 1999. He is currently pursuing the master's degree with Dalian Polytechnic University, Dalian, China. His research interests include object detection and person re-identification.

ZHANG NIANCHAO was born in Weifang, Shandong, China, in May 1999. He is currently pursuing the master's degree with Dalian Polytechnic University, Dalian, China. His research interest includes deep learning.

WANG YURU received the M.S. degree in engineering. Currently, she is a Lecturer with the Automation Department, School of Information Science and Engineering, Dalian Polytechnic University, Dalian, China. Her research interests include embedded applications and intelligent control.

LI PING received the M.S. degree in communication and electronic systems from Harbin Engineering University, Harbin, China. Since 2005, she has been an Assistant Professor with the Optical Communication Department, School of Optical Engineering, Dalian Polytechnic University, Dalian, China. She has published more than 20 papers in publications, such as the 4th International Conference on Manufacturing Science and Engineering and international conferences, including eight included by EI. Her research interests include optical engineering and intelligent systems.