

# Supplementary Materials: Dual-head Genre-instance Transformer Network for Arbitrary Style Transfer

Anonymous Authors

## 1 IMPLEMENTATION DETAILS

### 1.1 Datasets

We utilize MS-COCO dataset [5] as the content and WikiArt dataset [3] as the style in our training process. The COCO dataset is a well-established benchmark in the computer vision, providing a wealth of content information. The WikiArt dataset comprises 81,444 artworks from diverse artists sourced from WikiArt.org, each annotated with genre labels such as Impressionism, Ukiyo-e, Pro-Impressionism, Realism, etc. During the test, we utilize the Pikip dataset [7] as our test data. Please note that there is no overlap between this dataset and our training set, and the test images remain unlabeled.

### 1.2 Network Information

Each transformer encoder includes a multi-head self-attention mechanism, LayerNorm, and a multi-layer perceptron (MLP) to ensure robust feature encoding. The Transformer decoder has four style decoder layers. Two of these layers are dedicated to merging instance-wise features with content, while the remaining two layers are responsible for refining the fusion feature. Followed by [1], each layer of the Transformer is equipped with multi-head attention mechanisms, where the feature representations are 512-dimensional, and the model utilizes 8 attention heads. The obtained sequence  $F_{cs}$  from the transformer takes the form of  $\frac{H}{16} \times \frac{W}{16} \times C$ . We employ a three-layer CNN decoder to generate the transformer decoder's outputs. Each layer's scale is expanded through a sequence of operations that includes two  $3 \times 3$  Conv, ReLU, and an Upsample process. The final output achieves a resolution of  $3 \times H \times W$ .

## 2 EXPERIMENTS

### 2.1 Additionally User Study

To comprehensively evaluate the effectiveness of our approach, we conducted additional user studies to compare it with existing state-of-the-art methods. Inspired by the user study in [2, 7], we have arranged comparison experiments to evaluate which method can generate the most favored results by humans. Our study involved a diverse group of participants, including 25 males and 25 females, ranging in age from 16 to 60. Each participant was presented with two groups of stylized results generated by our proposed method and one of the existing methods, respectively. Participants were expected to choose the image which best captured the style features from the style images while preserving the content details from the original content images. Finally, we collected 2,500 votes from 50 participants. The percentage of votes for each method is reported in Fig. 1. It demonstrated the superiority of our approach, as it consistently garnered higher preferences from the participants compared to other methods.

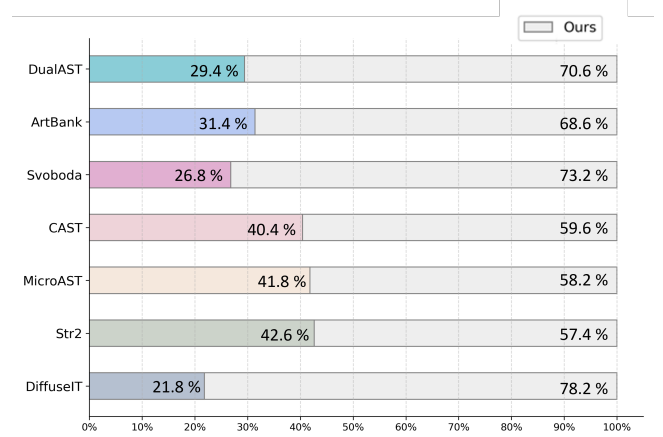


Figure 1: User study on the stylized results which exhibit the best performance between our method and one of the existing methods.

### 2.2 Comparison with Diffusion Models

In this section, we provide additional comparative results generated by the other diffusion-based method, namely, InST [6]. For your convenience, we herewith present more stylized results generated by our method, InST [6], DiffuseIT [4], and ArtBank [8] in Fig. 2. InST [6] proposes a textual inversion model to learn a single style image and transfer its style to the content image with a simple text prompt. ArtBank [8] presents an Implicit Style Prompt Bank to learn and store knowledge from the collection of artworks. In contrast, DiffuseIT [4] is an image-to-image translation approach that utilizes a pre-trained ViT model to guide the generation process of DDPM models in terms of preserving content structure. All the results in Fig. 2 are generated by the open-source codes provided by these authors.

Specifically, InST [6] needs to train different models for various topics, and requires users to provide texts to describe the style instead of directly using style images during the test. As depicted in Fig. 2 (d), the outcomes by InST [6] do not explicitly match the color distribution of the reference style images (e.g. the 1st, 2nd, and 4th rows), and there is a noticeable content change in the 3rd row. ArtBank [8] focuses on learning a range of artists and stores their features, which inherently limits the range of styles that users can choose from. When ArtBank [8] confronts an unseen artist's style during the test, the generated results towards the style of the artist that has been trained on (e.g. the 2nd, and 3rd rows). This behavior limits the model's generalization capability in arbitrary style transfer. DiffuseIT [4] is an image-to-image diffusion model, yet it still struggles with certain issues. As observed in Fig. 2 (f), DiffuseIT exhibits difficulties in retaining fine details, often leading

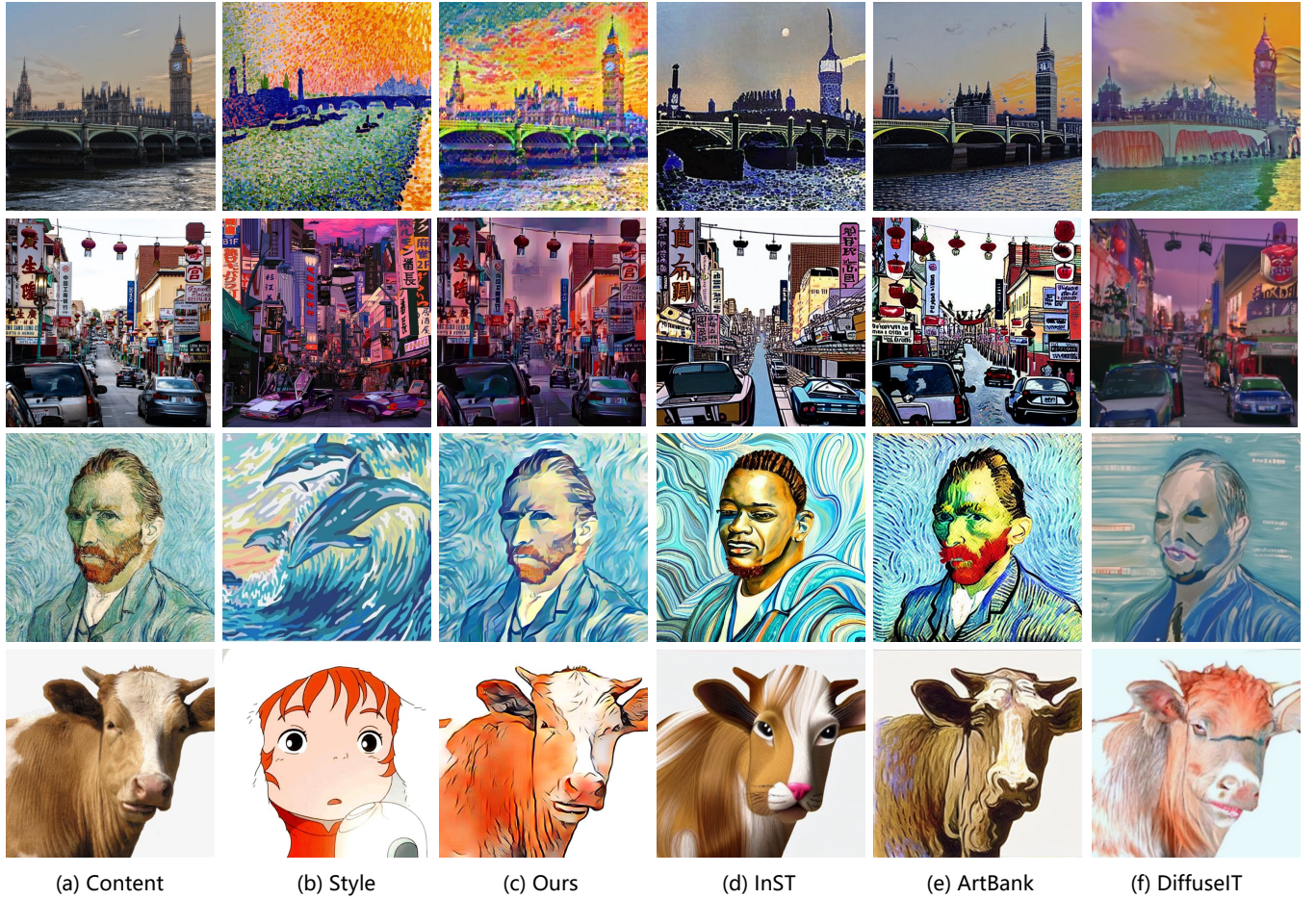


Figure 2: Qualitative comparisons. From the left to right: content, style, Ours, DiffuseIT [4], InST [6], and ArtBank [8].

Table 1: Quantitative comparison. Red indicates the best score.

Methods	Ours	InST [6]	ArtBank [8]	DiffuseIT [4]
SSIM $\uparrow$	<b>0.5849</b>	0.2022	0.3261	0.2487
LPIPS $\downarrow$	<b>0.4946</b>	1.3234	1.2076	1.9687
Style Loss $\downarrow$	<b>0.5714</b>	3.7631	3.8581	3.0719
Content Loss $\downarrow$	<b>0.9731</b>	2.4232	2.6591	3.0312
Inference Time	<b>1.78</b>	8.31	8.53	583.59

to blurriness (e.g. the 3rd row), and sometimes modifies the original features of the content (e.g. the 2nd and 4th rows).

Compared to these diffusion-based methods, our results can maintain higher fidelity color palettes and brushstrokes of the reference style image. As shown in Fig. 2 (c), our approach can ensure that the style elements are uniformly and coherently applied across the entire image while successfully preserving the overall outline and texture details of the content image. Moreover, we do not need users to provide text style descriptions, which allows arbitrary style transfer by directly utilizing style images, thereby offering

a more user-friendly and intuitive experience. The quantitative comparisons are shown in Table 1. Our method achieves the best performance in terms of SSIM, LPIPS, Style Loss, and Content Loss, demonstrating our superiority over the other methods. For inference, we measure the time for each method to process a single content-style pair using an NVIDIA A5000. Our method is faster than the other methods, especially surpassing the DiffuseIT [4] method.

### 2.3 More Results on Our Method

As illustrated in Figure 3, we present a series of additional results produced by our method, showcasing its adaptability across diverse styles. Even when presented with an artwork of an unknown genre, our method can still capture the artistic style of the piece and achieve arbitrary style transfer. These results emphasize our ability to preserve the structure and details of the content images while capturing the distinctive features of each style. Moreover, our method can still generate natural and coherent images even when faced with complex stylistic elements. This consistency is crucial for maintaining the visual integrity of the style transfer.





Figure 3: Our results on various styles. The reference style image is displayed in the lower right corner of the image.

## REFERENCES

- [1] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11326–11336.
- [2] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. 2020. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*. 2719–2727.
- [3] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. 2014. Recognizing Image Style. In *Proceedings of the British Machine Vision Conference*. BMVA Press.
- [4] Gihyun Kwon and Jong Chul Ye. 2023. Diffusion-based Image Translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. The International Conference on Learning Representations.
- [5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- [6] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10146–10156.
- [7] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–8.
- [8] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. 2024. ArtBank: Artistic Style Transfer with Pre-trained Diffusion Model and Implicit Style Prompt Bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7396–7404.