Article Information

Article Type:	research-article
Journal Title:	ACM Transactions on Multimedia Computing, Communications and Applications
Publisher:	ACM
ISSN (E):	1551-6865
DOI Number:	10.1145/3735137
Volume Number:	0
Issue Number:	0
First Page:	000
Last Page:	000

Multi-view Panoramic Image Style Transfer with Multi-scale Attention and Global Sharing

Left running head: W. Wang et al.

Right running head: Multi-view Panoramic Image Style Transfer

AQ1 AQ2 AQ3 AQ4 AQ5 AQ6

Weiyu Wang weiyu2021@gmail.com

Weiyu Wang weiyu2021@gmail.com

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China;

Pazhou Lab, Guangzhou, Guangdong, China

Pazhou Lab, Guangzhou, China

Alibaba Group, Hangzhou, China

Chunmei Qing qchm@scut.edu.cn

Chunmei Qing qchm@scut.edu.cn

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Pazhou Lab, Guangzhou, China

Junpeng Tan tjeepscut@gmail.com

Junpeng Tan tjeepscut@gmail.com

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Xiangmin Xu xmxu@scut.edu.cn

Xiangmin Xu xmxu@scut.edu.cn

School of Future Technology, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China

School of Future Technology, South China University of Technology, Guangzhou, China

Pazhou Lab, Guangzhou, China

Author Notes

This work is partially supported by the following grants: National Natural Science Foundation of China (61972163, U1801262), Natural Science Foundation of Guangdong Province (2022A1515011555, 2023A1515012568), Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004), Key Laboratory of Cognitive Radio and Information Processing (Ministry of Education, Guilin University of Electronic Technology) and Pazhou Lab, Guangzhou, China.

~~Authors' Contact Information: Weiyu Wang, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China; e-mail: weiyu2021@gmail.com; Chunmei Qing (corresponding author), School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China; e-mail: qchm@scut.edu.cn; Junpeng Tan, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China; e-mail: tjeepscut@gmail.com; Xiangmin Xu, School of Future Technology, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China; e-mail: xmxu@scut.edu.cn.~~

Authors' Contact Information: Weiyu Wang, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China; Pazhou Lab, Guangzhou, China and Alibaba Group, Hangzhou, China; e-mail: weiyu2021@gmail.com; Chunmei Qing (corresponding author), School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China; e-mail: qchm@scut.edu.cn; Junpeng Tan, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China; e-mail: tjeepscut@gmail.com; Xiangmin Xu, School of Future Technology, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China; e-mail: xmxu@scut.edu.cn.

Abstract

Style transfer for panoramic images is a challenging task, due to the problems associated with its unique structure, including edge discontinuities, pole distortion, fuzzy details, and memory limitation. In this article, we propose a novel Multi-view Transformation network for Panorama Style Transfer (MuTPST). First, this architecture has a multi-view panoramic transformation mechanism, which includes a multi-view cubic projection and a multi-view equirectangular re-projection of panoramic images. This can address pole distortion and edge discontinuity by skillfully applying multiple types of projections and transformations. To capture different levels of context and structure in the stylization stage, we carefully design a multi-scale attention content encoder, which can coordinate the distribution of visual attention across space and channels. Besides, by the sharing of global style features in thumbnails and patches, MuTPST can process ultra-high-resolution panoramic images (e.g., 10,000 × 5,000 pixels) with limited GPU memory. Extensive experiments illustrate that the proposed method outperforms the state-of-the-art with a discernible improvement in panoramic image style transfer. More results and interactive features can be found on https://weiyang001.github.io/MuTPST/.

CCS Concepts: • **Computing methodologies** → **Image processing;**

Additional Key Words and Phrases

Neural Style Transfer, Panoramic Images, Multi-view Transformation, Multi-scale Attention Content Encoder, Ultra-high Resolution

Funding

Funding support for this article was provided by the National Natural Science Foundation of China (61972163, U1801262) ,

Natural Science Foundation of Guangdong Province (2022A1515011555, 2023A1515012568) ,

Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004) ,

Key Laboratory of Cognitive Radio and Information Processing (Ministry of Education, Guilin University of Electronic Technology) ,

Pazhou Lab, Guangzhou, China.

1 Introduction

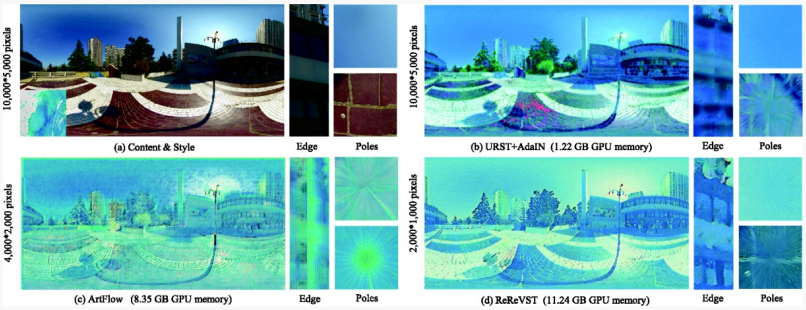
AQ7 AQ8 AQ9**Neural Style Transfer (NST)** is a very challenging task to automatically convert real-world scene images into high-quality artistic images [17]. Due to the broad application prospects, many efforts have been made to improve its performance. Current methods for NST can be roughly separated into two

categories: image-based optimization methods [10, 21] and model-based optimization methods [20, 24, 41, 47]. The former category is an image-based reconstruction technique that achieves style transfer by iteratively optimizing images. Benefiting from better time efficiency, model-based methods prevail in NST compared to image-based methods. The core idea of model-based methods is to apply feed-forward networks to the image reconstruction process. Due to the strong generalization capability, the instance normalization [38] and its variants [15, 27, 30] have significant performance. Some researchers have extended NST to video [42], portrait [50], text [49], and so on.

Recently, with the rapid development of **Virtual Reality (VR)**, panoramic images present attractive capabilities that cannot be achieved with traditional 2D images [26], such as capturing more scene information [14]. Normally, the panoramic image can be projected into 2D images for further processing. For example, the most common type of 2D projection for panoramic images is the equirectangular format. However, we cannot apply existing 2D NST methods directly to the equirectangular image because the resolution of equirectangular images is huge, which is necessary to produce the same visual impression as conventional images in all directions. In addition, the field of view in the equirectangular images is much larger than in the 2D images, which results in severe distortion of the area around the poles. Moreover, the equirectangular images also differ from traditional images in that the pixels on the left and right edges are continuous in content. However, the left and right edges may obtain different styles after stylization. These differences lead to the unsatisfying effectiveness of traditional NST on equirectangular images, as shown in Figure 1. From this figure, it can be seen that (1) it is difficult to solve the problems of pole distortion and edge discontinuities caused by style transfer because of the over-the-horizon boundary limit of panoramas; (2) the lack of sufficient consideration of different levels of semantic structure results in unpleasant localized content distortion; (3) it is expensive in terms of memory consumption and cannot handle real-world VR scenes stylization.

Fig. 1. Unsatisfying effectiveness of traditional NST on equirectangular images. It can be observed that due to the unique structure of the panoramic image, existing methods cause pole distortion, edge discontinuities, blurred details, and memory limitation.

Alternate Text:



Note: This figure has been annotated.

To solve these problems, this article proposes a novel **Multi-view Transformation Network for Panoramic Image Style Transfer (MuTPST)**. The pipeline of the MuTPST can be divided into three stages: projection, stylization, and re-projection. In the projection stage, **Multi-view Cubic Projection (MCP)** is proposed to obtain different view cube combination images. In the stylization stage, a **Multi-scale Attention Content Encoder (MACE)** and dynamic inter-channel filters of the **Global Feature-Sharing Module (GFSM)** are used to perform **High-resolution Style Transfer (URST)**. In the re-projection stage, stylized multi-view cube combination images are converted to panoramic images via **Multi-view Equirectangular Re-projection (MER)**. The contributions of this article can be summarized as follows:

(1) A novel MuTPST is proposed, which is an end-to-end network and can be divided into three stages: projection, stylization, and re-projection.
(2) For conversion of panoramic images, MCP and MER mechanisms are presented. This can eliminate over-the-horizon pole distortion and edge AQ10 discontinuities AQ11 without boundaries.
(3) Two novel modules (MACE and GFSM) are presented to reduce the impact of redundant information and obtain finer details to generate stylized high-resolution images with clear textures and appropriate colors.
(4) Extensive experiments illustrate that the proposed MuTPST can generate higher visual quality of stylized panoramic images compared with existing works.

2 Related Work

2.1 2D Image Style Transfer

The 2D image style transfer has been attracting the attention of many researchers. There are plenty of studies exploring how to automatically convert real scene images into artistic works [35, 40, 45, 54]. Inspired by Gatys et al. [10] who first proposed a AQ12CNN-based style transfer algorithm, Johnson et al. [18] designed a feed-forward network that directly generates artistic images and enables real-time style transfer. By expanding on the above approach, Li et al. [23] proposed multi-scale **Whitening and Coloring Transformation (WCT)** to apply feature modification globally. Huang and Belongie [15] proposed **Adaptive Instance Normalization (AdaIN)** in which the mean and variance of the style image feature are matched to the content image feature.

More recently, an **Adaptive Attention Normalization (AdaAttN)** module is presented in [27] to learn the spatial attention score about visual quality. An et al. [1] use an unbiased style transfer network and reversible neural flows and to mitigate content leaks. By taking into account the long-range dependencies of the content images, Deng et al. [9] offered a transformer-based method to mitigate biases by content representation. Hamazaspyan and Navasardyan [13] introduced the diffusion mechanism into image region matching and preserved the fine-grained features of the content image. Zhang et al. [52] believe that artwork is unique in that it cannot be adequately explained in normal language and offers a framework for style transfer based on reverse reasoning.

Nowadays, 2D image style transfer has demonstrated excellent results, but due to significant variations in space structure and resolution size, panoramic image style transfer is still challenging.

2.2 High-Resolution Image Style Transfer

Some research works have been studied on how to perform high-resolution image style transfer under GPU memory limits. An et al. [2] alleviate memory pressure using a lightweight network pruned from GoogLeNet [37]. Jing et al. [16] developed a lightweight network based on MobileNet, lowering computing difficulties significantly. Wang et al. [39] proposed a knowledge distillation framework based on the encoder-decoder pairs, significantly reducing the computation complexities. Recently, Wang et al. [44] proposed a lightweight framework for high-resolution image style transfer.

Previous efforts focused on increasing the processable resolution by compressing the network size and decreasing the number of parameters. However, Chen et al. [7] perform patch-wise style transfer with special thumbnail instance normalization and partly avoid the GPU memory problem caused by high-resolution images.

At present, there are some methods for high-resolution image style transfer. However, how to take into account the stylization effect while processing high-resolution images is still a problem worth investigating.

2.3 Panoramic Image Style Transfer

So far, many style transfer models on panoramic images have improved based on traditional 2D images. There are fewer researchers in this field than in traditional 2D images. Depending on how they are taken, panoramic images can be divided into stereoscopic panorama and monocular panoramic images.

For stereoscopic panoramic images, Chen et al. [4] proposed a disparity loss to penalize the bidirectional disparity based on a framework structure similar to their earlier video style transfer network. Gong et al. [11] proposed a method for photo-consistent stylization for stereo pairs, which produces more consistent strokes for different views.

Considering the wide range of applications, current image processing algorithms for panoramic images mainly focus on monocular panoramic images. Ruder et al. [32] studied the boundary constraints of the cubic projection of panoramic images and developed a spherical image processing network. Zhang [51] noticed the problem of equirectangular image boundary continuity and proposed a method to mitigate edge discontinuity. Noting the problem of the corresponding saliency area being scattered, Xiang et al. [48] proposed a panoramic image style transfer network by multi-attention fusion based on AdaAttN.

Although these methods for monocular panoramic images have achieved some progress, they lack a systematic analysis of the panoramic images and cannot simultaneously solve problems including pole distortion, edge continuity, and high-resolution image processing.

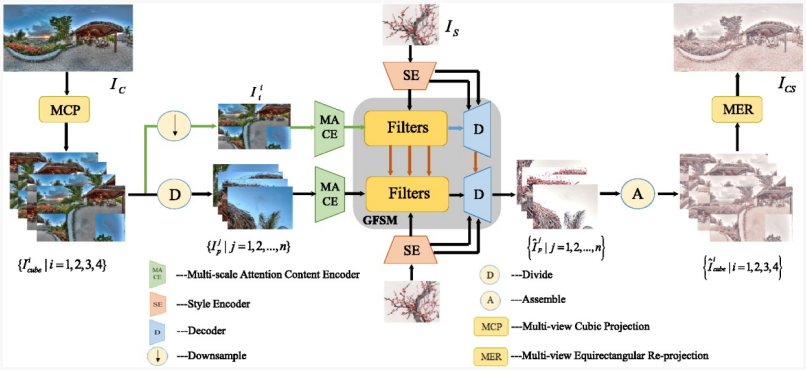
3 Method

3.1 The Network Architecture

The framework of MuTPST is illustrated in Figure 2, which is proposed to overcome the distortions and GPU memory limit caused by the special structure of panoramas. It consists of three key designs: (1) A series of flexible transformations named **Multi-view Panoramic Transformation (MPT)** to mitigate pole distortion and edge discontinuities. (2) A novel MACE that can capture refined features and reduce the appearance of artifacts. (3) A carefully designed GFSM that shares style features of thumbnail images and patches, overcoming memory constraints. Benefiting from these ingeniously designed structures, our MuTPST can efficiently handle panoramic images.

Fig. 2. The framework of proposed MuTPST. The pipeline is separated into three stages: projection, stylization, and re-projection. The core is that the panoramic image will perform a series of transformations, and then the thumbnail style transfer and the patch-wise style transfer will be carried out successively.

Alternate Text:



Note: This figure has been annotated.

As shown in Figure 2, MuTPST takes the high-resolution equirectangular images I_C as inputs. It consists of three key designs:

(1) In the projection stage, the content image I_C is executed MCP process and obtains a sequence of cube combination images $\{I_{cube}^i | i=1,2,3,4\}$.

(2) In the stylization stage, the cube combination image I_{cube}^i is downsampled to the thumbnail image I_t^i. I_t^i is fed to the MACE and GFSM, and the global style features are collected. Then, the cube combination image is divided into a series of small patches $\{I_p^j | j=1,2,\ldots,n\}$. Specifically, we employ a sliding window for segmentation with adjustable size and stride. The collected global style features are applied to GFSM, and produce the stylized patches $\{\tilde{I}_p^j | j=1,2,\ldots,n\}$. Finally, all stylized patches are assembled into stylized cube combination image $\{\tilde{I}_{cube}^i | i=1,2,3,4\}$.

(3) In the re-projection stage, the sequences of stylized cube combination images are executed by MER and generate the final stylized image. Inspired by [34], multi-skip connections are added between the style encoder and decoder to obtain a finer style.

3.2 MPT

In VR environments, panoramic images are projected onto a spherical surface centered at the viewer's position, thereby enabling a 360° viewing experience of the virtual content. For the convenience of compression and transmission, panoramic images are typically stored in the format of 2D image.

AS the most widely applied projection method, equirectangular projection maps the scene information from the spherical surface onto a rectangular plane. The formula for mapping a panoramic image from a Cartesian coordinate system to a cylindrical coordinate system using equirectangular projection is as follows:

$$\begin{cases} x = r(\varphi - \varphi_0)\cos\theta_0 \\ y = r(\theta - \theta_0) \end{cases} \quad (1)$$

where x and y represent the horizontal and vertical coordinates in the Cartesian coordinate system, r is the radius of the sphere, θ is the longitude coordinate, φ is the latitude coordinate, and θ_0 and φ_0 are the manually set reference coordinates for longitude and latitude, respectively. The mapped image encompasses 360° of content in the horizontal direction and only 180° in the vertical direction. It is crucial to note that while this calculation assumes coordinate values to be continuous

variables, pixel coordinates in actual computer storage are discrete values. Consequently, directly applying the formula will lead to an inability to completely fill the rectangular image, thus requiring the use of interpolation techniques to achieve complete coverage. Due to the presence of θ_0, the originally continuous content is interrupted at θ_0, resulting in the segmentation of complete objects. Furthermore, the polar regions are forcibly stretched to the upper and lower boundaries of the rectangular image, resulting in significant distortion and thereby increasing the complexity and challenges of panoramic image style transfer.

Cubic projection involves overlaying an outer cube that is circumscribed around the sphere, with its side lengths equal to the sphere's diameter. Subsequently, linear mapping is employed to project the content from the spherical surface onto the six patches of the cube. The formula for mapping the panoramic image from a Cartesian coordinate system to a cylindrical coordinate system in cubic projection is:

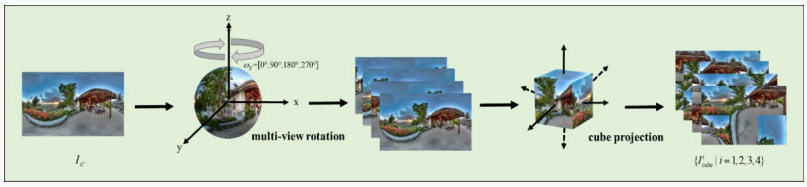
$$\begin{cases} x = \cos\varphi \sin\theta \\ y = \sin\varphi \cos\theta \\ z = \cos\theta \end{cases} \tag{2}$$

where all variables are defined in the same way as in equirectangular projection. The formulas show that each of the six faces of the cubic projection exhibits some certain degree of deformation, but the degree of distortion is significantly lower than that at the poles in equirectangular projection.

To eliminate the severe content distortion in the polar region and to mitigate the edge discontinuity after stylization, the MPT is proposed. MPT can be composed of MCP and MER, which can be easily inserted into most existing methods. As shown in Figure 3, MCP performs multi-view rotation and cubic projection before the beginning of the stylization stage, while in Figure 4, MER performs equirectangular projection and multi-view fusion after ending the stylization stage.

Fig. 3. The structure of proposed MCP. The sphere means projecting the panoramic image into the shape of a sphere, while the square means projecting the panoramic image into a cube.

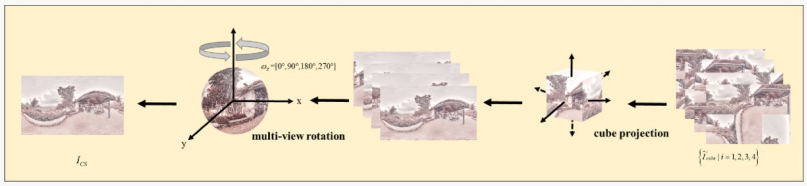
Alternate Text:



Note: This figure has been annotated.

Fig. 4. The structure of proposed MER. The final result was generated through a series of inverse projections, rotations, and transformations.

Alternate Text:



Note: This figure has been annotated.

The details of multi-view rotation in the MCP step are as follows. To mitigate edge discontinuities, multi-view rotation is designed delicately. ω_x is a group of angles, which means the angles that the sphere image rotates along x-axis of each rotation operation, e.g., $\omega_x = \{\omega_{x1}, \omega_{x2}, \ldots, \omega_{xk}\}$, where k is the number of the rotation. In this framework, we set $\omega_x = \{0°, 90°, 180°, 270°\}$. Taking an equirectangular image as input, MCP first projects it into the spherical format, and the sphere image rotates ω_x with x-axis $m \in \{1, 2, \ldots, k\}$ sequentially. After a series of rotation operations, k equirectangular images can be obtained. With reasonable settings, the content of the edges has a chance to appear in the center of the image, which greatly alleviates the discontinuity at the edges. Then, to reduce the severe content distortion in the polar region, equirectangular images are projected into a cubic format which uniformly maps the spherical panoramic images onto the six faces of a cube and

combines them into one image. Note that combining the cubic projection surfaces in the order of $[left, front, right; back, bottom, top]$ can ensure maximum pixel continuity. In contrast, in the MER step, ω_z is the same as MCP's, but the direction of rotation is opposite to MCP. Finally, these images are fused by linear combination, resulting in the stylized result.

3.3 MACE

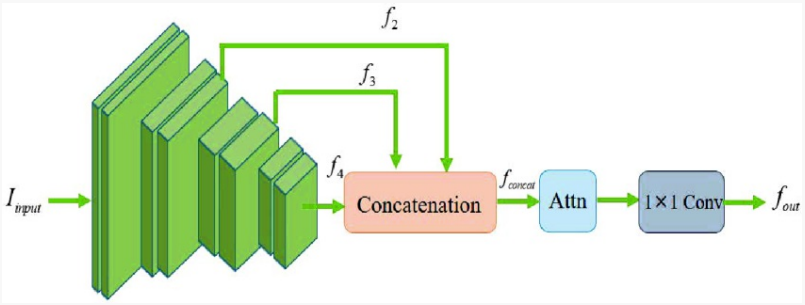
Due to its strong robustness, most researchers directly employ **Visual Geometry Group (VGG)** [36] as the encoder module. However, we have found that directly using the high-level features of $ReLU\_4\_1$ can lead to the loss of much semantic information. Furthermore, experiments have demonstrated that style transfer often amplifies noise.

Combining considerations into account, the structure of MACE is illustrated in Figure 5. First, three different feature maps $\{f_n | n = 2,3,4\}$ are obtained by convolutional layers of different scales. In general, convolutional layers of various scales can obtain different contextual information [6]. Then, for f_{concat}, the approach of upsample convolution concatenation is used to combine different scale features. The formula can be donated as follows:

$$f_{concat} = Concat(Down_4(f_2), Down_2(f_3), f_4), \quad (3)$$

Fig. 5. The structure of proposed MACE.

Alternate Text:



Note: This figure has been annotated.

where $Down_4$ and $Down_2$ are downsampled with factors 4 and 2, respectively. $Concat$ is the channel concatenation of different feature maps. f_{concat} denotes the mixed feature. To capture more feature information while suppressing the emergence of noise, we introduce an attention module inspired by [46], which enhances the important semantic information of channels and spatial regions. Here, the feature has 896 channels, and the direct computational complexity would be too high. Therefore, it is necessary to use $Conv$ to reduce the number of channels in the feature map after $Attn$.

$$f_{out} = Conv(Attn(f_{concat})), \quad (4)$$

where $Attn$ is the attention module composed of channel attention and spatial attention, $Conv$ is a 1×1 convolutional layer.

3.4 GFSM

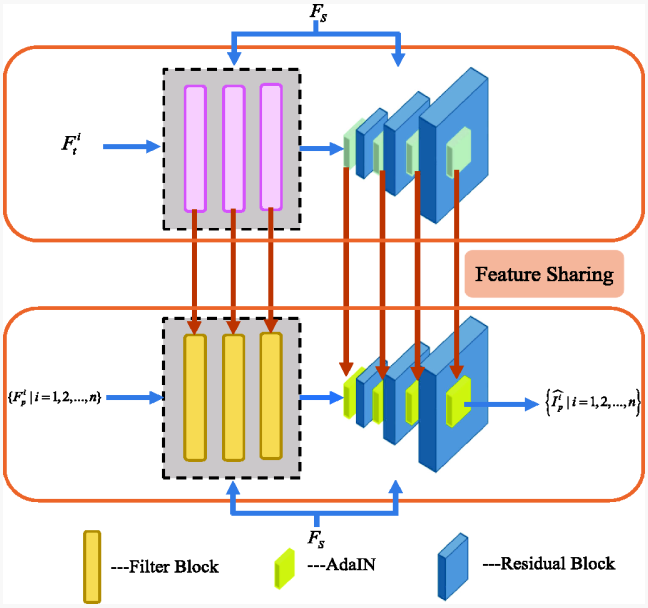
Due to the high performance on high-resolution images, the URST net [7] is introduced, which consists of three stages: dividing, stylization, and assembling. It is worth noting that this strategy is considered to be inserted only into the methods based on instance normalization [38] or instance whitening [23]. However, this article extends this conclusion by inserting the filter-based method into an URST net and obtaining higher-quality results.

As shown in Figure 6, we introduce three consecutive transfer modules and a decoder based on the residual structure for style transfer. Given its excellent performance in fine-grained image processing, we integrated the dynamic filter into the style transfer process. The formula for the filter is defined as follows:

$$Filter(F_s^i, F_c) = Conv(Concat(FC(Concat(GAP\_A(F_s^i), GAP\_A(F_c))), Conv(F_s^i))), \quad (5)$$

Fig. 6. The visualization process of proposed GFSM.

Alternate Text:



Note: This figure has been annotated.

where $GAP_{\downarrow 8}$ represents the operation of first downsampling the vector followed by applying global average pooling, and FC represents the fully connected layer. To reduce the computational complexity, we reduce the vector dimension to 32 before entering the filter. For the decoder, we employed a structure that embeds AdaIN into multi-layer residual blocks to map the image from the feature domain to the image domain. The formula for the decoder is presented as follows:

$$Decoder(\cdot) = AdaIN(Res(\cdot)), \quad (6)$$

where $AdaIN$ and Res represent the multi-layered instances of AdaIN and residual module, respectively. In addition, the multi-scale skip connections between the style encoder and the decoder are utilized to fuse features at different scales. The formula for a single patch stylization can be donated as follows:

$$\hat{I}_p^i = Decoder(Filters(F_p^i, F_S)), \quad (7)$$

where $Filters$ represents the continuous dynamic filters. During the training process, the network removes the feature-sharing step, focusing exclusively on the stylization process depicted in the lower half of Figure 6.

As is well known, the process of stylization can be simplified to removing style features from the content image and adding style features from the style image. During the inference process, the thumbnail image I_t^i is first processed by the network to extract the style feature removed from I_t^i. Subsequently, during patch-wise stylization, each patch is no longer normalized independently but instead has the style feature of I_t^i removed. This operation maintains stylistic consistency among different patches. A total of seven global style features are shared in our network, including three dynamic filter features representing inter-channel global information and four AdaIN [15] features representing intra-channel global information.

3.5 Loss Function

This network is trained by minimizing the loss function. The total loss function contains the content loss $\mathcal{L}_c$, the style loss $\mathcal{L}_s$, the reconstruction loss $\mathcal{L}_{recon}$, and the total variation loss $\mathcal{L}_{tv}$. which is defined as:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_{recon} \mathcal{L}_{recon} + \lambda_{tv} \mathcal{L}_{tv}, \quad (8)$$

where λ_c, λ_s, λ_{recon}, and λ_{tv} are hyper-parameters controlling weights of their corresponding loss terms.

The content loss and the style loss are defined as:

where $\Phi_i(.)$ denotes the pretrained VGG-19 feature map at layer i. μ denotes features mean, and σ denotes features variance. I_c denotes content image and I_s denotes style image. I_{cs} denotes the stylized result. For I_c, we use $ReLU\_4\_1$. while for I_s, we use $ReLU\_1\_1$, $ReLU\_2\_1$, $ReLU\_3\_1$, and $ReLU\_4\_1$. $\mathcal{L}_{tv}$ denotes total variation loss [33]. The reconstruction loss is defined as

$$\mathcal{L}_{recon} = \|F(I_c^{gray}, I_c^{color}) - I_c^{color}\|_2, \quad (11)$$

where $F(.)$ denotes the combination of MACE, SE, GFSM, and $F(I_c^{gray}, I_c^{color})$ represents colorizing gray image by using the style of the corresponding colorful image.

4 Experiments

4.1 Dataset and Implementation Details

Overview. Following previous research work, our model is trained on MS-COCO [25] as content image dataset and WikiArt [31] as style image dataset. λ_c, λ_s, λ_{recon}, and λ_{tv} are set as 1, 20, 20 and 10, respectively. Our model is trained using the Adam optimizer [19] with a batch size of 4 and a learning rate of 0.0001. In the training phase, all images are randomly cropped to 256×256 pixels in size. While in inference processing, our model can be applied to images with any resolution. The model was put into practice using PyTorch and trained for four epochs.

Noteworthy, MPT and feature-sharing operations only be used in the inference phase. Moreover, different from other panoramic image tasks [3, 29], only 2D images are used in the training phase and do not need to use panoramic image datasets for fine-tuning, which greatly simplifies the training process.

4.2 Comparison with State-of-the-Art (SOTA) Models

2D Image Style Transfer. As shown in Figure 7, our method is compared with seven SOTA 2D style transfer methods, including CAST [53], AdaIN [15], AdaAttN [27], CSBNet [28], IEST [5], MCCNet [8], and MicroAST [44]. CAST and IEST cannot completely transfer the style of the content image and retain a part of the style of the content image incorrectly (4th, 5th, and 6th rows). AdaIN directly changes second-order statistics of content features distorts the brush strokes and causes severe content details loss (1st, 4th, and 6th rows). AdaAttn performs well in detail. However, it prefers to amplify noise, which can lead to artifacts such as the fuzzy background (1st, and 3rd rows). CSBNet cannot capture textural patterns of style images adaptively (1st, and 2nd rows). Both MCCNet and MicroAST have higher content distortion and blurry textures, distorting the brush strokes (1st, 3rd, and 5th rows). As shown in the 3rd column, MuTPST can transfer style to each position of content images appropriately, and improve the balance between style transfer and semantic reconstruction.

Fig. 7. Comparison with other state-of-the-art methods in 2D image style transfer.

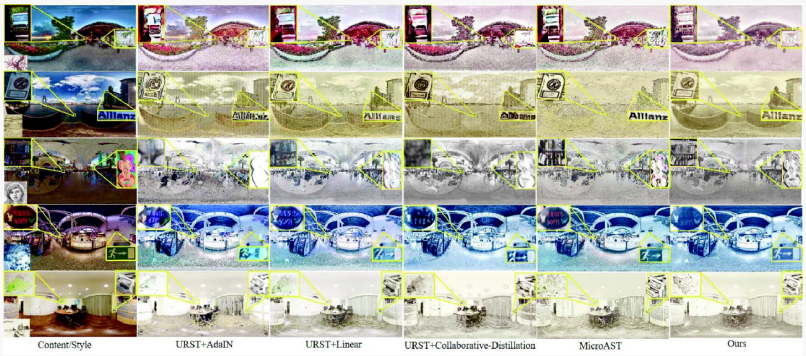
Alternate Text:

Note: This figure has been annotated.

Panoramic Image Style Transfer. Style transfer of images at the tens of millions of pixels typically requires extremely substantial computational resources. To compare stylization effects on a regular GPU, such as the NVIDIA 2080 Ti, we compared our method with the SOTA 2D image style transfer methods [15, 22, 39] combined with URST [7] and MicroAST [44]. mentions that URST can only be inserted into the methods based on instance normalization or instance whitening, therefore many latest methods cannot handle high-resolution images due to memory limitations. Figure 8 shows the comparison results of high-resolution panoramic images (e.g., 10,000 × 5,000 pixels) from the Salient360! dataset [12], which exhibits excellent performance in the task of saliency detection for panoramic scenes. Specifically, AdaIN [15] leaves a lot of stylized parts. Learning [22] retains some of its original color features. Collaborative-Distillation [39] results in blur stylization results. MicroAST [44] severely distorts the content structure. However, our model migrates style and preserves semantic details better compared with other methods. The enlarged parts show additional comparison results, and it can be observed that other methods lead to poor visual effects, proving the superiority of our proposed framework.

Fig. 8. Qualitative comparisons between the state-of-the-art panoramic image style transfer methods. The complete visualization presentation and additional interactive features are displayed on our Web site.

Alternate Text:



Note: This figure has been annotated.

Quantitative Comparison. **Peak Signal-to-Noise Ratio (PSNR)** is a measurement method based on the error between corresponding pixel points, so the evaluation results are often inconsistent with human subjective feeling. **Structural Similarity Index Measure (SSIM)** is an image quality evaluation metric suitable for human vision, which measures image similarity in terms of brightness, contrast, and structure, respectively. Therefore, we demonstrate the performance of MuTPST and the SOTA style transfer methods over PSNR and SSIM on Tables 1 and 2. The bold outcomes are the best results in each case and the second-place score with underlining.

Note: The table layout displayed in 'Edit' view is not how it will appear in the printed/pdf version. This html display is to enable content corrections to the table. To preview the printed/pdf presentation of the table, please view the 'PDF' tab.

Table 1. Quantitative Comparison of MuTPST and the State-of-the-Art 2D Image Style Transfer Methods over PSNR and SSIM

		CAST	AdaIN	AdaAttN	CSBNet	IEST	MCCNet	StyleFormer	MicroAST	Ours
Simple images	PSNR	17.84687	17.87340	18.12782	18.36692	18.28003	18.65237	16.53017	18.92669	**19.28344**
	SSIM	0.59098	0.15670	0.57949	0.57190	0.57488	0.57283	0.47201	**0.60344**	0.59748
Complex images	PSNR	10.57678	9.73117	9.93418	9.85217	**11.59242**	9.77647	9.22212	10.11261	10.48521
	SSIM	0.36302	0.24243	0.31828	0.34604	0.33523	0.38267	0.28928	0.42659	**0.47212**

PSNR, signal-to-noise ratio; SSIM, structural similarity index measure. The bolded results represent the *best* results in each category and the *second-place* score with underlining.

Note: The table layout displayed in 'Edit' view is not how it will appear in the printed/pdf version. This html display is to enable content corrections to the table. To preview the printed/pdf presentation of the table, please view the 'PDF' tab.

Table 2. Quantitative Comparison of MuTPST and the State-of-the-Art Panoramic Image Style Transfer Methods over PSNR and SSIM

		URST+AdaIN	URST+WCT	URST+Liner	URST+C-D	MicroAST	Ours
Indoor scenes	PSNR	9.51314	8.38857	9.52508	9.64917	10.21433	**10.48117**
	SSIM	0.39105	0.35347	0.46113	0.39994	0.52387	**0.53165**
Outdoor scenes	PSNR	9.39044	8.55585	**9.43677**	9.12024	8.69384	8.82955
	SSIM	0.24733	0.25947	0.37640	0.31375	0.35579	**0.42435**

PSNR, signal-to-noise ratio; SSIM, structural similarity index measure.

The bolded results represent the *best* results in each category and the *second-place* score with underlining.

Table 1 illustrates the comparative results of the 2D image style transfer methods. Based on the complexity of the texture within the content image structure, we

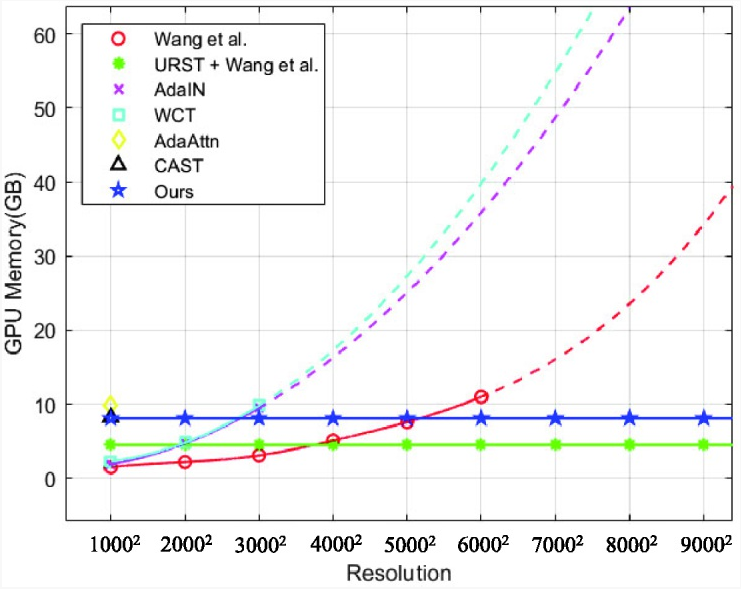
categorize 2D images as simple images and complex images. It should be noted that almost all methods have relatively low scores on PSNR and SSIM due to less image content. Since some approaches do not completely transform the style image's style, the features of the content image are kept, resulting in greater PSNR rates on some complex images. In the simple images category, our method outperforms the vast majority of methods in terms of both PSNR and SSIM scores. In the complex images category, Our method achieved the highest SSIM score.

In contrast to the 2D images, we categorized the test panoramic images into indoor and outdoor categories based on where the actual photographs were captured, the result is shown as Table 2. Our model achieved higher scores on both PSNR and SSIM for indoor scenes than previous methods. For outdoor scenes, our model got the highest score on SSIM and performed average on PSNR. The fact that the outdoor scenes had more targets which needed large brush strokes such as the sky and the ground, resulted in them achieving slightly higher scores on PSNR. However, we achieved the highest score on the SSIM, which is more consistent with human perception. Thus, our proposed MuTPST can produce more satisfactory results.

Memory Occupation. In addition, we provide visualization of the comparative memory consumption results in Figure 9. The dashed lines indicate that images of these resolutions cannot be rendered on an 11 GB GPU (NVIDIA 2080 Ti) due to GPU memory limitations. As the resolution of the input image proliferates, the memory cost of most methods increases dramatically and eventually exhausts GPU memory. In contrast, our MuTPST and URST can handle images with unconstrained resolutions with limited memory. It should be emphasized that URST combines the knowledge distillation network proposed by Wang et al. [39] to reduce the number of parameters. This operation yields the least memory footprint but produces poorer stylization as in the fifth column of Figure 8.

Fig. 9. GPU memory comparison of different style transfer methods.

Alternate Text:



Note: This figure has been annotated.

User Study. We conduct a user study involving 23 volunteers to subjectively evaluate the performance of various methods. For 2D image, each volunteer takes in 20 tests. In each test, we present a set comprising an input image, a style reference, and a stylized image generated by nine algorithms from a random combination of 20 content images and 20 style images. For panoramic image, each volunteer completes 10 trials. During the experiment, pairs of stylized images randomly generated from 10 panoramic images and 10 style images are displayed using VR head-mounted displays. Participants in the user study are asked to select the most visually pleasant one. As shown in Table 3, our method achieves the highest preference rate of 15.9% in the 2D image style transfer task. In Table 4, the preference rate of our method reached 35.7%, significantly outperforming existing technologies and thereby demonstrating the effectiveness of our approach for panoramic image style transfer.

Table 3. Results of the User Study for Panoramic Image Style Transfer Methods

Method	CAST	AdaIN	AdaAttN	CSBNet	IEST	MCCNet	StyleFormer	MicroAST	Ours
Favorite rate	13.48%	6.09%	10.22%	11.52%	9.35%	8.04%	10.65%	14.78%	**15.87%**

Place the cursor position on table column and click 'Add New' to add table footnote.

Table 4. Results of the User Study for 2D Image Style Transfer Methods

Method	URST+AdaIN	URST+WCT	URST+Liner	URST+C-D	MicroAST	Ours
Favorite rate	10.43%	6.96%	16.09%	9.57%	21.3%	**35.65%**

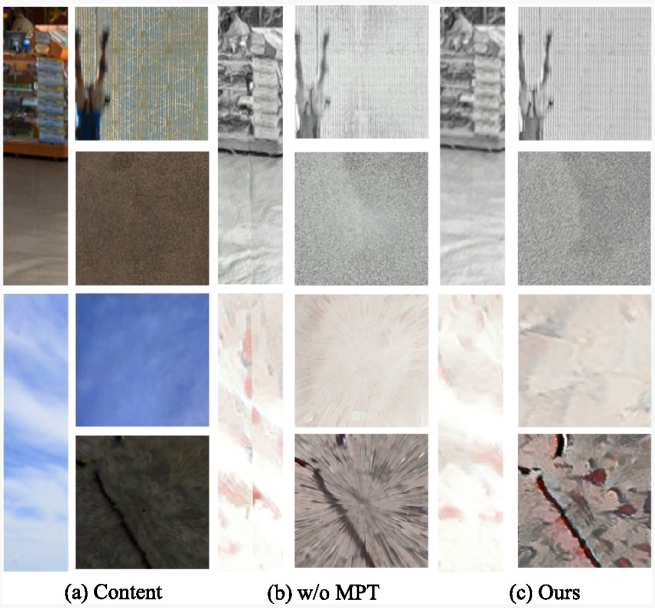
Place the cursor position on table column and click 'Add New' to add table footnote.

4.3 Ablation Study

MPT. Figure 10 illustrates the ablation AQ13 study of the MPT. To visually demonstrate the effect of MPT, the first row shows a high-resolution panoramic image of 10,000 × 5,000 pixels, and the second row shows a panoramic image of 2,000 × 1,000 pixels for a more visual comparison. The general method leads to the pixels in the polar regions distorting toward the poles, causing considerable visual discomfort. However, it can be observed that the pole distortion and edge discontinuity are greatly alleviated in our method, which demonstrates the effectiveness of the proposed MPT.

Fig. 10. Ablation study of MPT. “w/o MPT” denotes without the MPT.

Alternate Text:

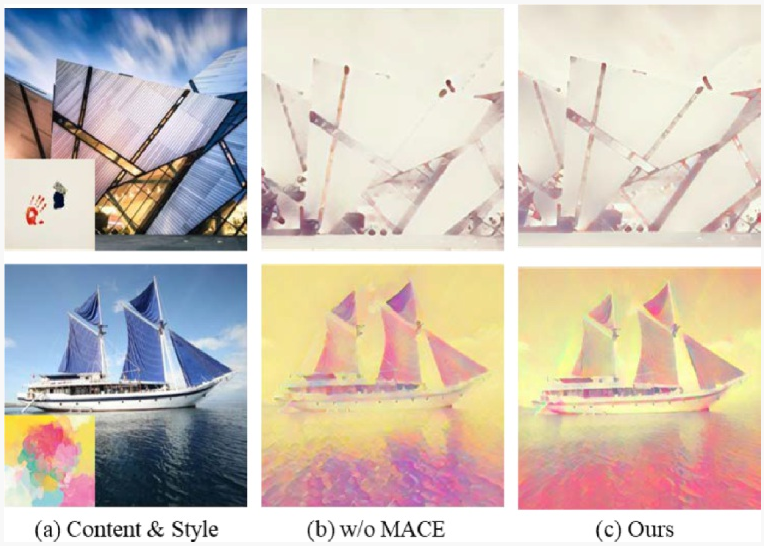


Note: This figure has been annotated.

MACE. To verify the effectiveness of MACE used in MuTPST, we remove MACE and replace it with a standard VGG encoder. Some local content damage and blurry textures can be observed (the texture of the building in the first row and the hull of the sailboat in the second row) in Figure 11. Nevertheless, MACE retains the texture of the building in the first row and the hull of the sailboat in the second row, essentially restoring the physical structure of the content image. Our MuTPST can efficiently utilize MACE to produce pleasing stylization results.

Fig. 11. Ablation study of MACE. “w/o MACE” denotes without the MACE.

Alternate Text:



Note: This figure has been annotated.

GFSM. As discussed, GFSM is important for URST. To verify this, we conduct experiments of high-resolution transfer with simple split and the proposed GFSM, respectively. From Figure 12, we can observe that a simple split leads to style inconsistency among different patches. In addition, this transformation amplifies local area noise, leading to a lot of heavy contamination in the input image in structurally smooth places such as the sky and the ground. Differently, GFSM avoids style inconsistencies and local noise amplification problems by sharing global features in each patch.

Fig. 12. Ablation study of GFSM. “w/o GFSM” denotes without the GFSM.

Alternate Text:

(a) w/o GFSM

(b) Ours

Note: This figure has been annotated.

Loss Function Hyper-parameters. In this section, we study the tradeoffs among the hyper-parameters of proposed loss function. The primary role of the $\mathcal{L}_{tv}$ is to fine-tune the stylized image and enhance the diversity of the generated image. Empirically, we set λ_{tv} to 10 to prevent over-smoothing. In addition, when comparing the same loss, we maintain the remaining hyper-parameters on the same scale in experiments. Table 5 presents the quantitative comparison results of the tradeoffs among λ_c, λ_s, and λ_{recon}. Evidently, with larger λ_c, the SSIM rate gets increased, while the PSNR rate decreases. This can be attributed to the excessive focus on content, leading to the amplification of noise. We can see that as λ_s increases, the PSNR score increases, indicating that larger λ_s correlates with stronger style similarity. The introduction of $\mathcal{L}_{recon}$ is intended to prevent color bias in the stylized image and lower λ_{recon} tends to reduce the SSIM of the stylized image. To achieve the best balance between content preservation and stylization, we set λ_c, λ_s, λ_{recon} and λ_{tv} to 1, 20, 20 and 10, respectively.

Note: The table layout displayed in ‘Edit’ view is not how it will appear in the printed/pdf version. This html display is to enable content corrections to the table. To preview the printed/pdf presentation of the table, please view the ‘PDF’ tab.

Table 5. Ablation Study of Loss Function Hyper-parameters

λ_c	λ_s	λ_{recon}	λ_{tv}	PSNR ↑	SSIM ↑
10	20	20	10	16.43371	0.50697
15	20	20	10	16.18597	0.52218
20	20	20	10	15.71864	0.51183
1	1	20	10	16.31236	0.48351
1	10	20	10	16.13940	0.47953
1	15	20	10	16.48893	0.50796
1	20	1	10	16.46871	0.48619
1	20	10	10	16.27651	0.51268
1	20	15	10	16.43371	0.50697
1	20	20	10	**16.67534**	**0.52301**

Place the cursor position on table column and click "Add New" to add table footnote.

Stylization Effects. To better understand each component in our model, we evaluate our method in comparison to three alternatives: without MACE, without GFSM, and without both, which is shown in Table 6. Comparing the changes in average content loss and average style loss, it is observed that MACE preserves the content structure and GFSM leads to better-stylized results. Thus, using all techniques can yield better performance.

Note: The table layout displayed in 'Edit' view is not how it will appear in the printed/pdf version. This html display is to enable content corrections to the table. To preview the printed/pdf presentation of the table, please view the 'PDF' tab.

Table 6. Ablation on MACE (I), GFSM (II)

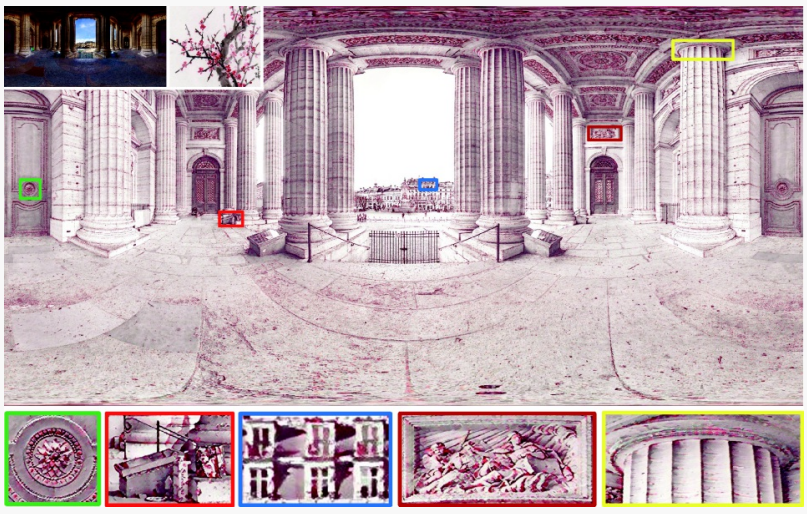
I	II	Content Loss	Style Loss
		11.734	0.526
✓		9.216	0.531
	✓	10.725	0.437
✓	✓	**8.645**	**0.409**

Bold indicates the best result.

To further demonstrate the effectiveness of MuTPST, we performed it on an ultra-high-resolution stylized panoramic image of 10,000 × 5,000 pixels (e.g., 50 megapixels), as shown in Figure 13. The first row of images shows the content image and style image in the upper left corner. The five close-ups are shown on the lower side of the style results.

Fig. 13. An ultra-high-resolution stylized panoramic image (10,000 × 5,000 pixels).

Alternate Text:



Note: This figure has been annotated.

5 Discussion

5.1 Limitation

Our framework still suffers from some limitations. The current model performs poorly in texture transfer. As indicated in Equations (1) and (2), this is primarily attributed to the distortion occurring in certain areas during the projection of panoramic images onto the sphere. Wang et al. [39] proposed that excessive texture transfer will lead to a significant amount of noise. As observed from the magnified regions in Figure 8 and in the virtual headset provided by our Web site, the SOTA methods result in blurring of detailed areas of the stylized panoramic images, such as the car license plate in the first row of Figure 8. Consequently, we believe that appropriately reducing the focus on texture transfer is vital for panoramic image style transfer. Moreover, although the proposed MPT method substantially alleviates edge discontinuities, minor cracks at the edges are still theoretically possible. By presenting and comparing visual effects, we believe that our model does not affect the viewer's experience.

Additionally, we present the runtime of a single panoramic image at several common resolutions without MPT in Table 7. Owing to the multiple transformations and stylizations of the panoramic images by MPT, the overall runtime of the full model is approximately five times longer than that without MPT. As shown in Section 4.2, most methods are unable to perform high-resolution panoramic image style transfer on regular GPUs, yet our method can reduce the total time consumption through parallel processing.

Table 7. Runtime of a Single Panoramic Image at Different Resolutions without MPT

Resolution	3,840 × 1,920	7,000 × 3,500	10,000 × 5,000
Time (s)	30.91	47.24	63.46

5.2 Future Work

In the future, we plan to investigate fast and lightweight methods for panoramic video style transfer. Existing methods [43] have highlighted that one of the keys to video stylization is maintaining temporal consistency, which demonstrates the necessity of appropriately reducing texture transfer. Moreover, enhancing the speed of stylization without weakening content preservation still needs to be explored in future work. We hope that this will help the panoramic scene style transfer algorithm to be applied to practical applications.

6 Conclusion

This article proposes a novel style transfer method for high-resolution panoramic images, named MuTPST, which includes a convenient and efficient panoramic image transformation operation that solves the pole distortion and edge discontinuity problem after stylization. Moreover, a MACE, which contains an attention module, is proposed to show details with finer strokes and improve the stylized performance. By considering memory limitation issues associated with high-resolution images, the global feature-share module is introduced. A series of experiments have demonstrated the effectiveness of our proposed method for high-resolution panoramic images.

References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. 2021. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 862–871.

[2] Jie An, Tao Li, Haozhi Huang, Li Shen, Xuan Wang, Yongyi Tang, Jinwen Ma, Wei Liu, and Jiebo Luo. 2020. Real-time universal style transfer on high-resolution images via zero-channel pruning. arXiv:2006.09029. Retrieved from https://arxiv.org/abs/2006.09029 AQ14

[3] Dongwen Chen, Chunmei Qing, Xiangmin Xu, and Huansheng Zhu. 2020. Salbinet360: Saliency prediction on 360 images with local-global bifurcated deep network. In *Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 92–100.

[4] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. 2018. Stereoscopic neural style transfer. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, 6654–6663.

[5] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu. 2021. Artistic style transfer with internal-external learning and contrastive learning. In *Advances in Neural Information Processing Systems*, Vol. 34, 26561–26573.

[6] Ruihan Chen, AQ15 Junpeng Tan, Zhijing Yang, Xiaojun Yang, Qingyun Dai, Yongqiang Cheng, and Liang Lin. 2024. DPHANet: Discriminative parallel and hierarchical attention network for natural language video localization. *IEEE Transactions on Multimedia* (2024).

[7] Zhe Chen, Wenhai Wang, Enze Xie, Tong Lu, and Ping Luo. 2022. Towards ultra-resolution neural style transfer via thumbnail instance normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 393–400.

[8] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 1210–1217.

[9] 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.

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

[11] Xinyu Gong, Haozhi Huang, Lin Ma, Fumin Shen, Wei Liu, and Tong Zhang. 2018. Neural stereoscopic image style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 54–69.

[12] Jesús Gutiérrez, Erwan J. David, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360 contents. In *Proceedings of the 2018 10th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–3.

[13] Mark Hamazaspyan and Shant Navasardyan. 2023. Diffusion-enhanced patchmatch: A framework for arbitrary style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 797–805.

[14] Jingwei Huang, Zhili Chen, Duygu Ceylan, and Hailin Jin. 2017. 6-DOF VR videos with a single 360-camera. In *Proceedings of the 2017 IEEE Virtual Reality (VR)*. IEEE, 37–44.

[15] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

[16] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. 2020. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 4369–4376.

[17] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* 26, 11 (2019), 3365–3385.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*. Springer, 694–711.

[19] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980

[20] Thi-Ngoc-Hanh Le, Ya-Hsuan Chen, and Tong-Yee Lee. 2023. Structure-aware video style transfer with map art. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3s (2023), 1–25.

[21] Shaohua Li, Xinxing Xu, Liqiang Nie, and Tat-Seng Chua. 2017. Laplacian-steered neural style transfer. In *Proceedings of the 25th ACM International Conference on Multimedia*, 1716–1724.

[22] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. 2019. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3809–3817.

[23] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, Vol. 30.

[24] Minxuan Lin, Fan Tang, Weiming Dong, Xiao Li, Changsheng Xu, and Chongyang Ma. 2021. Distribution aligned multimodal and multi-domain image stylization. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 3 (2021), 1–17.

[25] 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 *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[26] Xu Lin, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. 2023. Multi-scale transformer network for saliency prediction on 360-degree images. In *Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP)*, 1700–1704. DOI: 10.1109/ICIP49359.2023.10222683

[27] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6649–6658.

[28] Haofei Lu and Zhizhong Wang. 2022. Universal video style transfer via crystallization, separation, and blending. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, 23–29.

[29] Akito Nishiyama, Satoshi Ikehata, and Kiyoharu Aizawa. 2021. 360 single image super resolution via distortion-aware network and distorted perspective images. In *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1829–1833.

[30] Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5880–5888.

[31] Fred Phillips and Brandy Mackintosh. 2011. Wiki Art Gallery, Inc.: A case for critical thinking . *Issues in Accounting Education* 26, 3 (2011), 593–608.

[32] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2018. Artistic style transfer for videos and spherical images. *International Journal of Computer Vision* 126, 11 (2018), 1199–1219.

[33] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* 60, 1–4 (1992), 259–268.

[34] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. 2018. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8242–8250.

[35] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2019. Increasing image memorability with neural style transfer. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2 (2019), 1–22.

[36] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from https://arxiv.org/abs/1409.1556

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

[38] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022. Retrieved from https://arxiv.org/abs/1607.08022

[39] Huan Wang, Yijun Li, Yuehai Wang, Haoji Hu, and Ming-Hsuan Yang. 2020. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1860–1869.

[40] Quan Wang, Sheng Li, Xinpeng Zhang, and Guorui Feng. 2022. Multi-granularity brushstrokes network for universal style transfer. *ACM Transactions on Multimedia Computing, Communications and Applications* 18, 4, Article 107 (Mar. 2022), 17 pages. DOI: 10.1145/3506710

[41] Weiyu Wang, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. 2024. Consistent panoramic video style transfer via temporal-spatial cross perception. In *Proceedings of the International Conference on Intelligent Computing*. Springer, 265–277.

[42] Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. 2020. Consistent video style transfer via compound regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 12233–12240.

[43] Wenjing Wang, Shuai Yang, Jizheng Xu, and Jiaying Liu. 2020. Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing* 29 (2020), 9125–9139.

[44] Zhizhong Wang, Lei Zhao, Zhiwen Zuo, Ailin Li, Haibo Chen, Wei Xing, and Dongming Lu. 2023. MicroAST: Towards super-fast ultra-resolution arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, 2742–2750.

[45] Linfeng Wen, Chengying Gao, and Changqing Zou. 2023. CAP-VSTNet: Content affinity preserved versatile style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18300–18309.

[46] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

[47] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. 2021. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14618–14627.

[48] Xin Xiang, Wujian Ye, and Yijun Liu. 2022. Panoramic image style transfer technology based on multi-attention fusion. In *Proceedings of the 5th International Conference on Computer Science and Software Engineering*, 293–299.

[49] Gaoming Yang, Changgeng Li, and Ji Zhang. 2024. ConIS: Controllable text-driven image stylization with semantic intensity. *Multimedia Systems* 30, 4 (2024), 174.

[50] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. 2022. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7693–7702.

[51] Xin Zhang. 2020. Style Transfer for 360 images. Master's thesis. Trinity College Dublin.

[52] 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.

[53] 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 *Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings*, 1–8.

[54] Sizhe Zheng, Pan Gao, Peng Zhou, and Jie Qin. 2024. Puff-Net: Efficient style transfer with pure content and style feature fusion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8059–8068.

Author Query

1. **Query** [AQ1] : Please review the alt-text you provided for figures in your article and confirm that it is correct. If you have not provided alt-text, please note that ACM strongly recommends that you add alt-text to the proof. Alt-text is a brief description of the image that is read by screen readers or other assistive technology. It provides information about an image's purpose for the reader and provides context on how the image relates to the page content. If you want to add alt-text for any figure, please add the alt-text using the "+" button at the end of the figure caption in the proof. Please provide alt-text or let us know you don't want to add alt-text.
   Response by Author: "Answered within text"

2. **Query** [AQ2] : If any changes are needed to the author list, please provide them, explain the reasons for the changes, and per journal policy, please provide emails from all authors noting that all changes have their approval.
   Response by Author: "I would like to request an addition to the author affiliation list in the page proofs. Specifically, we would like to add "Alibaba Group" as the third affiliation for the first author (Weiyu Wang). This addition is necessary because: 1. He joined Alibaba Group before the revision period of our paper 2. The additional experiments conducted during the major revision utilized Alibaba Group's GPU resources and infrastructure We understand the importance of proper documentation for such changes. We can provide email confirmations from all co-authors indicating their approval of this affiliation addition if needed. Could you please advise if it's possible to make this addition in the page proofs? We will ensure to follow any required procedures and provide all necessary documentation."

3. **Query** [AQ3] : Weiyu Wang's and Chunmei Qing's affiliation is listed in information we received from ACM as School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China, but the manuscript lists the authors' affiliation as South China University of Technology, China and Pazhou Lab, China. Please let us know what is correct to use for publication.
   Response by Author: "Answered within text"

4. **Query** [AQ4] : Junpeng Tan's affiliation is listed in information we received from ACM as School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, but the manuscript lists the author's affiliation as South China University of Technology, China. Please let us know what is correct to use for publication.
   Response by Author: "Answered within text"

5. **Query** [AQ5] : Xiangmin Xu's affiliation is listed in information we received from ACM as School of Future Technology, South China University of Technology, Guangzhou, China and Pazhou Lab, Guangzhou, China, but the manuscript lists the author's affiliation as South China University of Technology, China and Pazhou Lab, China. Please let us know what is correct to use for publication.
   Response by Author: "Answered within text"

6. **Query** [AQ6] : The funding information has been added in the proof as provided in the source manuscript for this article, but the funding information is missing from information we received from ACM. Please let us know the correct funding information to use for publication.
   Response by Author: "Answered within text"

7. **Query** [AQ7] : If your proof includes supplementary material, please confirm that all supplementary material is cited in the text.
   Response by Author: "Accepted"

8. **Query** [AQ8] : If your proof includes links to Web sites, please verify that the links are valid and will direct readers to the correct Web page.
Response by Author: "Accepted"

9. **Query** [AQ9] : Please confirm that all of the abbreviations and expansions are correct as they appear and used consistently throughout. Also, please ensure that terms other than names of models, networks, algorithms, and similar terms are expanded on first mention and then abbreviated throughout your article.
Response by Author: "Accepted"

10. **Query** [AQ10] : Is the short title (Multi-view Panoramic Image Style Transfer) appropriate as it appears in the proof? if not, please provide a short title of 75 characters or fewer, including spaces.
Response by Author: "Accepted"

11. **Query** [AQ11] : If your proof includes figures, please check the figures in your proof carefully. If any changes are needed, please provide a revised figure file.
Response by Author: "Accepted"

12. **Query** [AQ12] : Please spell out CNN on first mention.
Response by Author: "Please revise "who first proposed a CNN-based style transfer algorithm" to "who first proposed a style transfer algorithm based on Convolutional Neural Network (CNN)"."

13. **Query** [AQ13] : Please add a table legend explaining what the bold in Tables 3, 4, and 5 denotes or advise that the bold can be changed to roman.
Response by Author: "Please help add "Bold indicates the best results in each comparison" as the table legend."

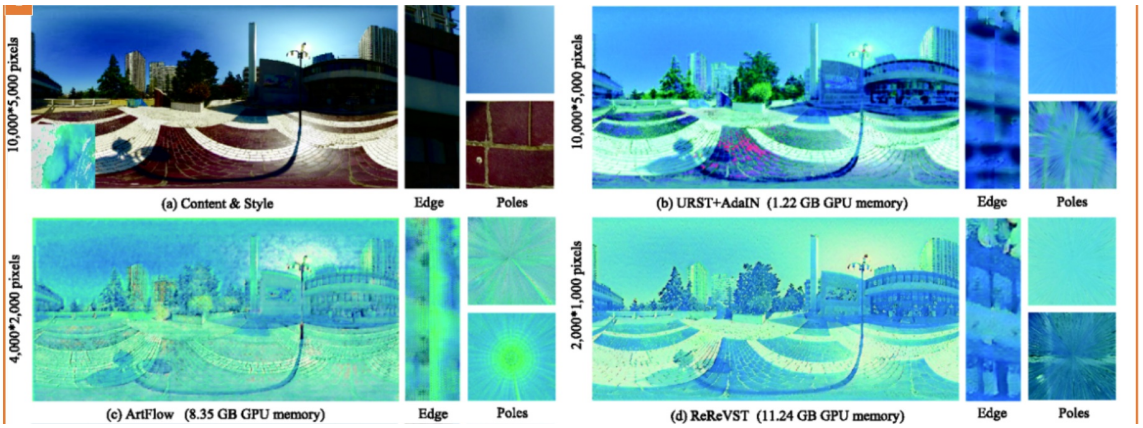
14. **Query** [AQ14] : Please confirm whether the URLs added in arXiv references are appropriate.
Response by Author: "Accepted"

15. **Query** [AQ15] : Please provide the volume and page numbers for Refs. [6].
Response by Author: "The detailed reference for this article is as follows:@ARTICLE{10517423, author={Chen, Ruihan and Tan, Junpeng and Yang, Zhijing and Yang, Xiaojun and Dai, Qingyun and Cheng, Yongqiang and Lin, Liang}, journal={IEEE Transactions on Multimedia}, title={DPHANet: Discriminative Parallel and Hierarchical Attention Network for Natural Language Video Localization}, year={2024}, volume={26}, number={}, pages={9575-9590}, keywords={Location awareness;Semantics;TV;Natural languages;Correlation;Glass;Electronic mail;Cross-modal retrieval;natural language video localization;video moment localization;video understanding}, doi={10.1109/TMM.2024.3395888}} "

Figure Replacement

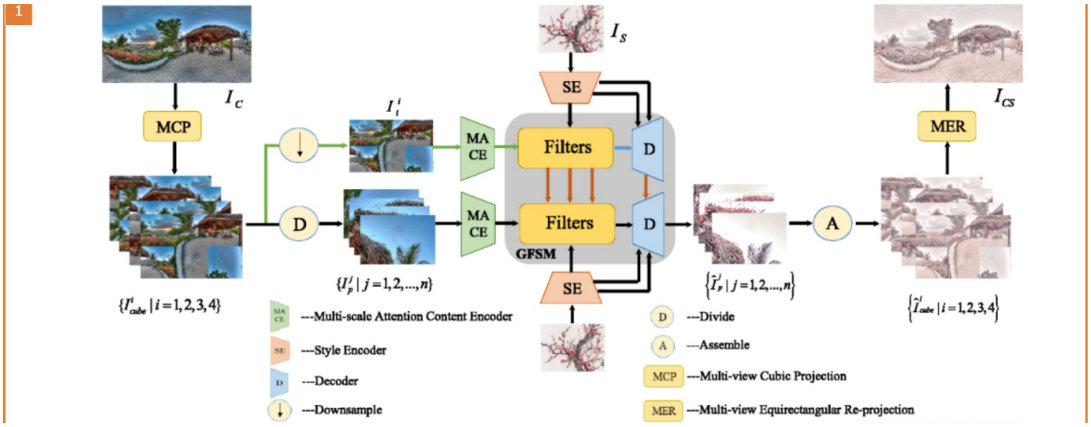
Figure Annotation

1. **Figure:** F001.png



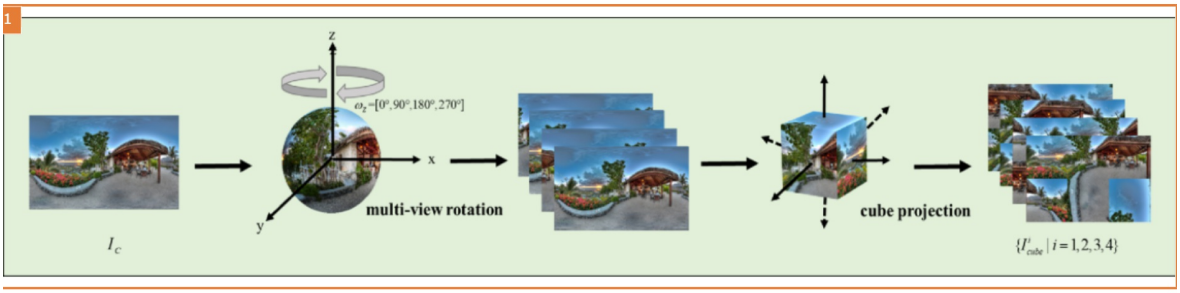
Annotation: (1) Fig. 1. Unsatisfying effectiveness of traditional NST on equirectangular images. It can be observed that due to the unique structure of the panoramic image, existing methods cause pole distortion, edge discontinuities, blurred details, and memory limitation.

2. **Figure:** F002.png



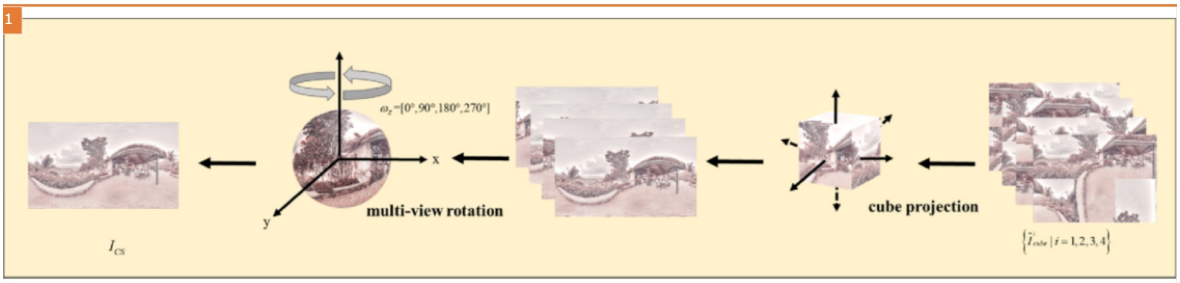
Annotation: (1) Fig. 2. The framework of proposed MuTPST. The pipeline is separated into three stages: projection, stylization, and re-projection. The core is that the panoramic image will perform a series of transformations, and then the thumbnail style transfer and the patch-wise style transfer will be carried out successively.

3. **Figure:** F003.png



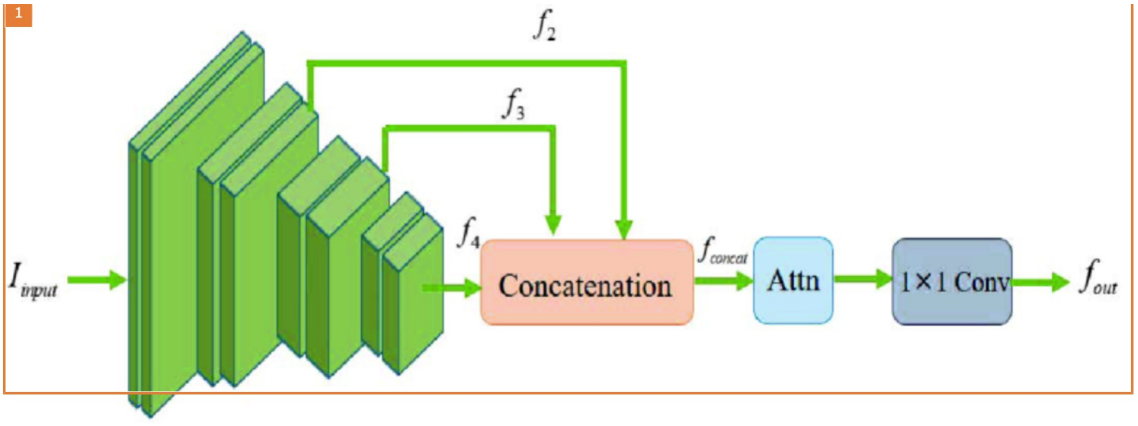
Annotation: (1) Fig. 3. The structure of proposed MCP. The sphere means projecting the panoramic image into the shape of a sphere, while the square means projecting the panoramic image into a cube.

4. **Figure:** F004.png



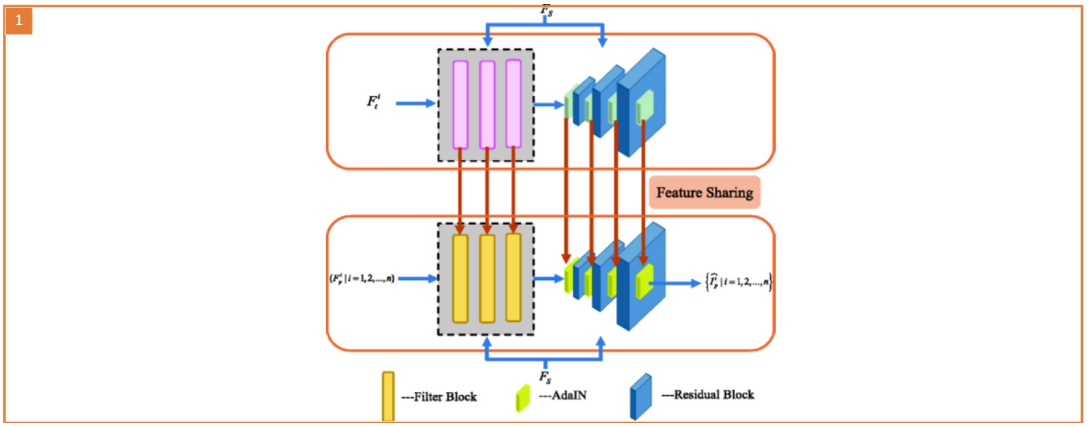
Annotation: (1) Fig. 4. The structure of proposed MER. The final result was generated through a series of inverse projections, rotations, and transformations.

5. **Figure:** F005.png



Annotation: (1) Fig. 5. The structure of proposed MACE.

6. **Figure:** F006.png

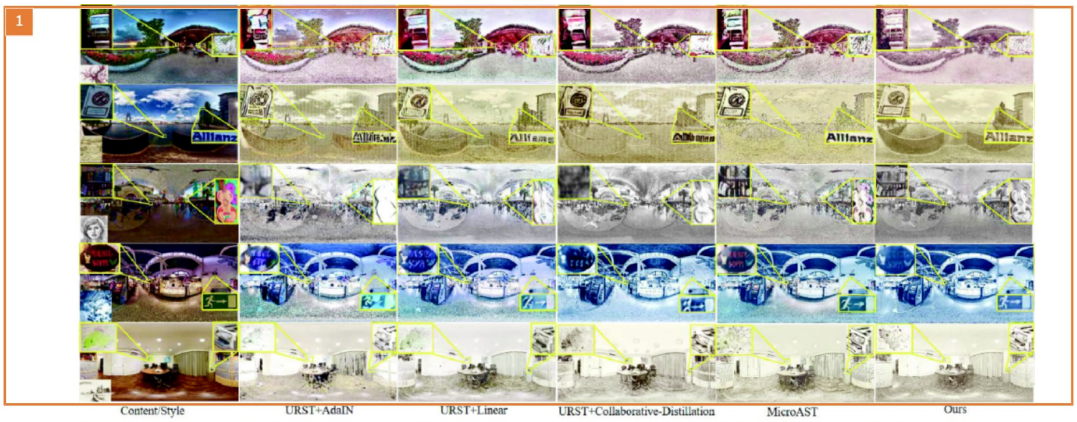


Annotation: (1) Fig. 6. The visualization process of proposed GFSM.

7. **Figure:** F007.png

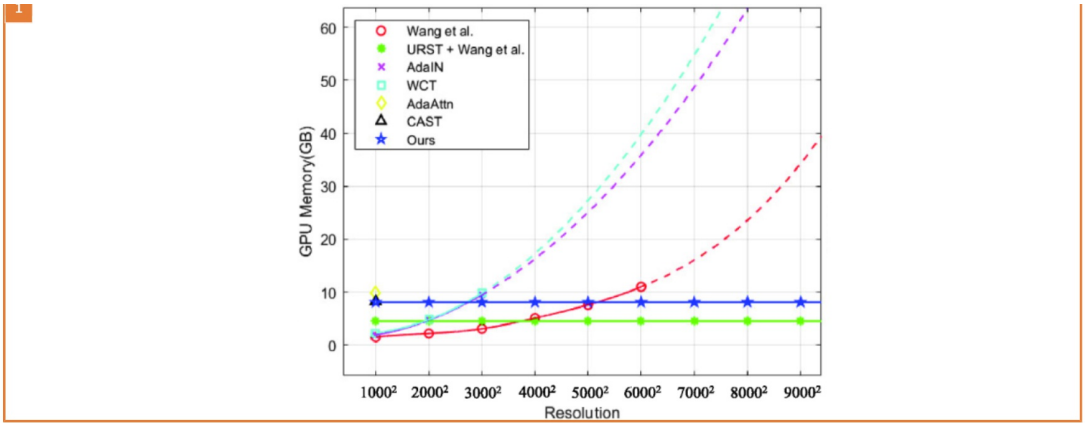
Annotation: (1) Fig. 7. Comparison with other state-of-the-art methods in 2D image style transfer.

8. **Figure:** F008.png



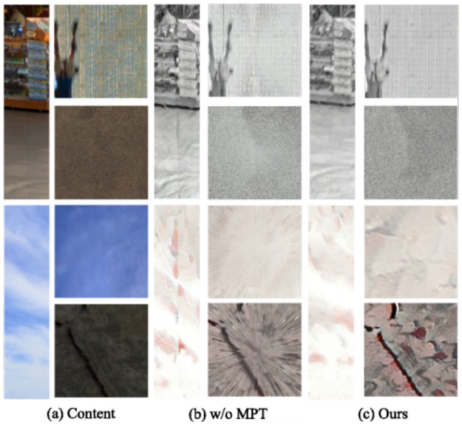
Annotation: (1) Fig. 8. Qualitative comparisons between the state-of-the-art panoramic image style transfer methods. The complete visualization presentation and additional interactive features are displayed on our Web site.

9. **Figure:** F009.png



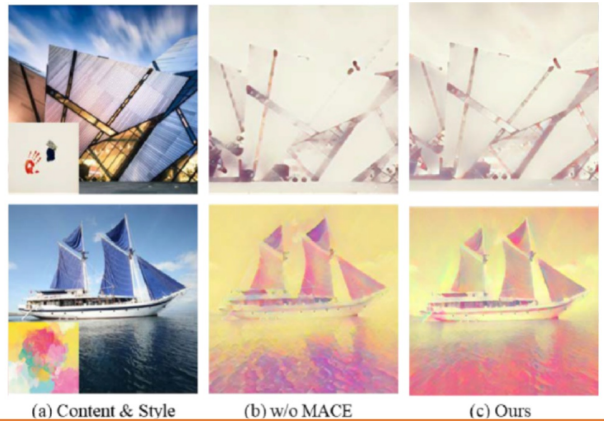
Annotation: (1) Fig. 9. GPU memory comparison of different style transfer methods.

10. **Figure:** F010.png



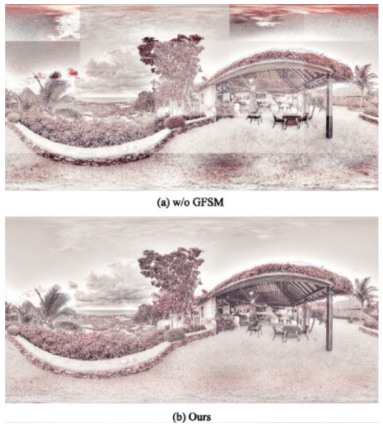
Annotation: (1) Fig. 10. Ablation study of MPT. “w/o MPT” denotes without the MPT.

11. **Figure:** F011.png



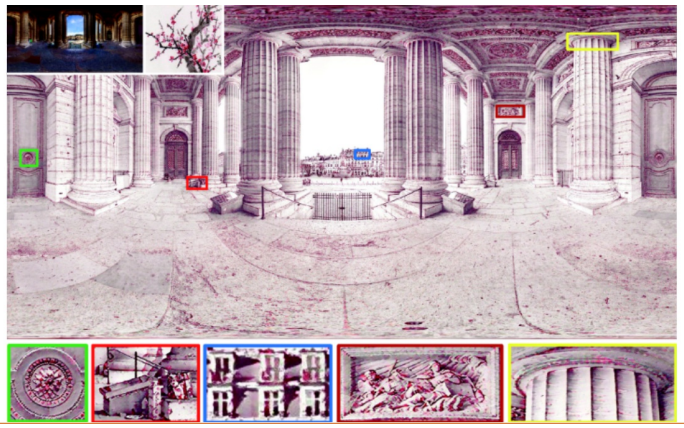
Annotation: (1) Fig. 11. Ablation study of MACE. “w/o MACE” denotes without the MACE.

12. **Figure:** F012.png



Annotation: (1) Fig. 12. Ablation study of GFSM. “w/o GFSM” denotes without the GFSM.

13. **Figure:** F013.png



Annotation: (1) Fig. 13. An ultra-high-resolution stylized panoramic image (10,000 × 5,000 pixels).