Progressive Attentional Manifold Alignment for Arbitrary Style Transfer

Xuan Luo[1], Zhen Han *[2], and Linkang Yang[1]

[1] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, China
[2] School of Computer Science, Wuhan University, Wuhan, Hubei, China
https://github.com/luoxuan-cs/PAMA

Fig. 1. The style degradation problem. Stylization results of AdaIN [13], WCT [22], SANet [29], AdaAttN [25], StyleFormer [38], IEC [2], and MAST [14] are shown. Existing methods may fail to transfer the brushstrokes (the first row) or the color distribution of the style image (the second row).

Abstract. Arbitrary style transfer algorithms can generate stylization results with arbitrary content-style image pairs but will distort content structures and bring degraded style patterns. The content distortion problem has been well issued using high-frequency signals, salient maps, and low-level features. However, the style degradation problem is still unsolved. Since there is a considerable semantic discrepancy between content and style features, we assume they follow two different manifold distributions. The style degradation happens because existing methods cannot fully leverage the style statistics to render the content feature that lies on a different manifold. Therefore we designed the progressive attentional manifold alignment (PAMA) to align the content manifold to the style manifold. This module consists of a channel alignment module to emphasize related content and style semantics, an attention module to establish the correspondence between features, and a spatial interpolation module to adaptively align the manifolds. The proposed PAMA can alleviate the style degradation problem and produce state-of-the-art stylization results.

Keywords: Style Transfer · Manifold Alignment · Image Synthesis.

* Corresponding author, hanzhen_2003@hotmail.com.

1 Introduction

Neural style transfer aims at rendering a content image with style patterns from a style image. The pioneering style transfer algorithms rather needs online optimization [6, 7, 18, 9], or be constrained to a few styles [16, 33, 34, 21, 37, 19]. Arbitrary style transfer methods [3, 13, 22, 20, 29] enable real-time stylization with arbitrary styles by leveraging statistical information. These flexible yet efficient approaches have received widespread attention from academics and industries. However, arbitrary style transfer methods suffer from the content distortion problem and the style degradation problem. The content distortion problem can be alleviated using the high frequency components [10], salient map guidance [26], and low-level features [25]. This paper focuses on the style degradation problem, which is less studied but significantly influences stylization quality.

Fig.1 demonstrates the style degradation problem. The first row shows examples generated using the style with thick colorful lines. The AdaIN [13], AdaAttN [25], StyleFormer [38], IEC [2] stylized the content structure insufficiently. Only a limited set of style patterns are used for rendering. Although the WCT [22], SANet [29], and AdaIN [13] brings sufficient style patterns, the patterns are mixed chaotically (WCT), distorted locally (SANet), or blurred and overlapped (MAST). These methods struggle to migrate the brushstroke information. The second row demonstrates that existing methods may also damage the color distribution. The AdaIN, WCT, and SANet mix the colors of the style image, producing a tainted appearance. Moreover, the AdaAttN, StyleFormer, IEC, and MAST merely adopt a few colors for rendering.

We identified that the semantic discrepancy between content and style images brings the style degradation problem. Since the content images are natural images but the style images are artificial images, the content and style features follow different manifold distributions [14]. Utilizing statistics of the data from different distributions is inherently challenging. For instance, the AdaIN [13] use the mean and variance of the style feature to transform the content feature. However, directly applying the style feature statistics distorts the style semantics and brings the style degradation problem. The MAST [14] learns a projection matrix to align the content manifold to the style manifold. The WCT [22] adopts whitening and coloring transformations to adjust the correlation of the content feature. Nevertheless, the learning-free linear transformations of MAST and WCT have limited expressiveness. For patch-based methods like SANet [29], AdaAttN [25], and StyleFormer [38], it is difficult to measure complex relations between content and style feature vectors. The patch-based methods cannot reorganize the style feature vectors to form complicated patterns.

To alleviate the style degradation problem, we proposed the progressive attentional manifold alignment (PAMA) to align the content manifold to the style manifold. The PAMA consists of a channel alignment module, an attention module, and a spatial interpolation module. In the channel alignment module, we adopt the squeeze and excitation operation [12] to extract channel weights of the content and style features. Then the content channel weights are used to

re-weight the style feature while the style channel weights are used to re-weight the content feature. This cross-manifold channel alignment emphasizes the related semantics (the related channels) of content and style features, helping the attention module parse complex relations between them. The attention module is the style attentional network [29] that computes pair-wise similarity (attention map) between content and style feature vectors. The style feature is redistributed according to the attention map, building the correspondence between the content and style feature vectors. The spatial interpolation module summarizes spatial information to adaptively interpolate between the content feature and the redistributed style feature. The content and style feature vectors with similar semantics are linearly fused, forming an intermediate manifold between the content manifold and style manifold. By repeating the whole process multiple times, the content manifold is gradually aligned to the style manifold along a geodesic between them, and thus the semantic gap between content and style features is filled.

We designed a multistage loss function with decreasing weight of content loss to train the progressive alignment procedure. The content loss is the self-similarity loss [18, 23, 30] to preserve the original manifold structure of the content feature. The style loss is the relaxed earth mover distance [18, 23, 30] which optimizes along the style manifold surfaces [30]. A momentum loss [18, 23, 30] is used together with the style loss to preserve the magnitude information. We also adopt a color histogram loss [1] to align the color distributions explicitly. Finally, an auto-encoder loss is used to maintain the common space for manifold alignment. Our contributions can be summarized as follows:

- We proposed a new arbitrary style transfer framework named PAMA, which gradually aligns the content manifold to the style manifold with a channel alignment module, an attention module, and a spatial interpolation module.
- A multistage loss function is designed to enable progressive manifold alignment. We also adopt an auto-encoder loss to maintain the shared space for manifold alignment.
- Experiments show that the proposed framework can generate fine-grained stylization results in real-time (100 fps for 512px images on Tesla V100 GPU). The style degradation problem is alleviated significantly.

2 Related Works

2.1 Arbitrary Style Transfer

The goal of arbitrary style transfer is to generate stylization results in real-time with arbitrary content-style pairs. The mainstream arbitrary style transfer algorithms can be divided into two groups: the global transformation based and local patch based. The global transformation based methods utilize global statistics for feature transformation. One of the representative methods is AdaIN [13], which forces the mean and variance of content features to be the same as the style features. To reduce the memory consumption, DIN [15] substitutes the VGG [32]

with the MobileNet [11] and adopts dynamic convolutions for feature transformation. Although practical, using the global statistics from another manifold distribution to transform the content feature brings degraded style patterns. There are also manifold alignment based style transfer algorithms performing global transformations in a common subspace. The WCT [22] changes the covariance matrix of content features with whitening and coloring transformation, which aligns the self-similarity structure of the content and style manifolds. The MAST [14] learns to project the content manifold to a subspace where the correspondence between the two manifolds can be found. However, the WCT and MAST use learning-free transformations for manifold alignment, limiting their ability to transfer complicated style patterns.

For the local patch based methods, they manipulate the feature patches for stylization. The style swap proposed by Chen *et al.* [3] is the earliest patch based method, which swaps the content patches with the most similar style patches. The DFR [9] and AAMS [39] further extend this method with global hints. Recently, the SANet [29] is proposed to matching content and patches using the attention mechanism. Then MANet [4] disentangles the content and style representation for the attention mechanism. AdaAttN [25] enhances the attention mechanism with multi-layer features to reduce content distortion. IEC [2] uses internal-external learning and contrastive learning to refine the feature representation of SANet. However, the semantic discrepancy between content and style features makes the nearest neighbor search and attentional matching struggle to parse complex semantic relations, triggering the style degradation problem. With the manifold alignment process, the proposed PAMA can parse complex relations like regional correspondence and high-order relations.

2.2 Manifold Alignment

Manifold alignment aims at revealing the relationship between two datasets from different manifold distributions. These algorithms learn to transform the two datasets in a shared subspace, establishing the correspondence of the two datasets while preserving their manifold structures. Existing manifold alignment methods for domain adaptation can be divided into subspace learning methods [5, 8], distribution alignment methods [27, 36]. Huo *et al.* [14] introduce manifold alignment methods to the style transfer community, which assumes that the content and style features follow different manifold distributions. This subspace learning method aligns the content manifolds to style manifolds with a global channel transformation. The proposed PAMA is a distribution alignment method that uses linear interpolation to align the content manifold to the style manifold. The alignment is performed along a geodesic between the two manifolds.

3 Method

3.1 Overall Framework

Fig.2 shows the architecture of the proposed progressive attentional manifold alignment (PAMA). Our method uses a pre-trained VGG [32] network to encode

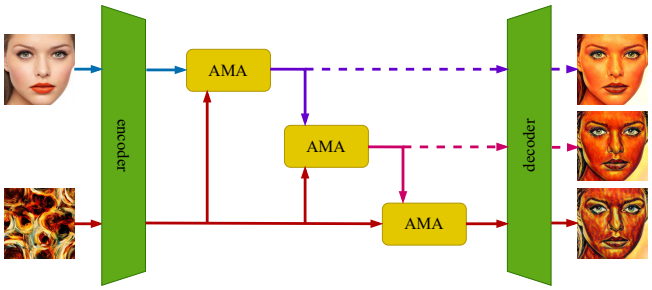

Fig. 2. The overall framework of the progressive attention manifold alignment (PAMA). The content manifold is gradually aligned to the style manifold with three independent attentional manifold alignment (AMA) blocks. The dash lines are only forwarded during training to generate the intermediate results for loss calculation.

the content image I_c and style image I_s, resulting the $ReLU4\_1$ features F_c and F_s. The features are transformed by the attentional manifold alignment (AMA) block for stylization, which consists of a channel alignment module, an attention module, and a spatial interpolation module (Fig.3). Passed through three AMA blocks, the aligned content feature will be fed into the decoder to generate the stylized image. Following the setting of [13], the structure of the decoder is symmetric to the encoder.

3.2 Channel Alignment Module

The channel alignment module aims at emphasizing the related semantic aspects (related channels) between content and style features. The alignment is achieved by manipulating cross-manifold information to re-weight feature channels (Fig.3). We adopted the squeeze-and-excitation operation of SENet [12]:

$$W = MLP(GAP(F)) \tag{1}$$

where a global average pooling operation (GAP) pools $F \in \mathbf{R}^{H\times W\times C}$ into $\mathbf{R}^C$, and a multilayer perceptron (MLP) is used to embed the channel feature to obtain the channel weights. The H, W denotes the height, width, and channels of the feature F. We applied Eq.1 on the content feature F_c and the style feature F_s to obtain the channel weights $A_c \in \mathbf{R}^C$ and $A_s \in \mathbf{R}^C$. As demonstrated in Fig.3, the features F_c and F_s are crossly re-weighted with A_s and A_c to, resulting the aligned features $\hat{F}_c$ and $\hat{F}_s$. The related feature channels (or related semantic aspects) between the content and style features are enhanced. This global channel re-weighting operation can help the attention module to parse cross-manifold semantics.

3.3 Attention Module

The middle part of Fig.3 is the attention module, which redistribute the style feature according to the content structure. This module builds the spatial cor-

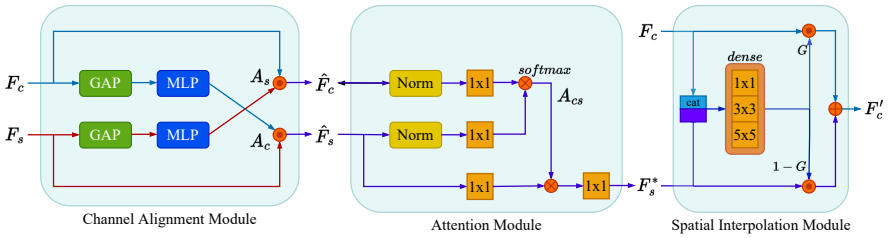

Fig. 3. The details of a single attentional manifold alignment (AMA) block. It consists of a channel alignment module, an attention module, and a spatial interpolation module.

respondence of the content and style feature vectors. The attention map is computed using normalized features:

$$A_{cs} = softmax(f(Norm(\hat{F}_c)) \otimes g(Norm(\hat{F}_s))^T) \tag{2}$$

where the $f(\cdot)$ and $g(\cdot)$ denote 1x1 convolution blocks for feature embedding, the $Norm(\cdot)$ refers to the mean-variance normalization, and the $\otimes$ is the matrix multiplication. With the attention map A_{cs} containing pair-wise similarities, the style feature vectors are redistributed according to the content feature:

$$F_s^* = \theta(A_{cs} \otimes h(\hat{F}_s)) \tag{3}$$

where the $h(\cdot)$ and $\theta(\cdot)$ are 1x1 convolution blocks for feature embedding. The F_s^* denotes the redistributed style feature. The architecture of the attention module is the same as the ones of [29, 4, 2].

3.4 Spatial Interpolation Module

The right part of Fig.3 shows the structure of the spatial interpolation module. The spatial interpolation module summarizes spatial information to adaptively interpolate between the content feature F_c and the redistributed style feature F_s^*. Concretely, the dense operation applies multi-scale convolution kernels on the concatenated feature to compute the interpolation weights $G \in \mathbf{R}^{H \times W}$:

$$G = \frac{1}{n} \sum_{i=1}^{n} \psi_i([F_c, F_s^*]) \tag{4}$$

where $\psi_i(\cdot)$ represent the i-th convolution kernel, and the $[\cdot, \cdot]$ denotes the channel concatenation operation. The concatenated feature can help us to identify the local discrepancy between the corresponding content and style feature, figuring out the appropriate interpolation strength. The interpolation weights G is then applied for linear interpolation:

$$F_c' = G \odot F_c + (1 - G) \odot F_s^* \tag{5}$$

where the $\odot$ refers to the Hadamard production. For the reason that the style feature has been redistributed by the attention module, the spatial interpolation module actually fuses the most similar content and style feature vectors. The manifold alignment is achieved by linearly redistributing the style feature and interpolating its linear components to the content feature.

3.5 Loss Functions

Multistage losses are applied for the proposed progressive attentional manifold alignment (PAMA). For each stage, the loss is a weighted summation of the self-similarity loss [31, 18] L_{ss}, the relaxed earth mover distance loss [18, 23, 30] L_r, the momentum loss [18, 30] L_m, and the color histogram loss [1] L_h. The overall loss is the weighted sum of the multistage losses and the auto-encoder loss L_{ae}:

$$L = \sum_i (\lambda_{ss}^i L_{ss}^i + \lambda_r^i L_r^i + \lambda_m^i L_m^i + \lambda_h^i L_h^i) + L_{ae} \tag{6}$$

where the λ_x^i denotes the weight for L_x ($x \in \{ss, r, m, h\}$) of the i-th stage. In the first stage, a high initial value of the self-similar loss weight λ_{ss} is set to preserve the content manifold structure. In the following stages, the weight of the self-similarity loss decreases gradually to relax constraints and align the content manifold to the style manifold.

Our content loss is based on the L_1 distance between the self-similarity matrices of the content feature F_c and the VGG [32] feature of the stylized image F_{cs}:

$$L_{ss} = \frac{1}{H_c W_c} \sum_{i,j} \| \frac{D_{ij}^c}{\sum_i D_{ij}^c} - \frac{D_{ij}^{cs}}{\sum_j D_{ij}^{cs}} \|_1 \tag{7}$$

where D^c and D^{cs} are the pairwise cosine distance matrices of the F_c and F_{cs}, and the subscript ij denotes the element from the $i-th$ row $j-th$ column. For simplicity, the superscript of λ and L indicating the manifold alignment stage are omitted.

Following the setting of [18, 30] we adapts the relaxed earth mover distance (REMD) to align the content manifold to the style manifold:

$$L_r = \max(\frac{1}{H_s W_s} \sum_i \min_j C_{ij}, \frac{1}{H_c W_c} \sum_j \min_i C_{ij}) \tag{8}$$

where the C_{ij} denotes the pair-wise cosine distance matrix between F_{cs} and F_s. We also added the moment matching loss [18, 30] to regularize the magnitude of features:

$$L_m = \|\mu_{cs} - \mu_s\|_1 + \|\Sigma_{cs} - \Sigma_s\|_1 \tag{9}$$

where the μ and Σ denote the mean and covariance matrix of feature vectors. The subscript s denotes the statistic of style feature and the subscript cs denotes the statistic of the VGG feature of stylization result.

The color distribution plays a central role in solving the style degradation problem. We adopt the differentiable color histogram loss proposed in HistoGAN [1] to learn to match the color distribution explicitly:

$$L_h = \frac{1}{\sqrt{2}} \| H_s^{1/2} - H_{cs}^{1/2} \|_2 \tag{10}$$

where the H refers the color histogram feature [1], the $H^{1/2}$ denotes the element-wise square root.

The target of the auto-encoder loss is maintaining the common subspace for manifold alignment. It learns how to reconstruct the content and style images using the encoded content and style features. We optimize pixel level reconstruction and semantic level reconstruction simultaneously:

$$\begin{aligned} L_{ae} = \lambda_{ae}(\|(I_{rc} - I_c)\|_2 + \|(I_{rs} - I_s)\|_2) + \\ \sum_i (\|\phi_i(I_{rc}) - \phi_i(I_c)\|_2 + \|\phi_i(I_{rs}) - \phi_i(I_s)\|_2) \end{aligned} \tag{11}$$

where I_{rc} and I_{rs} are the content and style images reconstructed from the encoded features, and the λ_{ae} is a constant weight. The $\phi_i(I)$ refers to the $ReLU\_i\_1$ layer VGG feature of image I. This common subspace is critical for the attention module to avoid matching features from different subspaces. The attention operation cannot accurately measure the similarities of features from different subspaces, leading to the shifted matching of feature vectors (Fig.7). The common subspace also helps the spatial interpolation module interpolate between features from the same subspace, thus reducing the information loss of the pipeline.

4 Experiments

4.1 Implementation Details

Our proposed PAMA is trained with content images from COCO [24] and style images from wikiart [28]. VGG [32] features from $ReLU\_3\_1$, $ReLU\_4\_1$, and $ReLU\_5\_1$ are used to compute the self-similarity loss L_{ss}, the REMD loss L_r, and the moment matching loss L_m. For self-similarity loss, we set its weights λ_{ss}^1, λ_{ss}^2, λ_{ss}^3 to 5, 4, 3 respectively. All weights of the style loss terms L_r and L_m are set to 1. The weights of color histogram loss λ_h^1, λ_h^2, λ_h^3 are set to 1, 2, and 4. The λ_{ae} for the reconstruction loss in Eq.11 is 50. We use the Adam [17] optimizer with learning rate of 0.0001 and momentum of 0.9. We take 8 content-style image pairs as a batch. The smaller dimension of content and style images are rescaled to 512, and then we randomly crop a 256x256 patch for efficient training. In the testing phase, our fully convolutional network can tackle images with any size.

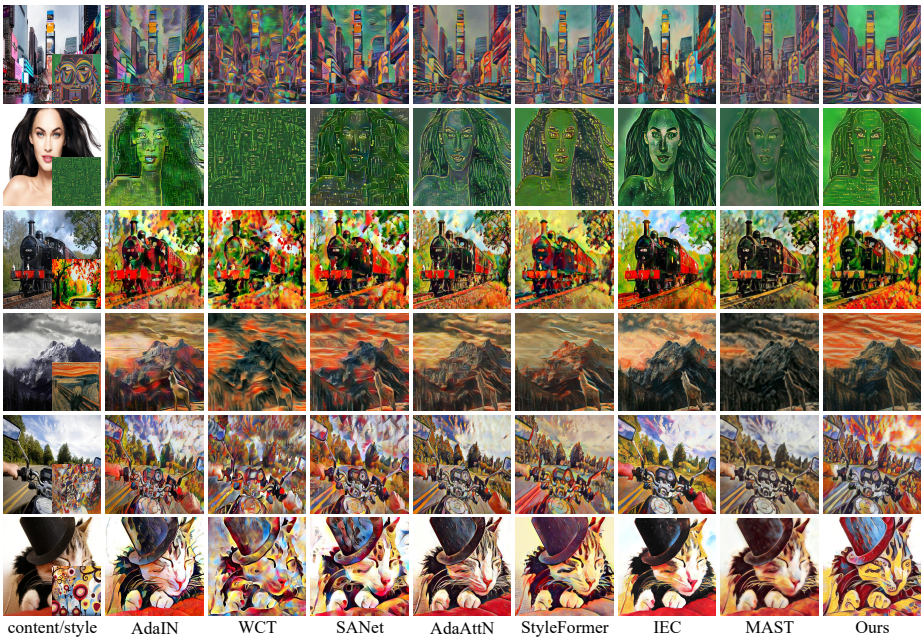

Fig. 4. Qualitative Comparison.

4.2 Comparison with Prior Arts

To evaluate the proposed progressive attentional manifold alignment (PAMA), we compared it with other arbitrary style transfer methods, including global statistic based methods AdaIN [13], WCT [22], and MAST [14], local patch based methods SANet [29], AdaAttN [25], StyleFormer [38], and IEC [2].

Qualitative Comparison. Fig.4 shows the comparison results. The AdaIN [13] often renders with crack-like style patterns (the 1st, 2nd, 3rd, 5th rows) or shifted color distributions (the 1st, 4th, 5th, 6th rows). Since the content and style features lie on different manifolds, directly using the mean and variance of style features is suboptimal. The WCT [22] performs the whitening and coloring transformations to adjust the covariance of the content feature to be the same as the style feature. The covariance based transformations of WCT may mix the style patterns chaotically (the 1st, 4th, 5th, 6th rows) or distort the content structures (the 2nd, 3rd, 5th rows). The SANet [29] is a patch based method that uses the attention mechanism to match content and style patches, but mixes the style patterns (the 1st, 2nd, 3rd, 6th rows) and brings arc-shaped artifacts (1st, 5th rows). The AdaAttN [25] further introduced low-level features to improve the content preservation ability of SANet, but suffers from understylization (the 3rd, 4th, 5th, 6th rows) and color distribution shift (the 1st, 2nd, 4th, 6th rows). The StyleFormer [38] adopts a transformer [35] inspired network for stylization, but

produces blurred (the 2nd, 3rd, 4th rows) and understylized (the 2nd, 5th, 6th rows) results. The IEC [2] suffers from the same problems of AdaAttN, which are the understylization problem (the 2nd, 3rd, 4th, 5th, 6th rows) and the color distribution shift problem (1st, 2nd). These patch based approaches cannot parse the complex relations between the content and style patches. The MAST [14] aligns the content manifold to the style manifold with subspace projection. This learning-free manifold alignment method has limited model capacity, thus not free from the blurry problem (the 3rd, 4th, 5th rows) and the understylization problem (the 2nd, 5th, 6th rows). The proposed PAMA aligns the content manifolds to the style manifolds, enabling the attention mechanism to establish complex relations like regional correspondence and high-order dependencies. The regional correspondence suggests that content semantic regions (e.g., the sky of the 4th row) are consistently rendered by continuous style patterns (the 1st, 4th rows). The high-order dependencies refer to the multi-level dependencies formed by progressive attention, which is essential to express complicated style patterns (the 6th, 7th rows). The proposed PAMA can alleviate the style degradation problem significantly.

Table 1. User Study.

method	content quality	style quality	overall quality	total
AdaIN [13]	273	165	252	690
WCT [22]	231	234	207	672
SANet [29]	374	351	453	1178
AdaAttN [25]	1382	541	956	2879
StyleFormer [38]	626	370	514	1510
IEC [2]	1042	589	934	2565
MAST [14]	286	242	235	763
Ours	786	2508	1449	4743

User Study. As shown in Tab.1 We perform user study on eight arbitrary style transfer methods: AdaIN [13], WCT [22], SANet [29], AdaAttN [25], StyleFormer [38], IEC [2], MAST [14], and the proposed PAMA. In the study, a single sample consists of a content image, a style image, and eight corresponding stylization results generated by the eight methods. We use 25 content images and 25 style images to generate 625 samples and randomly draw 20 samples for each user. For each sample, users are asked to choose the best stylization result according to the evaluation indicators: content quality, style quality, and overall quality. Each user judges the quality according to their subjective perception. We collect 5000 votes from 250 users for each indicator. The results are shown in Table A. The result demonstrates that our method produces results with better style quality and overall performance.

Table 2. Stylization time comparison.

Method	Time (256px)	Time (512px)
AdaIN [13]	2.813ms	2.912ms
WCT [22]	1210ms	4053ms
SANet [29]	4.792ms	6.351ms
AdaAttN [25]	19.76ms	22.52ms
StyleFormer [38]	7.154ms	9.790 ms
IEC [2]	4.936ms	6.851ms
MAST [14]	1488ms	2743ms
Ours	8.726ms	9.937ms

Efficiency. In Tab.2, we compare the stylization time of our method with other baselines at 256px and 512px. All of the methods are evaluated on a server with an NVIDIA V100 PCIe 32G GPU. The results are the average running time of 3600 image pairs. The proposed PAMA can generate artworks in real-time (100 fps at 512px).

4.3 Ablation Study

Loss Analysis. In this part, we explore the effect of different loss weights. Firstly, we change the weight of relaxed earth mover distance loss λ_r and the weight of momentum loss λ_m simultaneously to verify the influence of stylization strength. As shown in Fig.5 (b), when the style loss is comparably low, the proposed PAMA only transfers the color information of the style image. Meanwhile, the content structure is well preserved. As the value of λ_r and λ_m increase, the delicate triangle like texture (the first row of Fig.5) and water lily patterns (the 2nd row of Fig.5) are introduced. In conclusion, higher λ_r and λ_m suggest more global and complicated style patterns. We also adjust the weight of the color histogram loss λ_h to demonstrate its effectiveness. The first row Fig.6 shows that higher λ_h brings richer color patterns. The second row shows that the color histogram loss can even help the proposed PAMA match the distribution of the greyscale image. With these loss functions, the proposed PAMA can not only migrate the complicated style patterns but also render with the style palette.

Subspace Constraint. In this part, we remove the subspace constraint (Eq.11) to explore its influences. The color distributions of the results from Fig.7 (b) and (c) are completely opposite. Without the subspace constraint, the subspace changes whenever the attention module is performed. The attention modules (except the first one) need to compute the pair-wise similarities with features from different subspaces, leading to the shifted matching of content and style feature vectors, which further influences the color distribution. The common subspace maintained by Eq.11 is necessary for the manifold alignment process.

Channel Alignment. To verify the effectiveness of the channel alignment module, we remove the channel alignment module of PAMA for ablation study. As

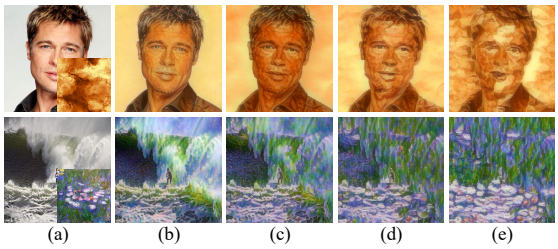

Fig. 5. The influence of the relaxed earth mover distance loss λ_r and the momentum loss λ_m. (a) the content and style images; (b) 0.5x weights; (c) 1x weights (the original weights); (d) 2x weights; (e) 4x weights.

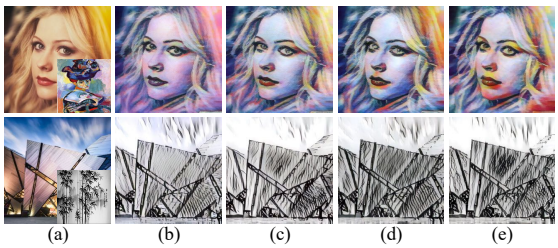

Fig. 6. The influence of the color histogram loss λ_r. (a) the content and style images; (b) 0.5x weights; (c) 1x weights (the original weights); (d) 2x weights; (e) 4x weights.

shown in Fig.8 (c), the style patterns generated without the channel alignment module overlap with each other (the first row) or form a blurred appearance (the second row). The attention module cannot establish a stable and consistent measurement. In the contrary, a consistent regional correspondence can be find in Fig.8 (b). The trees, buildings, and sky are consistently matched with the blue, red, and yellow patterns (the first row). The bridge, ocean, and sky are rendered with distinguishable brown, blue-green, and blue patterns(the second row). The channel alignment module is critical to parse the semantic relations between the content and style images.

Spatial Interpolation In this section, we perform ablation study about the spatial interpolation module. Since the redistributed feature is directly passed to the next stage without interpolating with the content feature, the content structure information gradually vanishes during the manifold alignment process. Compared to Fig.8 (b), the content structures of Fig.8 (d) are blurry and distorted. By removing the spatial interpolation module, the stylized feature of the first stage is re-rendered multiple times without the guidance of content information. This makes the manifold deviates from the geodesic between the content and style manifolds, producing collapsed results. The spatial interpolation module is indispensable for the proposed PAMA.

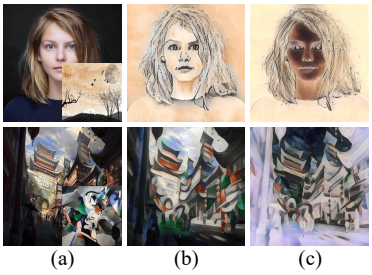

Fig. 7. The influence of the subspace constraint. (a) the content and style images; (b) the results generated with the subspace constraint; (c) the results generated without the subspace constraint.

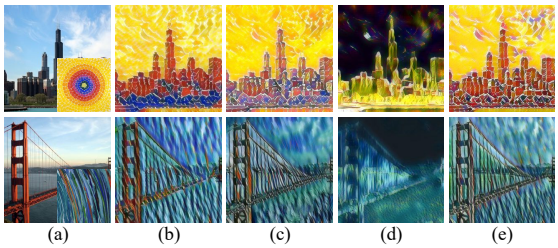

Fig. 8. Ablation studies about the architecture. (a) the content and style images; (b) the original PAMA; (c) PAMA without the channel alignment module; (d) PAMA without the spatial interpolation module; (d) PAMA with single-stage manifold alignment.

Progressive Alignment In this part, we demonstrate the effectiveness of the progressive alignment mechanism of the proposed PAMA. Fig.8 (e) shows the results generated with single-stage manifold alignment. The single-stage manifold alignment fails to establish the regional correspondence as Fig.8 (b) but renders with uniform and mixed style patterns. In the first row of Fig.8 (e), the blue patterns overlap with the yellow and orange pattern. In the second row, the result is rendered by uniform blue-green patterns without apparent variance. The attention module cannot distinguish between different content structures to render them differently. Although single-stage manifold alignment can provide acceptable results, it cannot transfer complicated style patterns and suffers from the understylization problem. Uniform and repetitive pattern are adopted for rendering. The capacity of the single-stage manifold alignment is limited.

4.4 Multi-style Transfer

Multi-style Interpolation In this part, we linearly interpolate between stylized features of different style images. Specifically, we linearly interpolate between the stylized features from the third stage, and decode them to obtain the final

Fig. 9. Linear interpolation between stylized features.

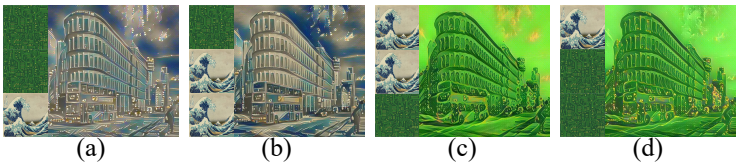

Fig. 10. Multi-style alignment. The content manifold is aligned with different style manifolds from top to bottom.

stylization results. Fig.9 illustrates the results interpolated with the five different weights. The blue and white wave patterns gradually shift to the yellow and green circuit patterns by changing the interpolation weights.

Multi-style Alignment The proposed PAMA consists of three attentional manifold alignment (AMA) stages (Fig.3). We align the content manifold to different style manifolds (different style feature for the three stages) to produce stylization results with textures similar to one set of styles but color distributions similar to another. Fig.10 shows the results of multi-style alignment. In Fig.10, example (a) and (b) have circuit-like pattern in blue and grey, while example (c) and (d) consists of waves with a greenish appearance. The proposed PAMA can render the texture information and color distribution separately by aligning to multiple style manifolds.

5 Conclusion

We proposed the progressive attention manifold alignment module (PAMA) to solve the style degradation problem. The proposed PAMA consists of a channel alignment module to emphasize related semantics, an attention module to establish correspondence between features, and a spatial interpolation module to align the content manifold to the style manifold. The proposed PAMA can produce high-quality stylization results in real-time.

Acknowledgements This work was supported in part by the National Nature Science Foundation of China under Grant 62072347.

References

1. Afifi, M., Brubaker, M.A., Brown, M.S.: Histogan: Controlling colors of gan-generated and real images via color histograms. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 7941–7950 (2021)
2. Chen, H., Zhao, L., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D.: Artistic style transfer with internal-external learning and contrastive learning. In: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 26561–26573 (2021)
3. Chen, T.Q., Schmidt, M.: Fast patch-based style transfer of arbitrary style. CoRR **abs/1612.04337** (2016)
4. Deng, Y., Tang, F., Dong, W., Sun, W., Huang, F., Xu, C.: Arbitrary style transfer via multi-adaptation network. In: MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. pp. 2719–2727 (2020)
5. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. 2013 IEEE International Conference on Computer Vision pp. 2960–2967 (2013)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 262–270 (2015)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2414–2423 (2016)
8. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. 2012 IEEE Conference on Computer Vision and Pattern Recognition pp. 2066–2073 (2012)
9. Gu, S., Chen, C., Liao, J., Yuan, L.: Arbitrary style transfer with deep feature reshuffle. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8222–8231. Computer Vision Foundation / IEEE Computer Society (2018)
10. Hong, K., Jeon, S., Yang, H., Fu, J., Byun, H.: Domain-aware universal style transfer. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 14589–14597 (2021)
11. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv **abs/1704.04861** (2017)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
13. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 1510–1519 (2017)
14. Huo, J., Jin, S., Li, W., Wu, J., Lai, Y.K., Shi, Y., Gao, Y.: Manifold alignment for semantically aligned style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14861–14869 (2021)
15. Jing, Y., Liu, X., Ding, Y., Wang, X., Ding, E., Song, M., Wen, S.: Dynamic instance normalization for arbitrary style transfer. In: The Thirty-Fourth AAAI

Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 4369–4376 (2020)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II. pp. 694–711 (2016)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
18. Kolkin, N.I., Salavon, J., Shakhnarovich, G.: Style transfer by relaxed optimal transport and self-similarity. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 10051–10060 (2019)
19. Kotovenko, D., Sanakoyeu, A., Lang, S., Ommer, B.: Content and style disentanglement for artistic style transfer. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 4421–4430 (2019)
20. Li, X., Liu, S., Kautz, J., Yang, M.: Learning linear transformations for fast image and video style transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 3809–3817 (2019)
21. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.: Diversified texture synthesis with feed-forward networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 266–274 (2017)
22. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.: Universal style transfer via feature transforms. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 386–396 (2017)
23. Lin, T., Ma, Z., Li, F., He, D., Li, X., Ding, E., Wang, N., Li, J., Gao, X.: Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 5141–5150 (2021)
24. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. pp. 740–755 (2014)
25. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 6629–6638 (2021)
26. Liu, X., Liu, Z., Zhou, X., Chen, M.: Saliency-guided image style transfer. In: IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2019, Shanghai, China, July 8-12, 2019. pp. 66–71 (2019)
27. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. 2013 IEEE International Conference on Computer Vision pp. 2200–2207 (2013)
28. Nichol, K.: Painter by numbers (2016)

29. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 5880–5888 (2019)
30. Qiu, T., Ni, B., Liu, Z., Chen, X.: Fast optimal transport artistic style transfer. In: International Conference on Multimedia Modeling. pp. 37–49 (2021)
31. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
33. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks: Feed-forward synthesis of textures and stylized images. In: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. pp. 1349–1357 (2016)
34. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 4105–4113 (2017)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017)
36. Wang, J., Chen, Y., Hao, S., Feng, W., Shen, Z.: Balanced distribution adaptation for transfer learning. 2017 IEEE International Conference on Data Mining (ICDM) pp. 1129–1134 (2017)
37. Wang, X., Oxholm, G., Zhang, D., Wang, Y.: Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 7178–7186 (2017)
38. Wu, X., Hu, Z., Sheng, L., Xu, D.: Styleformer: Real-time arbitrary style transfer via parametric style composition. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 14598–14607 (2021)
39. Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y.J., Wang, J.: Attention-aware multi-stroke style transfer. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1467–1475 (2019)