

## Attention-aware Multi-stroke Style Transfer

Yuan Yao<sup>1</sup>, Jianqiang Ren<sup>2</sup>, Xuansong Xie<sup>2</sup>, Weidong Liu<sup>1</sup>, Yong-Jin Liu<sup>1</sup>, Jun Wang<sup>3</sup>  
<sup>1</sup>Tsinghua University, <sup>2</sup>Alibaba Group, <sup>3</sup>University College London

### Abstract

Neural style transfer has drawn considerable attention from both academic and industrial field. Although visual effect and efficiency have been significantly improved, existing methods are unable to coordinate spatial distribution of visual attention between the content image and stylized image, or render diverse level of detail via different brush strokes. In this paper, we tackle these limitations by developing an attention-aware multi-stroke style transfer model. We first propose to assemble self-attention mechanism into a style-agnostic reconstruction autoencoder framework, from which the attention map of a content image can be derived. By performing multi-scale style swap on content features and style features, we produce multiple feature maps reflecting different stroke patterns. A flexible fusion strategy is further presented to incorporate the salient characteristics from the attention map, which allows integrating multiple stroke patterns into different spatial regions of the output image harmoniously. We demonstrate the effectiveness of our method, as well as generate comparable stylized images with multiple stroke patterns against the state-of-the-art methods.

### 1. Introduction

Style transfer is a powerful technique for art creation and image editing that enables recomposing the images in the style of other images. Recently, inspired by the power of Convolutional Neural Network (CNN) in visual perception tasks, Gatys *et al.* [4, 5] opens up a new field named Neural Style Transfer, which first introduces neural representations to separate and recombine content and style of arbitrary images. They propose to extract the content features and style correlations (Gram matrix) along the processing hierarchy of a pretrained network classifier. Based on this work, several algorithms have been proposed to accelerate the development in terms of the generalization and efficiency issues, through optimization-based methods [5, 14, 19] and feed-forward networks [10, 12, 23]. The success of style transfer makes it possible to deploy service upon mobile applications (*e.g.*, Prisma, Artify), allowing users to create an art-



Figure 1. Existing methods are unable to coordinate spatial distribution of visual attention between content image and stylized image, misleading the stylized rendering on diverse content regions in an indiscriminate way, resulting in insufficient stylization (Style-Swap), unexpected patterns (AdaIN, StrokePyramid) or distorted attention regions (WCT, Avatar-Net). Due to attention mechanism and multi-stroke fusion strategy, our method can achieve faithful transfer of style and attention consistency with content image simultaneously.

work out of a picture they took with their phones.

In spite of significant progress already achieved, these methods suffer from the restricted binding between the model and specific styles. Recently, *Arbitrary-Style-Per-Model* Fast Style Transfer Methods (ASPM) [9] are proposed to conquer this dilemma. One possible solution is to coordinate high-level statistical distribution between content features and style features. Although visual quality and efficiency can be greatly improved, they unexpectedly introduce unexpected or distorted patterns to the stylized result because of treating diverse image regions in an indiscriminate way, such as AdaIN [7] and WCT [13] in Figure 1. Another solution is to swap the content feature patch with the closest style feature patch at the intermediate layer of a trained autoencoder. However, this method may generate insufficient stylized results when huge difference exists between content and style images, such as the Style-Swap [2] method shown in Figure 1. Compared with Style-Swap, Avatar-Net [21] further dispels the domain gap between content and style features, leading to better stylized results, but it still maintains inconsistent spatial distribution

of visual attention with the content image and thus manifests distortion in terms of semantic perception.

Stroke textons [33], which are referred to fundamental micro-structures in natural images, reflect perceptual style patterns. Methods such as [8, 25] are dedicated to learn stroke control in transfer process. Jing *et al.* [8] first proposes to achieve continuous stroke size control by incorporating multiple stroke sizes into one single StrokePyramid model. The StrokePyramid result shown in Figure 1 is produced by mixing two different stroke sizes. However, due to the lack of local awareness of content image, they perform the stroke interpolation in a holistic way regardless of region diversity, leading to lacking of level of details. In addition, these methods are inflexible to handle arbitrary styles in one feedforward pass.

To address the aforementioned problems, we propose an attention-aware multi-stroke (AAMS) model for arbitrary styles transfer. Our model encourages *attention consistency* (which refers to spatial consistency of visual attention distribution) for corresponding regions between content image and stylized image, and it achieves both scalable multi-stroke fusion control and automatic spatial stroke size control in one shot. Specifically, we introduce self-attention mechanism as complementary to the autoencoder framework. The self-attention module calculates the response at a position as a weighted sum of the features at all positions, which helps to capture long-range dependencies across image regions. By performing a reconstruction training process for the self-attention assembled autoencoder, the attention map could grasp salient characteristics within any content images. As shown in Figure 1, the attention map of the content image highlights the salient parts while enabling consistency of the attention degree for long-range features. Based on the correlation between receptive field and stroke size, a multi-scale style swap module is proposed to blend distinct stroke patterns via swapping the content features with multi-scale style features in high-level representations. We inject the attention map into a multi-stroke fusion module to synthesize distinct stroke patterns harmoniously, which achieves automatic spatial stroke size control. Comprehensive experiments have been conducted to demonstrate the effectiveness of our method, and the model is capable of generating comparable stylized images with multiple stroke patterns against the state-of-the-art methods. The main contributions of this work are:

- We introduce self-attention mechanism to an autoencoder network, allowing capturing critical characteristics and long-range region relations of the input image.
- We propose multi-scale style swap to break the limitation of fixed receptive field in high-level feature space and produce multiple feature maps reflecting different stroke patterns.

- By combining with attention map, we present a flexible fusion strategy to integrate multiple stroke patterns into different spatial regions of the output image harmoniously, which enables the attention consistency between the content image and stylized image.

## 2. Related Work

**Neural Style Transfer.** Arbitrary-Style-Per-Model methods (ASPM) [9] have recently been proposed to transfer arbitrary styles through only one single model. The backend idea is to formulate the style transfer as an image reconstruction process, with feature statistics fusion between content and style features at intermediate layers. Chen *et al.* [2] first proposes to swap the content feature patch with the best matching style feature patch using a Style-Swap operation. Huang *et al.* [7] introduces adaptive instance normalization (AdaIN) to adjust the mean and variance of the content feature to match those of the style feature. Li *et al.* [13] performs the style transfer by integrating the whitening and coloring transforms (WCT) to match the statistical distributions and correlations between the features of content and style. Avatar-Net [21] elevates the feature transfer ability by matching the normalized counterparts features and applying a patch-based style decorator. However, the above methods either locally exchange the closest feature patches or transfer feature statistics globally, which tends to display a uniform stroke pattern without attention aware guaranty. In comparison, our methods perform style transfer to render visually plausible result with multi-stroke patterns integrated within the same stylized image. Another related work is [8], which proposes a StrokePyramid module to incorporate multiple stroke sizes into one single model, it empowers distinct stroke sizes in different spatial regions within the same transformed image. Our approach can automatically manipulate multiple stroke fusion through the guidance of attention feature in one shot within ASPM framework, while [8] achieves spatial stroke size control by feeding masked content image and need to be retrained for each new style.

**Attention Models.** One of the most promising trends in research is the emergence of incorporating attention mechanism into deep learning framework [11, 17]. Rather than compressing an entire image or a sequence into a static representation, attention allows the model to focus on the most relevant part of images or features as needed. Such mechanism has been proved to be very effective in many vision tasks including image classification [26, 32], image captioning [28, 30] and visual question answering [27, 29]. In particular, self-attention [16, 24] has been proposed to calculate the response at a position in a sequence by attending to all positions within the same sequence. Shaw *et al.* [20] proposes to incorporate relative position information for se-

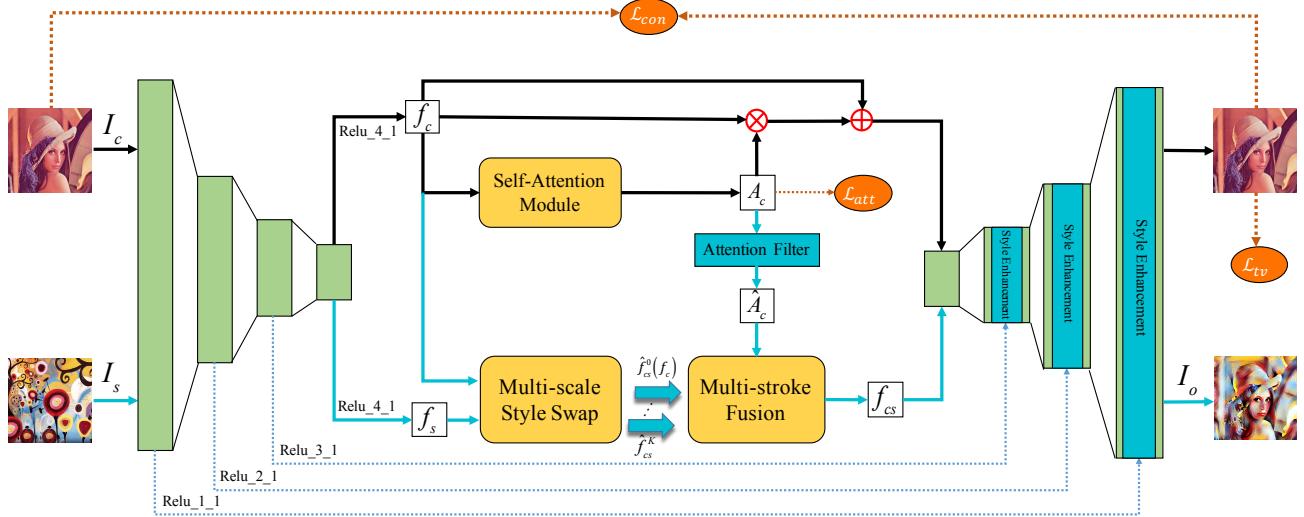


Figure 2. An overview of our network architecture. We train a self-attention autoencoder for image reconstruction in a style-agnostic manner. Given a content image  $I_c$  and a style image  $I_s$  in the testing phase, we perform multi-stroke style transfer via style swap and fusion modules under the guidance of attention map  $\hat{A}_c$ . Skip connections are introduced to enhance the stylization effects during the reconstruction process. The black and green lines represent the training and testing phases respectively, while the orange dotted line denotes the flow of our loss functions.

quences in the self-attention mechanism of the Transformer model, which improves the translation quality on machine translation tasks. Zhang *et al.* [31] demonstrates that the self-attention model can capture the multi-level dependencies across image regions and draw fine details in the context of GAN framework. Compared with [31], we adapt the self-attention to introduce a residual feature map to catch salient characteristics within content images.

### 3. Proposed Approach

For the task of style transfer, the goal is to generate a stylized image  $I_o$ , given the content image  $I_c$  and style image  $I_s$ . To satisfy arbitrary style transfer in one feed-forward pass while integrating images with multiple stroke patterns, we develop three modules( *multi-scale style swap*, *multi-stroke fusion* and *self-attention module*) in the bottleneck layer for feature manipulation. The three modules cooperate with each other and form a coupling framework. The network architecture of our proposed approach is depicted in Figure 2.

Assume that  $f_c$  and  $f_s$  denote the feature maps extracted from the encoder correspond to  $I_c$  and  $I_s$  respectively. At the core of our network, a *self-attention module* is proposed to learn the saliency properties by feeding with  $f_c$ , which will generate self-attention feature map  $A_c$  for any content  $I_c$  after reconstruction training process. During the intermediate process of transferring  $f_c$  into the domain of  $f_s$  via a WCT transformation in the testing phase, a *multi-scale style swap module* is first designed to synthesize features of multiple stroke sizes  $\hat{f}_{cs}^k(k = 1, \dots, K)$ , taking both the con-

tent feature  $f_c$  and style features  $f_s$  as input. The module conducts a style-swap procedure for the content feature and  $K$  style features simultaneously. To perform a flexible integration of the features, a *multi-stroke fusion module* is presented to handle controllable blending. The attention map  $\hat{A}_c$  filtered from  $A_c$  is incorporated to guide fusion among the content feature  $f_c(\hat{f}_{cs}^0)$  and the  $K$  swapped stroke features  $\hat{f}_{cs}^k(k = 1, \dots, K)$ , where  $K$  is the user provided clustering number. After the two steps, the synthetic feature  $f_{cs}$  is fed into the trained decoder to generate the stylized image  $I_o$  in one feed-forward pass. We further introduce skip connections to enhance the stylization effects by adapting multiple level of synthetic features with style features. In this section, we introduce the three components in details.

#### 3.1. Self-Attention Autoencoder

We employ a self-attention mechanism [31] to catch the relationships between separated spatial regions. In detail, let  $\Phi_e(\cdot)$  and  $\Phi_d(\cdot)$  represent the encoder and decoder respectively. The encoder produces  $f_x = \Phi_e(x) \in \mathbb{R}^{H \times W \times C}$  by mapping the input image  $x$  into the intermediate activation space, with  $H, W, C$  denoting height, width and channel number respectively.

Let  $flat(\cdot)$  denote the flatten operation along the channel dimension. We calculate the self-attention feature of  $f_x$  in the  $i^{th}$  location as:

$$A_x^i = \sum_{j=1}^{HW} \alpha_{ij} * flat(f_x \otimes \Theta_h)_j, \quad (1)$$

where  $\otimes$  denotes the convolution operation and  $\alpha_{ij}$  is the

weight coefficient.  $\alpha_{ij}$  indicates the dependencies between two regions, not just for neighboring positions. It is computed using a softmax function:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{HW} \exp(e_{ik})}, \quad (2)$$

where  $e$  is obtained by a compatibility function that compares elements by a scaled dot product:

$$e = \text{flat}(f_x \otimes \Theta_u) * \text{flat}(f_x \otimes \Theta_g)^T, \quad (3)$$

The self-attention feature map decoded from the self-attention module is  $A_x = (a_x^1, a_x^2, \dots, a_x^N)^T \in \mathbb{R}^{N \times C}$ ,  $N = HW$ , which will be reshaped as the same dimensions of  $f_x$ . In the above formulation,  $\Theta_h, \Theta_u, \Theta_g \in \mathbb{R}^{C \times \hat{C}}$  are learned parameter matrices within our self-attention autoencoder framework. We implement them as  $1 \times 1$  convolution [31].

Given the self-attention feature map  $A_x$ , we propose to obtain a self-attention residual  $R_x$  by multiplying the hidden feature map  $f_x$  with  $A_x$ . The output is then given by appending the residual to the feature map  $f_x$ .

$$O_x = R_x + f_x = A_x \odot f_x + f_x, \quad (4)$$

where  $\odot$  denotes the element-wise multiplication operator. We then feed  $O_x$  into the decoder and reconstruct the input image  $\hat{x} = \Phi_d(O_x)$ . In this manner, the degree of variance on each spatial position can be revealed by  $R_x$ , exhibiting the salient regions in synthesizing the image.

Similar to [10, 13, 21], we define the semantic content loss as the sum of perceptual loss and pixel reconstruction loss to generate visually indistinguishable images with the input images:

$$\mathcal{L}_{con} = \sum_{l \in l_c} \|\phi_l(\hat{x}) - \phi_l(x)\|_2^2 + \lambda_p \|\hat{x} - x\|_2^2, \quad (5)$$

where  $\phi_l(x)$  is the activations of the  $l^{th}$  layer of the pre-trained VGG-19 network when processing image  $x$ ,  $\lambda_p$  is the weight to balance the two losses. Both losses are calculated using normalized Euclidean distance. In addition, we introduce a sparse loss on self-attention feature map  $A_x$  to encourage the self-attention autoencoder to pay more attention to small regions instead of the whole image:

$$\mathcal{L}_{att} = \|A_x\|_1. \quad (6)$$

With total variational regularization loss  $\mathcal{L}_{tv}$  [10] added to encourage the spatial smoothness in the generated images, we obtain our total loss function as:

$$\mathcal{L} = \lambda_{con} \mathcal{L}_{con} + \lambda_{att} \mathcal{L}_{att} + \lambda_{tv} \mathcal{L}_{tv}, \quad (7)$$

where  $\lambda_{con}$ ,  $\lambda_{att}$  and  $\lambda_{tv}$  are the balancing factors.

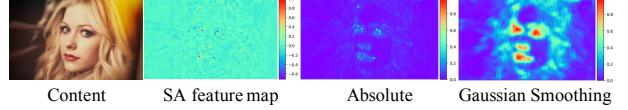


Figure 3. *Attention filter* conducts Absolute and Gaussian smoothing transformation upon the self-attention feature map, the generated attention map can exhibit evident saliency distribution.

### 3.2. Multi-scale Style Swap

Style swap is the process of substituting content feature with the closest style feature patch-by-patch [2]. Given a specific patch size, the style swap procedure can be implemented efficiently as two convolutions operations and a channel-wise argmax operation. The filters of the convolutional layers are derived from the extracted style patches. Based on the analysis of how receptive field influences stroke size in [8], we note that larger patch size leads to larger receptive field and style stroke accordingly. However, the increase of the scale for patch size is strictly limited by the network structure and easily saturated when the patch size is larger than the fixed receptive field of network.

To resolve the above issue in an efficient manner, we propose to fix the patch size while changing the scale of activation feature map of styles, such that we introduce multi-scale stroke patterns after swapping with the same content feature. Specifically, we first adopt whitening transform [13] on content feature  $f_c$  and style feature  $f_s$  to peel off their style information while preserving global structure, results are denoted as  $\hat{f}_c$  and  $\hat{f}_s$ . We then obtain a series of multi-scale style features by casting the whitened style feature into multiple scales:

$$\hat{f}_s^k = \mathcal{T}_{\beta_k}(\hat{f}_s), \quad (8)$$

where  $\mathcal{T}$  denotes the scaling operation,  $\beta_k (k = 1, 2, \dots, K)$  is the scale coefficient controlling different stroke sizes. Finally, the multiple swapped features are produced by performing style-swap procedure between  $\hat{f}_c$  and multiple  $\hat{f}_s^k$  simultaneously.

$$\hat{f}_{cs}^k = \mathcal{F}_{ss}(\hat{f}_c, \hat{f}_s^k), \quad (9)$$

where  $\mathcal{F}_{ss}$  denotes the parallelizable style swap process [2].

### 3.3. Multi-stroke Fusion

Equipped with self-attention mechanism, the residual  $R_c$  in eq. (4) is able to capture critical characteristics and long-range region correlations of the feature map  $f_c$ . In detail, during the reconstruction process, the residual learns to fine tune intrinsically crucial parts of content feature by adding variation to it, therefore the non-trivial (zero) parts of the residual deserve special attention. We utilize an *attention filter* by first performing an absolute operation to highlight

these non-trivial parts in  $A_c$ , following by a Gaussian kernel convolution layer to enhance the regional homogeneity of the features. The variance of the Gaussian kernel can further be used to control the salient region’s proportion in content image. We obtain our *attention map*  $\hat{A}_c$  after normalizing into the range  $[0, 1]$ . Figure 3 shows the visualization image of the intermediate results. We notice that the attention map can enlarge the attending influence of saliency regions while maintaining the correlations among distant regions.

In addition to the  $K$  swapped features  $\hat{f}_{cs}^k$  obtained from the multi-scale style swap, we introduce the whitened content feature  $\hat{f}_c$  as another entry to characterize the most significant regions, which is known as the fine-grained stroke  $\hat{f}_{cs}^0$ . Therefore, we have in total of  $K+1$  features for multi-stroke fusion.

To integrate arbitrary strokes in a scalable framework, we propose a flexible fusion strategy by first dividing the attention map  $\hat{A}_c$  into multiple clusters according to the stroke number provided. We apply k-means method to cluster our attention map, the goal is to iteratively find  $K+1$  intensity centers, minimizing the Euclidean distance between center and elements among each clusters.

$$\arg \min \mathcal{G} = \sum_{k=0}^K \sum_{\hat{a}_c^i \in S_k} \|\hat{a}_c^i - m_k\|^2, \quad (10)$$

where  $K+1$  clusters are generated, and  $m_k \in [0, 1]$  denotes the mean intensity value of all the attention points  $\hat{a}_c^i$  in cluster  $S_k$ .

The multi-stroke fusion can then be formulated as integrating the content feature with multiple swapped features under the guidance of attention map.

$$\hat{f}_{cs} = \sum_{k=0}^K \hat{A}_c^k \hat{f}_{cs}^k, \quad (11)$$

where  $\hat{A}_c^k$  is the weight map assigned to the  $k^{th}$  stroke size, and  $k = 0$  denotes the fine-grained stroke. The weight map is computed by a softmax function:

$$\hat{A}_c^k = \frac{e^{\gamma \cdot (1 - D_k)}}{\sum_{i=0}^K e^{\gamma \cdot (1 - D_i)}}. \quad (12)$$

We define  $D_k = |\hat{A}_c - m_k|$  to measure the absolute distance with regard to the center  $m_k$ , thus  $1 - D_k$  can be used to indicate how much extent each stroke size contributes to synthesis the feature. The smoothing factor  $\gamma$  is used to control the smoothing degree for fusion.

Before feeding into the decoder to generate the stylized result, we derive  $f_{cs}$  by performing coloring transform with the syncretic feature  $\hat{f}_{cs}$  to match feature statistics to the style feature  $f_s$  following [13].

## 4. Experiments

### 4.1. Implementation Details

We train our self-attention autoencoder for reconstruction using MS-COCO dataset [15], which contains roughly 80,000 training examples. We preserve the aspect ratio of image and rescale the smaller dimension to 512 pixels, and then randomly crop to  $256 \times 256$  pixels. Our encoder contains the first few layers from VGG-19 model [22] pre-trained on the ImageNet dataset [3]. The decoder is symmetric to the encoder structure. The Relu\_X\_1 ( $X = 1, 2, 3, 4$ ) in the encoder are used to compute the perceptual loss and  $\lambda_{con}, \lambda_p, \lambda_{att}, \lambda_{tv}$  are set as 1, 10, 6, 10 to balance each loss. We use three  $1 \times 1$  convolutions for  $\Theta_h, \Theta_u, \Theta_g$  in our self-attention module, and set  $\hat{C} = C$  for  $\Theta_h$  and  $C/2$  for  $\Theta_u, \Theta_g$ . We train our network using the Adam optimizer and a batch size of 8 for 10k iterations. Although our method can handle arbitrary number of stroke fusions, we select three stroke scenario as a default setting in the following experiments, including the fine-grained stroke size and two coarse stroke sizes ( $K = 2$ ) whose scale coefficients  $\beta_k$  are 0.5 and 1 respectively. We set the smoothing factor  $\gamma$  as 50 to reveal discriminative patterns for different stroke. To balance the stylization effects and the attention consistency, we set the kernel size, mean and deviation in our *Attention filter* as 3, 6, 5 respectively.

To obtain abundant stylized information, it is advantageous to match features across different levels in the VGG-19 encoder to fully capture the characteristics of the style as suggested in [13, 21]. We hence adopt a similar strategy for our reconstruction process. We use skip connections to perform style enhancement using adaptive instance normalization [7], feeding with style features extracted from Relu\_X\_1 ( $X = 1, 2, 3$ ) outputs and the reconstruction features in corresponding deconvolution layers.

### 4.2. Qualitative Evaluation

**Comparison With Prior Methods.** We evaluate four state-of-the-art methods for arbitrary style transfer: Style-Swap [2], AdaIN [7], WCT [13] and Avatar-Net [21]. The stylized results for various content/style pairs are shown in Figure 4. Style-Swap simply relies on the patch similarity between content and style feature, so as to strictly preserve the content feature, which is validated in its results that only low-level style patterns (*e.g.*, colors) are transferred. AdaIN presents an efficient solution by directly adjusts the content feature to match the mean and variance of the style feature. But it usually brings similar texture patterns for all the stylized images (*e.g.*, the crack pattern in all stylized results) due to the style-dependent training. WCT holistically adjusts the feature covariance of content feature to match that of style feature, which inevitably brings unseen information and unconstrained patterns, for example the missing circu-

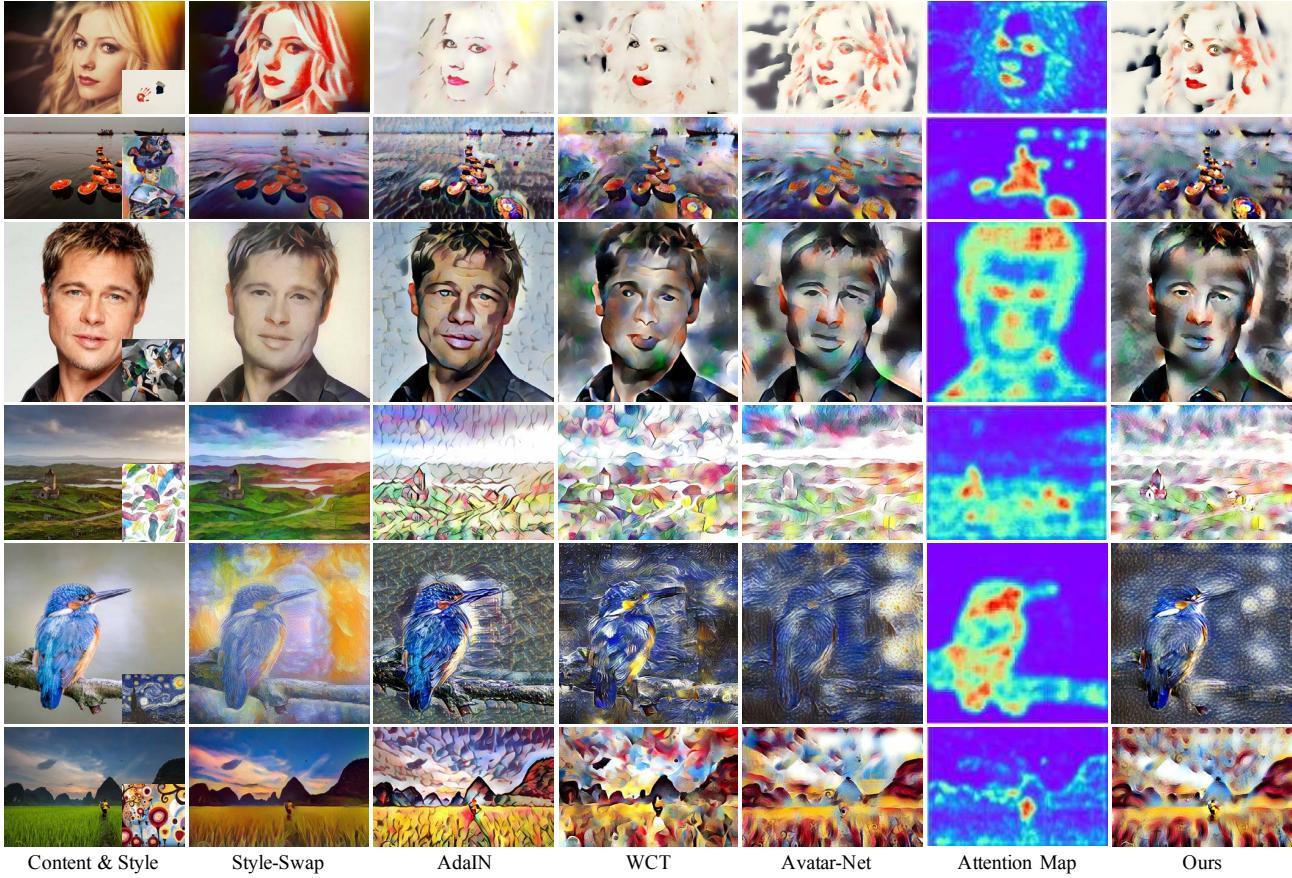


Figure 4. Results from different ASPM style transfer methods. We evaluate various content images including portraits, animals, landscapes with distinctive styles. The results of compared methods are obtained by running their codes with default configurations.

lar patterns in the 6<sup>th</sup> row and the unexpected textures in the 5<sup>th</sup> row. Avatar-Net shrinks the domain gap between content and style features and enhances propagation for feature alignment. Although concrete style patterns are reflected, it still cannot handle attention-aware feature adaption and manifest distortion from semantically perception, for example the eyes on the first row and the farmer on the last row.

By contrast, our method can produce visually plausible stylized results against the previous methods. The attention map enables the seamlessly synthesis among multiple stroke sizes, while demonstrating superior spatial consistency of visual attention between content image and stylized image. In the last column of Figure 4, the salient regions in content images such as eyes, candles, house and farmer still maintain the focus in the stylized image. This validates the effectiveness of the attention map in the 6<sup>th</sup> column, where salient regions are mainly attributes to the fine-grained stroke size. In addition, the attention map exhibits similarity measurement between distant regions, which allows detailed features in distant portions of the image are consistent with each other.



Figure 5. Comparison results to present the effectiveness of attention guidance. Regions marked by bounding boxes are zoomed in for a better visualization.

Method	AUC_Judd $\uparrow$	SIM $\uparrow$	NSS $\uparrow$	CC $\uparrow$	KL $\downarrow$
AAMS(-SA)	0.479	0.677	3.627	0.732	0.458
AAMS	0.484	0.744	4.172	0.834	0.396

Table 1. Quantitative measurements on saliency maps predicted by SalGAN [18]. The attention consistency is dramatically improved with our method for all the evaluation metrics [1].

**Ablation Studies.** We discuss in Section 3 about the contributions of the proposed modules. The self-attention module is responsible for capturing attention characteristics of content image, such that the multi-scale style swap module and the multi-stroke module can work together to integrate multiple stroke patterns. Several evaluations are performed to verify the effectiveness of the coupling framework.



Figure 6. Stylized result via different stroke sizes.

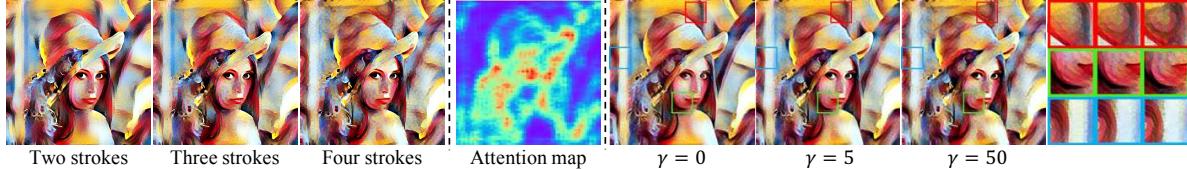


Figure 7. Left: fusion strategy with varying number of stroke sizes, Right: level of detail control via regulating the smoothing factor  $\gamma$ .

We first train an autoencoder by removing the self-attention module. Since no guidance from the attention map, we apply an average fusion strategy in the multi-stroke fusion module during the testing phase. We name this method as AAMS(-SA). Figure 5 shows the comparison with our method AAMS. Our stylized result with self-attention module demonstrates dramatical improvement on visual effect compared with AAMS(-SA), mainly in two aspects: 1) Our attention-aware method emphasizes the salient regions by painting with fine-grained stroke. 2) The multi-stroke fusion allows abundant integration among stroke patterns and presents discriminative style patterns locally without sacrificing the holistic perception.

Apart from the cooperation between self-attention and multi-stroke fusion module, the multi-scale style swap also form a compact affiliation with the fusion module. Without the multi-scale style swap module, the style transfer will be restricted to single stroke pattern. We present the final stylized results in Figure 6 with different stroke patterns. The stroke size is controlled by changing the scaling coefficient  $\beta$  in eq. (8). From left to right, the larger the stroke size, the coarser stylized pattern emerges. When multiple stroke patterns are generated together, the multi-stroke fusion module can integrate them into one stylized image.

### 4.3. Quantitative Evaluation

**User Studies.** As artistic style transfer is a highly subjective task, we resort to user studies to better evaluate the performance of the proposed method. Since Style-Swap [2] only transfers low-level information, resulting in insufficient stylization effects. We compare the proposed method to the other three competing methods, i.e., AdaIN [7], WCT [13] and Avatar-Net [21]. We use 10 content images and 15 style images collected from the aforementioned methods and synthesize 150 images for each method, from which we show 20 randomly chosen content and style combinations to each subject.

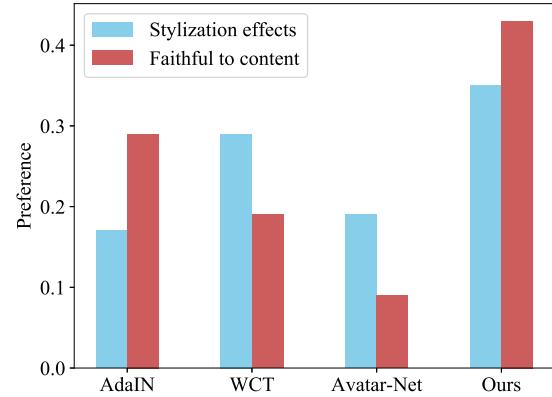


Figure 8. User study results.

We conduct two user studies on the results in terms of the stylization effects and faithfulness to content characteristics. In the first study, we ask the participants to select which stylized image better transfers both color and texture patterns of style image to content image. In the second study, we ask the participants to select which stylized image better preserves the content characters with less distorted artifacts. In each question, we display the content/style pair and randomly ordered stylized images by 4 compared methods and ask the participants to vote the best one according to the questions. We finally collect totally 600 votes from 30 subjects and demonstrate the preference results in Figure 8. The studies show that our method receives most votes for both stylization and attention consistency. The better preservation for visual attention enables perceptual promotion for stylization.

**Consistency Evaluation.** To further evaluate the effectiveness on preserving the attention consistency of our method, we propose to adopt saliency metrics [1] to measure the saliency similarity between content and stylized image pairs. These metrics are widely used to evaluate a saliency model's ability to predict ground truth human fixa-

Method	Execution Time (sec)	
	256×256	512×512
AdaIN	0.09	0.17
WCT	0.92	1.05
Style-Swap	1.96	5.77
Avatar-Net	0.78	0.86
Ours (2-stroke)	0.74	0.81
Ours (3-stroke)	0.80	0.94
Ours (4-stroke)	0.92	1.15
Ours (8-stroke)	0.94	1.18

Table 2. Run time performance. We show runtime test of different stroke numbers with our method and other feedforward methods.

tions. We generate the saliency maps of stylized images and their corresponding content images using SALGAN [18], a state-of-the-art saliency prediction method. We then employ five metrics to evaluate the degree of visual attention consistency for 400 content/stylized image pairs. Comparison results between the proposed method and AAMS(-SA) are shown in Table 1. With our AAMS framework, the attention consistency is dramatically improved for all the evaluation metrics. The results further verify the ability in maintaining saliency for style transfer.

**Speed Analysis.** We show the run time performance of our method and prior feed-forward methods in Table 2. Results are obtained with a 12G Tesla M40 GPU and averaged over 400 transfers. Among the patch based methods (Style-Swap, Avatar-Net), our method achieves comparable speed even with multi-scale feature processing. Given three strokes, it takes averagely 0.80 and 0.94 seconds to transfer images with size  $256 \times 256$  and  $512 \times 512$  respectively. Since the running time increases with stroke numbers in a controllable trend, the tradeoff between efficiency and details diversity should be considered.

#### 4.4. Runtime Control

Given the learnt attention map, our method is flexible to provide different controls on the multi-stroke stylization.

**Multi-stroke Fusion Control.** One of the advantage towards the prior methods is that our method can effectively integrate multiple stroke patterns with different control strategies, including the number of stroke sizes and level of detail control. As discussed in the previous section, the different scale of high-level style feature will lead to different statistics after a style swap process with the same content feature. Given arbitrary number of stroke sizes, our method allows flexible fusion among these strokes. The left part of Figure 7 demonstrates three fusion results with different number of stroke sizes. Note that by integrating with more stroke sizes, the stylized result presents more varying patterns with different stroke boundaries, the reason for this is that our clustering algorithm produces different attention distributions for different number of clusters, which results in varying weight maps. The right part of Figure

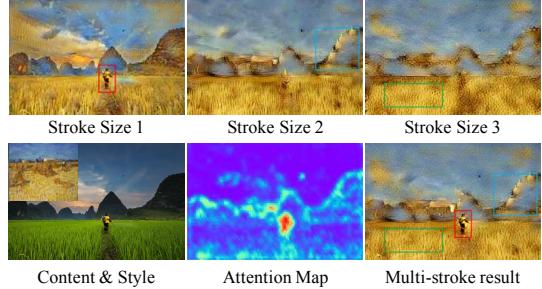


Figure 9. Our method enables automatic spatial stroke size control.

7 shows our ability in regulating level of detail control for four-stroke fusion. According to eq. (12), the smoothing factor  $\gamma$  determines the weight distribution for all stroke sizes. As shown in Figure 7, the larger it is, the more discriminative for each stroke within corresponding regions.

**Automatic Spatial Stroke Size Control.** Spatial stroke size control refers to properly utilizing strokes of different sizes to render regions with different levels of detail in an image. The gist is that for a visually plausible stylized artwork, we usually hope to use stroke of small size for salient regions (e.g., objects, humans), while large size for less salient ones (e.g., sky, grassland). To this end, previous method [8] generally employs several hand-crafted masks and then stylize corresponding regions separately. Our method, however, empowers spatial control in an automatic way due to the self-attention mechanism. As shown in Figure 9, three stroke sizes are adopted to render regions of different saliency, i.e., the fine stroke patterns on the farmer, the middle stroke patterns on the mountain and coarse stroke patterns on the grass. In particular, these regions are partitioned and integrated automatically under the guidance of content attention map during our multi-stroke fusion procedure, thus providing automatic solution for spatial control (which refers to transferring each region with different styles in the content image) instead of mask controlling methods [6, 13].

#### 5. Conclusion

In this paper, we propose an attention-aware multi-stroke style transfer model for arbitrary styles. We employ self-attention as a residual to obtain the attention map, and then introduce multi-scale style swap and a flexible stroke fusion strategy to adaptively integrate multiple style patterns into one stylized image. Experimental results demonstrate the effectiveness of our method in generating favorable results in terms of stylization effects and visual consistency with content image. As a future direction, one may further explore the relation between the stroke assignment and aesthetic factors for enhanced stylization.

**Acknowledgements.** This work was supported by the Natural Science Foundation of China (61725204).

## References

- [1] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 2018.
- [2] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [4] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, 2015.
- [5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.
- [6] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *CVPR*, 2017.
- [7] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.
- [8] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. Stroke controllable fast style transfer with adaptive receptive fields. In *ECCV*, 2018.
- [9] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *arXiv preprint arXiv:1705.04058*, 2017.
- [10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- [11] Hugo Larochelle and Geoffrey E Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In *NIPS*, 2010.
- [12] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016.
- [13] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NIPS*, 2017.
- [14] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [16] Zhouhan Lin, Minwei Feng, Cicero Nogueira Dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *ICLR*, 2017.
- [17] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014.
- [18] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O’Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.
- [19] Eric Risser, Pierre Wilmot, and Connelly Barnes. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv preprint arXiv:1701.08893*, 2017.
- [20] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.
- [21] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *CVPR*, 2018.
- [22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [23] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, 2016.
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- [25] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *CVPR*, 2017.
- [26] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015.
- [27] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.
- [28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- [29] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.
- [30] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.
- [31] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
- [32] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.
- [33] Song-Chun Zhu, Cheng-En Guo, Yizhou Wang, and Zijian Xu. What are textons? *International Journal of Computer Vision*, 2005.