This WACV 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Exploiting Spatial Relation for Reducing Distortion in Style Transfer

Jia-Ren Chang[1,2] Yong-Sheng Chen[1]
[1]National Chiao Tung University, Taiwan [2]aetherAI
{followwar.cs00g, yschen}@nctu.edu.tw

Abstract

The power of convolutional neural networks in arbitrary style transfer has been amply demonstrated; however, existing stylization methods tend to generate spatially inconsistent results with noticeable artifacts. One solution to this problem involves the application of a segmentation mask or affinity-based image matting to preserve spatial information related to image content. The main idea of this work is to model spatial relation between content image pixels and thus to maintain this relationship in stylization for reducing artifacts. The proposed network architecture is called spatial relation-augmented VGG (SRVGG), in which long-range spatial dependency is modeled by a spatial relation module. Based on this spatial information extracted from SRVGG, we design a novel relation loss which can minimize the difference of spatial dependency between content images and stylizations. We evaluate the proposed framework on both optimization-based and feedforward-based style transfer methods. The effectiveness of SRVGG in stylization is demonstrated by generating stylized images of high quality and spatial consistency without the need for segmentation masks or affinity-based image matting. The quantitative evaluation also suggests that the proposed framework achieve better performance compared with other methods.

1. Introduction

Transferring style from one image to another is an image editing task that draws a lot of attention recently. The process takes as input the content and style of two images in synthesizing a new image that preserves semantic content but carries over style characteristics. The key challenge is extracting an effective representation of the style and then mapping it with the content image. Gatys *et al*. [4] demonstrate that convolutional neural networks (CNNs) can be used to encode the semantic content of an image as well as its texture-related information. In that framework, VGG-19 [19] is used as a feature extractor, and a simple statistic (a Gram matrix) is used to capture texture. However, their style transfer method tends to generate spatially inconsistent stylizations with noticeable artifacts.

A number of methods have been proposed to remedy spatial inconsistencies in neural style transfer. To reduce distortion, Luan *et al*. [16] propose a photorealism regularization term, which constrains stylizations by using locally affine color transformation of the input. Gatys *et al*. [5]seek to avoid content mismatch by introducing a scheme to guide the transfer of styles based on semantic segmentation masking of the inputs. Li *et al*. [12] propose affinity-based smoothing to ensure spatially consistent style transfer.

The problem of spatial distortion also arises in image generation tasks. Even high-resolution images generated using state-of-the-art generative adversarial networks (GANs) lack perceptual appeal [6, 17]. Furthermore, GANs are unable to provide guidance in preventing spatial distortion. Several methods have been proposed to address this issue. Karras *et al*. [10] propose a novel training methodology involving the iterative addition of layers to enhance the level of detail as training progresses. Although this so-called progressive GAN scheme is able to generate very high-quality images, considerable training time is required. The recently developed self-attention GAN [23] introduces attention-driven, long-range dependency modeling for image generation tasks. It reduces spatial distortion and improves the quality of generated images.

The basic assumption of style transfer using CNN is that features from CNN encode both content and texture information. This assumption also suggests that semantically related pixels share similar texture information. To follow this suggestion, one simple way to provide the semantic relationship among pixels is to exploit a semantic segmentation mask [12, 16]. In the proposed framework, this semantic relationship is modeled and learned by using a neural network. This is achieved by developing a *spatial relation-augmented VGG* (SRVGG) network, which incorporates a spatial relation module to model the long-range dependency of encoded VGG features. To VGG network, we add spatial relation modules and retrain the network on ImageNet. We show that the retrained SRVGG reduces the top-1 validation error with only a slight increase in computational over-

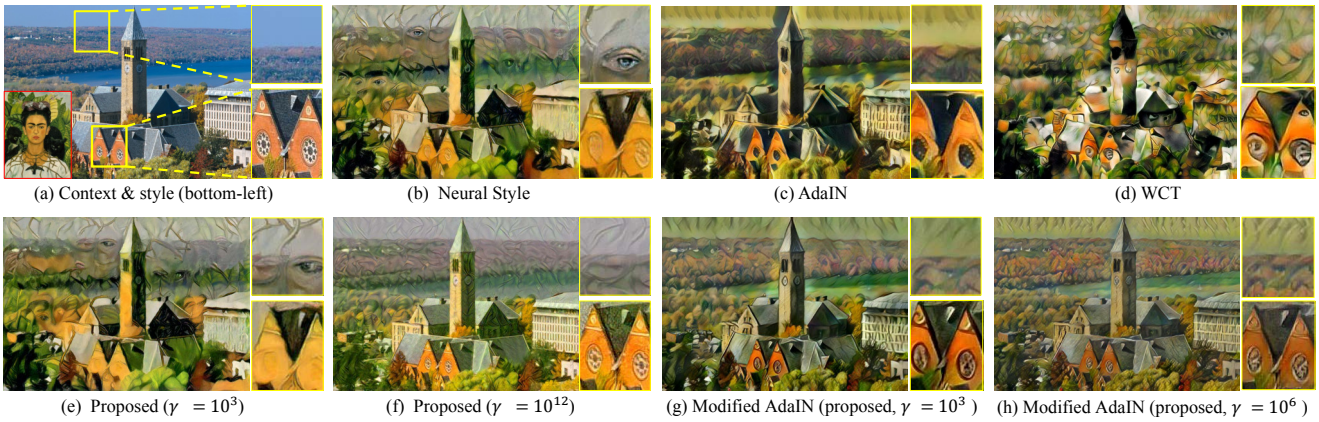


Figure 1. (a) Content and style images; (b) output image obtained using [4], showing artifacts (eyes) and loss of spatial information (rose windows) in the right panel; (c) and (d) output images obtained using state-of-the-art feedforward style transfer methods: AdaIN [8] and WCT [11]; (e) and (f) output images obtained using relation loss weights set to $\gamma = 10^3$ and $\gamma = 10^{12}$. (g) and (h) output images obtained using modified feedforward models that trained with relation loss weights $\gamma = 10^3$ and $\gamma = 10^6$. Increases in relation loss lead to decreases in the degree of spatial distortion of rose windows.

head. Based on SRVGG, we propose a novel relation loss to limit spatial distortion. We apply the proposed framework for both optimization-based and feedforward-based style transfer. The effectiveness of the proposed SRVGG is demonstrated by generating high-quality stylized images that provide a high degree of spatial consistency without the need for segmentation masks or affinity-based image matting. Fig. 1 shows examples of artistic image stylization as well as the effect of relation loss.

The main contributions of this study are the following:

- We develop SRVGG, a novel feature extraction network which incorporates a spatial relation module.
- We develop a relation loss that controls the scope of spatial distortion.
- We demonstrate that the proposed framework can be applied to both optimization-based and feedforward-based style transfer.
- We demonstrate that artifacts in stylization can be alleviated without the need for segmentation masks or affinity-based image matting.

2. Related Work

Recently, Gatys *et al*. [4] employ the feature maps of trained deep CNNs (VGG-19 in their work) to achieve groundbreaking performance in artistic style transfer. They subsequently introduce style loss to match correlations (Gram matrices) between feature maps extracted using VGG-19. This method is based on an optimization framework that iteratively updates the input image in order to minimize content loss and style loss. Subsequent studies have focused on improving speed [8, 11], user control [5], and photorealism [16].

The main drawback of [4] is the inefficiency of the optimization process. The speed issue is dealt with in [20, 9], wherein stylization involves training a feedforward neural network under the same objective pursued in [4]. With other techniques, CycleGAN [25] treats style transfer as a domain transfer problem, and trains a generator for feedforward-based style transfer. However, these feedforward methods are limited by the fact that each network is tied to a fixed style. Several studies have addressed this problem using a single network for the transfer of multiple styles. Dumoulin *et al*. [3] introduce a novel conditional normalization layer to deal with the scale and shift parameters of normalization for each style. Li *et al*. [13] propose a feedforward architecture based on binary selection units. Nonetheless, neither of those methods can be used to transfer arbitrary styles that have not undergone extensive training.

To achieve arbitrary style transfer using the feedforward method, Chen and Schmidt [2] introduce a style swap layer that matches content features with the closest style features in a patch-by-patch manner. Deep Image Analogy [14] was developed to extend patch-matching and reconstruction from a single layer of features to a hierarchy. Other methods align the mean and variance of content features with those of style features. Huang and Belongie [8] introduce an adaptive instance normalization (AdaIN) layer, which simply adjusts content features to match the mean and variance of the style features. Li *et al*. [11] introduce whitening and coloring transforms (WCT) to match the covariance of content features to those of style features. Shen *et al*. [18] propose a meta-style transfer scheme which generates network weights from the mean and variance of given style features.

The loss function is also critical to style transfer. Li *et al*. [13] demonstrate that several other loss functions can be used as an alternative to the Gram matrix, such as maximum mean discrepancy loss and the distance between channel-wise mean and variance.

It is also important to preserve the spatial structure of content in stylizations. Gatys *et al*. [5] propose maintaining perceptual control using a spatial mask. Luan *et al*. [16] introduce a segmentation mask and a matting Laplacian matrix for the suppression of distortion. However, their scheme adds considerably to the computational overhead. In this work, we tackle the problem of spatial inconsistency by modifying the fundamental architecture of the network and by introducing a relation loss.

3. Optimization-based Style Transfer

In this section, we summarize the Neural Style Transfer [4], the objective of which is to generate a stylized output image (I_O) given a content image (I_C) and a reference style image (I_S) by minimizing the objective function

$$\mathcal{L} = \alpha\mathcal{L}_C + \beta\mathcal{L}_S\,, \tag{1}$$

where α and β are the weights for content and style loss functions for balancing the content and style in the stylization. Content loss $\mathcal{L}_C$ is defined as the squared error of the feature maps extracted from I_O and I_C:

$$\mathcal{L}_C = \sum_{l\in l_C} \|\mathcal{F}^l(I_O) - \mathcal{F}^l(I_C)\|^2\,, \tag{2}$$

where $\mathcal{F}^l \in \mathbb{R}^{c\times h\times w}$ comprises feature maps in layer l, c is the number of feature channels, and h and w represent the height and width of the feature maps. Style loss is defined by the squared error between feature correlations expressed using Gram matrices of I_O and I_S:

$$\mathcal{L}_S = \sum_{l\in l_S} \omega_l \|\mathcal{G}(\mathcal{F}^l(I_O)^{'}) - \mathcal{G}(\mathcal{F}^l(I_S)^{'})\|^2\,, \tag{3}$$

where ω_l is the weight of the loss in layer l. The Gram matrix $\mathcal{G}(\mathcal{F}^l(\cdot)^{'}) \in \mathbb{R}^{c\times c}$ can be calculated over the vectorized feature maps $\mathcal{F}^l(\cdot)^{'} \in \mathbb{R}^{c\times hw}$ as

$$\mathcal{G}(\mathcal{F}^l(\cdot)^{'}) = [\mathcal{F}^l(\cdot)^{'}][\mathcal{F}^l(\cdot)^{'}]^{\mathrm{T}}\,. \tag{4}$$

4. Feedforward-based Style Transfer

Neural Style Transfer [4] is slow because it involves an optimization procedure. Many feedforward-based style transfer methods have been proposed to improve the efficiency [8, 11, 18]. In this section, we summarize the AdaIN [8] method for arbitrary style transfer. The objective of AdaIN is to align the channel mean and variance of content features to match those of style features in layer l:

$$\begin{aligned}&\mathrm{AdaIN}(\mathcal{F}^l(I_C), \mathcal{F}^l(I_S)) = \\ &\sigma(\mathcal{F}^l(I_S))(\frac{\mathcal{F}^l(I_C) - \mu(\mathcal{F}^l(I_C))}{\sigma(\mathcal{F}^l(I_C))}) + \mu(\mathcal{F}^l(I_S))\,.\end{aligned} \tag{5}$$

As with instance normalization [21], the statistics of AdaIN are computed across spatial locations. The affined features are further inverted to the stylized image (I_O) with a feedforward decoder trained using the pre-trained VGG as a loss network. Similar to [4], the loss function of the decoder defined in Eq. 1 is a weighted combination of the content loss and the style loss. The content loss $\mathcal{L}_C$ is defined as the squared error of the feature maps extracted from I_O and the AdaIN output:

$$\mathcal{L}_C = \|\mathcal{F}^l(I_O) - \mathrm{AdaIN}(\mathcal{F}^l(I_C), \mathcal{F}^l(I_S))\|^2\,. \tag{6}$$

The style loss $\mathcal{L}_S$ of AdaIN matches the mean and standard deviation of the style features, computed across spatial locations:

$$\begin{aligned}\mathcal{L}_S = &\sum_{l\in l_S} \|\sigma(\mathcal{F}^l(I_O)^{'}) - \sigma(\mathcal{F}^l(I_S)^{'})\|^2 \\ &+ \sum_{l\in l_S} \|\mu(\mathcal{F}^l(I_O)^{'}) - \mu(\mathcal{F}^l(I_S)^{'})\|^2\,.\end{aligned} \tag{7}$$

5. Proposed Method

In this work, we propose a spatial relation module, SRVGG, and present how to train SRVGG to be a feature extraction network. The relation loss is further introduced to control the extent of spatial distortion in the stylizations by utilizing the spatial relation obtained from SRVGG.

5.1. Spatial Relation Module

We introduce a spatial relation module as a means to calculate the relationship between a position in an image and all other positions within the same image. The spatial relation is defined as follows:

$$R_{ij} = SR(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i)^{\mathrm{T}} g(\mathbf{x}_j)\,, \tag{8}$$

where $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^c$ are image feature vectors from the previous layer at spatial positions i and j. Here $f(\mathbf{x}_i)$ and $g(\mathbf{x}_j)$ are implemented by a linear layer that can be learned during training. The relation matrix $\mathbf{R} \in \mathbb{R}^{hw\times hw}$ represents the spatial relationship among locations.

To augment features for style transfer, we signify the *important relation* $\mathcal{E}_{ij}$ from other positions by:

$$\mathcal{E}_{ij} = \frac{exp(R_{ij})}{\sum_{k=1}^{hw} exp(R_{ik})}\,. \tag{9}$$

Thus, the output of the important relation layer is:

$$o = h(x)\mathcal{E}^{\mathrm{T}}\,, \tag{10}$$

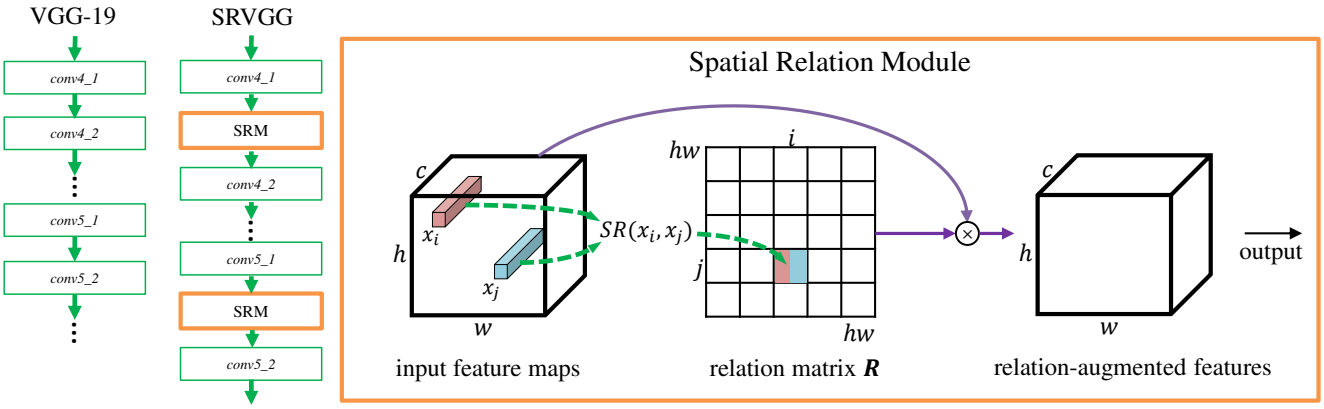


Figure 2. The proposed spatial relation module is added to follow the convolutional layers `conv4_1` and `conv5_1` of VGG-19 [19].

where $o \in \mathbb{R}^{c \times hw}$, h is a linear layer that can be learned during training, and the output of h is also vectorized. The output is reshaped back to $\mathbb{R}^{c \times h \times w}$ followed by a summation with x to obtain the relation-augmented feature maps

$$y = o + x\,. \tag{11}$$

Fig. 2 presents the spatial relation module. It is worth noting that the proposed spatial relation module can also be regarded as a non-local operation in [22].

5.2. Spatial Relation-augmented VGG

Pioneers in this field developed methods by which to use the features extracted using VGG-19 [19] for style transfer tasks. In this study, representative features are extracted using the proposed SRVGG, which captures long-range spatial dependency through spatial relation module and thereby remedies the problem of spatial inconsistency. SRVGG is modified from a pre-trained version of VGG-19 provided by *torchvision*. A spatial relation module is added after convolutional layers `conv4_1` and `conv5_1`. Thus the added spatial relation modules are referred to as `sr_conv4_1` and `sr_conv5_1`.

5.3. SRVGG Training

We finetune SRVGG using the ImageNet dataset through 70 epochs. The learning rate starts at 0.001 and decays by a factor of 10 for every 30 epochs. The retrained SRVGG is shown to reduce the top-1 validation error from 27.62% to 25.80%, thereby demonstrating the efficacy of the proposed spatial relation-augmented mechanism for image recognition. Notice that our objective in this work was to use this relationship to enhance the style transfer function in preserving spatial information.

5.4. Relation Loss

In SRVGG, spatial relation is used to model dependency between locations. To make better use of this property, we propose a novel relation loss function defined as

$$\mathcal{L}_R = \|R_O - R_C\|^2\,, \tag{12}$$

where **R** is relation matrix in Eq. (8). To preserve spatial consistency, the objective of this loss function is to keep the spatial relationships in the output image close to those in the content image. The total loss function is written as

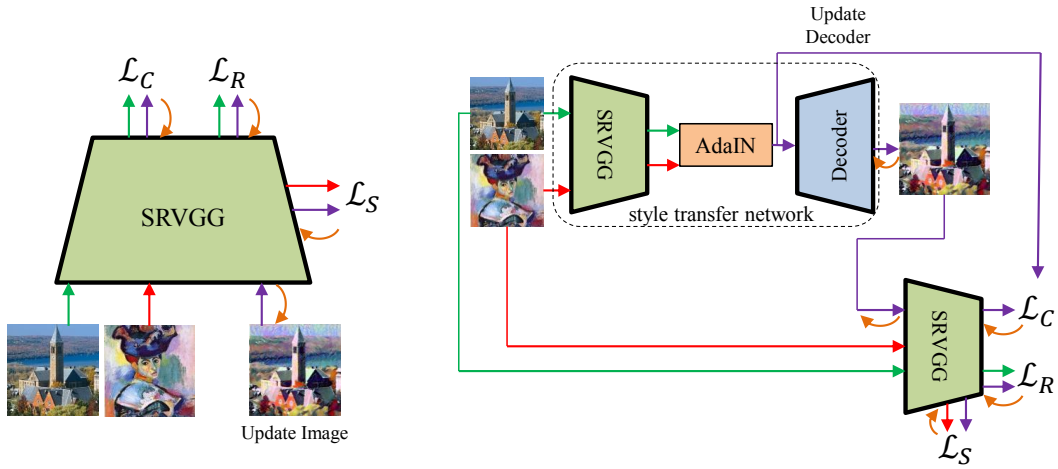
$$\mathcal{L} = \alpha\mathcal{L}_C + \beta\mathcal{L}_S + \gamma\mathcal{L}_R\,, \tag{13}$$

where α, β,and γ are the weights used for content loss, style loss, and relation loss, respectively. The choice of content loss and style loss is according to the task. Fig. 1 illustrates the influence of relation loss, without which stylization produces artifacts, as reported in [4, 8]. By applying relation loss, stylization produces results that are visually appealing while mitigating the artifacts.

5.5. Implementation Details

Optimization-based style transfer We employed the trained SRVGG as a feature extraction network for the style transfer task. Layer `sr_conv4_1` was used for content representation; `conv1_1`, `conv2_1`, `conv3_1`, `conv4_1`, and `conv5_1` were used for style representation; and the relation matrix in layer `sr_conv4_1` was used for relation loss computation. The content loss weight was set at $\alpha = 1$ and the style loss weight was set at $\beta = 10^{12}$ for all results. The weight of relation loss γ varied in each example.

We adopted the framework proposed by [4] for the optimization of output images. LFBGS [24] was used as the optimizer and the content image was used as the initial output image. An overview of optimization-based style transfer methods is shown in Fig. 3 (a). The number of iterations in the optimization process was set to 500 epochs for artistic style transfer, and 1000 epochs for photorealistic style transfer. To produce the results of Neural Style Transfer method [4], we used the same hyper-parameters as mentioned above.



(a) Optimization-based style transfer

(b) Feedforward-based style transfer

Figure 3. Overviews of modified (a) optimization-based and (b) feedforward-based style transfer methods. We employ SRVGG layers for feature extraction in both algorithms. Relation loss $\mathcal{L}_R$ is added to the objective functions of both (a) and (b). In (a), the method is modified from [4]. In (b), we follow the AdaIN algorithm [8] in which a decoder is trained to invert the AdaIN output to an image.

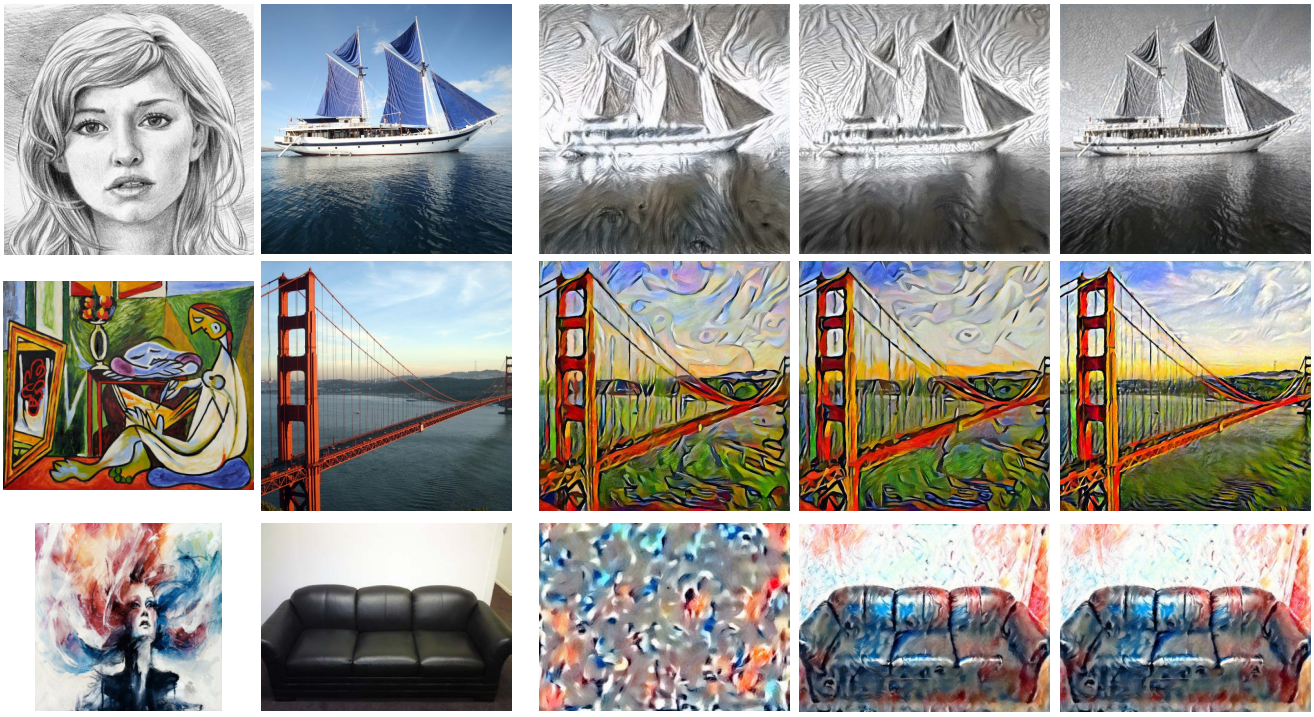


Figure 4. Comparison of results obtained using the proposed method and Neural Style Transfer [4] method. Our method achieves greater spatial consistency, resulting in images that are visually more appealing.

Feedforward-based Style Transfer We modified AdaIN [8] to incorporate SRVGG and relation loss. First, we replaced VGG-19 with SRVGG as an encoder network and a loss network. Layer `sr_conv4_1` was used for content representation; `conv1_1`, `conv2_1`, `conv3_1`, and `conv4_1` were used for style representation; and the relation matrix in layer `sr_conv4_1` was used for loss computation. The content loss weight was set at $\alpha = 1$, the

style loss weight was set at $\beta = 10$, and the relation loss weight was set at $\gamma = 10^3$ for training the decoder. We set the content-style tradeoff parameter in AdaIN to 1 as in [8], thus pushing the network to synthesize the most stylized images. An overview of optimization-based style transfer methods is shown in Fig. 3 (b).

We modified the feature extraction in encoder and the loss function in the network architecture, while retaining the decoder architecture as well as its training procedure as that in [8]. MS-COCO [15] dataset was used as content images and WikiArt [1] dataset was used as style images. Adam optimizer was employed and a batch size of 8 was chosen. Data augmentation and the learning rate were also the same as in [8], except for the number of training iterations, which was set to 100,000.

6. Results

6.1. Optimization-based Artistic Style Transfer

A number of experiments were conducted to verify the efficacy of the proposed method. Fig. 4 presents the comparison results of stylization obtained using the proposed method and the Neural Style Transfer method introduced by [4]. The proposed method controlled spatial consistency via relation loss with its weight set as $\gamma = 10^3$ and 10^6. Increasing the weight of relation loss enabled the preservation of structure-related information. Note that Neural Style Transfer produced artifacts in stylization. As shown in the first row in Fig. 4, the distortions were from the shape of the hair rather than the style of the image. In contrast, results obtained using the proposed method are more visually appealing with fewer artifacts.

6.2. Optimization-based Photorealistic Style Transfer

The proposed method was also shown to significantly improve the results obtained using photorealistic style transfer without the need for segmentation masks. Fig. 5 presents a comparison of the results obtained using our method and those obtained using Neural Style Transfer [4] and Deep Photo Transfer [16]. As shown in the third column in Fig. 5, Neural Style Transfer produced strong distortion. As shown in the fourth column in Fig. 5, Deep Photo Transfer improved on those results through the use of photorealism regularization and segmentation masks (also shown). As shown in the fifth column in Fig. 5, the proposed method achieved undistorted results without the need for segmentation masks. The relation loss weights used in rendering the images in Fig. 5 were as follows: $\gamma = 10^7, 10^{12}, 10^{14}$ (from top to bottom). Noted that as with Deep Photo Transfer, our results are post-processed using guided filtering [7].

6.3. Feedforward-based Artistic Style Transfer

We have shown that the proposed method can also be used to train feedforward-based style transfer models and to remedy the problem of spatial distortion. We modified the AdaIN [8] method for fast style transfer. Fig. 6 presents a comparison of the results obtained using modified AdaIN and those obtained using the original AdaIN [8] and WCT [11]. For generating stylizations with AdaIN and WCT, we set the content-style tradeoff to 1 for AdaIN and WCT. These results suggest that our method preserves the shape of details. For example, our result in first row in Fig. 6 demonstrates that the shape of the windows can be preserved; however, results of other methods are highly distorted. Our method preserves more spatial information of the content image, generating satisfying stylizations.

Table 1. Quantitative ablation analysis of different settings in terms of the style loss ($\mathcal{L}_S$), content loss ($\mathcal{L}_C$), and relation loss ($\mathcal{L}_R$) in the proposed method. When SRVGG is used for loss network, we adopted relation loss with its weight (γ) shown in this table.

Encoder	Loss net		$\mathcal{L}_S$	$\mathcal{L}_C$	$\mathcal{L}_R$
VGG	VGG		6.97	0.55	0.93
SRVGG	VGG		6.95	0.52	0.85
VGG	SRVGG	$\gamma = 0$	6.85	0.43	0.85
VGG	SRVGG	$\gamma = 10^3$	6.88	0.43	0.84
SRVGG	SRVGG	$\gamma = 0$	**6.81**	0.43	0.88
SRVGG	SRVGG	$\gamma = 10^3$	6.87	0.41	0.85
SRVGG	SRVGG	$\gamma = 10^6$	7.02	**0.32**	**0.64**

6.4. Ablation Study

A quantitative ablation study was conducted to demonstrate that the relation loss can improve style transfer. We varied the combination of encoder and loss network to train the decoder as in AdaIN [8]. We generated 300 stylizations from 20 content images and 15 style images, and computed its losses by SRVGG. As shown in Table 1, the decoder trained with relation loss can achieve lower loss value compared to that of AdaIN. The results show that the increased γ can lead to higher style loss, but lower content and relation loss, suggesting a trade-off between style and content. We adopted $\gamma = 10^3$ as mentioned in Sec. 5.5. The combination of SRVGG encoder and VGG loss network indicates that SRVGG can provide better features for style transfer than VGG. We suggest that the relation loss can help to maintain spatial consistency and the relation-augmented feature maps from SRVGG can capture high-level representation of the style.

6.5. Visualization of Important Relation

We visualized the important relation of content images produced by the spatial relation module. The important

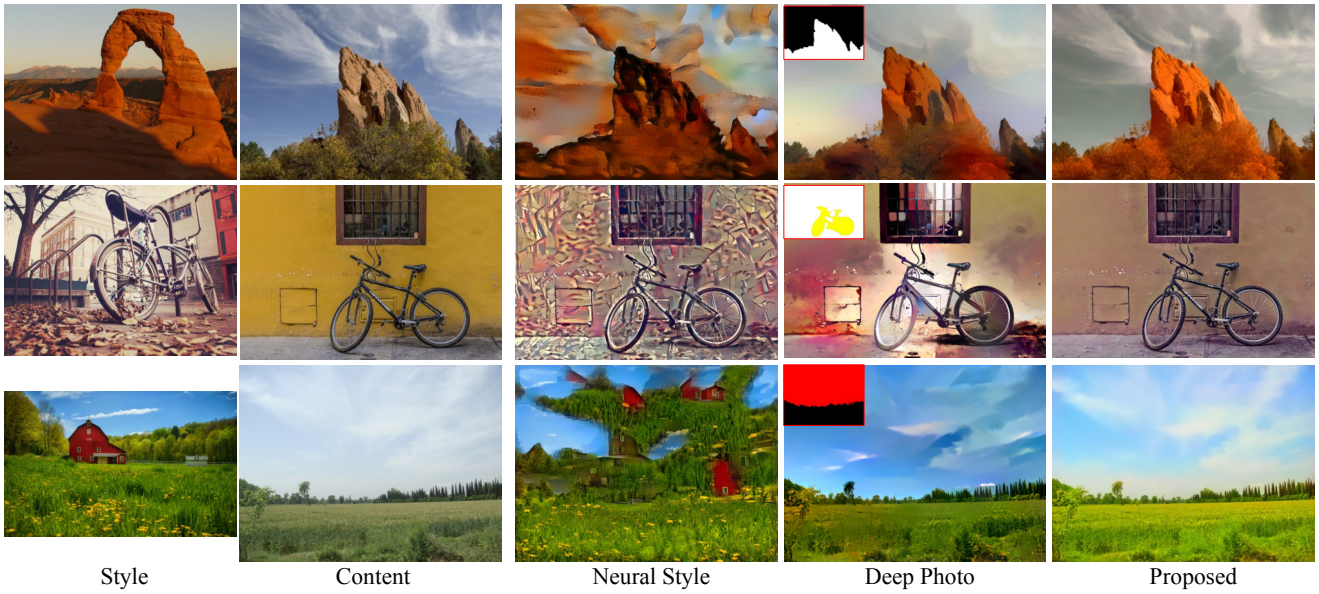


Figure 5. Comparison of Neural Style Transfer [4], Deep Photo Transfer [16], and the proposed method. Neural Style Transfer causes the formation of strong artifacts in the output images. Deep Photo Transfer produces satisfying results at the expense of complex optimization and segmentation masks. The proposed method achieves results that are visually appealing and less-distorted without the need for segmentation masks.

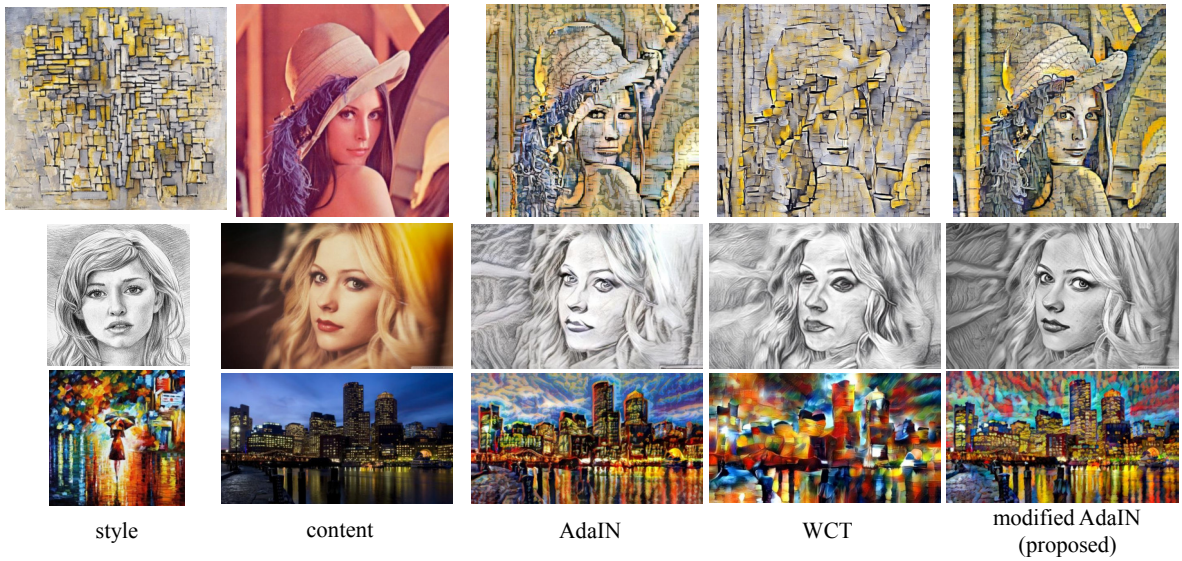


Figure 6. Comparison of feedforward-based methods: original AdaIN [8], WCT [11], and the modified AdaIN proposed in this work. Compared to the original AdaIN and WCT, our methods can obtain results with much less spatial distortion.

relation were extracted from the `sr_conv4_1` layer. The green, red, and blue visualization points in Fig. 7 are highly correlated with the corresponding semantic regions. This supports our assumption that SRVGG can encode semantic relationships. For example, the green points in the sky show a strong correlation with cloud regions; the red points indicate that foreground objects are correlated with foreground objects; and the blue point indicate that the lawn (bottom row of Fig. 7 is correlated only with the ground. The visualization of important relations clearly demonstrates the ability of SRVGG to capture long-range spatial dependency.

Figure 7. Visualization of important relations corresponding to labeled points (green, red, and blue). The important relations demonstrate the ability of SRVGG to preserve spatial relationship.

Table 2. Quantitative evaluation between different stylization methods in terms of the style loss ($\mathcal{L}_S$), content loss ($\mathcal{L}_C$), relation loss ($\mathcal{L}_R$), and user preference (visual appealing and spatial consistency). The run time was clocked on an NVIDIA 1080Ti with 512×512 images.

	NST [4]	AdaIN [8]	WCT [11]	Ours (NST)	Ours (AdaIN)
$\log(\mathcal{L}_S)$	6.27	6.97	7.00	**5.45**	6.87
$\log(\mathcal{L}_C)$	0.61	0.55	0.65	0.63	**0.41**
$\log(\mathcal{L}_R)$	0.97	0.92	1.09	1.19	**0.85**
Visual appealing (%)	16.2	15.5	6.7	22.7	**39.0**
Spatial consistency (%)	14.4	13.4	5.3	21.3	**45.6**
Run time (s)	47	0.065	56.5	47	**0.06**

6.6. Quantitative Evaluation

We conducted a user study for quantitative evaluation of the results of NST [4], AdaIN [8], WCT [11], and our methods (for both modified AdaIN and NST). For each method, 300 stylized images were generated from 20 content images and 15 style images, which were not used during training. While using each method to generate the stylized images, we computed relation loss, content loss, and style loss via SRVGG for each method for a fair comparison. The losses of optimization-based methods are averaged from last ten iterations. We also conducted a preference study for above-mentioned methods. For each of the 43 participants, twenty combinations of content and style images were randomly chosen. For each combination, the results of five methods were presented in random order and participants were asked to select the most visually appealing one and the most spatially consistent one. We collected 860 votes for both questions and the percentages of votes are shown in Table 2. Compared to other state-of-the-art methods, the proposed modified AdaIN method achieves the lowest relation loss, content loss, and style loss, as shown in Table 2. These results demonstrate the generalization capability of the proposed approach, considering that the proposed network has never seen the test style and content images during training. Furthermore, these results also indicate that the proposed method can maintain the content structure during stylization. Our user study also support this evaluation. Participants thought that the modified AdaIN can produce the best visually appealing and spatially consistent results.

7. Conclusions

We present a novel spatial relation-augmented VGG, that is SRVGG, and a novel relation loss function to facilitate style transfer. Spatial relation module is incorporated within the VGG-19 framework to enrich representations by including global spatial relationships. The relation loss function then utilizes this information to preserve spatial consistency. We apply the proposed framework to both optimization-based and feedforward-based style transfer. In experiments, the proposed method is shown to produce images with high visual appeal and no obvious distortion without the need for a segmentation mask. The quantitative evaluation also supports the effectiveness of the proposed method. Visualizations of important relations suggest that SRVGG captures long-range spatial dependency. We demonstrate that the proposed method can achieve spatially consistent style transfer without using segmentation masks or matting affinity.

References

[1] wikiart. https://www.wikiart.org/.

[2] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016.

[3] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *Proc. of ICLR*, 2017.

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

[5] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[7] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis & machine intelligence*, (6):1397–1409, 2013.

[8] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017.

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

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[11] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, pages 386–396, 2017.

[12] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. *arXiv preprint arXiv:1802.06474*, 2018.

[13] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[14] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[16] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6997–7005. IEEE, 2017.

[17] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017.

[18] Falong Shen, Shuicheng Yan, and Gang Zeng. Meta networks for neural style transfer. *arXiv preprint arXiv:1709.04111*, 2017.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, page 4, 2016.

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor S Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, volume 1, page 3, 2017.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[23] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[24] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[25] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2242–2251. IEEE, 2017.