The Visual Computer
https://doi.org/10.1007/s00371-018-1609-4

ORIGINAL ARTICLE

Photographic style transfer

Li Wang[1] · Zhao Wang[1] · Xiaosong Yang[1] · Shi-Min Hu[2] · Jianjun Zhang[1]

Abstract

Image style transfer has attracted much attention in recent years. However, results produced by existing works still have lots of distortions. This paper investigates the CNN-based artistic style transfer work specifically and finds out the key reasons for distortion coming from twofold: the loss of spatial structures of content image during content-preserving process and unexpected geometric matching introduced by style transformation process. To tackle this problem, this paper proposes a novel approach consisting of a dual-stream deep convolution network as the loss network and edge-preserving filters as the style fusion model. Our key contribution is the introduction of an additional similarity loss function that constrains both the detail reconstruction and style transfer procedures. The qualitative evaluation shows that our approach successfully suppresses the distortions as well as obtains faithful stylized results compared to state-of-the-art methods.

Keywords Photographic style transfer · Deep learning · Photorealism preservation · Image processing

1 Introduction

Image style transfer has shown a promising future for new forms of image manipulation. A neural artistic style transformation method proposed by Gatys et al. [7] has achieved great success with convolutional neural networks, which is followed by many works [2,3,8,11,15,29–31,34] recently. They produce convincing visual results by transferring artistic features from reference painting onto the content photograph. However, these artistic style transfer methods suffer from visual distortion problem, even when both of the content and reference style images are photographic. The stylized results always contain visually intricate distortions which make them have a painting-like looking. Luan et al. [17] point out that the distortions appear only at style transformation process, and they thus propose a photorealism regularization term based on locally affine colour transformations to reconstruct fine content details. To avoid the unexpected geometric matching problem, Luan et al. [17] integrate semantic segmentation masks to Gatys et al.'s method [7]. Although the content spatial structures are preserved in many situations, details and exact shapes of structures are erased when semantic segmentation is inaccurate or contains overlapping areas. And the computation of matting Laplacian matrix and semantic segmentation consumes much extra time for high-quality output. After investigating the style transformation procedure, we discover the distortions occur at two stages: the spatial structures of content image may be lost during content-preserving process and the unexpected geometric matching can be introduced during style transformation process. Figures 1 and 2 illustrate the distortions occur at both content-preserving and style transformation processes. For example, shown as zoom-ins (c-ii), the buildings of content image are obviously distorted by content-preserving process. Moreover, shown as zoom-ins (c-iii), the buildings are also distorted after style transformation process. However, buildings of (c-iii) hold different shapes and edges from (c-ii) from content-preserving process, which means the zoom-in buildings are distorted twice.

To improve the photorealism, this paper introduces an additional similarity layer with the corresponding loss function to constrain both content preservation and style transformation processes. This similarity layer is added into several places of the convolutional neural networks to prevent dis-

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00371-018-1609-4) contains supplementary material, which is available to authorized users.

✉ Li Wang
lwang@bournemouth.ac.uk

[1] National Centre for Computer Animation, Bournemouth University, Poole, UK

[2] Tsinghua University, Beijing, China

Published online: 10 November 2018

tortions by minimizing a similarity loss function and other loss functions proposed in *fast neural style* algorithm [13].

The entire proposed method consists of two stages: detail reconstruction process and style transfer process. Our system has two key components: a dual-stream deep convolution network as Loss Network and edge-preserving filters as style fusion model (SFM). The edge-preserving filter is used to extract details and colour information of the outputs generated from the loss network, which means our scheme combines the details without colour from content and the colour without details from reference style. During the optimization process, the content and style features are captured first by the additional layers in loss network, and then a random white noise image X is passed through both detail reconstruction and style transfer networks. The final output of SFM is the stylized result.

The main contributions of this paper are as follows: we investigate into the problem of Gatys et al.'s method and find out that the lost photorealism of stylized result is caused by distortions occurring at both content preservation and style transformation stages; we propose a photographic style transfer method which is capable of improving the photorealism of stylized results. A similarity loss function using L1-norm is applied for reconstructing finer content details and preventing geometric mismatching problem. And a style fusion model using edge-preserving filter is utilized to reduce artefacts.

2 Related work

Global colour transfer methods Global colour transfer methods tend to utilize spatial-invariant objective functions to transfer images. Input images with simple styles can be processed well by these algorithms [9,12,22,23]. For example, a colour shift technique proposed by Reinhard et al. [23] can extract global features in a decorrelated colour space from reference style image and transfer them onto content input. Pitié et al. [22] propose an approach that also achieves the goal of global style transfer by matching full 3D colour histograms between images with a series of 1D histogram transformation. Although these methods can handle several simple situations like tone curves (e.g., low or high contrast), they are limited in the ability to match complex areas with corresponding colour styles.

Local colour transfer methods. Local colour transfer researches propose to use spatial colour mapping technique like semantic segmentation [10,16,17,21,27,28,32] to handle various applications such as semantic colour gradient transfer (dark and bright) [10,21,27,32], transfer of artistic edits [1,24,26,33], and painting stylistic features [3,4,7,13,15,31,34]. Many of them [7,10,13,15,17,28,31,34] are using convolutional neural network to achieve this goal. Gatys et al. [7] achieve groundbreaking performance of painterly style transfer [15,34] by using the responses of activation layers to represent features from input images. This work focuses mainly on the photographic style transfer, especially the preservation of photorealistic attribute, which is distinguished from their painting-like style transformation [3,7,13,15]. The artistic stylized results are compelling; however, because of distortion problem, the photorealism is lost when their artistic style methods are naively applied to photographic style transfer. To improve the photorealism, recently, Luan et al. [17] propose a photographic style transfer method which uses semantic segmentation and post-processing step to solve the distortion problem. Mechrez et al. [19] propose to use Screened Poisson Equation to replace Luan et al.'s post-processing step and preserve more precise content details than Luan et al.'s results. Liao et al. [16] propose a photorealistic style transfer method for sophisticated images, which are based on finding the nearest neighbour field on deep features extracted from CNN. Our work follows from the neural style algorithm [7] and presents better results than aforementioned methods.

3 Method

This section presents the architecture of our approach and the key loss functions to constrain both detail reconstruction and style transfer processes.

3.1 Architecture

Gatys et al. [7] propose an image transformation network with convolutional neural networks to accomplish the task that an input image is transformed into an output image. The network architecture of Gatys et al. [7] includes a pre-trained VGG-19 network [25] and two loss layers. The layers learn feature representations of input images and compute the representation differences between a generated image and inputs. Their algorithm adds two additional layers: content layer and style layer, which capture and store feature representations of inputs. Then, a random white noise image initialized as the same size of content input is fed into the network. The loss functions compute the distance of feature representations between the generated image with respect to content and reference style inputs separately. The derivatives of loss terms are propagated back to the loss network for next iteration until the maximum iteration number is reached. Similar to this optimization-based approach, our basic network uses the pre-trained VGG-16 network [25] as our loss network. The content loss function and perceptual loss functions in [13] are used in our network. In addition, we add another additional layer with pixel-level loss function into our network. Moreover, we also add a style fusion model as

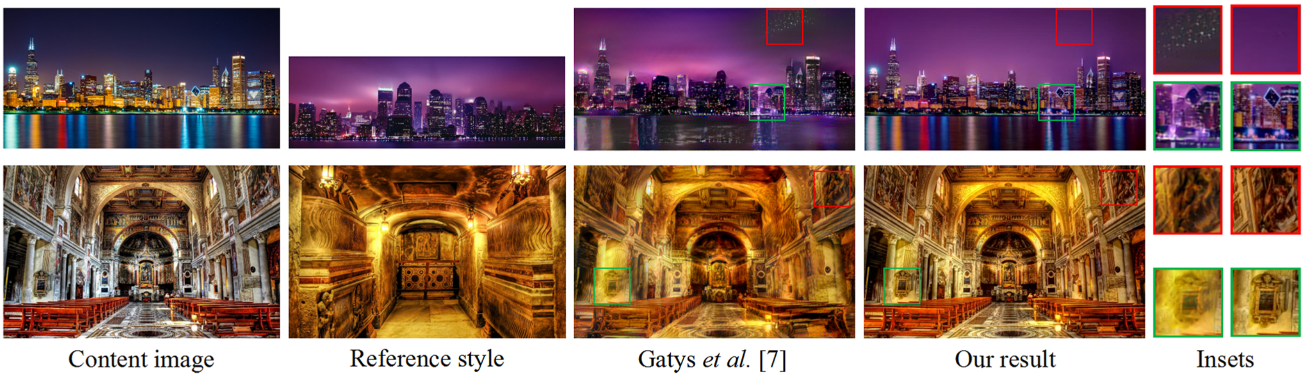


Fig. 1 Given a reference style image and a content image as inputs, photographic style transfer seeks to generate output with photorealistic attribute, which should preserve both the context of content and style of reference. Gatys et al. [7] succeed in transferring style colour but introducing distortions to the context of output. In comparison, our method transfers faithful style colour; meanwhile, it also preserves the photorealistic attribute

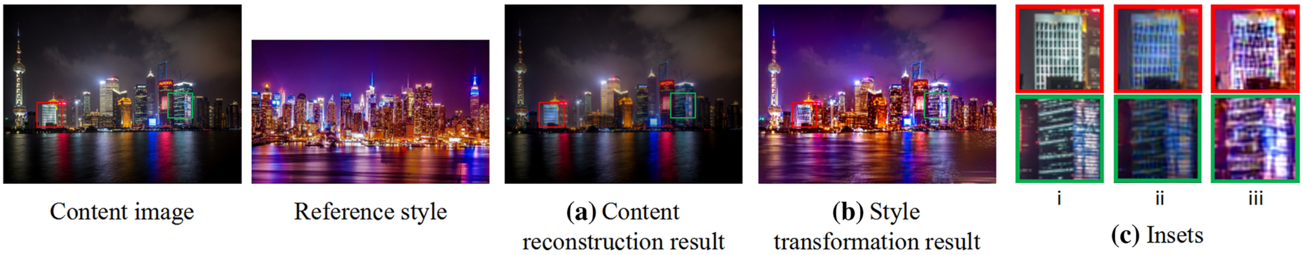
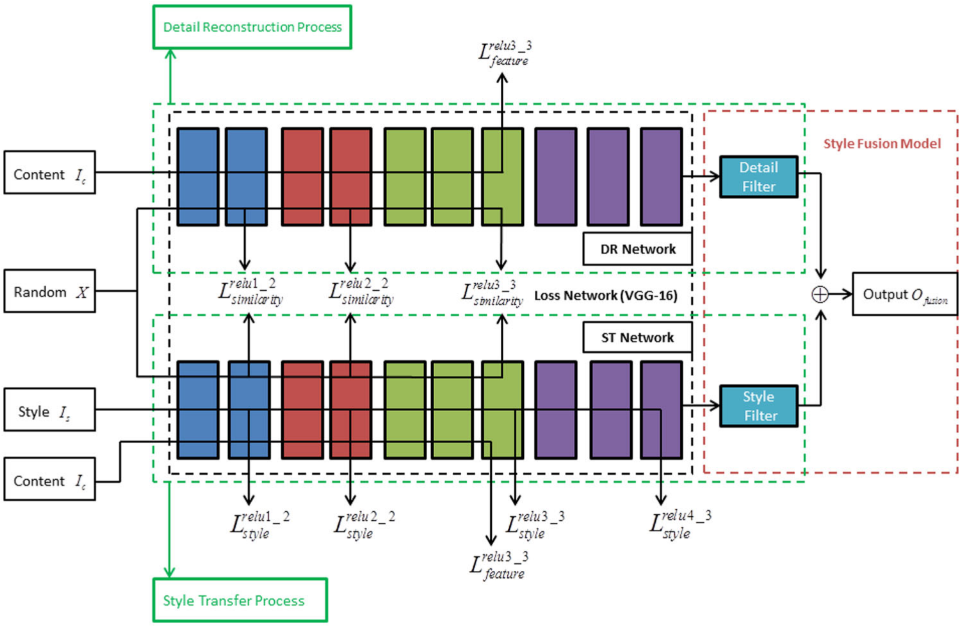

Content image
Reference style
(a) Content reconstruction result
(b) Style transformation result
i
ii
iii
(c) Insets

Fig. 2 Distortions occur at both content-preserving and style transformation processes. **c** contains the zoom-in insets of input content, **a** and **b**. (c-ii) shows that **a** introduces distortions into reconstructed content details, and (c-iii) shows that **b** distorts details of **a**

Fig. 3 Framework overview. We use the Loss network to preserve content and transfer style from inputs to outputs. The loss functions are added into the pre-trained VGG-16 network [25], which are computed at certain layers and backpropagated to the loss network during optimization process. For example, $L^{\text{relu1\_2}}_{\text{style}}$ computes the feature representation differences between random white noise image X and style image I_s, where relu1_2 denotes the placement for style layer in VGG-16 network. Then, the deviation of $L^{relu1\_2}_{style}$ is propagated back to ST network



our post-processing step to reduce artefacts. Our network is an optimization-based approach which is designed for arbitrary style and content image pairs. Thus, it does not need a training process.

As shown in Fig. 3, our framework consists of two components: a dual-stream convolution network consisting of a *loss network* and a *style fusion model*. The loss network is composed by two parallel deep convolution networks and several additional layers. A scalar value $Ł^i(y, y_t)$ of loss function at layer i is computed to measure the Euclidean distance between the output image y and target image y_t (y_t can be content image and reference style image). For the dual-stream loss network, we refer to the upper deep convolution network as *detail reconstruction network* (DR network), which is designed for preserving the content details. Meanwhile, the lower convolution network is referred to as *style transfer network* (ST network), which aims to transfer style information, mainly colour, from reference style image to content input. As shown in the right side of Fig. 3, the *style fusion model* (SFM) also has two components: a detail filter and a style filter, which take the outputs of two parallel deep networks as their inputs separately.

Inputs and outputs For the DR network, the inputs are one photograph as content image I_c and one random white noise image X_{DR} with the same size of I_c, and the output is one image O_c. For the ST network, the inputs are one photograph as content image I_c, one random white noise image X_{ST} with the same size of I_c and one photograph as style image I_s. The output is one image O_s. The X_{DR} and X_{ST} are initialized by random white noise image X. For the detail filter, the input is the output O_c of DR network, and the input of style filter is the output O_s of ST network. The output of entire SFM is one image O_{fusion}.

Additional layers: There are three different layers in total: content layer, style layer and similarity layer. The content and similarity layers carry loss functions for the purpose of preserving content features from I_c onto O_c. And the style layers hold the loss functions to transfer stylistic features from I_s to O_s.

3.2 Loss functions

In general, we define three different loss terms for two purposes: 1 preserve the content feature information F as structure details and reconstruct them on X_{DR}; 2. learn the reference style features and correctly match them to X_{ST}.

Layers in convolutional neural network define nonlinear filter banks to encode input image. Hence, the representations of features in a neural network actually are the filter responses to input image [18]. We assume that a layer has D different filters, and each filter has a size M, where M is height times width. For the reconstruction of feature, let F_i be the feature representations captured at ith activation layer of the DR network when I_c is on processing. Then, F_i is a feature map with the size of $D_i \times M_i$. The feature reconstruction loss is the squared and normalized Euclidean distance between the feature representations of X and target I_c:

$$L_{feat}(X, I_c) = \sum_{i \in L} \frac{1}{D_i \times M_i} \|F_i(X) - F_i(I_c)\|_2^2 \qquad (1)$$

where L denotes the set of activation layers containing feature loss. This term helps to minimize the visual distinguishability between the random image X and target image I_c. However, as this reconstruction is from high layers [18], the rough spatial structure of content image can be preserved, but details especially exact shapes of the structure are lost.

For the same convolutional neural network architecture, Zhao et al. [35] demonstrate using L1-norm loss in the spatial constraint better preserves the spatial structures as compared to using L2-norm. Hence, we introduce another similarity preserved loss L_{simi} based on mean absolute error (L1-norm) into loss network. We found that the L1-norm loss employed outside of the network makes the style transformation output lose the colour information from style image. Hence, we attempt to add L1-norm loss inside the network. Let MAE be the mean absolute error of the feature representations of X and I_c at jth activation layer of the loss network, and then the similarity preserved loss is defined as:

$$L_{simi}(X, I_c) = \sum_{j \in L} \mathrm{MAE}(F_j(X), F_j(I_c)) \qquad (2)$$

where L denotes the set of activation layers added as similarity layers. The purpose of this loss term is how much information of target I_c is lost by X, which contributes to reconstruct exact pixels of I_c into X as many as possible by minimizing this term.

As mentioned above, reconstructing content features with only L_{feat} is not enough to preserve precise details, especially the exact edges inside structures. Figures 4 and 5 demonstrate the effect of L_{simi}.

For the transformation of style, we need to obtain an effective representation of style in the reference image. According to [6], we use correlations of feature space to be the representation of style. And these feature correlations can be given by Gramian Matrix. Let G_k be the Gramian Matrix of vectorized feature map F_k at kth activation layer of ST network when the input x is on processing, and the vectorized feature map F_k is reshaped to $D_k \times H_k W_k$. We define the Gramian Matrix as:

$$G_k(x) = \frac{1}{N} F_k(x) \cdot F_k(x)^T \qquad (3)$$

where N is the total number of pixels of $F_k(x)$. The Gramian Matrix is the inner product between feature maps at kth activation layer, which gives the feature correlations. Then, the style loss is the squared Frobenius norm of the difference between the Gramian Matrices of the random image X_{ST} and the target I_s:

$$L_{style}(X_{ST}, I_s) = \sum_{k \in L} \|G_k(X_{ST}) - G_k(I_s)\|_F^2 \qquad (4)$$

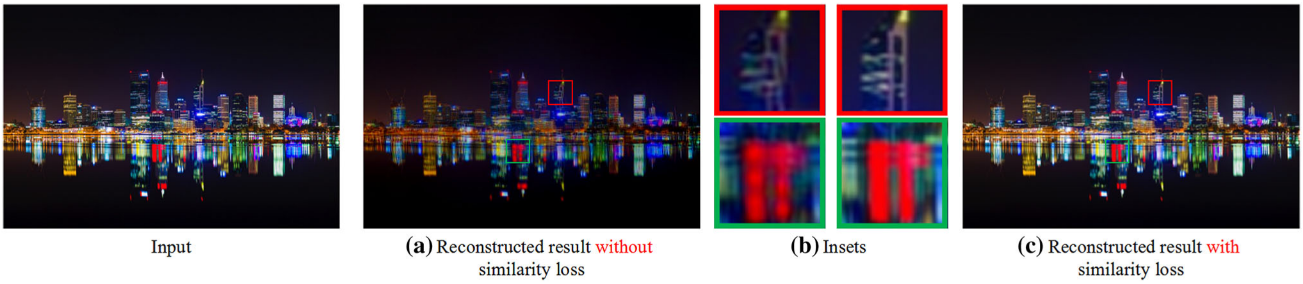

Input
(a) Reconstructed result without similarity loss
(b) Insets
(c) Reconstructed result with similarity loss

Fig. 4 The similarity function for reconstructing finer content details. Left: the input content image. **a** and **c** are the reconstructed results through our DR network without and with similarity loss, respectively. **b** shows two insets of **a** and **c** (in that order), respectively. We may notice that **c** preserves more precise context of input than **a**

where L denotes the set of activation layers holding style loss. The style loss is well defined even for different sizes of X_{ST} and I_{s} since the $G_k(x)$ always has the same $D_k \times D_k$ size. As demonstrated in [6], the generated output will only preserve the stylistic feature from target image, which means the spatial structure of target image cannot be preserved by minimizing the style loss.

In this paper, the L_{feat} and L_{simi} are used to constrain the detail reconstruction procedure, which preserves the spatial structures, exact details like shapes and edges inside content image onto output O_{c} [shown as (c) in Fig. 4]. These two loss terms forms L_{DR}, the joint loss of DR network. The L_{style}, L_{feat} and L_{simi} constrain the style transformation procedure, which generates the output O_{s} with stylistic features mainly colour information from reference image and detailed features from content image. The combination of three loss terms forms L_{ST}, the joint loss of ST network. Therefore, the two final joint loss terms are defined as:

$$L_{\mathrm{DR}} = \alpha_f L_{\mathrm{feat}} + \alpha_d L_{\mathrm{simi}} \tag{5}$$

and

$$L_{\mathrm{ST}} = \beta_f L_{\mathrm{feat}} + \beta_d L_{\mathrm{simi}} + \beta_{\mathrm{s}} L_{\mathrm{style}} \tag{6}$$

where α_f and α_d denote the weights of content layers and similarity layers in DR network, and β_f, β_d and β_{s} denote the weights of three corresponding layers in ST network. All the implementation details of these parameters are introduced in Sect. 4.

In previous researches [12,22], the output of prior process contains stylistic features from reference style, and these features are distributed according to the semantic structures of content input. Hence, the style transformation procedure in our ST network learns stylistic features and also distributes them into the semantic structures, which needs both style loss term and detail reconstruction loss terms. One example result from ST network is shown as (c) in Fig. 6.

3.3 Style fusion model

In Sect. 1, we mention that the distortions are introduced by both detail preservation and style transformation procedures. We use L_{simi} to prevent geometric mismatching; however, the output of ST network may still *exhibit* distortion and noise artefacts due to the content-style trade-off (shown in Fig. 8). To reduce the artefacts, we apply a refinement technique *style fusion model* (SFM) into our approach. The edge-preserving filter (recursion filter) proposed by Gastal et al. [5] is capable of effectively smoothing always noise or textures while retaining sharp edges, which is a suitable technique for reducing artefacts. We thus use the edge-preserving filter (recursion filter) [5] to smooth both output image O_{c} and O_{s} with guidance O_{c}. In this paper, we refer to detail filter and style filter as the smooth process of O_{c} and O_{s}, respectively. The final result O_{fusion} is defined as:

$$O_{\mathrm{fusion}} = (O_{\mathrm{c}} - \mathrm{RF}(O_{\mathrm{c}}, \sigma_{\mathrm{s}}, \sigma_{\mathrm{r}}, O_{\mathrm{c}})) + \mathrm{RF}(O_{\mathrm{s}}, \sigma_{\mathrm{s}}, \sigma_{\mathrm{r}}, O_{\mathrm{c}}) \tag{7}$$

where σ_{s} denotes the spatial standard deviation and σ_{r} denotes the range standard deviation for the edge-preserving filter [5]. Shown as (e) in Fig. 6, the clear stylized result O_{fusion} obtained by our SFM is free to the artefacts.

4 Implementation details

This section describes the implementation details for our approach. We choose pre-trained VGG-16 network [25] as the basic architecture of our DR network and ST network. The content layer with L_{feat} is added into the activation layer of **relu3_3**, and the style layers with L_{style} are added into **relu1_2**,**relu2_2**,**relu3_3** and **relu4_3** activation layers. The similarity layers are added into **relu1_2**,**relu2_2**, **relu3_3** activation layers. For the DR network, we add content and similarity layers into the pre-trained VGG-16 network and

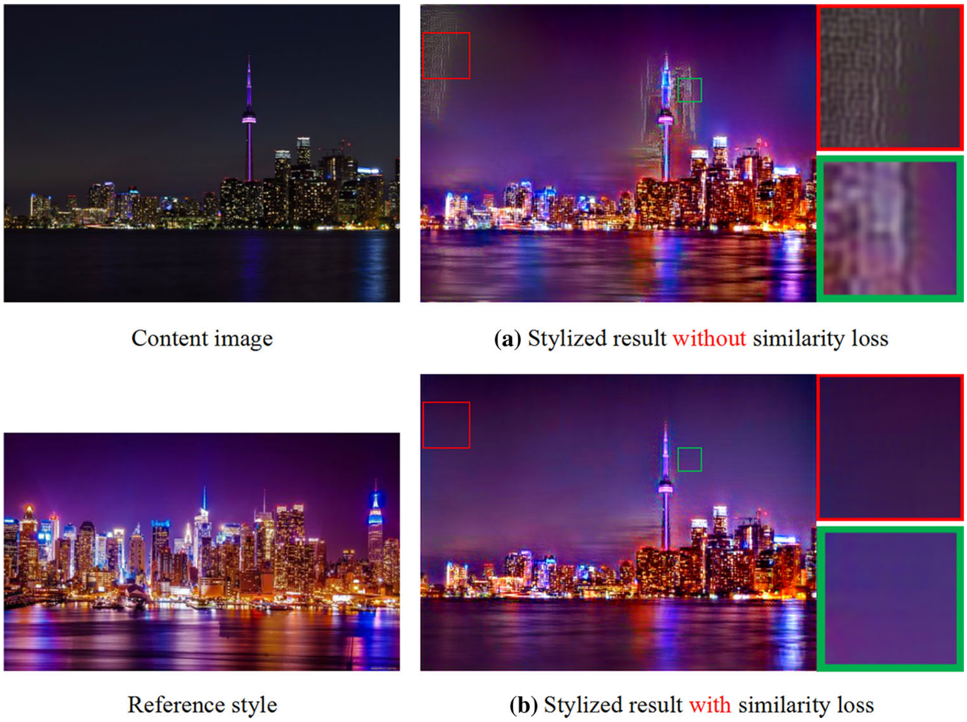

Content image
(a) Stylized result without similarity loss
Reference style
(b) Stylized result with similarity loss

Fig. 5 The similarity function for preventing geometric mismatching problem. **a** is the stylized result without similarity loss, and **b** is the stylized result with similarity loss. Note that the zoom-in regions show that the similarity loss effectively prevents the unexpected geometric matching

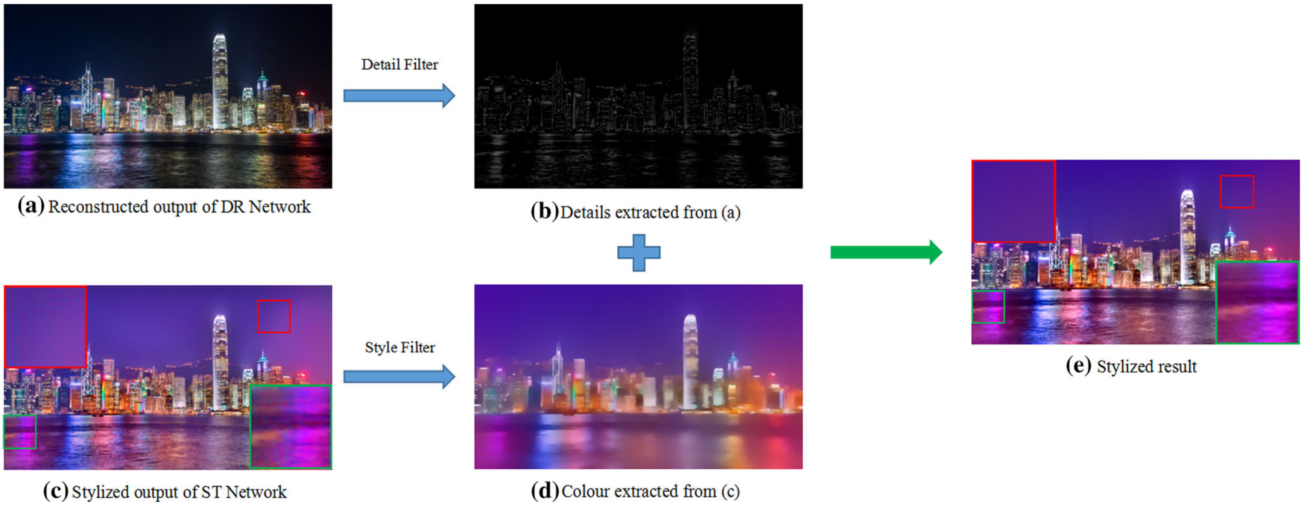

Detail Filter
(a) Reconstructed output of DR Network
(b) Details extracted from (a)
Style Filter
(c) Stylized output of ST Network
(d) Colour extracted from (c)
(e) Stylized result

Fig. 6 The style fusion model for reducing noise artefacts and avoiding distortions. **a** is the reconstructed content output of our DR network, and **b** is the extracted details (white points) of content without colour from **a**. **c** is the stylized output of our ST network, and **d** is the extracted colour without details from **c**. **e** Is the fusion stylized result from SFM. We may notice that **c** still exhibits noise (red rectangles) and distortion (green rectangles) artefacts due to content-style trade-off (please refer to Fig. 8). However, the final stylized result (**e**) is free of noise and distortion artefacts. We recommend readers to view the electronic version

choose parameters $\alpha_f = 5$ and $\alpha_d = 10^3$ for the detail reconstruction. For the ST network, we add content, similarity and style layers into the pre-trained VGG-16 network and choose $\beta_f = 5$, $\beta_d = 10$ and $\beta_s = 100$ for the style transformation. We use $\sigma_s = 60$ (default in the public source code) and $\sigma_r = 1$ for the edge-preserving filter [5] in SFM. The effect of parameter α_d, β_d and σ_r is illustrated in Figs. 7, 8 and 9, respectively.

We use a random white noise image X (X_{DR} and X_{ST} represent X for DR network and for ST network, respectively) with the same size of content image as our initialized input and choose Adam [14] optimization algorithm with learning rate 1 and iteration 1000 in the optimization process for all our experiments in this paper. All the inputs including I_c, I_s and X are scaled into $width = 512$ if their widths are over 512; otherwise, they remain original resolution. The dual-

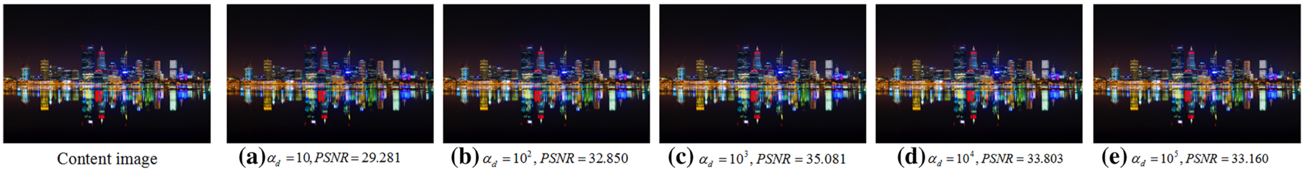

Content image
(a) $\alpha_d = 10, PSNR = 29.281$
(b) $\alpha_d = 10^2, PSNR = 32.850$
(c) $\alpha_d = 10^3, PSNR = 35.081$
(d) $\alpha_d = 10^4, PSNR = 33.803$
(e) $\alpha_d = 10^5, PSNR = 33.160$

Fig. 7 The effect of parameter α_d for our DR network. Note that the reconstructed content result achieves the highest PSNR at $\alpha_d = 10^3$. The lower and larger values decrease the accuracy of reconstructed result. Hence, we find the best parameter $\alpha_d = 10^3$ for our DR network and use it to produce all the other results in this paper

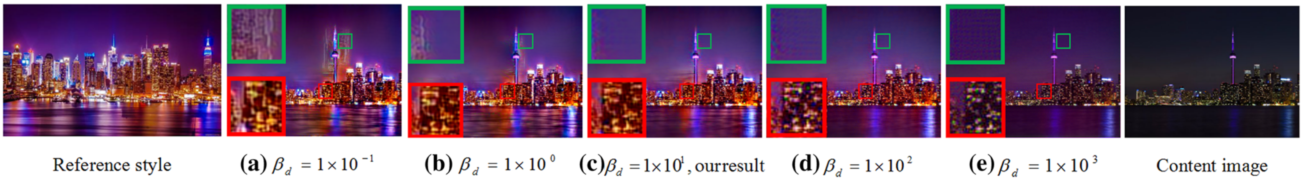

Reference style
(a) $\beta_d = 1 \times 10^{-1}$
(b) $\beta_d = 1 \times 10^{0}$
(c) $\beta_d = 1 \times 10^{1}$, ourresult
(d) $\beta_d = 1 \times 10^{2}$
(e) $\beta_d = 1 \times 10^{3}$
Content image

Fig. 8 The effect of parameter β_d for content-style trade-off. A lower β_d value cannot prevent unexpected geometric matching. For example, the regions of tower tops (green rectangles) in **a** and **b**. A larger β_d value loses the style of reference image. For example, the buildings (red rectangles) in **d** and **e** have undesired dark colour style, which should be in the golden light style. Note that the stylized result at $\beta_d = 1 \times 10^1$ still exhibits some distortion and noise artefacts but they will be eliminated by SFM. We thus choose $\beta_d = 1 \times 10^1$ to produce our style transformation result of the ST network and all the other results in this paper. We recommend readers to view the electronic version

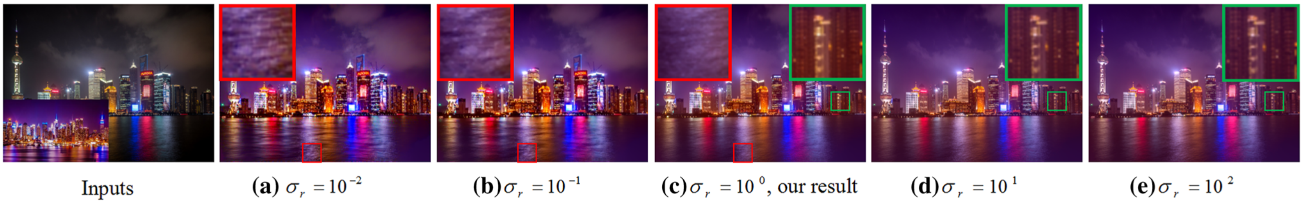


Fig. 9 The effect of parameter σ_r for SFM. Note that a lower σ_r value cannot prevent noise artefacts, for example, red rectangles in **a** and **b**, and a larger σ_r value suppresses the transferred style, for instance, green rectangles in **d** and **e**. We found the best parameter $\sigma_r = 1$ to produce our result and all the other results in our paper

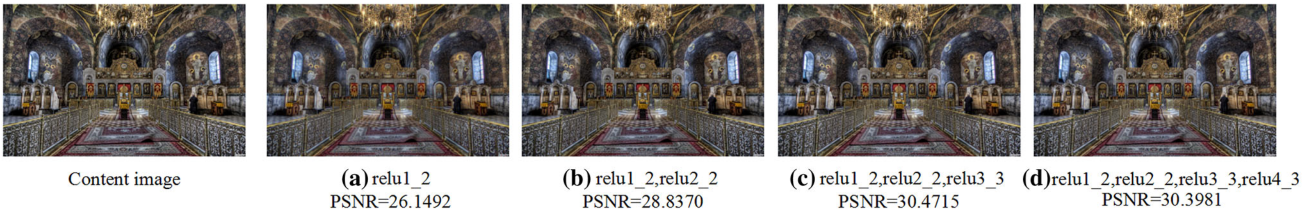
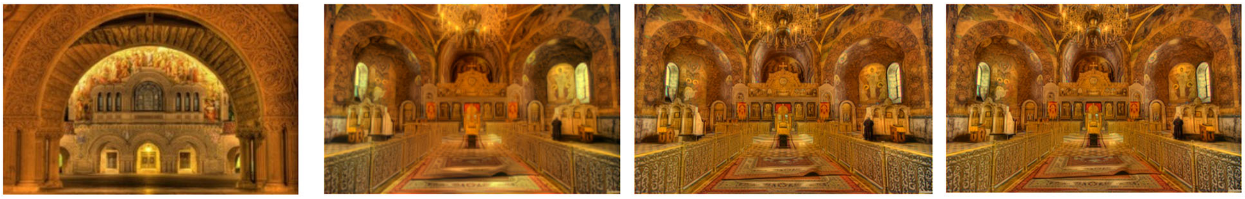


Fig. 10 Placements for similarity layers in DR network. **a**–**d** Show the reconstructed content results with similarity layers at different places in our DR network. Note that the reconstructed result achieves the highest PSNR score at relu1_2,relu2_2,relu3_3. Hence, we place similarity layers at relu1_2,relu2_2,relu3_3 in our DR network for all the experiments in this paper

stream convolution networks run the optimization process at the same time, and the optimization time is around 2.5 min. by running on our GPU card (NVIDIA GeForce GTX 1060, 6G GDDR5). The whole optimization process only needs one content image and one reference style image without any limitation on resolution.

5 Results

This section discusses the selection for hyperparameters, placement for similarity layer, comparisons between our methods and state-of-the-art methods in terms of global and local colour transfer.



Reference style **(a)** relu1_2, relu2_2 **(b)** relu1_2, relu2_2, relu3_3 **(c)** relu1_2, relu2_2, relu3_3, relu4_4

Fig. 11 Placements for similarity layers in ST network. **a**–**c** show the stylized results with similarity layers at different places in our ST network. Note that **a** presents a worse stylized result than **b** and **c** as the centre area of blanket and walls upside are not in golden style colour. It is difficult to tell that either **b** or **c** outperforms better style transformation as they achieve a very similar style transfer result (conducting a series of other experiments, please refer to our supplemental materials for more details). We thus choose to place similarity layers at relu1_2,relu2_2,relu3_3 in our ST network, which keeps the same placements as the DR network

Table 1 Additional layers and VGG-16 network

Additional layers	Layers of VGG-16	Size	Activation
Similarity, style	conv1_1	64 × 3 × 3	relu1_1
	conv1_2	64 × 3 × 3	relu1_2
	Maxpooling	2 × 2	
Similarity, style	conv2_1	128 × 3 × 3	relu2_1
	conv2_2	128 × 3 × 3	relu2_2
	Maxpooling	2 × 2	
Content, similarity, style	conv3_1	256 × 3 × 3	relu3_1
	conv3_2	256 × 3 × 3	relu3_2
	conv3_3	256 × 3 × 3	relu3_3
	Maxpooling	2 × 2	
Style	conv4_1	512 × 3 × 3	relu4_1
	conv4_2	512 × 3 × 3	relu4_2
	conv4_3	512 × 3 × 3	relu4_3
	Maxpooling	2 × 2	
	conv5_1	512 × 3 × 3	relu5_1
	conv5_2	512 × 3 × 3	relu5_2
	conv5_3	512 × 3 × 3	relu5_3
	Maxpooling	2 × 2	
	fully connection	4096 × 1 × 1	relu
	fully connection	4096 × 1 × 1	relu
	fully connection	1000 × 1 × 1	relu

5.1 The effect of hyperparameters

Figures 7 and 8 demonstrate the effect of parameters α_d and β_d, respectively. As shown in Fig. 7, the content reconstructed result achieves the highest PSNR (peak signal-to-noise ratio) value when $\alpha_d = 10^3$. We thus choose $\alpha_d = 10^3$ to reconstruct content details in our DR network. In Fig. 8, a lower β_d value still produces stylized result with geometric mismatching problem. Conversely, a too larger β_d value produces less style result. Hence, we find the best value $\beta_d = 10$ to produce our stylized result and all the other results in this paper. Figures 10 and 11 illustrate the choices of similarity layers in our DR network and ST network, respectively. For DR network, we choose to place similarity layers at relu1_2,relu2_2,relu3_3 as it achieves the highest PSNR score. For ST network, the stylized results (b) and (c) have very similar style transformation appearance, and we thus choose to place similarity layers at relu1_2,relu2_2,relu3_3 in our ST network, which keeps the same placements as the DR network. The implementation details of our networks are described in Tables 1, 2, and 3.

Table 2 Implementation details of DR network

Loss	Parameters	Placements in VGG-16
L_{feat}	$\alpha_f = 5$	relu3_3
L_{simi}	$\alpha_d = 10^3$	relu1_2,relu2_2,relu3_3

Table 3 Implementation details of ST network

Loss	Parameters	Placements in VGG-16
L_{feat}	$\beta_f = 5$	relu3_3
L_{simi}	$\beta_d = 10$	relu1_2,relu2_2,relu3_3
L_{style}	$\beta_s = 100$	relu1_2,relu2_2,relu3_3,relu4_3

5.2 Comparisons

This section presents several comparisons between state-of-the-art methods and ours.

Comparison between representative artistic style transfer methods and ours We compare Gatys et al. [7], Ghiasi et al. [8] with ours across great differences among content images in Fig. 12. Our results preserve content structures with more precise details than other artistic prior methods. For example, our results contain all details of ceiling lamp, frescoes, carpets, and railings which are not reconstructed well by Gatys et al. [7] and Ghiasi et al. [8]. To illustrate the ability of preserving precise details, we compare content and reference style image with great details to prior artistic style transfer methods in third row. Our method reconstructs almost every detail in content image and transfers the colour style faithfully, while Gayts et al. and Ghiasi et al. [8] lose great details. The detail representations on other examples also show our strong ability to reduce distortions and preserve content spatial structures as well.

Comparison between representative global colour transfer methods and ours. In Fig. 13, we compare our method with representative global colour transfer algorithms such as Reinhard et al. [23] and Pitié et al. [22]. A global colour mapping technique is applied by both of them to match the colour statistics of content input and reference style image. However, they cannot obtain faithful colour transformation results when the inputs have spatial-varying contents, which limits their applications. For example, in the second row of Fig. 13, Reinhard et al. and Pitié et al. methods cannot transfer light style in reference style image to buildings.

Comparison between representative local photographic style transfer methods and ours In Fig. 14, we compare our method ([7]+ours) with the state-of-the-art methods, Luan et al. [17] and Liao et al. [16]. The approaches proposed by Luan et al. [17] and Liao et al. [16] are the latest methods which effectively avoids the distortion problem. Our method preserves more precise content details than Luan et al. For

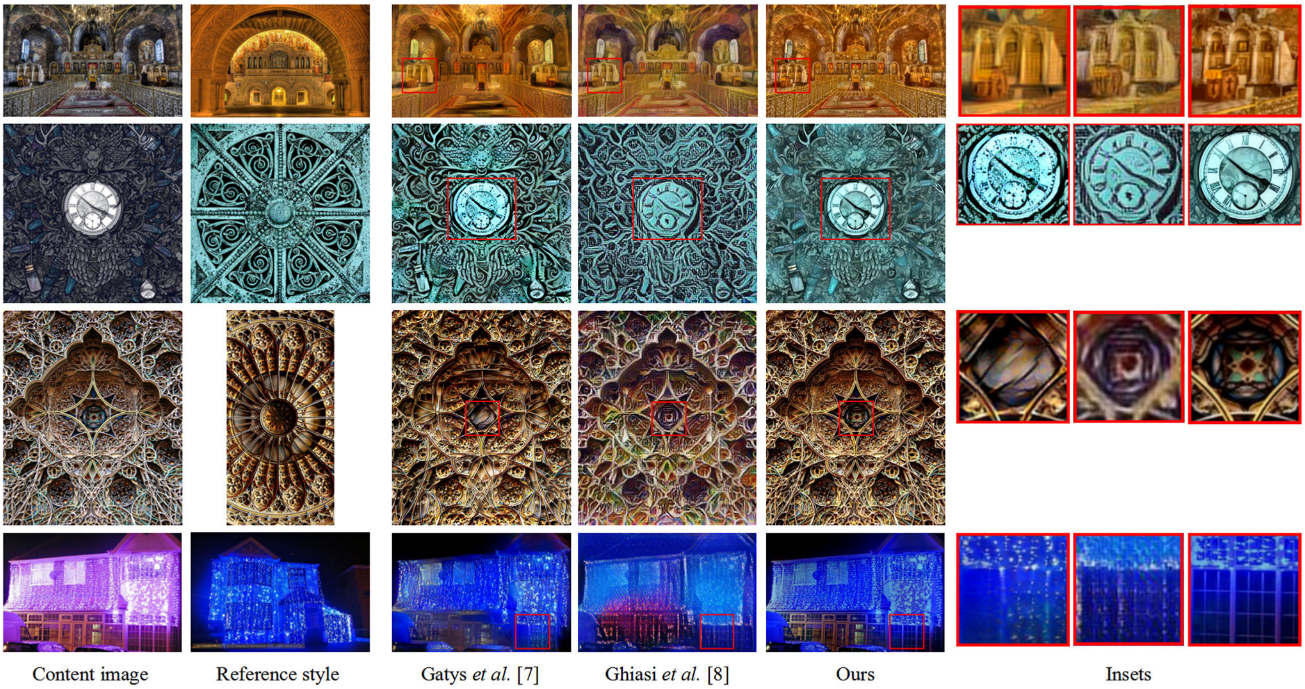


Fig. 12 Comparison between Gatys et al. [7], Ghiasi et al. [8] and ours. Gatys et al. [7] and Ghiasi et al. [8] produce a larger amount of distortions in their results while ours are free of distortions. The stylized results of Ghisai et al. [8] method use the interpolation weight of 0.8 and other default parameter values in their paper

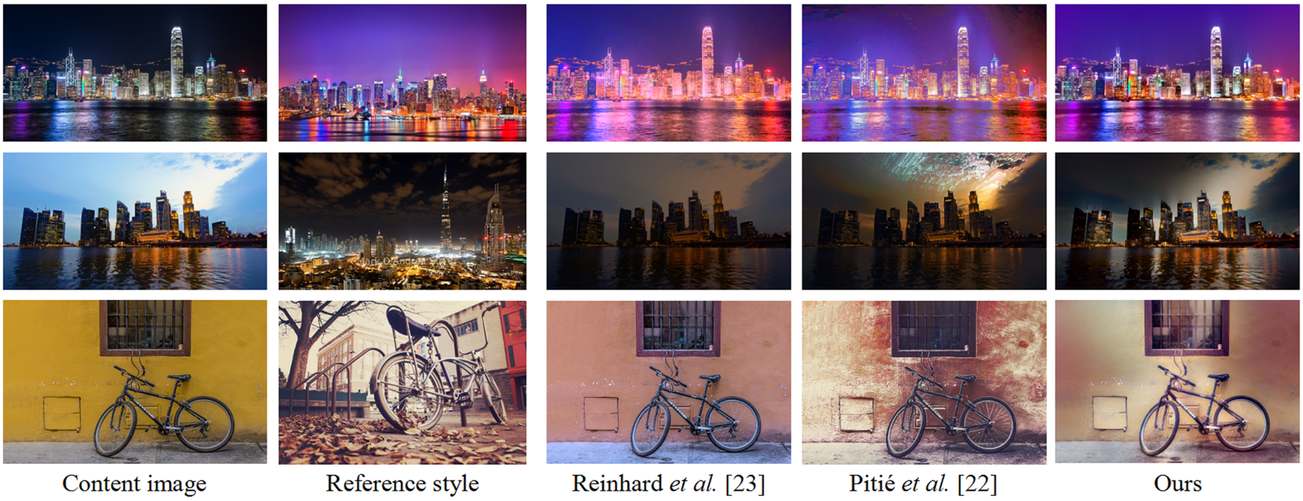


Fig. 13 Comparison between representative global colour transfer methods Reinhard et al. [23], Pitié et al. [22] and ours

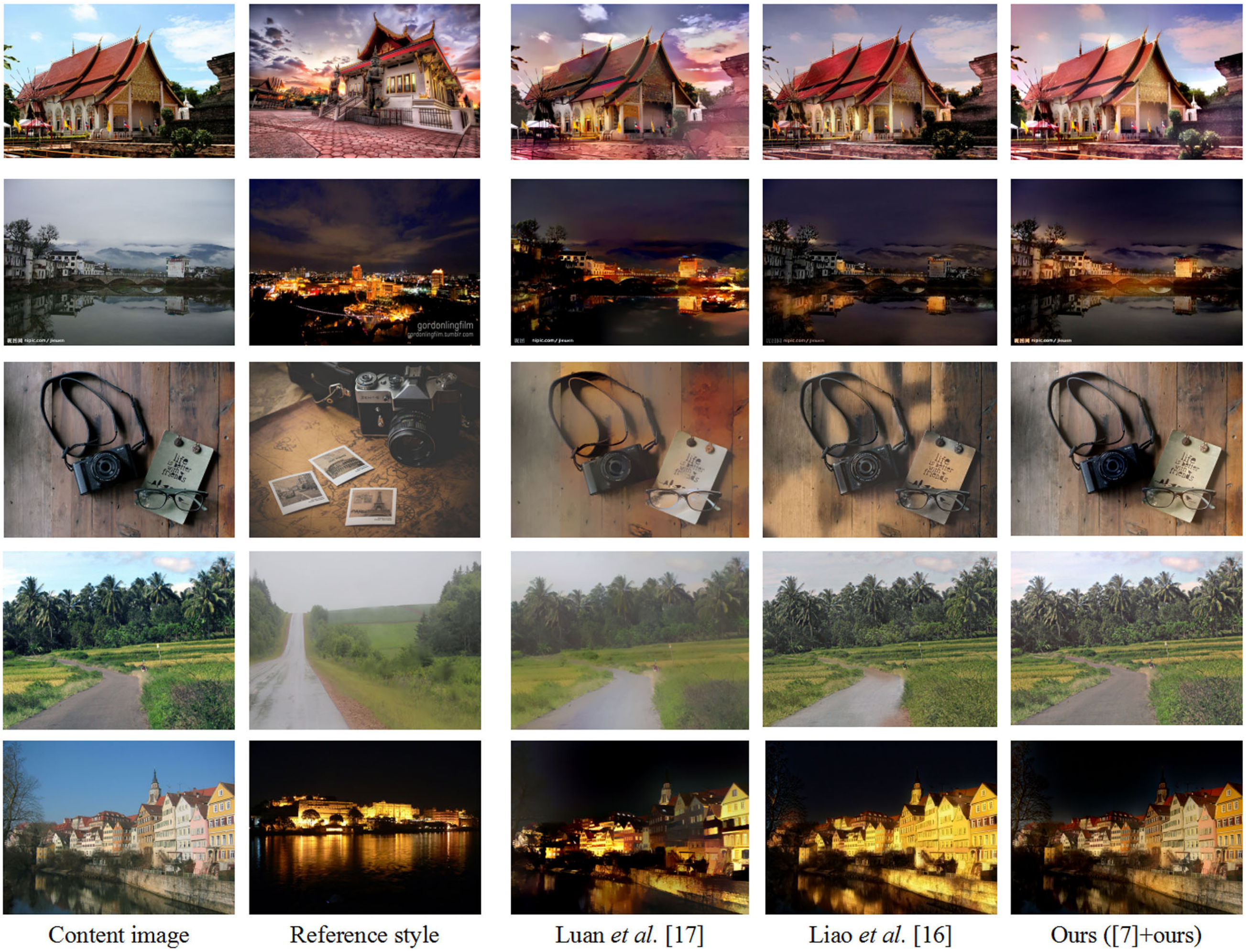


Fig. 14 Comparison between Luan et al. [17], Liao et al. [16] and ours. All examples from Luan et al. [17] dataset

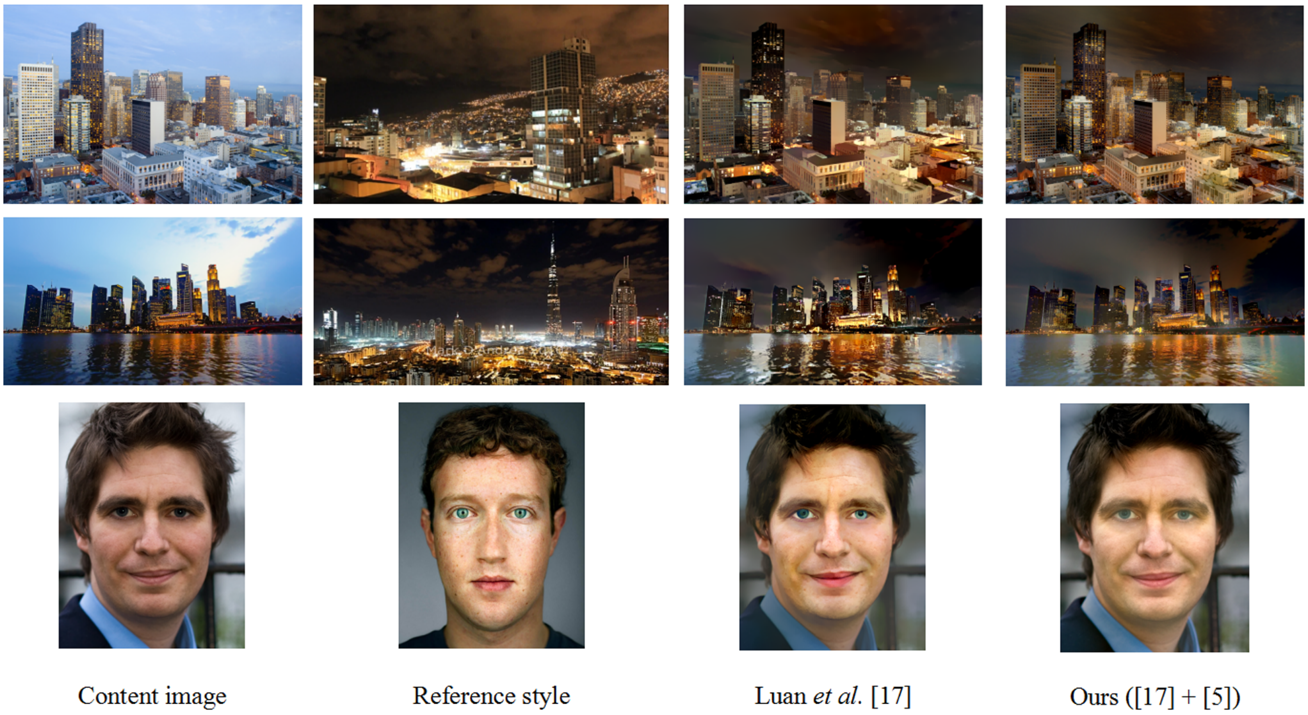


Fig. 15 Comparison between Luan et al. [17] and ours ([17]+[5]). Our method effectively handles the posterization effect of Luan et al. [17]. All examples from Luan et al. [17] dataset. We recommend readers to view the electronic version

example, the plants in the first row, the characters of postcard in the third row and the windows in the bottom row. Our method may not obtain better faithful transformation results but our method achieves the highest score on the photorealism. Please refer to user study for more details in Sect. 5.3 and more scores in our supplemental materials. All the stylized results (including user study) of Luan et al. [17] use manually semantic segmentation masks provided by the authors and parameter $\lambda = 10^4$ (default value in Luan et al.'s paper). We further compare our method with Luan et al. using different λ values on the images in Fig. 12, please refer to our supplemental materials for more details.

Luan et al. [17] propose a two-stage photograph style transfer method which expands Gatys et al.'s artistic style transfer method. Their first stage integrates semantic segmentation into neural style [7] method for object-to-object colour transfer, and their second stage applies a post-processing step to improve the photorealism of stylized result obtained from the first stage. In terms of local object-to-object colour transfer, our similarity loss function may not transfer colour for object-to-object as faithful as manually semantic segmentation. However, our edge-preserving filter [5] used in SFM may help Luan et al.'s results avoid the posterization artefacts. In Fig. 15, we show the stylized results that we apply edge-preserving filter [5] to process the results obtained from Luan et al.'s first stage. For example, our method effectively prevents the posterization artefacts on buildings in the first row, water in the second row and forehead in the third row.

In Fig. 16, we compare our method ([17]+[5]) with state-of-the-art photographic style transfer methods Luan et al. [17] and Liao et al. [16]. Note that our method ([17]+[5]) preserves more precise content details than Luan et al. [17] while transferring style more faithfully than Liao et al. [16]. Please see more details in the our supplemental materials.

In Fig. 17, we compare our method ([17]+[5]) with Mechrez et al. [19] which proposes to apply Screened Poisson Equation (SPE) [20] to improve the photorealism of result obtained from Luan et al.'s first stage. Note that Mechrez et al. [19] method cannot remove artefacts introduced by Luan et al.'s first stage. For example, the unexpected blue colour and inconsistent colour in the first and third rows, respectively.

In Fig. 18, we demonstrate that our method is robust on preserving content spatial details and achieving faithful style transformation results. Note that (c) still preserves well the details of first content input even through two style transformation process with different reference style images, and the photorealism of (c) is also preserved well.

Limitation Our method is unable to transfer faithful colour between images which have semantic similarity for human observers but with much complex spatial-varying. In Fig. 19, we show some failure cases. For example, the blanket and floor in first row fail to be transferred into brown and white colour style.

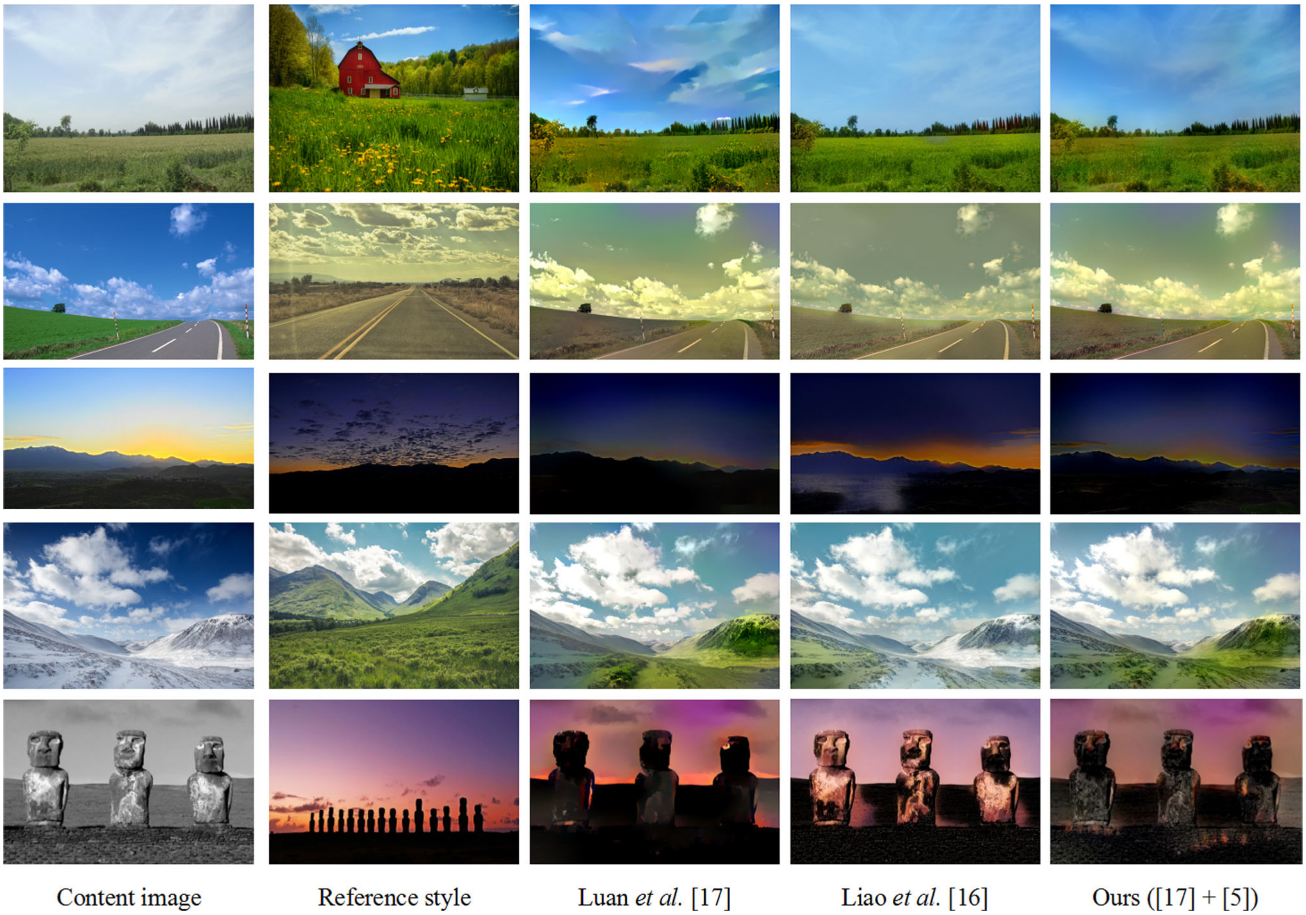


Fig. 16 Comparison between Luan et al. [17], Liao et al. [16] and ours ([17]+[5]). Our method preserves finer content details than Luan et al. [17] and transfer style more faithful than Liao et al. [16]. All examples from Luan et al. [17] dataset. We recommend readers to view the electronic version

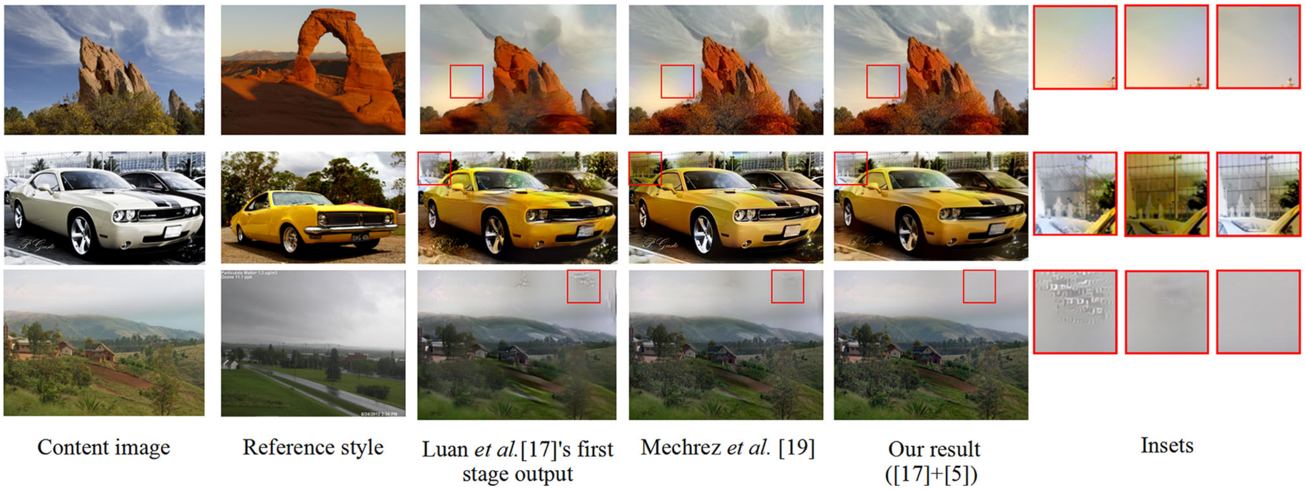
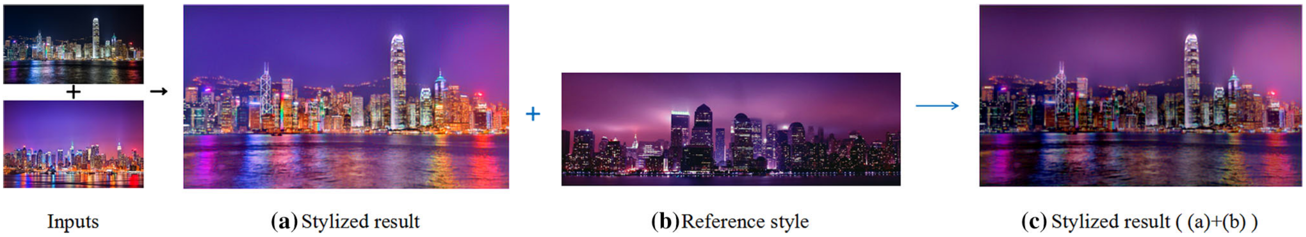


Fig. 17 Comparison between Mechrez et al. [19] and ours ([17]+[5]). The zoom-ins show the *insets* of Luan et al.'s first stage output, Mechrez et al. [19] and ours ([17]+[5]) (in that order). We recommend readers to view the electronic version



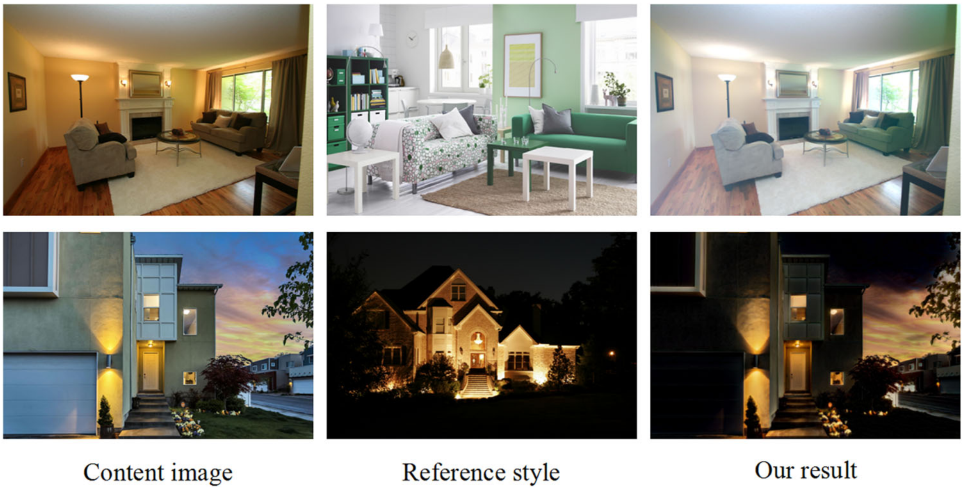
Inputs

(a) Stylized result

(b) Reference style

(c) Stylized result ((a)+(b))

Fig. 18 Inputs: Content image (upper) and Reference Style image 1 (lower). **a** represents the stylized result combining context of content image and style of Reference Style image 1. **b** is the Reference Style image 2. **c** is the stylized result transferring style of Reference Style image 2 to context of **a**

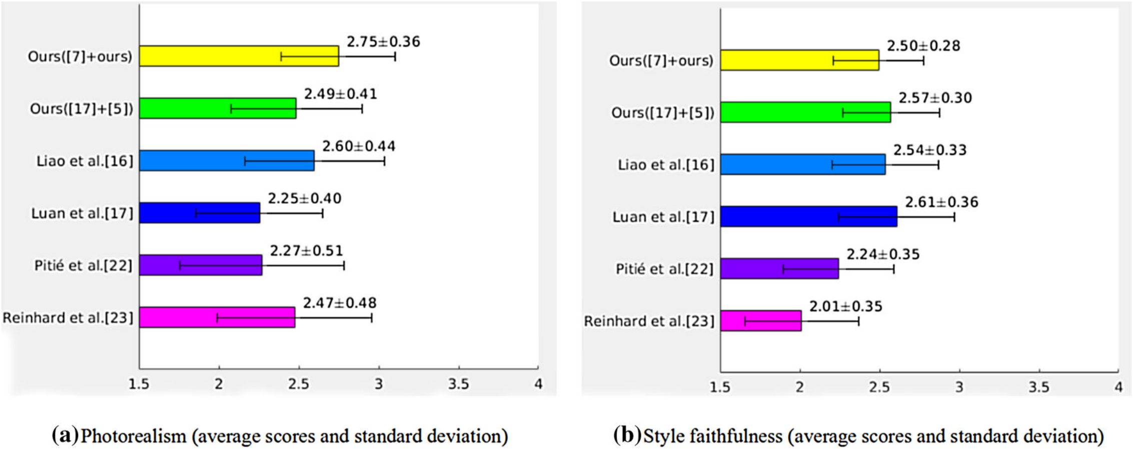


Content image

Reference style

Our result

Fig. 19 Some failure cases



(a) Photorealism (average scores and standard deviation)

(b) Style faithfulness (average scores and standard deviation)

Fig. 20 User study results for photorealism and style faithfulness

5.3 User study

We conduct a user survey to verify several colour transfer methods on photorealism and style faithfulness. There are six different methods considered in this survey, which include Reinhard et al. [23], Pitié et al. [22], Luan et al. [17], Liao et al. [16], our methods ([7]+ours, [17]+[5]). We ask 26 human participants to score stylized results on 1-to-4 scale. For the photorealism, the score ranges from “1: definitely not photorealistic” to “4: definitely photorealistic”. For the style faithfulness, the score ranges from “1: definitely not style faithful to reference style” to “4: definitely style faithful to reference style”. For each person, he or she is asked to score the stylized results of 6 methods in a random order. There are totally 44 different scenes (excluding unrealistic and repeated scenes) selected from Luan et al. [17] dataset.

In Fig. 20, we show the average score and standard deviation of each method. For the photorealism, our method ([7]+ours) and Liao et al. [16] rank the 1st and 2nd, respectively. Luan et al. [17] and Pitié et al. [22] have the worst performance regarding the photorealism as their results *exhibit* some artefacts. For the style faithfulness, Luan et al. [17] and our method ([17]+[5]) rank 1st and 2nd, respectively. Our edge-preserving filter [5] used in SFM slightly declines the style faithfulness score of Luan et al. [17], but it still achieves a higher score than Liao et al. [16]. Moreover, it significantly improves the photorealism score of Luan et al.'s results. Reinhard et al. [23] and Pitié et al. [22] perform the worst in the style faithfulness as their limitations for sophisticated images.

6 Conclusions

We investigate into the reason why the photorealism of stylized results is lost even when the photographic images are input to Gatys et al.'s method [7]. And we knowledge that both content preservation and style transformation stages distort images to lose the photorealistic attribute. Hence, we propose a photographic style transfer method that constrains detail reconstruction and style transformation processes by introducing a similarity loss function. This similarity loss function not only preserves exact details and structures of content image but also prevents the mismatch of texture patches between reference style and content image. The qualitative evaluation on Luan et al.'s [17] dataset shows that our proposed approaches are capable of preventing the distortions effectively and obtaining faithful stylized results as well.

Acknowledgements This work was supported by the European Commission under FP7 Marie Curie IRSES project AniNex (612627) and the China Scholarship Council. Li Wang appreciated support from Visual Media Research Centre at Tsinghua University, during his Visiting in Beijing.

References

1. Bae, S., Paris, S., Durand, F.: Two-scale tone management for photographic look. ACM Trans. Gr. (TOG) **25**, 637–645 (2006)
2. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stylebank: an explicit representation for neural image style transfer. arXiv preprint arXiv:1703.09210 (2017)
3. Chen, T.Q., Schmidt, M.: Fast patch-based style transfer of arbitrary style. arXiv preprint arXiv:1612.04337 (2016)
4. Chi, M.T., Liu, W.C., Hsu, S.H.: Image stylization using anisotropic reaction diffusion. Vis. Comput. **32**(12), 1549–1561 (2016)
5. Gastal, E.S., Oliveira, M.M.: Domain transform for edge-aware image and video processing. ACM Trans. Gr. (ToG) **30**, 69 (2011)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
8. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., Shlens, J.: Exploring the structure of a real-time, arbitrary neural artistic stylization network. arXiv preprint arXiv:1705.06830 (2017)
9. Gong, H., Finlayson, G.D., Fisher, R.B., Fang, F.: 3d color homography model for photo-realistic color transfer re-coding. Vis. Comput. pp. 1–11 (2017)
10. He, M., Liao, J., Yuan, L., Sander, P.V.: Neural color transfer between images. arXiv preprint arXiv:1710.00756 (2017)
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. arXiv preprint arXiv:1703.06868 (2017)
12. Hwang, Y., Lee, J.Y., So Kweon, I., Joo Kim, S.: Color transfer using probabilistic moving least squares. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3342–3349 (2014)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp 694–711. Springer, Berlin (2016)
14. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Li, C., Wand, M.: Combining markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2479–2486 (2016)
16. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. arXiv preprint arXiv:1705.01088 (2017)
17. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. arXiv preprint arXiv:1703.07511 (2017)
18. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5188–5196 (2015)
19. Mechrez, R., Shechtman, E., Zelnik-Manor, L.: Photorealistic style transfer with screened poisson equation. arXiv preprint arXiv:1709.09828 (2017)
20. Morel, J.M., Petro, A.B., Sbert, C.: Screened poisson equation for image contrast enhancement. Image Process. Line **4**, 16–29 (2014)
21. Oliveira, M., Sappa, A.D., Santos, V.: Unsupervised local color correction for coarsely registered images. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 201–208. IEEE, New York (2011)
22. Pitie, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005, vol. 2, pp 1434–1439. IEEE, New York (2005)
23. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Comput. Gr. Appl. **21**(5), 34–41 (2001)
24. Shih, Y., Paris, S., Barnes, C., Freeman, W.T., Durand, F.: Style transfer for headshot portraits. ACM Trans. Gr. (TOG) **33**, 148 (2014)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Trans. Gr. (TOG) **29**, 125 (2010)

27. Tai, Y.W., Jia, J., Tang, C.K.: Local color transfer via probabilistic segmentation by expectation-maximization. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, pp. 747–754. IEEE, New York (2005)
28. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. arXiv preprint arXiv:1703.00069 (2017)
29. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Multi-style generative network for real-time transfer. In: ICML, pp. 1349–1357 (2016)
30. Wang, X., Oxholm, G., Zhang, D., Wang, Y.F.: Multimodal transfer: a hierarchical deep convolutional neural network for fast artistic style transfer. arXiv preprint arXiv:1612.01895 (2016)
31. Wilmot, P., Risser, E., Barnes, C.: Stable and controllable neural texture synthesis and style transfer using histogram losses. arXiv preprint arXiv:1701.08893 (2017)
32. Wu, F., Dong, W., Kong, Y., Mei, X., Paul, J.C., Zhang, X.: Content-based colour transfer. Comput. Gr. Forum, Wiley Online Libr. **32**, 190–203 (2013)
33. Yi, Z., Li, Y., Ji, S., Gong, M.: Artistic stylization of face photos based on a single exemplar. Vis. Comput. **33**(11), 1443–1452 (2017)
34. Zhang, H., Dana, K.: Multi-style generative network for real-time transfer. arXiv preprint arXiv:1703.06953 (2017)
35. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Is l2 a good loss function for neural networks for image processing? arXiv preprint arXiv:1511.08861 (2015)

Li Wang is currently a PhD candidate at National Centre for Computer Animation, Bournemouth University, UK. He received his BS and MS degree in Computer Science from Jilin University (China) in 2013 and 2016, respectively. His research interests include deep learning, computer vision and computer graphics.

Zhao Wang received his PhD degree from National Centre for Computer Animation, Bournemouth University, UK. He currently works as a research fellow in Zhejiang Lab, China. His research interests include 3D vision, multi-task learning, edge computing and distributed AI.

Xiaosong Yang is currently an Associate Professor in the National Centre for Computer Animation, Bournemouth University, UK. He received his bachelor (1993) and master degree (1996) in computer science from Zhejiang University (P. R. China) and Ph.D. (2000) in computing mechanics from Dalian University of Technology (P. R. China). He worked as PostDoc (2000–2002) in the Department of Computer Science and Technology of Tsinghua University for two years and as Research Assistant (2001–2002) at Chinese University of Hong Kong. His research interests include deep learning, computer vision, computer animation, motion capture and synthesis, VR&AR, special effects and game development, digital health, data mining, medical visualization.

Shi-min Hu is currently a professor in the department of Computer Science and Technology, Tsinghua University, Beijing. He received the PhD degree from Zhejiang University in 1996. His research interests include geometry modeling and processing, video processing, rendering, computer animation, and Virtual Reality. He has published more than 100 papers in journals and refereed conference. He is Editor-in-Chief of Computational Visual media, and on editorial board of several journals, including IEEE Transactions on Visualization and Computer Graphics, Computer Aided Design and Computer & Graphics. He is a senior member of IEEE and ACM.

Jianjun Zhang is currently a Professor of Computer Graphics at the National Centre for Computer Animation, Bournemouth University and leads the Computer Animation Research Centre. His research focuses on a number of topics relating to 3D computer animation, including virtual human modelling and simulation, geometric modelling, motion synthesis, deformation and physics-based animation. He is also interested in virtual reality and medical visualization and simulation. Prof. Zhang has published over 200 peer-reviewed journal and conference publications. He has chaired over 30 international conferences and symposia and serves on a number of editorial boards.