Towards Deep Style Transfer: A Content-Aware Perspective

Yi-Lei Chen[1]
https://sites.google.com/site/fallcolor

Chiou-Ting Hsu[2]
http://www.cs.nthu.edu.tw/~cthsu/candy.html

[1] Pixart Imaging Inc.
Hsinchu, Taiwan

[2] Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan

Abstract

Modern research has demonstrated that many eye-catching images can be generated by style transfer via deep neural network. There is, however, a dearth of research on content-aware style transfer. In this paper, we generalize the neural algorithm for style transfer from two perspectives: *where to transfer* and *what to transfer*. To specify where to transfer, we propose a simple yet effective strategy, named *masking out*, to constrain the transfer layout. To illustrate what to transfer, we define a new style feature by high-order statistics to better characterize content coherency. Without resorting to additional local matching or MRF models, the proposed method embeds the desired content information, either semantic-aware or saliency-aware, into the original framework seamlessly. Experimental results show that our method is applicable to various types of style transfers and can be extended to image inpainting.

1 Introduction

Exampler-based image editing has been applied to a wide range of research, including image analogy, texture synthesis, color transfer, image retargeting, content/style separation, and so on. In this paper, we are interested in image style transfer, which aims to synthesize an image with its “style” similar to the exemplar from several aspects, e.g., by distorting texture, deviating color tone, adjusting visual contrast, or generating irregular patterns. Existing methods can be categorized into parametric [1, 2, 3] and non-parametric [4, 5, 6, 7] approaches. Most parametric methods were developed only for specific source images (e.g., faces or objects) because developing a comprehensive model is almost intractable. By contrast, non-parametric methods relying on patch-based synthesis is model-free, and they therefore usually outperform parametric models in terms of both quality and applicability.

Recently, it has been shown that one can invert a deep convolutional neural network originally trained for classification tasks to generate visually plausible images [8]. The impressive results inspire broader change in image generation [9, 10, 11] as well as image synthesis [12, 13, 14, 15, 16, 17, 18, 19]. Gatys et al. [12] first applied this technique to image style transfer and successfully reproduced famous painting styles on natural images. A series of following work was subsequently proposed to improve neural algorithms for style transfer. For example, [13] and [14] proposed to speed up the computation by training a one-pass feed-forward network to circumvent the iterative network evaluation. Gardner

doi:10.5244/C.110.###

Pages 8.1-8.11
DOI: https://dx.doi.org/10.5244/C.30.8

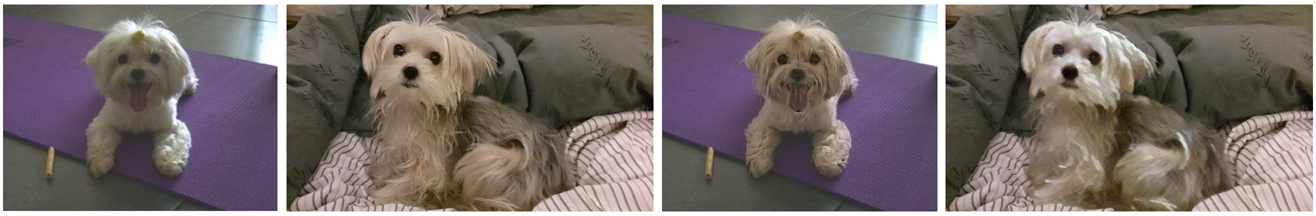

Fig. 1. An example of content-aware style transfer. The face appearance of two different breeds of dogs, *Maltese* and *Yorkshire terrier*, are exchanged by our method.

et al. [15] extended the original method to support the exemplar from one image to a data-driven manifold defined by two image sets; and Lin et al. [16] further trained a linear classifier to enable attribute-based transfer. Moreover, because the algorithm proposed by Gatys et al. transfers style features globally, the method tends to spoil local image plausibility. A few methods hence devoted to content-aware style transfer by either using region segmentation [17] or including patch-based MRF priors [18] and semantic maps [19].

In this paper, we delve into the style transfer algorithm [12] and propose a generalized formulation to address *content-aware style transfer* (see Fig. 1 for an example). In contrast to [17, 18, 19], we include no advanced image processing techniques and merely introduce a simple yet effective concept named *masking out* to achieve the content-aware style transfer. We will show that our method can seamlessly embed the desired content information, either semantic-aware or saliency-aware, into the original formulation. In order to better characterize content coherency, we propose a new style feature encoding high-order statistics, and further show that this new feature can also be unified by our formulation.

2 Method

2.1 Review of Neural Algorithms for Style Transfer

We first briefly review the style transfer algorithm proposed by Gatys et al. [12]. Given a source image (or content image) $\mathbf{c}$ and a target image (or style image) $\mathbf{s}$, the method aims to synthesize an image $\mathbf{x}$ which simultaneously shares the visual content of $\mathbf{c}$ and the style representation of $\mathbf{s}$. Specifically, the authors model the image rendering as an optimization problem by minimizing the difference between $\mathbf{c}$ and $\mathbf{x}$ and the difference between $\mathbf{s}$ and $\mathbf{x}$ in terms of content and style features, respectively. They characterize both features by the deep convolutional neural network VGG19 trained on the ImageNet dataset [20].

The VGG19 model was composed of a series of convolution layers, pooling layers, and activation layers. In [12], neither re-training nor fine-tuning was required to use the deep model for style transfer. Although the VGG19 model was trained for image classification, because of its top-down representability, it has also been adapted to many image generation tasks. Given an input image $\mathbf{x} \in \mathbb{R}^{W \times H}$, let $\mathbf{F}_{l(\mathbf{x})} \in \mathbb{R}^{W_l \times H_l \times N_l}$ be the set of N_l feature maps at the l^{th} layer. For simplicity's sake, we ignore the color channel here and reshape $\mathbf{F}_{l(\mathbf{x})}$ as a $W_l H_l \times N_l$ matrix in the following content. Gatys et al. proposed to represent the content feature by the upper layers' response $\mathbf{F}_{l(\mathbf{x})}$ and represent the style feature by the Gram matrix $\mathbf{G}_{l(\mathbf{x})} = \mathbf{F}_{l(\mathbf{x})}{}^{\mathrm{T}} \mathbf{F}_{l(\mathbf{x})}$ at multiple layers. The desired image was obtained by

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\min_{\mathbf{x}} \lambda \mathrm{f}_{\text{content}}(\mathbf{x}) + \mathrm{f}_{\text{style}}(\mathbf{x}) + \gamma\Gamma(\mathbf{x}) \\ &= \arg\min_{\mathbf{x}} \lambda \textstyle\sum_{l\in l_{\mathbf{c}}} \left\| \mathbf{F}_{l(\mathbf{x})} - \mathbf{F}_{l(\mathbf{c})} \right\|^2 + \sum_{l\in l_{\mathbf{s}}} \left\| \mathbf{G}_{l(\mathbf{x})} - \mathbf{G}_{l(\mathbf{s})} \right\|^2 + \gamma\Gamma(\mathbf{x}),\end{aligned} \tag{1}$$

where $\Gamma(\mathbf{x})$ denotes any regularization term to enforce smoothness constraints between neighbouring pixels. With a reasonably initialized $\mathbf{x}$, Equation (1) can be minimized by gradient descent with back-propagation to generate a style-transfer output.

The impressive results in [12] show that the success of this method mainly benefits from the style features $\mathbf{G}_{l(\mathbf{x})}$, which successfully (i) encodes cross-feature dependencies and (ii) captures robust statistics by aggregating pixel-wise responses globally. Recently, this bilinear feature (i.e., Gram matrix) has also been proven to achieve state-of-the-art accuracy in texture recognition [16]. Despite the success in capturing global statistics, the style feature $\mathbf{G}_{l(\mathbf{x})}$ tends to produce visible artifacts in synthesized images because it only matches the global style without imposing any spatial layout constraints.

To include spatial layout constraints, Li et al. [18] proposed a patch-based MRF prior and defined the style loss function in terms of a new style feature $\Phi_i(\mathbf{F}_{l(\mathbf{x})})$ by

$$\mathrm{f}_{\text{style}}(\mathbf{x}) = \textstyle\sum_{l\in l_{\mathbf{s}}} \sum_i \left\| \Phi_i(\mathbf{F}_{l(\mathbf{x})}) - \Phi_{\mathrm{NN}(i)}(\mathbf{F}_{l(\mathbf{s})}) \right\|^2, \tag{2}$$

where $\Phi_i(\cdot)$ denotes a sampling function of the i^{th} local patch, and $\Phi_{\mathrm{NN}(i)}(\mathbf{F}_{l(\mathbf{s})})$ denotes the most similar patch of $\Phi_i(\mathbf{F}_{l(\mathbf{x})})$ matched in $\mathbf{F}_{l(\mathbf{s})}$ by nearest neighbor search. Although including the patch-based MRF prior improves the content-aware synthesis at the expense of extra complexity, this method, like most patch-based synthesis, does not always yield plausible results unless $\mathbf{c}$ can be well-reconstructed by the MRF prior defined by $\mathbf{s}$. Therefore, Equation (2) is less adaptive to the content image in comparison with Equation (1). In addition, the quality of style transfer now depends on the quality of patch matching, which would inevitably lose large-scale feature statistics because of the limited patch size.

In [19], Champandard noticed that the MRF prior implemented by naïve patch matching may distort original contents due to mismatched patches; the author then proposed to use semantic maps to guide nearest neighbour search. With the aid of semantic maps, their method is able to find similar patches from semantic-aware regions. However, the rendered images shown in [19] usually looks oversmoothed under the strong MRF prior.

2.2 Generalized Style Transfer

Inspired by the idea of semantic-guided transfer [19], we propose to develop a generalized style transfer method which involves spatial constraint in the transfer process. Unlike [18] and [19], we mean to formulate the style transfer based on Equation (1) under two additional constraints: *where to transfer* and *what to transfer*.

We first clarify that the style features $\mathbf{G}_{l(\mathbf{x})}$ [12] capture global feature statistics through:

i. Inner products of cross-feature correlation to aggregate information across the whole image plane (or feature map); and
ii. Back-propagation of the gradients in each layer to every pixel of the previous layer.

Therefore, to constrain *where to transfer*, we propose a masking out process to specify the spatial correspondence during the cross-feature aggregation and the layerwise back-propagation. To constrain *what to transfer*, we propose new high-order feature statistics to better capture and match the style representation. Below we detail how to unify both the two constraints by a general formulation.

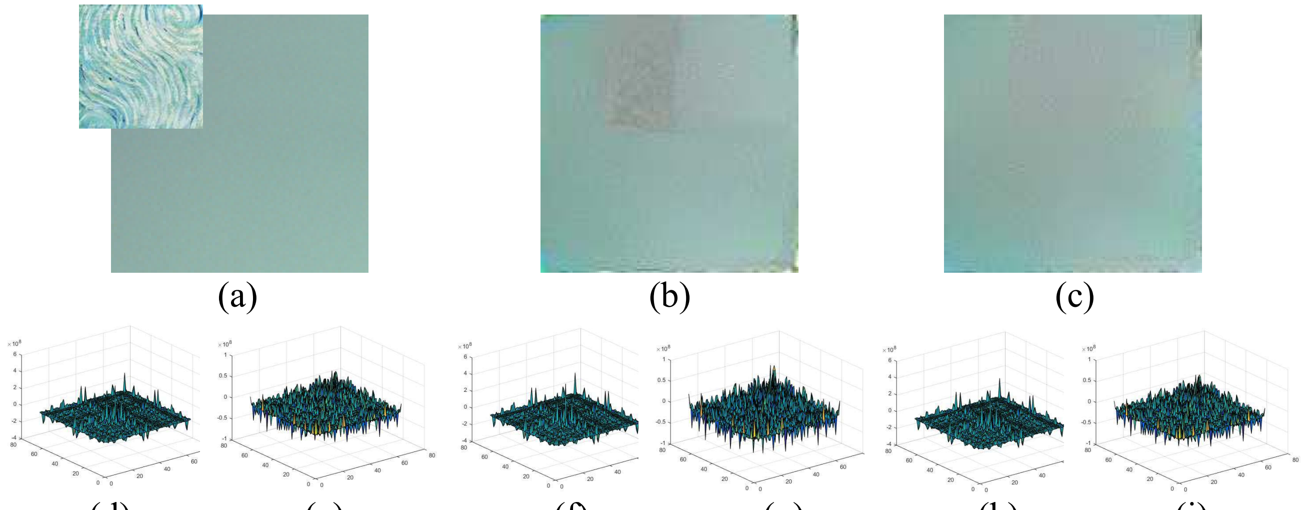

Fig. 2. Given in (a), a desired style image (a flat background) and its initial content (left corner), we show the transferred results without and with high-order feature statistics in (b) and (c), respectively. The zero-order and high-order Gram matrices at conv1_1 layer of (a)-(c) are given in (d)-(e), (f)-(g), and (h)-(i) for visualization.

A. *Spatially Constrained Transfer by Masking Out*

Assume a pair of predefined masks (e.g., semantic maps) is given at the l^{th} layer, we introduce a *masking out* process into the original formulation to constrain the spatial correspondence between source and target images. Let $\mathbf{M}_{l(\mathbf{x})} \in \mathbb{R}^{W_l H_l \times W_l H_l}$ be a diagonal matrix whose $(i,i)^{\text{th}}$ entry m_i $(0 \leq m_i \leq 1)$ is a soft indictor of aggregation; for example, $m_i = 1$ indicates that the i^{th} pixel of $\mathbf{F}_{l(\mathbf{x})}$ should be fully aggregated in the cross-feature dependency and then back-propagated to previous layers, and $m_i = 0$ implies that the i^{th} pixel should be filtered out from the transfer process. We then define the style loss by:

$$\mathrm{f}_{\text{style}}(\mathbf{x}) = \sum_{l\in l_{\mathbf{s}}} \left\| \widehat{\mathbf{G}}_{l(\mathbf{x})} - \widehat{\mathbf{G}}_{l(\mathbf{s})}/N_l \right\|^2, \tag{3}$$

where $\widehat{\mathbf{G}}_{l(\mathbf{x})} = (\mathbf{M}_{l(\mathbf{x})}\mathbf{F}_{l(\mathbf{x})})^{\mathrm{T}}(\mathbf{M}_{l(\mathbf{x})}\mathbf{F}_{l(\mathbf{x})})$, and N_l is a normalization factor to balance the effective fields of the two masks:

$$N_l = \mathrm{trace}(\mathbf{M}_{l(\mathbf{s})}^{\mathrm{T}}\mathbf{M}_{l(\mathbf{s})})/\mathrm{trace}(\mathbf{M}_{l(\mathbf{x})}^{\mathrm{T}}\mathbf{M}_{l(\mathbf{x})}). \tag{4}$$

In Equation (3), the correspondence between two image styles are now associated with their predefined layerwise masks $\mathbf{M}_{l(\mathbf{x})}$ and $\mathbf{M}_{l(\mathbf{s})}$. Note that $\widehat{\mathbf{G}}_{l(\mathbf{s})}$ can be pre-calculated and only the pixelwise product between $\mathbf{F}_{l(\mathbf{x})}$ and $\mathbf{M}_{l(\mathbf{x})}$ is additionally required during optimization. Accordingly, we derive the layerwise gradient of style loss:

$$\nabla_{\mathbf{F}_{l(\mathbf{x})}} = \mathbf{M}_{l(\mathbf{x})}^{\mathrm{T}}(\mathbf{M}_{l(\mathbf{x})}\mathbf{F}_{l(\mathbf{x})})(\widehat{\mathbf{G}}_{l(\mathbf{x})} - \widehat{\mathbf{G}}_{l(\mathbf{s})}/N_l). \tag{5}$$

The masking out term $\mathbf{M}_{l(\mathbf{x})}$ in Equation (5) provides a simple way to guide *where to transfer* with the aggregated feature statistics $\widehat{\mathbf{G}}_{l(\mathbf{x})}$.

Equation (5) can be further extended to general scenarios when there are$K$ region correspondences between images $\mathbf{c}$ and $\mathbf{s}$. In this case, each region has its own semantic label, either an object class or a scene segment. Hence we have $2K$ binary masks at the original image resolution $W \times H$. To apply the masking out process, we downsample the binary masks to $W_l \times H_l$ by bilinear interpolation at the l^{th} layer and aggregate the K objective costs of style matching. The value of layer-wise masks thus ranges from 0 to 1 and blends the rendering effects of style transfer at object boundaries. Consequently, we derive the layerwise gradient of style feature:

$$\nabla_{\mathbf{F}_{l(\mathbf{x})}} = \sum_{k=1}^{K} {\mathbf{M}_{l(\mathbf{x})}^{(k)}}^{\mathrm{T}} \left(\mathbf{M}_{l(\mathbf{x})}^{(k)} \mathbf{F}_{l(\mathbf{x})}\right)\left(\widehat{\mathbf{G}}_{l(\mathbf{x})}^{(k)} - \widehat{\mathbf{G}}_{l(\mathbf{s})}^{(k)}/N_l^{(k)}\right), \text{ where}$$
$$\widehat{\mathbf{G}}_{l(\mathbf{x})}^{(k)} = (\mathbf{M}_{l(\mathbf{x})}^{(k)} \mathbf{F}_{l(\mathbf{x})})^{\mathrm{T}} (\mathbf{M}_{l(\mathbf{x})}^{(k)} \mathbf{F}_{l(\mathbf{x})}). \quad (6)$$

B. *High-order Statistics of Style Feature*

Although the proposed masking out process effectively constrains the transfer layout, we may still obtain implausible results if the source and target images are extremely different. We use Fig. 2 to demonstrate the limitation of the original style features. In Fig. 2 (a), we attempt to smooth the texture by transferring the style from a flat background. Undesired patterns (Fig. 2(b)) are shown after the style transfer, although both the style features look very similar (Fig. 2(d) and (f)). This example shows that cross-correlation of feature maps alone is far from enough to match the style of images.

Rather than using zero-order feature maps $\mathbf{F}_{l(\mathbf{x})}$ only, we resort to high-order feature statistics to characterize consistent style representation. We propose a new style feature $\widehat{\mathbf{G}}_{l(\mathbf{x})} = (\mathbf{P}_{l(\mathbf{x})} \mathbf{F}_{l(\mathbf{x})})^{\mathrm{T}} (\mathbf{P}_{l(\mathbf{x})} \mathbf{F}_{l(\mathbf{x})})$ by introducing a convolutional matrix $\mathbf{P}_l$, where $\mathbf{P}_{l(\mathbf{x})}$ denotes a Toeplitz matrix constructed by high-order filter coefficients. When $\mathbf{P}_l$ is an identity matrix, the new style features is the same as the original representation. In the following experiments, we use the Laplacian of Gaussian (LoG) filter to construct high-order $\mathbf{P}_l$ on account of its noise-resilient property. As shown in Fig. 2(c), a consistent and more plausible result is obtained when including the proposed high-order statistics; the high-order Gram matrices also shows striking similarity between Fig. 2 (e) and (i) than that between Fig. 2 (e) and (g).

Moreover, we can represent the proposed high-order style features together with the masking out process into one single formulation. Assuming that we utilize J filters to characterize feature statistics, we combine the new style features into Equation (6) and derive the layerwise gradient in a general form:

$$\nabla_{\mathbf{F}_{l(\mathbf{x})}} = \sum_{j=1}^{J} \sum_{k=1}^{K} {\mathbf{P}_l^{(j)}}^{\mathrm{T}} {\mathbf{M}_{l(\mathbf{x})}^{(k)}}^{\mathrm{T}} \left(\mathbf{M}_{l(\mathbf{x})}^{(k)} \mathbf{P}_l^{(j)} \mathbf{F}_{l(\mathbf{x})}\right)\left(\widehat{\mathbf{G}}_{l(\mathbf{x})}^{(k)} - \widehat{\mathbf{G}}_{l(\mathbf{s})}^{(k)}/N_l^{(k)}\right), \text{ where}$$
$$\widehat{\mathbf{G}}_{l(\mathbf{x})}^{(k)} = (\mathbf{M}_{l(\mathbf{x})}^{(k)} \mathbf{P}_l^{(j)} \mathbf{F}_{l(\mathbf{x})})^{\mathrm{T}} (\mathbf{M}_{l(\mathbf{x})}^{(k)} \mathbf{P}_l^{(j)} \mathbf{F}_{l(\mathbf{x})}). \quad (7)$$

Equation (7) formally describes our method.

2.3 Extension to Image Inpainting

The proposed masking out process is essentially one of the *layout priors* in many other image editing problems. For example, we may consider image inpainting as transferring image features (either content or style) from the surrounding regions into the target region. Most state-of-the-art inpainting approaches are based on patch-synthesis framework [21, 22]. In this section, we aim to investigate whether our method can generalize to this application as well.

In image inpainting, the single input image serves as both the content and style images; for example, we can assign the style masks of $\mathbf{F}_{l(\mathbf{x})}$ and $\mathbf{F}_{l(\mathbf{s})}$ to be pixels inside and outside the target region, respectively. However, before we recast the inpainting problem as style transfer, we need to address three major differences between the two problems. First, because there is no layout constraint (i.e., to exclude the target region) on the content image, the inpainting result tends to be random texture. Second, only the pixels inside the target region should be updated during the optimization process, whereas the other pixels should not be changed. Third, the surrounding pixels which are closer to the target region should contribute more in the inpainting process. Considering these issues, we propose three modifications to generalize the proposed method to image inpainting:

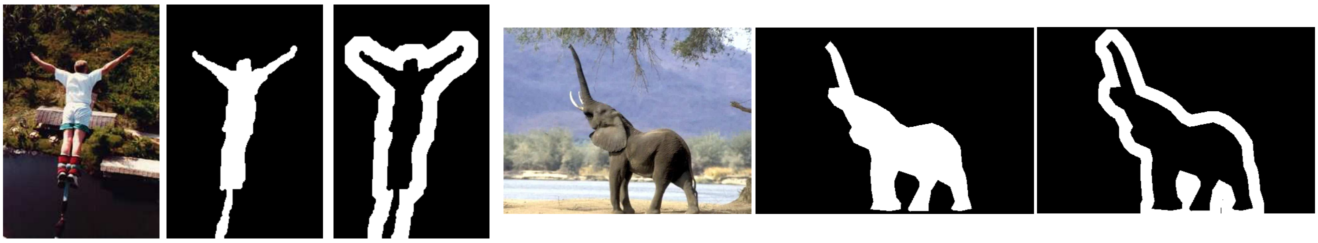

Fig. 3. Two images *bungee* and *elephant* as well as their content masks (missing pixels) and style masks (reference pixels) used for our image inpainting.

i. We use existing inpainting baseline method to initialize the content image, and update the content features every twenty iterations.

ii. We additionally mask out the gradient derived in the original image space using the content mask smoothed by a 7x7 average filter.

iii. As shown in Fig. 3, we restrict the style mask to the surrounding boundary of target region with twenty-pixel width.

3 Experiments

We first detail our parameter settings. Following [12], we use one single layer (conv4_2) and five layers (conv1_1, conv2_1, conv3_1, conv4_1, conv5_1) to characterize content and style features, respectively. Total variation (TV) with quadratic penalty is used for the regularization function $\Gamma(\mathbf{x})$ as widely used in image generation tasks [10]. In all the experiments, we set (λ, γ) as $(0.02, 5 \times 10^{-8})$ for style transfer and $(0.2, 5 \times 10^{-6})$ for image inpainting. For optimization, we minimize Equation (3) by Adam [23] with learning rate equal to two and stop Adam until 100 iterations. We implement the proposed algorithm using the public matlab library *MatConvNet* [24]. Without GPU support, a 230x300 image requires about eight minutes to conduct the style transfer.

3.1 Semantic-Aware Style Transfer

We use two images *Gogh* and *Seth* released by [19] for style transfer. Note that both images share similar semantic contents (portrait) but contain very different artistic styles (painting vs. photo). If no spatial layout is constrained, undesired background features could be transferred to foreground objects and vice versa.

We show the results in Fig. 4 and also compare our algorithm with the original style transfer method [12] and the one combining MRF prior and semantic maps [19]. We implement the method of [12] as a special case of our algorithm. On the other hand, the results of [19] as well as the two semantic maps are both downloaded from the author's project website. As shown in Fig. 4 (e) and (i), [12] propagates the style features globally across the whole image. From Fig. 4 (f) and (j), although [19] preserved semantic correspondence between the input images, the rendered results look oversmoothed under the strong MRF prior. In contrast, by leveraging the proposed masking out process, our method achieves content-aware style transfer in accordance with the semantic mapping.

Moreover, the undesired background patterns (in Fig. 4(g)), which result from utilizing zero-order style features only, are readily removed once we include the high-order statistics as style features (Fig. 4(h)). Note that how the blurry eyes of *Gogh* in Fig. 4(g) become sharper in Fig. 4(h) and how the semantic structures are better characterised. Another example for real-life photo transfer is shown in Fig. 1. Using the semantic masks

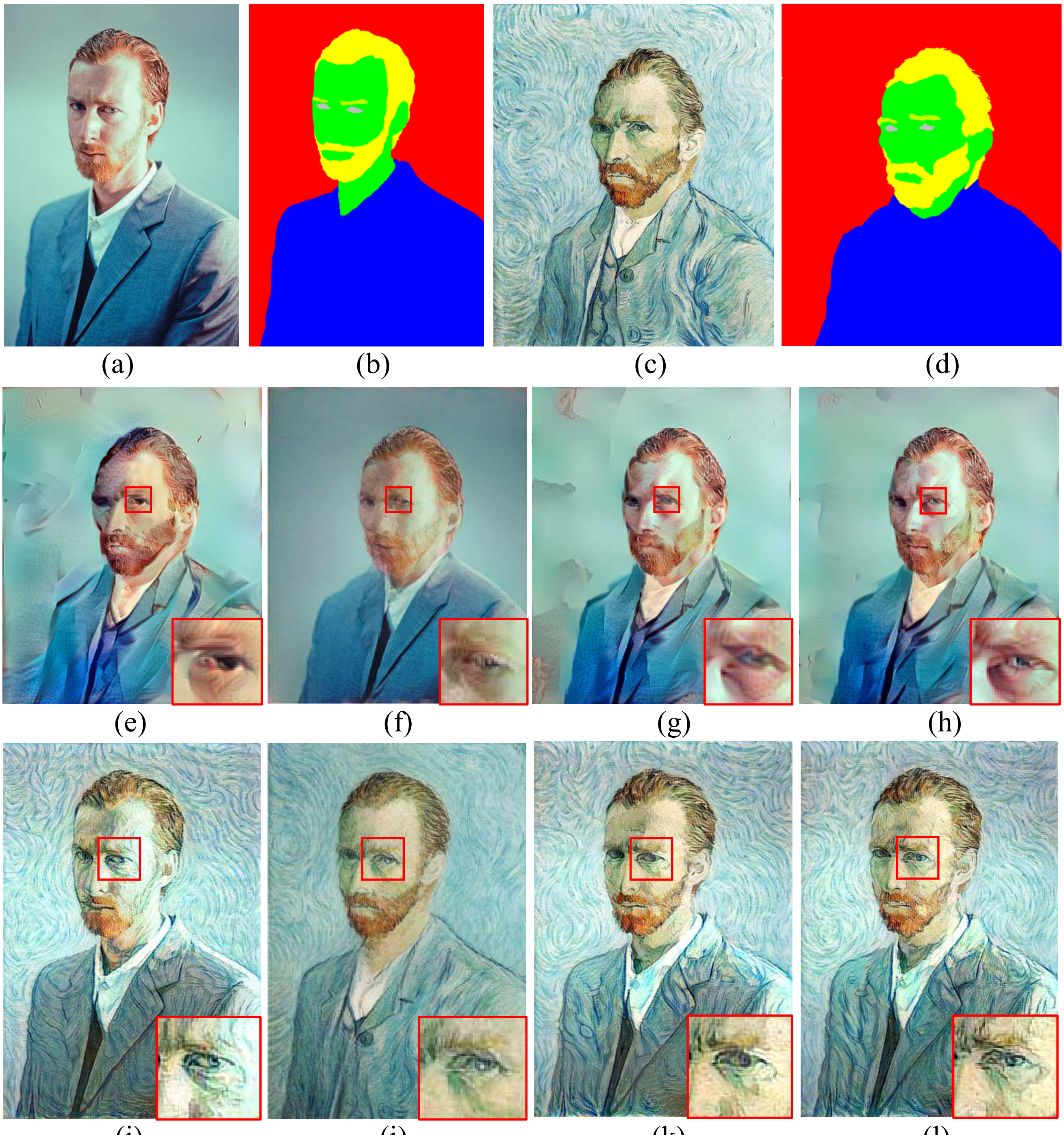

Fig. 4. Semantic-aware style transfer, where (a)-(d) shows the source images and the corresponding semantic maps. We display the transferred results in (e)-(h) and (i)-(l), which are obtained by [12], [19], and our method without and with high-order feature statistics, respectively. See the enlarged regions for detailed comparison.

estimated by image matting [25], our method successfully transfers the dogs' appearance without noticeable artifacts.

3.2 Saliency-Aware Style Transfer

To investigate the significance of salient features in style transfer, we study to include a saliency-aware mask into the proposed method. Recent advance in deep learning-based vision tasks have shown that larger responses at upper convolutional layers relate to salient objects. This property has been utilized for fast object classification and object detection [26]. We therefore propose to determine the saliency-aware mask by preserving the top $p\%$

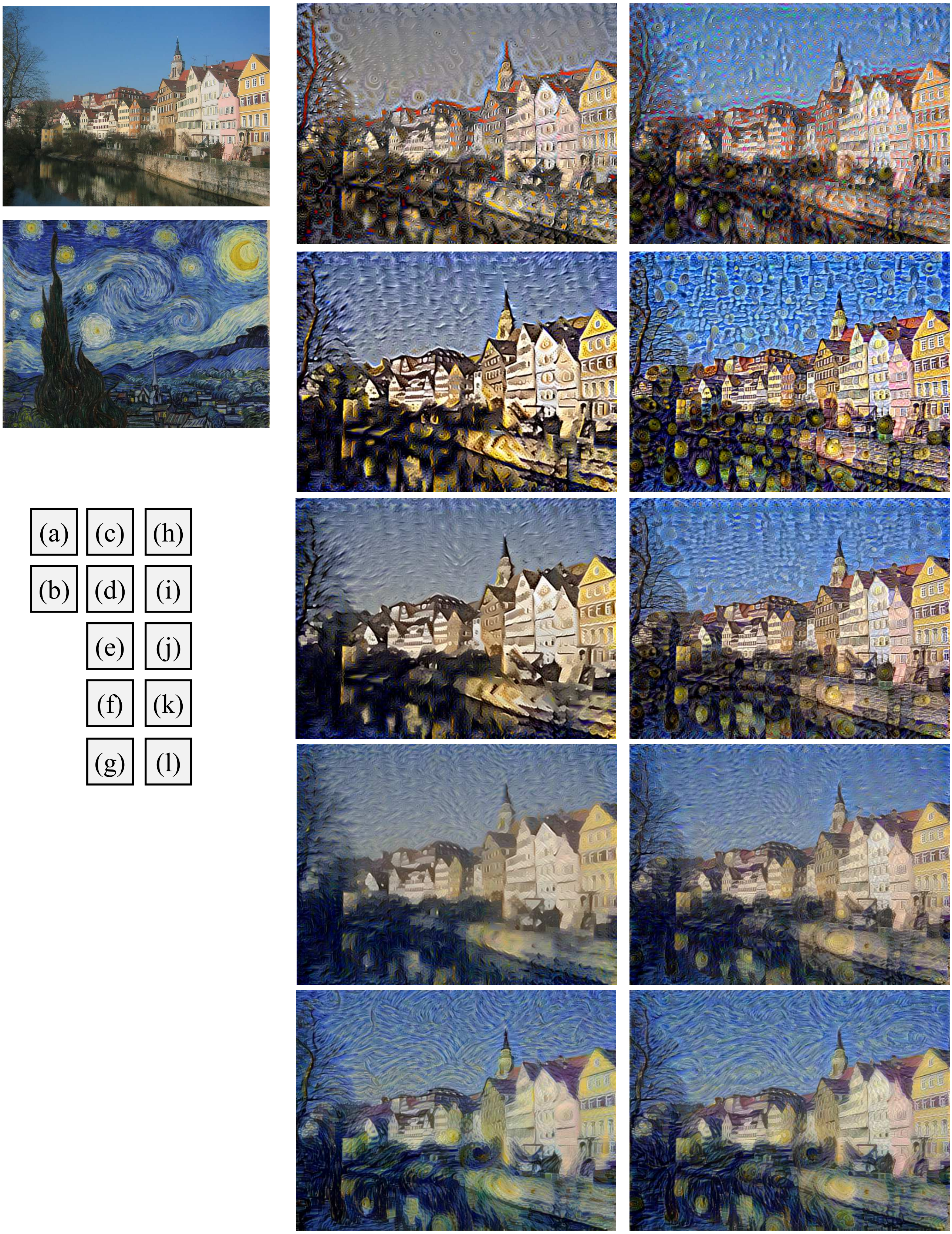

Fig. 5. Saliency-aware style transfer, where (a) and (b) shows the content and style images. We display the transferred results without and with high-order feature statistics in (c)-(g) and (h)-(l), when keeping 0.1%, 1%, 10%, 50%, 100% salient feature response from top to bottom, respectively.

feature responses in terms of the absolute values at each layer and eliminating all the other responses. Before passing this mask to the masking out process, we conduct Gaussian smoothing on the saliency-aware mask, which results in a soft binarized mask with values ranging from 0 to 1, to suppress noisy feature responses.

Fig. 5 shows the results of saliency-aware style transfer. Because our goal here is to analyze what salient features are transferred, we only constrain the style image with salient-aware mask but allow the style features to transfer to the whole resultant image. In Fig. 5, we keep 0.1%, 1%, 10%, 50%, and 100% feature responses with or without using high-order feature statistics for comparison. As shown in Fig. 5(c)-(g), the zero-order style features first characterize color tone, then distort textures and edges, and finally render high-level patterns. When including high-order style features, we observe that textures and small patterns are transferred in priori, while color tone changes smoothly along with the increased percentage of feature responses (see Fig. 5(h)-(l)). The rendered results empirically demonstrate what the so-called *visual style* is composed of. In addition, this example also illustrates how different features reveal different visual saliency in a photo or an artistic painting.

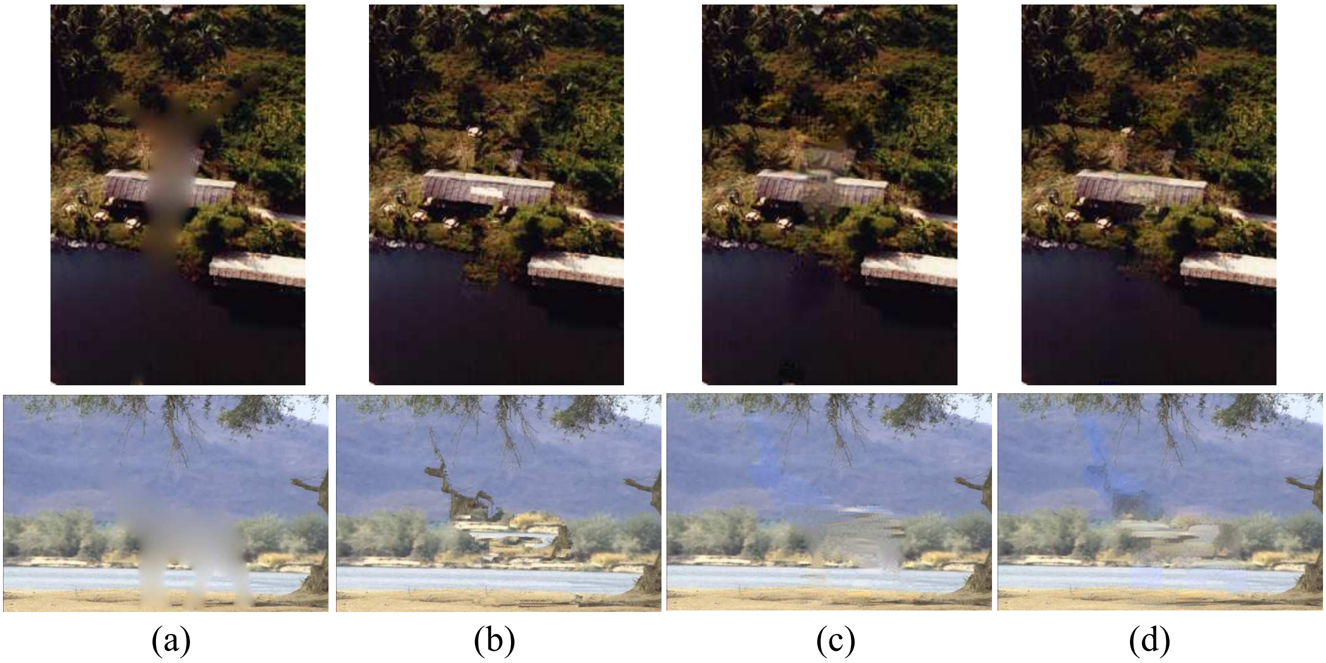

Fig. 6. Extension to image inpainting. The results are obtained by (a) [27], (b) [28], (c) [27] + our method, and (d) [28] + our method.

3.3 Image Inpainting by Style Transfer

We use two images *bungee* and *elephant* (see Fig. 3) to validate our method. We apply two baseline methods, PDE-based inpainting [27] and exemplar-based inpainting [28], to initialize the target region of content image. As shown in Fig. 6(a)-(b), exemplar-based method obtains plausible region filling but generates visible artifacts due to imperfect filling priority; by contrast, PDE-based method fails to recover the large hole but simply diffuses pixels from region boundary. Nevertheless, after conducting the proposed style transfer, our method characterizes the underlying scene structure more accurately. Note that how our method unveils the rock in Fig. 6(c) and removes the grass in Fig. 6(d) in comparison with Fig. 6(a) and Fig. 6(b) for the image *bungee*. We attribute the success in structure transfer to the hierarchical scene understanding inherited from the VGG19 model.

Note that we do not claim that the proposed framework outperforms state-of-the-art image inpainting methods; in fact, as shown in Fig. 6(c)-(d), there remain some artificial patterns produced during the optimization process. Instead, we expect the preliminary results can stimulate new thinking for image editing problems.

4 Conclusion

In this paper, we propose to generalize the recently impressive work *style transfer* [12] and focus on addressing *content-aware style transfer*. Instead of combining any complex technique (e.g., MRF prior) with extra costs, we seamlessly embed two key components, *masking out process* and *high-order feature statistics*, into the original formulation and show that this unified formulation enables different variations of content-aware style transfer. The proposed masking out process is adaptive to different applications, objectives, and user constraints. In addition, transferring high-order feature statistics significantly improves content coherency as well as subjective quality. The experiments demonstrate that our method is widely applicable to various style transfers and show its potential in image inpainting. The preliminary results suggest some interesting directions. For future work, we plan to investigate the relation between: 1) the salient feature maps and the semantic content mapping so as to automate semantic-aware style transfer; and 2) the choice of high-order filter and the subjective image quality so as to improve the visual plausibility.

References

[1] S. Zhu, Y. Wu, and D. Mumford. Filters, Random fields and maximum entropy (FRAME): towards a unified theory for texture modeling. *IJCV*, 27(2): 107-126, 1998.

[2] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *IJCV*, 40(1): 49-70, 2000.

[3] J. B. Tenenbaum, W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6): 1247-1283, 2000.

[4] A. Hertzmann, C. Jacobs, N. Oliver, B. Curless, and D. Salesin. Image analogies. In *Proc. SIGGRAPH*, 2001.

[5] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proc. SIGGRAPH*, 2001.

[6] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra. Texture optimization for example-based synthesis. In *Proc. SIGGRAPH*, 2005.

[7] W. Zhang, C. Cao, S. Chen, J. Liu, and X. Tang. Style transfer via image component analysis. *IEEE Trans. Multimedia*, 15(7): 1594-1601, 2013.

[8] Deep Dream. http://deepdreamgenerator.com/.

[9] A. Dosovitskly, J. T. Springenberg, M. Tatarchenko, and T. Brox. Learning to generate chairs, tables, and cars with convolutional networks. In *Proc. CVPR*, 2015.

[10] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proc. CVPR*, 2015.

[11] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *arXiv preprint arXiv*: 1511.06434, 2015.

[12] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, 2016.

[13] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: feed-forward synthesis of texture and stylized images. In *arXiv preprint arXiv*: 1603.03417, 2016.

[14] J. Johnson, A. Alahi, and F.-F. Li. Perceptual losses for real-time style transfer and super-resolution. In *arXiv preprint arXiv*: 1603.08155, 2016.

[15] J. R. Gardner, P. Upchurch, M. J. Kusner, Y. Li, K. Q. Weinberger, K. Bala, and J. E. Hopcroft. Deep manifold traversal: changing labels with convolutional features. In *arXiv preprint arXiv*: 1511.06421, 2015.

[16] T.-Y. Lin and S. Maji. Visualizing and understanding deep texture representations. In *Proc. CVPR*, 2016.

[17] R. Yin. Context aware neural style transfer. In *arXiv preprint arXiv*: 1601.04568, 2016.

[18] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proc. CVPR*, 2016.

[19] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine networks. In *arXiv preprint arXiv*: 1603.01768, 2016.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv*: 1409.1556, 2014.

[21] N. Komodakis and G. Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Trans. Image Processing*, 16(11): 2649-2661, 2007.

[22] K. He and J. Sun. Statistics of patch offsets for image completion. In *Proc. ECCV*, 2012.

[23] D. Kingma and J. Ba. Adam: a method for stochastic optimization. In *arXiv preprint arXiv*: 1412.6980, 2014.

[24] MatConvNet. http://www.vlfeat.org/matconvnet/.

[25] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(2): 228-242, 2008.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proc. NIPS*, 2015.

[27] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proc. SIGGRAPH*, 2000.

[28] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Processing*, 13(9): 1200-1212, 2004.