IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 31, 2018, accepted September 25, 2018, date of publication October 8, 2018, date of current version October 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2874203

Exemplar-Based Portrait Style Transfer

MING LU[1], FENG XU[2], HAO ZHAO[1], ANBANG YAO[3], YURONG CHEN[3], AND LI ZHANG[1]
[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]School of Software, Tsinghua University, Beijing 100084, China
[3]Cognitive Computing Laboratory, Intel Labs China, Beijing 100080, China

Corresponding author: Li Zhang (chinazhangli@mail.tsinghua.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant U1533132, Grant 61871248, Grant 61822111, Grant 61727808, and Grant 61671268.

ABSTRACT Transferring the style of an example image to a content image opens the door of artistic creation for end users. However, it is especially challenging for portrait photos since human vision system is sensitive to the slight artifacts on portraits. Previous methods use facial landmarks to densely align the content face with the style face to reduce the artifacts. However, they can only handle the facial region. As for the whole image, building the dense correspondence is difficult and may easily introduce errors. In this paper, we propose a robust approach for portrait style transfer that gets rid of dense correspondence. Our approach is based on the guided image synthesis framework. We propose three novel guidance maps for the synthesis process. Contrary to former methods, these maps do not require the dense correspondence between content image and style image, which allows our method to handle the whole portrait photo instead of facial region only. In comparison with recent neural style transfer methods, our method achieves more pleasing results and preserves more texture details. Extensive experiments demonstrate our advantage over former methods on portrait style transfer.

INDEX TERMS Portrait style transfer, semantic segmentation, patch match, texture synthesis.

I. INTRODUCTION

Style transfer is a hot topic in image processing. It aims to transfer the artistic style of an example painting to a content image. Traditionally, creating artistic styles requires tedious manual works of skilled artists. On the other hand, achieving style transfer automatically or with minimum user interaction enables the end users to generate visually pleasing effects. Thus style transfer draws close attentions from both industry and academia. In industry, the popular app *prisma* transfers attractive effects to users' daily photos. In academia, with the development of deep learning, style transfer techniques based on deep neural network (DNN) [1], [2], [6] achieve appealing results.

Portraits take a large proportion of daily photos and portrait style transfer is strongly demanded. However, it is a more difficult task in comparison with general image stylization since human vision system is especially sensitive to the visual quality of portraits. Previous works [3], [4], [29] can handle the facial region by detecting reliable landmarks in the face. They build the facial correspondence to align the content face with the style face in order to reduce the artifacts. However, these methods can not handle the full image since their methods based on facial landmarks can not build the dense correspondence between the two images. Besides, precise alignment may not even exist when the contents of the input and example images are quite different (e.g. long hair vs. short hair).

Although plenty of works on neural style transfer have been proposed, they can not robustly handle portrait. Global style transfer methods [1], [2] only transfer the overall style to the content image, which will cause obvious artifacts for portrait. References [6], [11] combine Markov Random Field (MRF) with deep neural features to regularize the texture synthesis and achieve more plausible results compared with [1] and [2]. However, they can not achieve consistently pleasing results for portrait. Reference [28] uses deep neural features to find the dense correspondence and reconstruct the stylized images. Although it achieves amazing bidirectional results, it still often fails for portrait since dense correspondence without any guidance is difficult even using deep neural features. References [27] and [30]—[32] allow the user to provide the semantic mask as a spatial guidance to improve the visual effect. However, these methods transfer the style in feature space and fail to preserve the texture details of the style image.

This paper proposes a novel method for portrait style transfer that gets rid of the dense alignment step. Our method is based on the guided image synthesis framework [33].

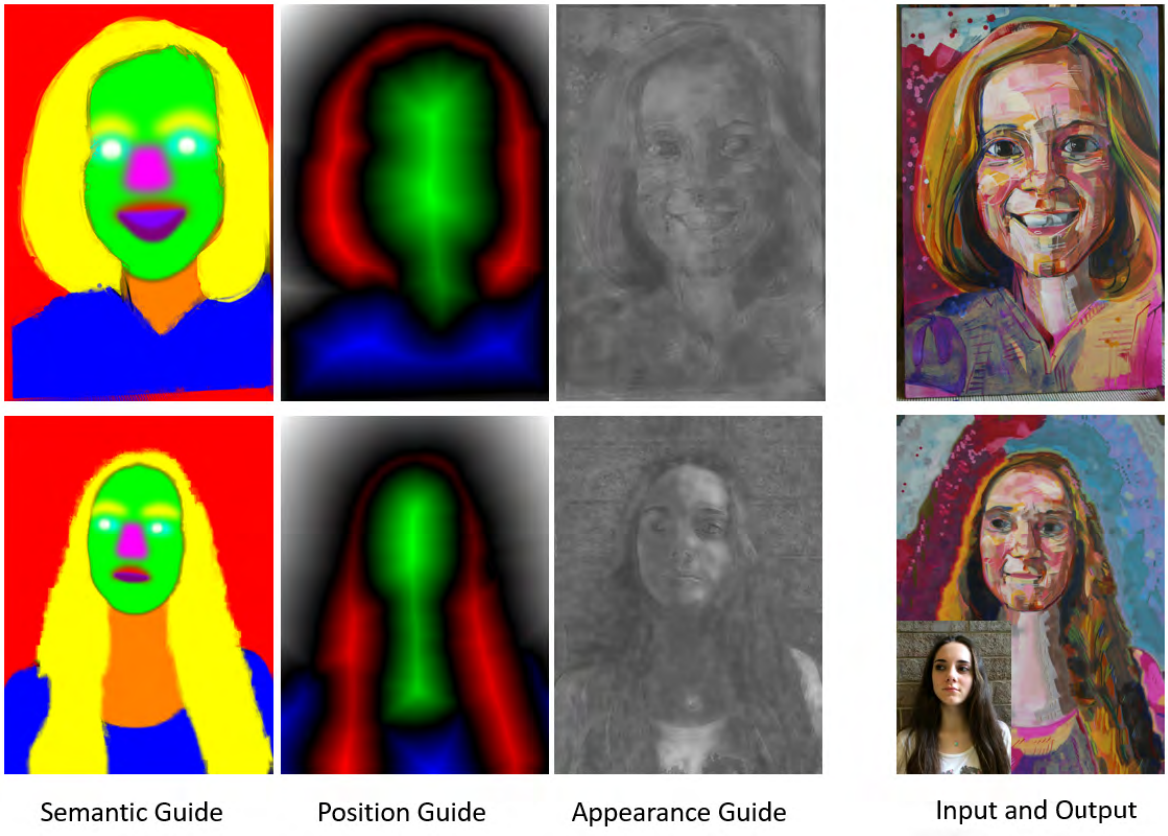


FIGURE 1. The pipeline of our method. We first construct the semantic, position and appearance maps from the style image and the content image. In this figure, the bottom row is the maps of content image and the top row is the maps of style image. The semantic map builds a loose correspondence between the input images. The position map further characterizes the pixels in the same region. The appearance map removes the style from the input image and provides the appearance guidance. None of these guidance maps requires dense correspondence and the final style transfer effect is achieved by the guided image synthesis process with these maps.

Given a content image and a style image, we propose three guidance maps to regularize the image synthesis process. Unlike former work [29], our guidance maps are not restricted to the facial region, thus we can stylize the whole image instead of facial region only.

We first use semantic map to build a loose correspondence since style is highly related to its semantic meaning. We collect a large dataset for portrait parsing and train a deep neural network to obtain a coarse semantic map. This coarse semantic map is further refined by the detected facial landmarks to deliver the final semantic map. We also propose a position map to characterize the pixels in the same semantic region. Representing a pixel's location in an irregular region is a hard problem. In this work, we calculate the distance between a specific pixel and its nearest pixel in the boundary. We then normalize the distance and use it as the position guidance for the pixels in this region. Besides semantic and position maps, we propose a method to separate the content and style from an image and use the content part as the appearance map. Former work [29] generates the appearance map for the facial region by matching the global intensity levels and local contrast values with the method proposed by [4]. Thus [29] needs to align the content face and style face. While, our method directly separates the content and style of an image. Therefore, we do not require any alignment when generating the appearance map. An illustration of the proposed guidance maps is shown in Figure 1.

The contributions of this paper are summarized as follows:

- First, we propose a robust method for portrait style transfer. Our method achieves more appealing results compared with recent neural style transfer methods.
- Second, we propose three novel guidance maps for portrait image synthesis. Compared with former works, these maps do not rely on the dense correspondence.
- Third, as an additional contribution, we construct a high-quality dataset for portrait image parsing, which we believe is a good supplementary to current face parsing datasets. We will release the dataset to facilitate future works.

II. RELATED WORK

A. STYLE TRANSFER VIA DNN

Recently, style transfer has been revisited by DNN-related techniques. Reference [5] uses the VGG-19 network to extract features of the input and example images. Then the

content loss and the style loss are formulated and used for reconstructing the final stylized result. Reference [6] extends this technique by using a local representation of the style to replace the original global one. Thus local style patterns of the example image can be well preserved. Reference [7] further investigates the DNN-based style transfer by exploring different nets, different initializations and different layers. Reference [8] proposes to control the perceptual factors such as color in order to improve the style transfer quality. However, these works are slow as they need to solve the optimization problem by back propagation to reconstruct the output image.

Fast DNN-based style transfer techniques are achieved by training convolution neural networks, which infer the final image directly from the input image [2], [10], [11]. In these techniques, the losses, originally used for reconstructing the output image, are used for training a feed-forward transfer network. Thus the time-consuming optimization step is removed. However, these techniques require training a new network for each example image. Reference [12] overcomes this drawback by using conditional instance normalization which makes single transfer network learn multiple styles. Recently, many methods [31], [32], [35], [36] are proposed to solve the arbitrary style transfer problem, however, they can not handle the portrait style transfer problem.

B. IMAGE PROCESSING FOR PORTRAITS

As portrait is a very important category of images and the tolerance of artifacts on portraits is very low, various image processing tasks utilize special techniques to handle portraits. Portrait deblur is achieved by transferring information from reference images [13]. Similarly, make-up can also be transferred from examples to an input image [14], [15]. Attractiveness of human faces can also be increased by removing blemishes from faces [16] or changing the distance of a variety of facial features [17].

Style transfer, the topic of this paper, is also investigated. A series of works are proposed to transfer a real image into a sketch painting [18]–[20]. Reference [21] improves the transfer by using a parsing graph. For handling more vivid painting styles, [22] learns how an artist make a painting from an example pair (a source image and a painting of that image), and the learned information is transferred to input photos to generate corresponding paintings. Reference [23] uses Markov Random Field to generate the painting of an input image from an example. Reference [4] uses Laplacian Pyramid to transfer local statistics from the example portrait to a new one. Recently, inspired by the DNN-based style transfer techniques, [3] uses the features obtained by the VGG network to perform transfer, which demonstrates painting style transfer for portraits for the first time. However, to avoid style transfer across different semantic regions, these techniques require pixel-wise alignment between the example and the input, which causes two major drawbacks. First, the face pose of the example is required to be similar to that of the input. Otherwise, it is impossible to calculate a precise alignment. Second, the alignment takes a lot of computation time which dramatically decreases the performance of the system.

III. METHOD

Our method is based on the image synthesis framework [33]. Compared with recent methods [1], [2] based on deep neural networks, our method uses image patches to synthesize the stylized output. However, image synthesis without guidance can not achieve pleasing results for portraits. Therefore, we introduce three novel guidance maps in Section III-A. After obtaining the guidance maps, we briefly describe the guided image synthesis process in Section III-B.

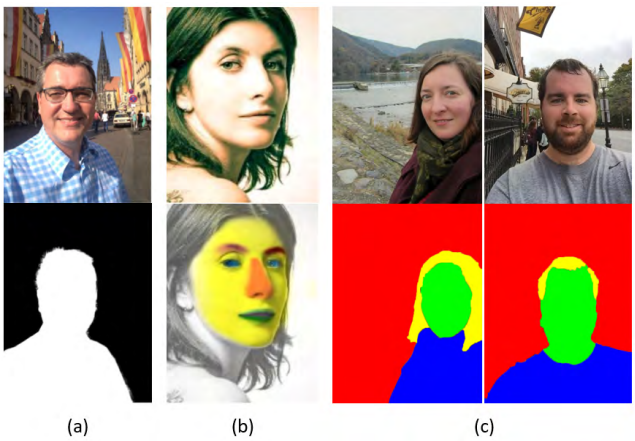


FIGURE 2. Illustration of our portrait dataset. (a) The portrait segmentation dataset introduced by [25]. (b) The face parsing dataset of [37]. (c) Our dataset for portrait parsing is a good supplementary to the portrait segmentation and face parsing datasets.

A. GUIDANCE MAPS

1) SEMANTIC GUIDANCE MAP

Painting style is highly related to its semantic meaning, for example, painters usually use different textures to describe different regions. So it is straightforward to transfer the style within semantically corresponding regions. Compared with pixel-wisely dense correspondence, semantic correspondence is loose, while it can still alleviate common failure cases like applying hair texture to facial skin. In order to obtain the semantic guidance map, we can train a deep neural network for portrait parsing. However, current published datasets are not suitable for this task. Reference [25] introduces a dataset for portrait segmentation, however, this dataset lacks further annotations of useful components like hair. Reference [37] proposes a dataset for face parsing, while it can not parse the portrait photo apart from facial region. In this paper, we contribute a dataset for portrait parsing. As shown in Figure 2, we label four regions (skin, hair, cloth, background) for each portrait photo using Photoshop. The dataset consists of 3800+ portrait photos (1700+ are from [25] and the rest are from the internet). Then we use a deep segmentation technique [24] to train a network for

automatically parsing. We will release the dataset and our trained models for future works on portrait.

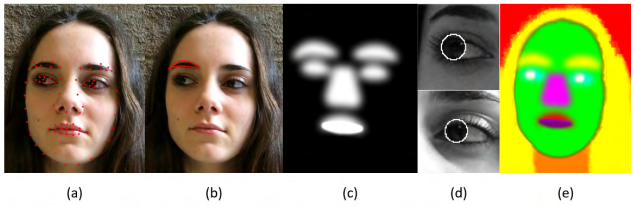


FIGURE 3. Illustration of parsing facial region. (a) The automatically detected facial landmarks. (b) The fitted upper and lower third order polynomial curves for left eyebrow. (c) The marked facial components by our method. (d) The tracked eyeballs. (e) The final composed facial parsing result.

After obtaining the coarse semantic map, we further refine it using the detected landmarks. As illustrated by Figure 3, we first automatically detect several facial landmarks. Then we divide the facial region into eyes, eyebrows, eyeballs, nose, upper lip, middle lip, bottom lip and other region. For the eyes, eyebrows, upper lip, middle lip and bottom lip, we fit two third-degree polynomial curves to the upper and lower bounds of each region. We mark the pixels between the upper curve and lower curve as the valid region. As for the nose region, we find the convex hull of nose landmarks. We further use the method mentioned in [4] to detect the eyeballs. Finally, we smooth the mask of each region to deliver the soft boundary and compose the final semantic guidance map.

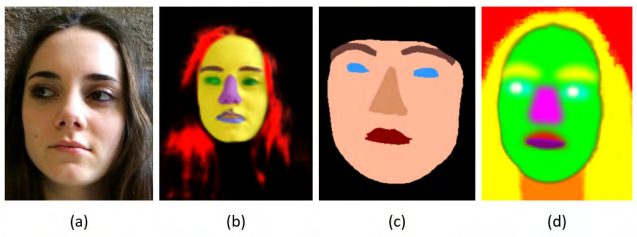


FIGURE 4. Comparison of our parsing method with former works. (a) The input portrait photo. (b) The face parsing result of [38]. (c) The face parsing result of [39]. (d) Our result. The three methods use different palettes to represent different regions in this figure. Our method obtains accurate parsing results compared with [38], [39].

We compare our semantic guidance map with related portrait parsing methods in Figure 4. Reference [38] trains a parsing network with multiple objective losses. However, their method contains obvious errors in some regions like hair. [39] combines landmark detection with face parsing. They jointly train a landmark detection network and a face parsing network. However, this method can only parse the region enclosed by the landmarks. Besides, its result also contains obvious errors. Apart from errors, both the above methods can not handle the full portrait photo, which limits their usages. Our method combines deep neural network with detected facial landmarks and can obtain accurate results for the whole image.

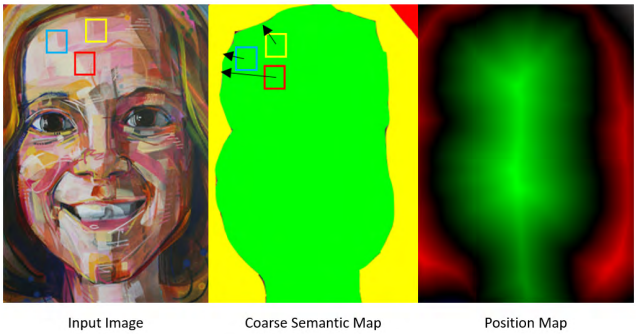


FIGURE 5. Illustration of the position map. (a) Three image patches in the facial region are different even from the appearance. (b) We calculate the distance between the image patch and its nearest pixel in boundary to represent the relative position in the region. (c) We normalize the distance to deliver the position guidance map for each region.

2) POSITION GUIDANCE MAP

Since semantic guidance map only represents the semantic property of each pixel, it can not characterize the pixels in the same region. For example, as shown in Figure 5, the three image patches all belong to the facial region. However, they are actually different even from the perspective of local appearance. Therefore, we need to generate a position guidance map to represent the position difference within the same region. Nevertheless, it is difficult to represent the relative position of a pixel in an irregular region. For some regular regions, such as triangle, we can represent the relative position by its barycentric coordinates. As for irregular regions, we are unable to find stable anchor points to obtain the barycentric coordinates. Therefore, we use distance transformation to represent the relative position. To be specific, for each pixel within a semantic region, we calculate the distance between this pixel and its nearest pixel in the boundary. We then normalize this distance in individual region and use it as the position guidance. We calculate this position guidance based on our coarse semantic map generated by the trained neural network.

3) APPEARANCE GUIDANCE MAP

Besides semantic guidance and position guidance, we also want to preserve the appearance during the guided image synthesis. However, for portrait style transfer, the content image and style image are quite different since they are from two different image domains, namely natural photo and artistic painting. Former method [29] only handles the facial region, thus the content face and style face can be densely aligned. They generate the appearance guidance map by modifying the global intensity levels and local contrast values of content face to match those in the style face. However, we can not densely align the content image with the style image since we need to handle the whole photo instead of facial region only.

Based on the research of neural style transfer, the neural features of different layers from VGG-19 network [40] perform as the encodings of the input image at different levels.

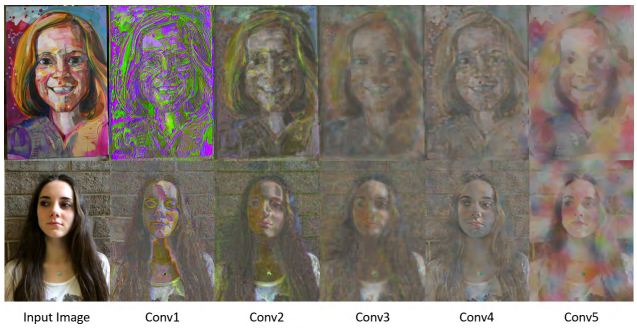


FIGURE 6. Illustration of the appearance maps. The appearance maps aim to preserve the content, thus they should not contain texture and color differences between content image and style image. We use the decoded images of whitened deep neural features as the appearance maps. As can be seen, the generated appearance maps greatly reduce the differences in color and texture.

Reference [1] formulates the style of the input image as the gram matrix of neural features. Recently, [30] proposes a neural style transfer method based on the whitening and coloring of neural features. Inspired by [30], we use the decoded image of whitened features as the appearance guide. We can obtain the appearance guidance from different layers as shown in Figure 6. As can be seen, the original content image and style image are quite different, therefore, they can not be used as the appearance guidance maps. However, after projecting them to the hidden space by whitening their deep neural features. They become more similar and the differences in texture and color are greatly reduced. In our implementation, we use the lightness channels of the projected content image and style image as the appearance maps. Different from [30], we simply use the feature whitening to generate the projected content image and style image, which reduce the texture and color difference for image synthesis. However, [30] directly transfers the style in feature space.

B. GUIDED IMAGE SYNTHESIS

In this section, we describe the guided image synthesis process used to generate the stylized output. We denote the semantic guidance, position guidance and appearance guidance of content image as G^c_{sem}, G^c_{pos} and G^c_{app}. Similarly, the corresponding guidance maps of style image are denoted as G^s_{sem}, G^s_{pos} and G^s_{app}. For each pixel p in content image, we need to find the corresponding pixel q in style image. We denote all pixels in content image as P and $p \in P$. All the corresponding pixels are denoted as Q and $q \in Q$. Q can also be called as nearest neighbor field (NNF) in other literatures. We denote the patch centered at p as $c(p)$. The guided image synthesis process can be formulated as follows.

$$d_{sem}(p, q) = ||G^c_{sem}(c(p)) - G^s_{sem}(c(q))|| \quad (1)$$

$$d_{pos}(p, q) = ||G^c_{pos}(c(p)) - G^s_{pos}(c(q))|| \quad (2)$$

$$d_{app}(p, q) = ||G^c_{app}(c(p)) - G^s_{app}(c(q))|| \quad (3)$$

$$\arg\min_{q \in Q} \sum_{p \in P} w_s d_{sem}(p, q) + w_p d_{pos}(p, q) + w_a d_{app}(p, q) \quad (4)$$

We solve Eq. 4 to obtain the NNF for image synthesis. The optimization is similar to former works [29], [33]. Then we use average voting based on the optimal NNF to get the final stylized output.

IV. RESULTS AND DISCUSSIONS

In this part, we first evaluate our method in Section IV-A. Then we compare our method with [3], [5], [6], [26], [28], and [29] qualitatively in Section IV-B. We show more results of our method in Section IV-C. We also provide a quantitative comparison in Section IV-D. Finally, we discuss our limitations and future works in Section IV-E.

Parameters For the synthesis process, we set w_s to 5.0, w_p to 5.0 and w_a to 1.0. We use 5×5 patches during the optimization and final voting. For all the results in this work, we fix these parameters.

Performance Our system takes 1 second for running the semantic segmentation network and about 0.486 second for refining these coarse semantic maps. It takes about 0.5 second to generate the appearance maps. As for the position maps, it takes 0.477 second. The guided image synthesis process takes about 4.9 seconds. The evaluation of running neural network is performed on a server with an NVIDIA TITAN graphics card. While the evaluations of other parts are performed on a laptop with a 2.8 GHz quad-core CPU. Therefore, the total time of our method is 7.36 seconds. This performance evaluation is under the image resolution of 600×800.

A. EVALUATION

1) INDIVIDUAL GUIDANCE

We evaluate the necessity of individual guidance map by setting the corresponding weight to zero. As shown in Figure 7, the forehead of the output will become more similar to the style image without appearance guidance, while our full guidance result maintains the highlight in forehead (red rectangle). Without the position guidance, the output fails to correctly stylize the left cheek of the content image. This is because the left cheek is brighter due to lighting, misleading the synthesis process. However, our full guidance result successfully stylizes this region (yellow rectangle). The semantic guidance map is obviously necessary and the output is wrong without it.

2) APPEARANCE MAPS FROM DIFFERENT LAYERS

In this part, we evaluate the appearance maps from different layers of VGG-19. We use the whitened features from five blocks of VGG-19 and decode them to images by the trained decoders. The bottom layer captures the low-level features of an image, while the top layer captures the high-level features. The whitening operation removes the style at a certain layer. Therefore, it is necessary to evaluate which layer we should choose to represent the appearance. As illustrated by Figure 8, the result using bottom layer fails to remove the texture and color differences. On the other hand, the result of top layer

FIGURE 7. Evaluation of individual guidance. We evaluate the necessity of individual guidance by setting the corresponding weight to zero. As can be seen, the appearance guidance maintains the highlight in the forehead (red rectangle). The position guidance corrects the error in the left cheek (yellow rectangle). The semantic guidance is important and the result is obviously wrong without it.

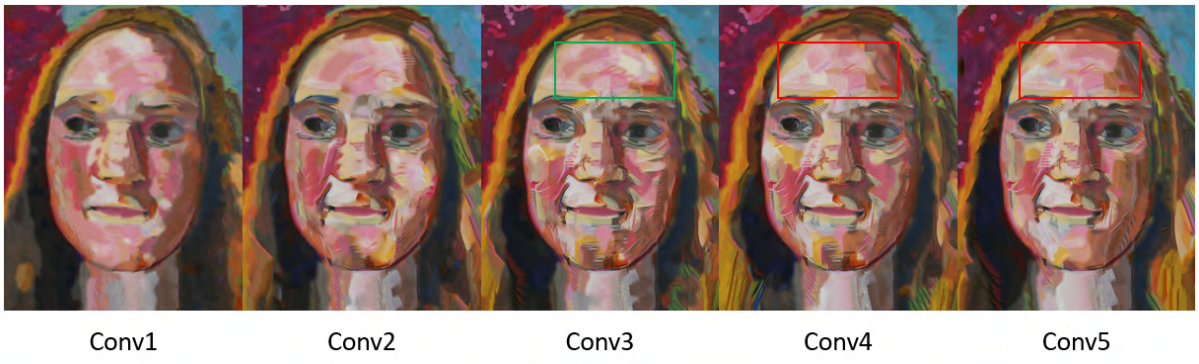


FIGURE 8. Evaluation of appearance maps from different layers. Since different layers of VGG-19 captures different features, we evaluate which layer we should use to generate the appearance map. The bottom layer fails to reduce the texture and color differences, while the top layer does not preserve important factors like illumination. Therefore, we empirically choose *Conv*3 to generate our appearance guidance map.

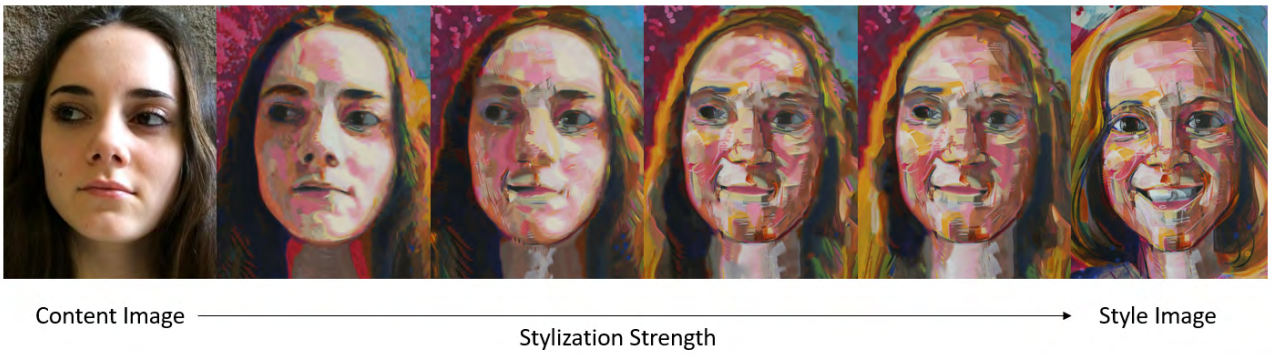


FIGURE 9. Illustration of style interpolation. We also evaluate the style interpolation as recent works on neural style transfer. We can not obtain an output same as the input content image even with large weight on appearance. Nevertheless, we demonstrate that the style interpolation can be achieved by simply adjusting the weight of appearance guidance.

fails to preserve some important factors like illumination. Therefore, we empirically use *Conv*3 to generate the appearance guidance since this layer makes a good balance.

3) STYLE INTERPOLATION

As recent works on neural style transfer [30], [32], we also evaluate the style interpolation by changing the weight of appearance guidance map. As shown in Figure 9, reducing the weight of appearance map generates more stylized output. However, since our appearance map removes the texture and color of input images, we can not obtain an output same as the input content image even with large weight on appearance. Nevertheless, we demonstrate that the style interpolation can be achieved by simply adjusting the weight of appearance guidance.

B. QUALITATIVE ANALYSIS

In this section, we will present the qualitative comparison of our method versus the state-of-the-art methods [5], [6], [26], [28]–[30]. Among them, [5], [6], [26], [28], [30]

are all based on deep neural networks and they kindly release the implementations. Therefore, we provide a side-by-side comparison with these methods on the same input in Figure 11. We manually tune the parameters of these methods to achieve the best results. As for [29], they do not publish their implementations, thus we compare our method with their online demo.

1) OUR METHOD VERSUS NEURAL STYLE TRANSFER

Reference [5] is the seminal work of neural style transfer. It defines the content loss and style loss by deep neural features. The style of an image is represented by the gram matrix of neural features. Back-propagation is used to optimize the content and style losses. However, [5] is slow and can not capture the local styles. As shown in Figure 11, [5] only transfers the overall appearance of style image to the content image.

Reference [30] first whitens the content features and then colorizes them by the style features. It uses multiple layers to transfer the style at different levels. The whitening is based on Zero-phase Component Analysis (ZCA), therefore the whitened features maintain the appearance of the original features. Reference [30] can use arbitrary content image and style image without network re-training. However, similar to [5], it can only transfer the global style. Although [30] can provide spatial control over the transfer process. It can not incorporate our soft semantic map to the transfer process.

Reference [6] combines MRF with deep neural features to regularize the transfer process. It uses neural patches to represent the style of an image instead of gram matrix. The style loss is defined as the distance between content image's patches and the corresponding nearest patches of style image. However, the patch matching will prefer to maintain the original appearance, which makes the output less artistic. Apart from this, the patch matching will introduce artifacts. For example, as shown in Figure 11 (row 5,col 6). The nostril is replaced by the eye since they are similar in appearance.

Reference [26] uses a semantic map to guide the neural patch matching of [6]. This can alleviate the obvious errors and achieve much better results. Although [26] uses convolution and finds the maximum response to improve the performance, it is still slow due to the back propagation optimization. Besides, [26] only finds the most similar patch, which means it is greedy. Recent work [36] explores the uniformity of patch usage to improve the results. We use our semantic guidance map as the input of [26] for a fair comparison. The performance is evaluated based on the author's published code with GPU acceleration.

Reference [28] uses deep neural features to find the nearest neighbor field between content image and style image. It follows the image analogy framework and generates the bidirectional transfer images. These images can be viewed as the stylized outputs. On the other side, they are used to solve the cross-domain matching problem. Reference [28] finds the NNF by features from multiple layers. In our comparison, we use their recommended parameters for portrait. As shown in Figure 11, [28] can not obtain consistently pleasing results since it does not use any guidance maps except the deep neural feature.

2) OUR METHOD VERSUS [29]

Reference [29] also uses guided image synthesis to generate the stylized output. They use semantic guidance, position guidance and appearance guidance as our method. However, It can only process facial region. Besides, we propose three different methods to generate the guidance maps because we can not densely align the content image with style image. Reference [29] can easily align the content face with style face based on the detected facial landmarks. As shown in Figure 10, the proposed guidance maps can achieve comparable results compared with [29], which only processes the facial region.

FIGURE 10. Comparison with [29]. (a,c) are our results and (b,d) are the results of [29]. [29] also uses guided image synthesis to generate the stylized output. However, it can only process facial region.

C. MORE RESULTS

In this section, we show more results of our method. We randomly choose seven content images and seven style images and show the pair-wise results in Figure 12. As can be seen, our method can robustly handle different styles and contents.

D. QUANTITATIVE ANALYSIS

In this section, we provide a perceptual user study for quantitative analysis as former works [30], [36]. We conduct this study to quantitatively compare our method with neural style transfer methods [5], [6], [26], [28], [30]. This user study is performed with 100 participants to evaluate our transfer technique. Each participant is shown eight images side-by-side at one time. The left two are the input and the example images while the right six are the results of our method and the results of [5], [6], [26], [28], and [30] which are randomly ordered. Then participants are asked to evaluate the six results how much they keep the content of the input image while taking the style of the example image, on a scale from 5 (the highest score) to 0 (the lowest score). We randomly choose 10 results for participants to score one by one and calculate the average score. Our method gets the score of 3.81. The scores of [5], [6], [26], [28], and [30] are 3.25, 3.1, 3.42, 2.98, 3.65 separately. Although artistic style is subjective, this study can kind of quantitatively demonstrate the advantage of our method over other methods.

FIGURE 11. Comparison with neural style transfer. (a) Results of [5]. (b) Results of [30]. (c) Results of [26]. (d) Results of [6]. (e) Results of [28]. (f) Our results. We also list the average running time of each method below. Our method achieves consistently pleasing results with acceptable computational cost. Especially, the texture details are well preserved by our method.

E. LIMITATIONS

Although our method can achieve better results compared with recent neural style transfer methods, there are still some limitations. For example, since our method is based on local image patches, we can not well capture the identity since identity is a global concept. The image synthesis process

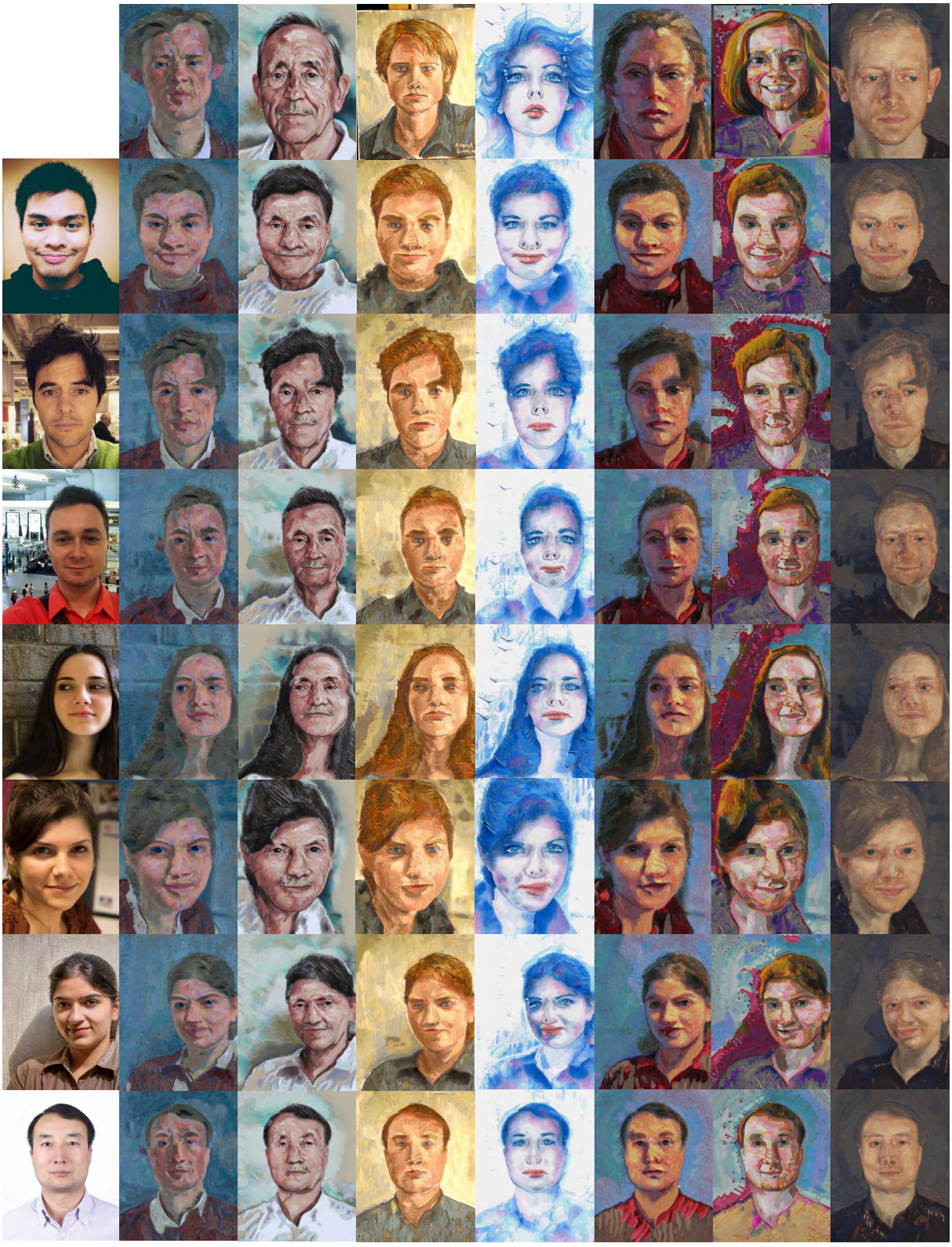


FIGURE 12. More results. We randomly choose seven content images and seven style images and show the pair-wise results. As can be seen, our method can robustly handle different styles and contents.

actually recomposes the content image with the patches of style image. Therefore, if we can not recompose a photo with similar identity as content image by the style image's patches, the stylized output can not preserve the identity. As illustrated by Figure 13, although the texture details are well-preserved, the identity is actually lost. In our future

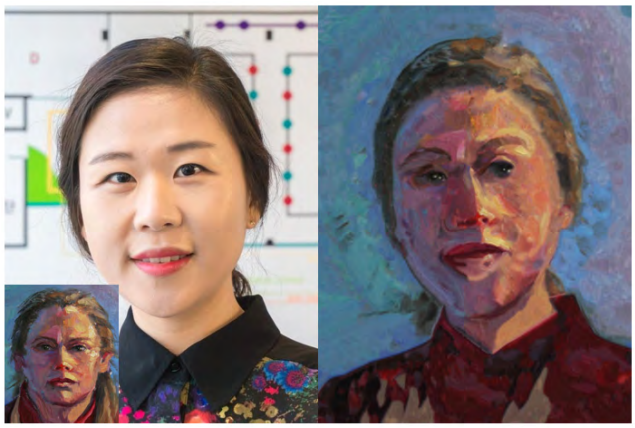


FIGURE 13. Limitation of our method. Our method sometimes fails to preserve the identity of content image. Our method is based local image patches, while identity is a global concept. In the future work, we will study how to better preserve the identity during the guided image synthesis process.

work, we will study how to better preserve the identity during the guided image synthesis process. One promising direction is training a destylization network [42] with facial perceptual loss [41]. Compared with whitening neural features in this work, we think training a destylization network with proper losses can help preserve the identity and achieve better results. Another possible solution is finding the doppelganger style image for content image from a large dataset of portrait paintings [43] before style transfer. Doppelganger has been used in other tasks like face replacement [44], [45]. However, it can not been used for arbitrary style image.

V. CONCLUSION

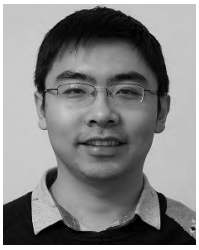
In this paper, we propose a robust portrait style transfer method. Our technique gets rid of the dense alignment and can process the full image instead of only facial region. We propose three novel guidance maps, which do not require dense correspondence. These guidance maps can regularize the image synthesis process and achieve visually pleasing stylized output. Compared with recent neural style transfer methods, our method achieves consistently appealing results for portrait and preserves more texture details. Besides, our method is also much more efficient. We hope our work can inspire future research on style transfer which aims for high-quality results.

REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge. (2015). "A neural algorithm of artistic style." [Online]. Available: https://arxiv.org/abs/1508.06576
[2] J. Johnson, A. Alahi, and L. Fei-Fei. (2016). "Perceptual losses for real-time style transfer and super-resolution." [Online]. Available: https://arxiv.org/abs/1603.08155
[3] A. Selim, M. Elgharib, and L. Doyle, "Painting style transfer for head portraits using convolutional neural networks," *ACM Trans. Graph.*, vol. 35, no. 4, p. 129, 2016.
[4] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand, "Style transfer for headshot portraits," *ACM Trans. Graph.*, vol. 33, no. 4, p. 148, 2014.
[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2414–2423.
[6] C. Li and M. Wand. (2016). "Combining Markov random fields and convolutional neural networks for image synthesis." [Online]. Available: https://arxiv.org/abs/1601.04589
[7] Y. Nikulin and R. Novak. (2016). "Exploring the neural algorithm of artistic style." [Online]. Available: https://arxiv.org/abs/1602.07188
[8] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. (2016). "Preserving color in neural artistic style transfer." [Online]. Available: https://arxiv.org/abs/1606.05897
[9] D. Ulyanov, A. Vedaldi, and V. Lempitsky. (2016). "Instance normalization: The missing ingredient for fast stylization." [Online]. Available: https://arxiv.org/abs/1607.08022
[10] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. (2016). "Texture networks: Feed-forward synthesis of textures and stylized images." [Online]. Available: https://arxiv.org/abs/1603.03417
[11] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 702–716.
[12] V. Dumoulin, J. Shlens, and M. Kudlur. (2016). "A learned representation for artistic style." [Online]. Available: https://arxiv.org/abs/1610.07629
[13] N. Joshi, W. Matusik, E. H. Adelson, and D. J. Kriegman, "Personal photo enhancement using example images," *ACM Trans. Graph.*, vol. 29, no. 2, p. 12, Apr. 2010.
[14] W.-S. Tong, C.-K. Tang, M. S. Brown, and Y.-Q. Xu, "Example-based cosmetic transfer," in *Proc. 15th Pacific Conf. Comput. Graph. Appl. (PG)*, Oct./Nov. 2007, pp. 211–218.
[15] D. Guo and T. Sim, "Digital face makeup by example," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 73–79.
[16] M. Brand and P. Pletscher, "A conditional random field for automatic photo editing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–7.
[17] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Data-driven enhancement of facial attractiveness," *ACM Trans. Graph.*, vol. 27, no. 3, p. 38, 2008.
[18] H. Chen, Y.-Q. Xu, H.-Y. Shum, S.-C. Zhu, and N.-N. Zheng, "Example-based facial sketch generation with non-parametric sampling," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 433–438.
[19] H. Chen, N.-N. Zheng, L. Liang, Y. Li, Y.-Q. Xu, and H.-Y. Shum, "PicToon: A personalized image-based cartoon system," in *Proc. 10th ACM Int. Conf. Multimedia*, 2002, pp. 171–178.
[20] H. Chen, Z. Liu, C. Rose, Y. Xu, H.-Y. Shum, and D. Salesin, "Example-based composite sketching of human portraits," in *Proc. 3rd Int. Symp. Non-Photorealistic Animation Rendering*, 2004, pp. 95–153.
[21] M. Meng, M. Zhao, and S.-C. Zhu, "Artistic paper-cut of human portraits," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 931–934.
[22] M. Zhao and S.-C. Zhu, "Portrait painting using active templates," in *Proc. ACM SIGGRAPH/Eurographics Symp. Non-Photorealistic Animation Rendering*, 2011, pp. 117–124.
[23] T. Wang, J. P. Collomosse, A. Hunter, and D. Greig, "Learnable stroke models for example-based portrait painting," in *Proc. BMVC*, 2013, pp. 36.1–36.11.
[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2016). "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1606.00915
[25] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 92–107.
[26] A. J. Champandard. (2016). "Semantic style transfer and turning two-bit doodles into fine artworks." [Online]. Available: https://arxiv.org/abs/1603.01768
[27] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. (2016). "Controlling perceptual factors in neural style transfer." [Online]. Available: https://arxiv.org/abs/1611.07865
[28] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. (2017). "Visual attribute transfer through deep image analogy." [Online]. Available: https://arxiv.org/abs/1705.01088
[29] J. Fišer *et al.*, "Example-based synthesis of stylized facial animations," *ACM Trans. Graph.*, vol. 36, no. 4, p. 155, 2017.
[30] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. (2017). "Universal style transfer via feature transforms." [Online]. Available: https://arxiv.org/abs/1705.08086

[31] M. Lu, H. Zhao, A. Yao, F. Xu, Y. Chen, and L. Zhang, "Decoder network over lightweight reconstructed feature for fast semantic style transfer," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2488–2496.
[32] X. Huang and S. Belongie. (2017). "Arbitrary style transfer in real-time with adaptive instance normalization." [Online]. Available: https://arxiv.org/abs/1703.06868
[33] J. Fišer *et al.*, "StyLit: illumination-guided example-based stylization of 3D renderings," *ACM Trans. Graph.*, vol. 35, no. 4, p. 92, 2016.
[34] T. Q. Chen and M. Schmidt. (2016). "Fast patch-based style transfer of arbitrary style." [Online]. Available: https://arxiv.org/abs/1612.04337
[35] L. Sheng, Z. Lin, J. Shao, and X. Wang. (2018). "Avatar-net: Multi-scale zero-shot style transfer by feature decoration." [Online]. Available: https://arxiv.org/abs/1805.03857
[36] S. Gu, C. Chen, J. Liao, and L. Yuan. (2018). "Arbitrary style transfer with deep feature reshuffle." [Online]. Available: https://arxiv.org/abs/1805.04103
[37] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang, "Exemplar-based face parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3484–3491.
[38] S. Liu, J. Yang, C. Huang, and M.-H. Yang, "Multi-objective convolutional learning for face labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3451–3459.
[39] A. S. Jackson, M. Valstar, and G. Tzimiropoulos, "A CNN cascade for landmark guided semantic part segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 143–155.
[40] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[41] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, p. 6.
[42] F. Shiri, X. Yu, P. Koniusz, and F. Porikli, "Face destylization," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov./Dec. 2017, pp. 1–8.
[43] E. J. Crowley, O. M. Parkhi, and A. Zisserman, "Face painting: Querying art with photos," in *Proc. BMVC*, 2015, p. 65.
[44] I. Kemelmacher-Shlizerman, "Transfiguring portraits," *ACM Trans. Graph.*, vol. 35, no. 4, p. 94, 2016.
[45] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: Automatically replacing faces in photographs," *ACM Trans. Graph.*, vol. 27, no. 3, p. 39, 2008.

MING LU is currently pursuing the Ph.D. degree with Tsinghua University. He is supervised by Prof. L. Zhang. His research interests include computer vision and computer graphics. He is particularly interested in 2-D/3-D object detection, 3-D face/body capture, and style/color transfer.



FENG XU received the B.S. degree in physics and the Ph.D. degree in automation from Tsinghua University, Beijing, China, in 2007 and 2012, respectively. He is currently an Assistant Professor with the School of Software, Tsinghua University. His research interests include face animation, performance capture, and 3-D reconstruction.

HAO ZHAO received the B.S. degree in electronic engineering from Tsinghua University in 2013. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Tsinghua University. His research interest focuses on 3-D scene understanding with multi-modal sensory inputs.

ANBANG YAO received the Ph.D. degree from Tsinghua University in 2010. He received 7+ scholarships including the 2009 Highest Honor from the School of Information Science and Technology. He joined Intel Labs China in 2012, where he has been leading to develop Intel 2-D and 3-D face analysis (landmark detection/tracking, head pose estimation, expression recognition, and 3-D face engine), deep learning-based scene understanding (object detection, semantic image segmentation, and room layout estimation), and efficient DNN design (network pruning, low-bit quantization, and deep compression) technologies.

YURONG CHEN is currently the Senior Research Director and a Principal Research Scientist with Intel Corporation and the Director of Cognitive Computing Lab, Intel Labs China, where he is responsible for driving cognitive computing, especially, visual understanding (VU), and machine learning research for smart computing on Intel platforms. He is also the co-owner of Intel Labs Visual Understanding and Synthesis Program, driving research innovation in smart visual data processing technologies on Intel platforms across Intel Labs. He led the team to win Intel China Award (Top Team Award of Intel China) 2016, Intel Labs Academic Awards (Top Award of Intel Labs)-Gordy Awards 2014/2015/2016 for delivering leading deep learning-based VU technologies, outstanding research achievement on multimodal emotion recognition and excellence in leading people and teams to deliver outstanding business results on advanced visual analytics.

Li Zhang received the B.S., M.S., and Ph.D. degrees in signal and information processing from Tsinghua University, Beijing, China, in 1987, 1992, and 2008, respectively. In 1992, he joined the Faculty of the Department of Electronic Engineering, Tsinghua University, where he is currently a Professor. His research interests include image processing, computer vision, pattern recognition, and computer graphics.