Noname manuscript No.
(will be inserted by the editor)

Automatic Semantic Style Transfer using Deep Convolutional Neural Networks and Soft Masks

Hui-Huang Zhao, Paul L. Rosin, Yu-Kun Lai and Yao-Nan Wang

Received: date / Accepted: date

Abstract This paper presents an automatic image synthesis method to transfer the style of an example image to a content image. When standard neural style transfer approaches are used, the textures and colours in different semantic regions of the style image are often applied inappropriately to the content image, ignoring its semantic layout, and ruining the transfer result. In order to reduce or avoid such effects, we propose a novel method based on automatically segmenting the objects and extracting their soft semantic masks from the style and content images, in order to preserve the structure of the content image while having the style transferred. Each soft mask of the style image represents a specific part of the style image, corresponding to the soft mask of the content image with the same semantics. Both the soft masks and source images are provided as multichannel input to an augmented deep CNN framework for style transfer which incorporates a generative Markov random field (MRF) model. Results on various images show that our method outperforms the most recent techniques.

Keywords Deep Neural Networks · Style Transfer · Soft Mask · Semantic Segmentation

Hui-Huang Zhao
College of Computer Science and Technology, Hengyang Normal University, Hengyang China
E-mail: happyday.huihuang@gmail.com
Paul L. Rosin, Yu-Kun Lai
School of Computer Science and Informatics, Cardiff University, Cardiff, UK
E-mail: RosinPL@cardiff.ac.uk, LaiY4@cardiff.ac.uk
Yao-nan Wang
College of Electrical and Information Engineering, Hunan University, Changsha, China
E-mail: yaonan@hnu.cn

1 Introduction

Style transfer is a process of migrating a style from a "style image" to a "content image". The goal is to be able to generate different renditions of the same scene according to different style images. Image style transfer has become a popular problem in computer vision and graphics, and can generate impressive results covering a wide variety of styles for both images [16] and videos [35]. It has also been widely employed to solve problems such as texture synthesis [9], inpainting [6], head portraits [36, 12], super-resolution [23, 7], font generation [1], and smoke simulations [25]. Moreover, a number of useful applications of image style transfer have been shown, such as: stylisation of 3D CAD models for more aesthetically pleasing presentation of design solutions [33], stylisation of maps to provide better visualisation [22], stylisation to provide seamless integration of content in augmented reality [28], and stylisation of 3D models for technical illustration [20].

When existing neural style transfer methods are applied to images with complex structures, visual elements from the style image are often transferred to semantically irrelevant areas of the content image. In order to achieve good results, users must pay attention to the composition and/or the selection of the style image, because for example the background colours or textures will often ruin the style transfer results, especially for portraits where the artefacts can be particularly off-putting. Addressing this problem, [4] (and subsequently [17]) recently proposed a method which uses a manually generated semantic map to help control the style transfer, and can achieve better results than some common methods.

In this paper, we specifically consider the problem of image style transfer guided by *automatically* extracted *soft* semantic masks. To achieve this, we adapt various semantic segmentation and labelling techniques to extract soft masks

associated with specific semantics. By deploying the semantic masks to control the transfer, it is possible to avoid errors such as those shown in figure 1(c) generated using the CNN-MRF method [29] in which stylised foreground objects are contaminated by the background texture, and vice verse.

The main contributions of the paper are as follows:

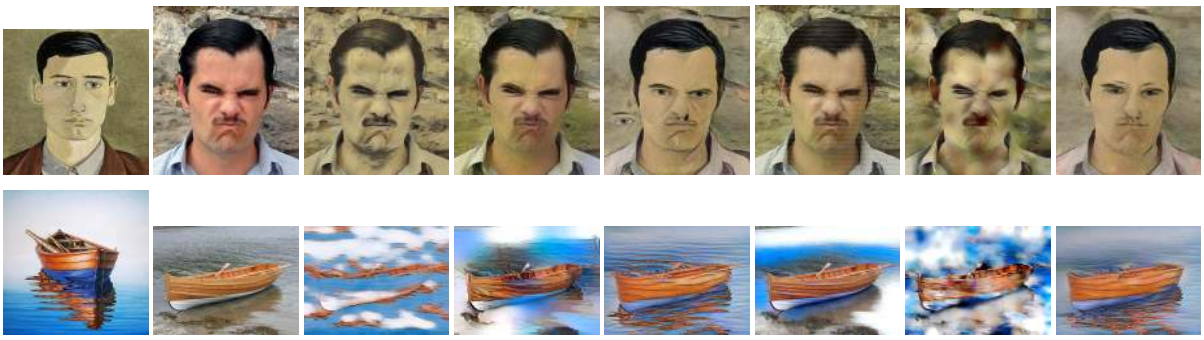
1. We adapt a state-of-the-art semantic segmentation method [48] to generate semantic masks automatically. Instead of using hard segmentation as [48], we propose to use soft masks containing the probabilities of occurrence of different objects in the image, since they preserve more information and are more robust when image regions have similar chances of belonging to multiple object categories. They are used to capture elements of the styles for objects in the style image and to preserve the structure of the content image. For the human face in particular we use a more detailed segmentation, in which different facial parts such as the nose, eyes and mouth are also automatically segmented, providing fine-grained control in perceptually crucial areas; these are also treated as semantic masks.
2. We augment a trained deep convolutional neural network by concatenating K soft mask channels and N channels of regular filters. This is further combined with a generative Markov random field (MRF) model [29] for image style transfer. Both the style and content images and their semantic maps are input into the augmented deep convolutional neural network. Extensive experiments show that such higher-level semantic information improves the quality of style transfer.

2 Related Work

Style transfer using deep networks. The success of deep CNNs (DCNNs) in image processing has also raised interest in image style transfer. [38] proposed a new style transfer method for headshot portraits. During their method, they presented a new multiscale technique based on deep networks to robustly transfer the local statistics of an example portrait onto a new one. [16, 15] showed remarkable results by using the VGG 19-layer network for style transfer. Their approach was employed in unguided settings and taken up by various follow-up papers. [17] in particular extended the Gram matrix method beyond the paradigm of transferring global style information between pairs of images, and they introduced control over spatial location, colour information and spatial scale. [46] built a Multi-style Generative Network (MSG-Net), which achieved real-time performance. [41] presented an alternative approach which trained compact feed-forward convolutional networks. The resulting networks are extremely light-weight and can generate images faster than [16]. By combining the benefits of training feed-forward convolutional neural networks and perceptual loss functions, [23] presented a novel approach for image style transfer. [29] suggested an approach to preserving local patterns of the style image. Instead of using a global representation of the style computed as a Gram matrix, they used patches of the neural activation from the style image. [35] presented an approach that transfers the style from one image (for example, a painting) to a whole video sequence. In order to solve the problem of distortions which made the result look like a painting in photographic style transfer, [32] developed a photographic style transfer method which constrains the transformation from the input to the output by using a photorealism regularization term. [24] proposed a new method named deep image analogy which is based on finding semantically-meaningful dense correspondences between two input images for image visual attribute transfer. Different from the image domain transfer problem, [5] proposed a new approach which involves two asymmetric functions, which are a forward function that encodes example-based style and a backward function that removes the style, for applying and removing makeup.

Two main types of methods are used in deep learning based style transfer: global approaches based on the Gram matrix or other global measures, and local approaches based on patch matching. Compared to the global methods, methods based on patch matching are more flexible and better cope with cases in which the visual styles or elements vary across the image. However, they could also produce visible artefacts when there are local matching errors. In order to control the region of application of the style image, [17] used several manually specified spatial guidance channels, containing values in [0,1], for both the content and style images. Their experiments showed that the guidance channels can ensure that the style is transferred between regions of similar scene content in the content and style images. It is however time-consuming to produce masks. As a result, for examples in their paper, they just used a mask to separate two parts of the image (e.g. sky and non-sky) for simple spatial control, and did not distinguish more detailed content in the images. For most methods not based on local matching such as [42, 30], they are prone to indiscriminately transfer different styles, which belong to specific objects in the style image, across the content image, thereby degrading the transfer result. Also, some methods use domain specific models and therefore cannot be applied to general images [5].

MRF-based image synthesis. Markov Random Fields (MRFs) are a famous framework for non-parametric image synthesis [10], [13]. [27] and [26] modelled the texture as an MRF and computed some approximation to the optimal solution. [47] formulated the patch mapping problem as a labelling problem modelled by a discrete MRF. Moreover, [14] proposed a novel unsupervised method for texture and colour



(a) style (b) content (c) [16] (d) [23] (e) [29] (f) [42] (g) [30] (h) our method

Fig. 1 Automatic semantic style transfer using deep convolutional neural networks.

transfer based on MRFs. In their approach an adaptive patch partition is used to capture the style of the example image and preserve the structure of the source image. MRF models suffer from a limitation that local image statistics are usually not sufficient for capturing complex image layouts at a global scale. [44] and [26] proposed a multi-resolution synthesis approach to improve this. We adapt this in our method. [29] presented a combination of generative Markov random field (MRF) models for image synthesis. Unlike other MRF-based texture synthesis approaches, their combined system can both match and adapt local features with considerable variability, and therefore our paper is based on this method.

Semantic segmentation. Recently, CNN architectures have been shown to be capable of providing semantic segmentation [19,40], generic object detection [31] and image completion [18]. [19] proposed a method called R-CNN, which combines region proposals with CNNs. In [18], a deep CNN is used to disentangle patch structure and style, and visual aesthetics (style) alongside structure and semantics are used to enhance image completion. [34] applied a trained network (VGG 16-layer net) to each proposal in an input image, and constructed the final semantic segmentation map by combining the results from all the proposals. [37] proposed a fully convolutional network for semantic segmentation. For producing accurate and detailed segmentations, they defined a skip architecture which combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer. In order to achieve better results, some existing face detection methods are also used in style transfer. By searching a database using Face++ [11] to find images with similar poses to a given source portrait image, [45] presented a novel colour transfer approach for portraits. [48] introduced a new form of convolutional neural network that combines the strengths of Convolutional Neural Networks and Conditional Random Fields based probabilistic graphical modelling. These models rely primarily on convolutional layers to extract high-level patterns, and then use deconvolution to label the individual pixels. Currently they have trained this model to recognise 20 classes, and our paper uses this method to obtain some semantic content from images.

Limitations of current methods. Although there has been considerable development on neural style transfer, the recent methods tend to have the same types of problems as the earlier works. Most existing methods either use Gram matrices (or equivalent measures) which treat images globally, or for methods based on local patch matching, can often match regions of one object in the style image to regions of a different object in the content image, causing artefacts such as those shown in figure 1. This is more critical for human faces as subtle mismatches can be detrimental to the quality of synthesised images. To address this, existing methods [4,17] use manual segmentation to improve style transfer. However, manual segmentation is time-consuming and laborious. In contrast, our method automatically performs a partial soft semantic segmentation of the content and style images. We augment the CNNMRF model used in [29] to further incorporate soft semantic masks, which can better capture features from the style image and preserve the structure of the content image. Another issue is that it is difficult to disentangle content and style, and so the transfer from one image to another tends to include some (undesired) elements of content in addition to style. Although we do not address this problem explicitly, by limiting the transfer of style to appropriate local regions the issue is mitigated.

We first briefly introduce our augmented DCNN architecture in section 3, followed by details for the style transfer algorithm in section 4. We then provide details for automatic semantic mask extraction in section 5. Experimental results and discussions are presented in section 6 and finally conclusions are drawn in section 7.

3 Architecture

We now discuss our augmented DCNN architecture which is based on VGG 19-layer network [39] for style transfer.

It takes as input a content image and a style image, both of which are fed into the VGG 19-layer network. The DCNN architecture combines pooling and convolution layers l with 3×3 filters (for example, the first layer after second pooling is named *Conv*3_1). Like common DCNNs, intermediate post-activation results denoted as x^l for the layer l consist of N channels, which capture patterns from the source images for each region of the image. Our augmented network is shown in figure 2.

Our augmented network also takes K semantic soft masks as input, which are down-sampled to produce semantic channels p^l at layer l with the same resolution as x^l.

We concatenate them to form the new output with $N+K$ channels, defined as d^l and labelled accordingly for each layer (e.g. *myConv*4_1). Before concatenation, the semantic channels are weighted by parameter β to balance their importance:

$$d^l = (x^l, \beta p^l). \tag{1}$$

We set $\beta = 20$ which we have found experimentally to provide interesting results.

4 Semantic Style Transfer Optimization function

Next, we introduce our style transfer model. We use an augmented loss function which is based on a patch-based approach [29] for style transfer, using optimisation to minimise content reconstruction error E_c and style remapping error E_s, which combines an MRF and a DCNN model, given a style image $x_s \in R^{3\times w_s \times h_s}$, a content image $x_c \in R^{3\times w_c \times h_c}$, and semantic maps $m_{c_k} \in R^{w_c \times h_c}$ and $m_{s_k} \in R^{w_c \times h_c}$ associated with the content and style images, respectively ($k = 1, 2, \ldots, K$). For simplicity, the semantic masks for the content and style images are also collectively represented as $m_c \in R^{w_c \times h_c \times K}$ and $m_s \in R^{w_s \times h_s \times K}$. The style transfer result image is denoted by $x \in R^{3\times w_c \times h_c}$. Since the synthesised image x is expected to have the same semantic layout as the content image, we treat m_c also as the semantic masks for the synthesised image. During our method, we make the high-level neural encoding of x similar to x_c and use the local patches similar to patches in x_s. As a result, the style of x_s is transferred onto the layout of x_c. Meanwhile, we penalise patch matches with inconsistent semantic masks. We define an energy function as follows and seek x that minimises it:

$$E(x) = \alpha_1 E_s(\Phi(x), \Phi(x_s), \Phi(m_c), \Phi(m_s)) + \alpha_2 E_c(\Phi(x), \Phi(x_c)). \tag{2}$$

E_s and E_c are defined as the style loss function and content loss function respectively, where $\Phi(x)$ is x's feature map (activation) that the network outputs in some layer, $\Phi(x_s)$ is the feature map (activation) of the style image x_s in the same layer, and $\Phi(m_c)$ and $\Phi(m_s)$ are the semantic masks of the content and style images downsampled to the same resolution as $\Phi(x)$ and $\Phi(x_s)$. For our method, E_s aims to penalise inconsistencies in neural activations and/or semantic masks between x and x_s. E_c computes the squared distance between the feature map of the synthesised image and that of the content source image x_c. Since x is assumed to have the same content layout as x_c, E_c does not involve the semantic masks.

Style loss function: We extract all the local patches from $\Phi(x)$, denoted as $\Psi(\Phi(x))$. For a given layer, assuming N is the number of channels, each patch in $\Psi_i(\Phi(x))$ has size $t \times t \times N$, where t is the width and height of the patch. Similarly, $\tilde{\Psi}(\Phi(m_{c_k}))$ and $\tilde{\Psi}(\Phi(m_{s_k}))$ are the down-sampled semantic masks of extracted patches, each of size $t \times t$. We define the modified energy function E_s incorporating semantic masks as

$$\begin{aligned} &E_s(\Phi(x), \Phi(x_s), \Phi(m_c), \Phi(m_s)) = \\ &\sum_{i=1}^{P} \| \Psi_i^*(\Phi(x)) - \Psi^*_{NN(i)}(\Phi(x_s)) \|^2 + \\ &\sum_{i=1}^{P}\sum_{k=1}^{K} \| \tilde{\Psi}_i(\Phi(m_{c_k})) - \tilde{\Psi}_{NN(i)}(\Phi(m_{s_k})) \|^2, \end{aligned} \tag{3}$$

where P is the number of patches in the synthesised image. For each patch $\Psi_i(\Phi(x))$ with semantic masks $\tilde{\Psi}_i(\Phi(m_{c_k}))$ we find its best matching patch $\Psi_{NN(i)}(\Phi(m_s))$ or $\Psi_{NN(i)}(\Phi(x_s))$ using normalised cross-correlation over all P_s example patches in $\Psi^*(\Phi(x_s))$:

$$NN(i) := \arg\max_{j=1,\ldots,P_s} \frac{\Psi_i^*(\Phi(x)) \cdot \Psi_j^*(\Phi(x_s))}{\mid \Psi_i^*(\Phi(x)) \mid \cdot \mid \Psi_j^*(\Phi(x_s)) \mid}, \tag{4}$$

where $\Psi_i^*(\Phi(x)) = \big(\Psi_i(\Phi(x)), \beta\tilde{\Psi}_i(\Phi(m_c))\big)$ is the concatenation of neural activation and semantic masks for the i^{th} patch of the synthesised image, and $\Psi_j^*(\Phi(x_s)) = \big(\Psi_j(\Phi(x_s)), \beta\tilde{\Psi}_j(\Phi(m_s))\big)$ is the concatenation of neural activation and semantic masks for the j^{th} patch of the style image. The nearest patch thus takes both style similarity and semantic consistency into account.

Content loss function: In order to control the content of the synthesised image, we define E_c as the squared Euclidean distance between $\Phi(x)$ and $\Phi(x_c)$:

$$E_c(\Phi(x), \Phi(x_c)) = \|(\Phi(x) - \Phi(x_c)\|^2. \tag{5}$$

Like method [29], we also minimise Equation 2 using backpropagation with L-BFGS. In Equation 2, α_1 and α_2 are weights for the style image and the content image constraints, respectively. According to our experiments, we set $\alpha_1 = 10^{-4}$ and $\alpha_2 = 20$, and these values can be fine tuned to interpolate between the content and the style preservation.

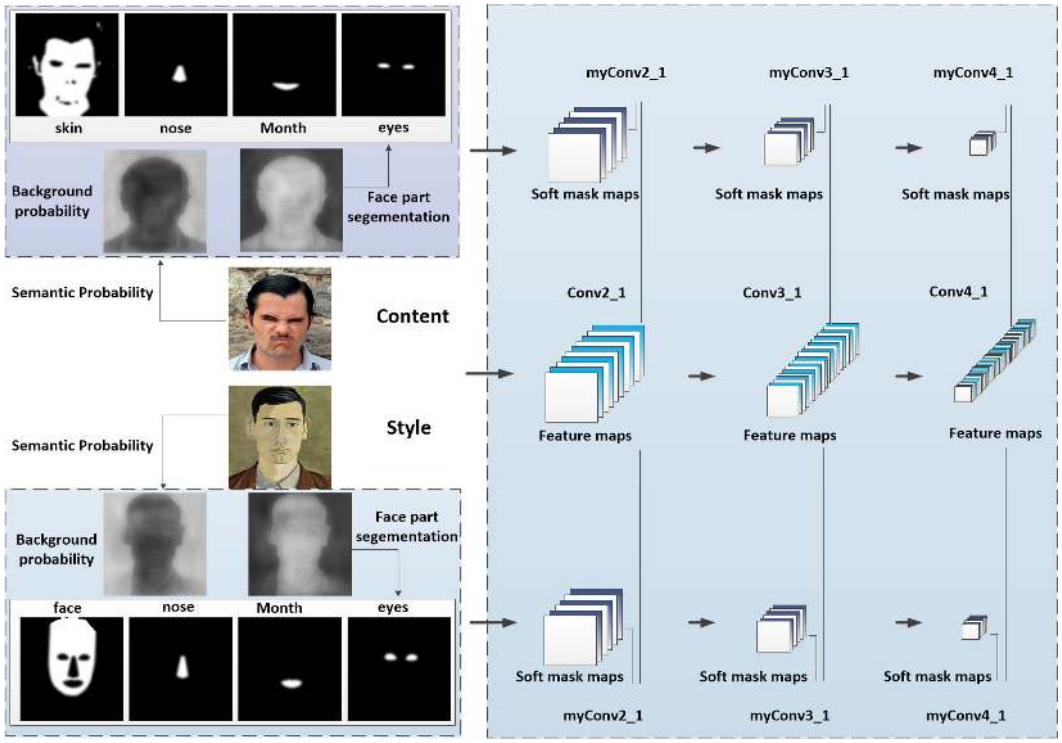


Fig. 2 Style transfer framework with deep neural networks and soft masks.

5 Automatic Soft Semantic Mask Extraction

[4] manually generated the semantic masks that they used in their work to control the style transfer. [12] proposed a stylised facial animation method by using facial segmentation. During [4], each image used one mask containing semantic labels, where each component (not necessarily connected) was indicated by a particular pixel value in the image. Often these values were carefully chosen so that components with similar appearance such as ear and nose would be assigned similar mask values. Not only is it tedious to manually segment the image, but for most images some parts cannot be partitioned accurately. Therefore, instead of using a single crisp mask to control an image stylisation, we propose to use a set of soft masks. Such soft masks provide more information than a single crisp mask, and do not require potentially unreliable boundaries to be set in the semantic mask, which is especially beneficial at ill-defined object boundaries.

In this paper we aim to automatically generate soft masks. Obviously this would make mask-based style transfer more convenient for the user. However, generating appropriate masks is challenging. Ideally, the segmentation of the style and content images should be consistent, e.g., using co-segmentation [43]. However, such approaches have not been developed for semantic segmentation. Moreover, the different appearance of photographs compared to artwork (typically used for style images) leads to the cross-depiction problem [21], so that semantic segmentation techniques trained on photographs will fail on paintings. In this paper we not only demonstrate our approach for the domain of portraits, which are a popular topic for style transfer [36], and non-photorealistic rendering in general, but also show stylisation of scenes containing other objects, such as cars and trains. Portrait style transfer allows us to leverage state-of-the-art techniques for face detection, which are more robust than general segmentation methods, and are effective even for many artworks. During our method, facial component masks are automatically extracted using a combination of semantic segmentation, facial landmark detection, and skin detection.

5.1 Semantic Image Prediction

[48] proposed a semantic segmentation method named CRF-RNN. CRF-RNN achieves a good result on the popular Pascal VOC segmentation benchmark. This improvement can be attributed to the uniting of the strengths of CNNs and CRFs in a single deep network. In our work, we use CRF-RNN to produce semantic probability maps. Instead of labelling each pixel with an object category, we skip over the max pooling stage and extract the neural activations before that and rescale them to [0, 1]. These are treated as probability maps predicting the chance of each pixel belonging to each object category. An example is shown in figure 3.

Using their pre-trained model with 20 categories provides 20 probability masks which represent different types

(a) content (b) style (c) probability map for (a) (d) probability map for (b)

Fig. 3 Probabilistic semantic segmentations using CRF-RNN [48] for person prediction.

(a) skin (b) nose (c) eyes (d) mouth (e) masks

Fig. 4 Segmentation of facial components for the images in figure 3.

of objects. Since most images only contain a small number of object types, rather than using all 20 semantic masks we just use a subset of five so as to reduce memory requirements and improve efficiency. For a given content and style image pair the five semantic masks are automatically selected as the five masks maximising their average probability.

We have found that the CRF-RNN is mostly reliable for photographs. For paintings its performance degrades, especially as the style of the artwork becomes more extreme. However, it is still capable of producing adequate extractions of people, cars, etc. for many paintings (used as style images) that we have tested.

5.2 Skin Detection

Skin detection is performed on the photographic images [3], using a rule-based analysis of pixels in YCbCr colour space. The skin mask is then intersected with the person mask provided by the CRF-RNN, so as to subdivide the person into skin and non-skin (e.g. hair, clothing). An example is shown in figure 4.

Since skin detection is primarily colour based, it is not in general effective on artwork due to the typical colour shifts, as well as distortions caused by strong brush stroke textures. Therefore, for paintings the facial region is detected using the face detector, rather than using skin detection.

5.3 Face and Facial Part Segmentation

Facial landmark detection aims to detect key-points in human faces, e.g. eyebrows, eyes, nose. There is an extensive literature on this topic. For example, [8] developed a style-aggregated network (SAN) to deal with the large intrinsic variance of image styles for facial landmark detection. One application of facial landmark detection is for face and facial part segmentation, which is used in our method.

During our method, the facial landmark detection is performed using OpenFace [2], which is based on Conditional Local Neural Fields, a version of the well known Constrained Local Model approach. Alternative facial landmark detection methods may be used instead. Sixty-eight facial landmarks are located, from which the eye, nose, inner and outer mouth regions are determined – see figure 4.

Since the facial landmarks only cover the lower half of the face, the outline of the face is extended upwards, and intersected with the person mask provided by the semantic segmentation to produce a good approximation to the head region. This mask is used for artwork. For photographs the skin mask is used instead of the extended facial region as it is more accurate (although prone to noise).

The above steps result in a set of masks that are blurred to produce soft masks identifying the following objects: face/skin, nose, eye, mouth, see figure 4 for an example. To provide a more compact visualisation we also combine the set of soft masks into a single colour image, see figure 4. The soft masks for body, background and face/skin are mapped to red, green, blue respectively, while the eyes, nose and mouth values are mapped to cyan, yellow, magenta respectively. (Note that when performing style transfer the multiple soft image masks are used instead.)

5.4 Background Masks

The individual foreground masks are combined by first applying a max operation to the set to produce a single foreground mask. This is inverted (subtracted from one) to generate a single background mask.

6 Results

We use the pre-trained VGG 19-layer network with the augmented layers *myConv*3_1 and *myConv*4_1. For layers *relu*3_1, *relu*4_1, *myConv*3_1 and *myConv*4_1 we use 3×3 patches, and we set the stride to one. Following the patch-based approach of [29], we synthesise at multiple increasing resolutions, and randomly initialise the optimisation. On a GTX Titan with 12GB of GPU RAM, synthesis takes from 5 to 30 minutes depending on the output quality and resolution.

We will now compare the proposed method with several popular methods: [16,29,42,30] which are representative global and local neural style transfer methods, and [17,

4] which use manual segmentation to improve style transfer. Note that for our method multiple soft masks were used; the single colour mask is just shown for illustrative purposes. For [4], we set the content weight to 10, style weight to 25, semantic weight to 100, and we use the masks from [4] when available and otherwise manually draw them ourselves. For [17], we used two image maps of values in the range [0,1] for content and style images like figures 3(c, d), similar to the examples used in their paper, which are also used in our method. To partially overcome orientation and scale differences between the style and the content images, we also allow a range of rotations and scalings to be considered in the CNNMRF, following the settings in [29].

We use figure 3(a) with several different backgrounds as the content image, and choose figure 3(b) as the style image. Style transfer results obtained by the different methods are shown in figure 5. Considering the six existing methods and by comparing the results in figure 5, it seems that [16] , [23] and [29] cannot transfer the background texture well. [4] achieves better background texture transfer, comparable to our method, but some key facial parts (nose and mouth) are lost. [17] can control the spatial texture very well, but the human style transfer is not so good. It also generates errors in rows 1 and 2 of figure 5 (d). Because both our method and [29] are based on the MRF regulariser, and [23, 16, 42, 30] have also demonstrated some good results, we mainly compare our results with those methods.

Given content and style images in figure 6, figure 7 shows style transfer applied separately to photographs of men and women. We transfer the style of each style image to each content image. We can see from figure 7 that our method can achieve better results than the other methods and avoid errors in applying style transfer to inappropriate parts. The style images contain a range of simple and more complicated textures. In both cases, our method achieves effective results, and preserves the content of the images. [29] can also achieve interesting results, but only for simple texture images.

For style images that contain a mixture of textures – e.g. rows 2 and 6 in figure 6 – the results of [29] have many errors in which styles are misapplied. [16] cannot transfer enough style to content image. [23], [42] and [30] also generate some imperfect results. Our method can achieve better results in specific parts in mouth and eyes. For examples in rows 7 and 8 of figure 6, our method can achieve better results in specific parts in eyebrows, mouth and eyes and mouth area.

Detection of skin and facial parts is affected by different skin colour, and by significant variations in illumination. However, our pipeline is reasonably robust, and is demonstrated on the content images shown in figure 8 which provide an example with harsh lighting and another with dark skin. The visualisations of their soft masks in figure 8 reveals some minor errors, but our method produces robust style transfer results (figure 9). In comparison we see many existing methods have difficulty stylising such images.

Style transfer of different object types. More examples of style transfer for objects like train, car, bus and boat are shown in figures 10 and 11. In these examples, in the mask images the green part shows the background probability mask, and the red part shows the object probability mask. Our method produces better results in all these examples. The comparison results are shown in figure 11. Our method achieves effective results, and preserves the content of the images. [29] can also achieve interesting results, but the results of [29] still have many errors in which styles are misapplied. [16] cannot transfer enough style to content images. [23], [42] and [30] generate some imperfect results too. Our method can achieve better results in specific parts in background area.

Comparison of soft masks and binary masks. Figure 12 shows probability masks of trains, and in the mask images the green part shows the background probability mask, and the red part shows the object probability mask. We compare our method using soft masks (shown in the second row) with alternative binary masks (shown in the third row). In comparison, the results with the soft masks in figures 13(b, d) not only avoid choosing thresholds but also are visually better than with hard masks (a, c) since more information is preserved.

Automatic multi probability map selection. Not only will probability maps provide a richer feature vector that will benefit the style transfer, but avoiding the need for thresholding or winner-take-all selection has the potential to improve robustness. Figures 14 and 15 show an example in which our automatic semantic mask selection effectively chooses relevant object types (person and dog). It demonstrates style transfer using our method when multiple object categories are present. Note that even though the irrelevant 3rd–5th masks contain very little response, it is not a problem to include them.

Figure 16 shows a further example of style transfer using multiple classes along with greater variation of pose and image composition.

One can see in figure 16 that the red color in the style image is propagated into the background in figure 16(g) when the soft mask weight $\beta = 20$. When the soft mask weight is set as $\beta = 25$ (i.e., with a higher semantic mask weight), it effectively constrains style transfer, and no red color is propagated into the background, as shown in figure 16(h).

Our method also allows styles to be transferred from multiple style images to a single content image. In this case, the semantic masks are essential to direct the method to choose suitable patches. Some interesting style transfer results are shown in figure 18 with input images and their masks presented in figure 17.

(a) Content (b) [16] (c) [23] (d) [17] (e) [4] (f) [29] (g) [42] (h) [30] (i) Our

Fig. 5 Style transfer results from several methods using versions of a content image with varied backgrounds.

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 6 Style (a,b,e,f) and content (c,d,g,h) images along with visualisations of their soft masks.

Colour control. Our method better preserves objects and their styles, thanks to soft semantic masks. However, similar to existing neural style transfer methods, when the same object from the content and style images has substantially different colours, our method tends to produce stylised images with colours mixed. An example is shown in figure 19 where the red car in the content image and the blue car in the style image lead to the purple car in the stylised result. [17] proposed two methods to control colour information in style transfer which can better preserve colour in the content image during stylisation, and we demonstrate applying their approach to our method.

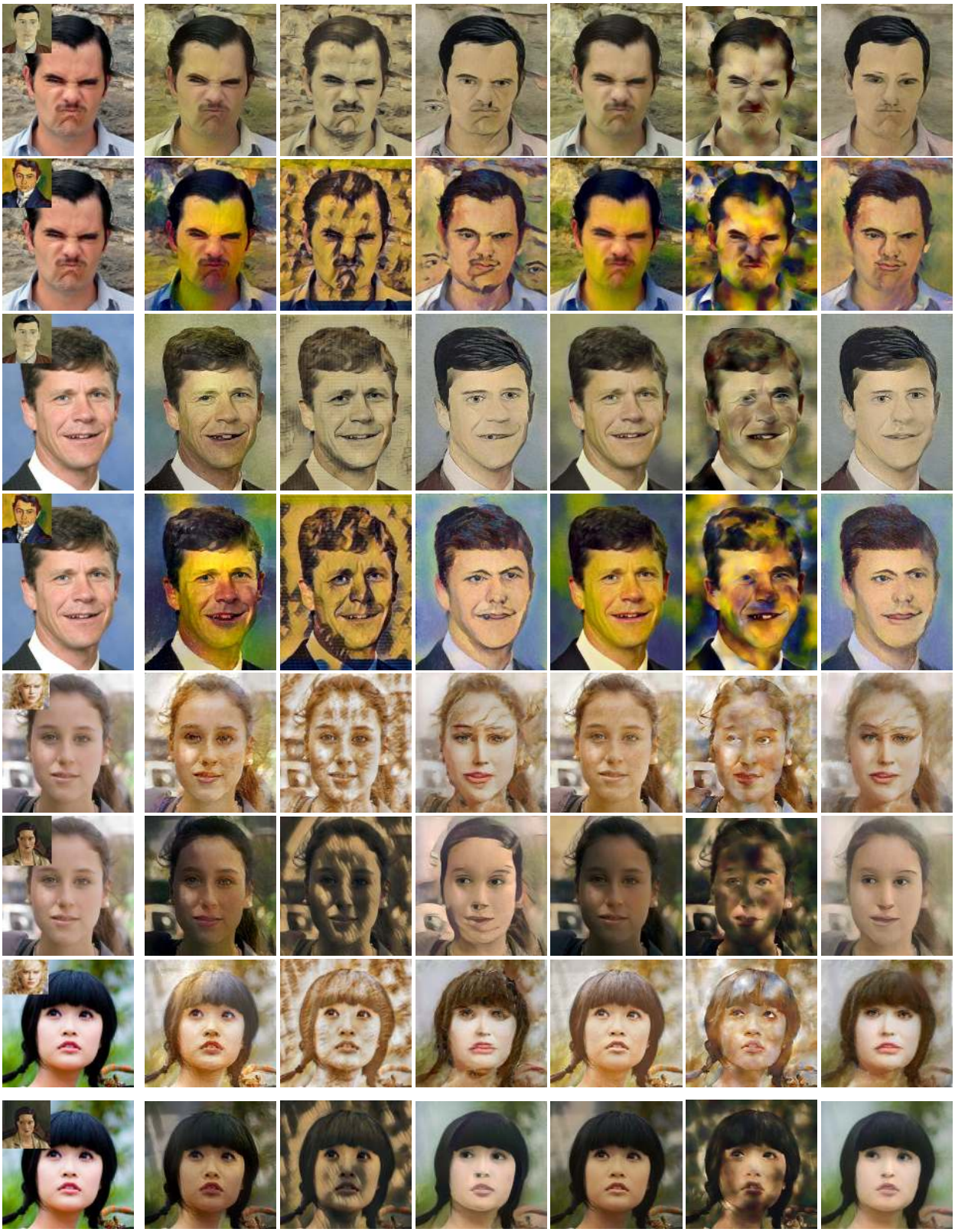
In the first approach we perform style transfer only in the luminance channel, while keeping the chromiance channels from the content image unchanged. To improve matching, before style transfer, we also use intensity mapping for each luminance pixel L_s in the style image to obtain L_{s^*}

$$L_{s^*} = \frac{D_c}{D_s}(L_s - u_s) + u_c, \tag{6}$$

where u_c and u_s are the mean luminances of the content and style images, respectively, and D_s and D_c are their standard deviations. For the second method, we apply colour histogram matching before style transfer. Each RGB pixel P_s in the style image is transformed as

$$P_s^* = AP_s + a, \tag{7}$$

where A is a 3×3 matrix and a is a 3-vector such that after the transformation the mean and covariance of the style image matches the content image. Colour control results based on the two methods are shown in figure 19. The standard result is shown in figure 19(a). We can see that, it can achieve a good colour control result by using the luminance channel, and is shown as figure 19(b). However, by using colour histogram matching method, there is some wrong colour transferred in background in figure 19(c). This is because a global transformation is not sufficient to capture colour differences between the content and style images. The method may work better if the transformation is applied separately for individual objects. However, since our method does not produce hard segmentation, it is not obvious how this can be achieved.



(a) input (b) [16] (c) [23] (d) [29] (e) [42] (f) [30] (g) Our

Fig. 7 Human style transfer comparison.

(a) (b) (c) (d)

Fig. 8 Content (a,c) images along with visualisations of their soft masks(b,d).

Failure cases and Limitations. Different categories of segmentation failures can occur, and affect the style transfer results in different ways. If an object present in the content image does not exist in the style image then the transfer result for that object will be similar to the results for the baseline CNNMRF method that we use [29]. If an object in the style image is mis-recognised as another object then there are two possibilities. If this object does not exist in the content image then again the result will be similar to the baseline method. Alternatively, the faulty classification will result in faulty style transfer.

An example of the effect of a segmentation failure is shown in figure 20. The content image is reasonably well segmented, but the more challenging task of segmenting the artwork used as the style image contains greater errors. While the skin (generated using the face fitted by OpenFace) is detected reasonably well in the style image, the head (generated using the CRF-RNN) is missing the left-hand part of the hair. This has caused our method to match part of the content image's background to hair, and consequently stylise it as hairy. Despite this flaw, the result is still considerably better than the baseline CNNMRF method which contains many instances of inappropriate style transfer.

Our method relies on automatically calculated semantic masks. When some objects, either in the content image or in the style image, are not correctly detected, the method cannot find the correct matching for the missing objects. Figures 21 and 22 show some failure examples of this kind. One can see from figure 21 that, because the dog in the style image is not correctly detected in semantic segmentation (figure 21 (f)), the style of the dog in figure 21(d) is not transferred to the dog in the output image. In the example shown in figure 22, the dog in the content image in figure 22(a) is not segmented correctly where part of it is considered to belong to the man, the style (red clothing) from the man in the style image is erroneously transferred to the dog in the output image.

Modifying the number of masks. The semantic segmentation significantly affects the style transfer results. For some style images, for example some paintings of portraits, it is difficult to automatically segment the face, skin, month, eyes, etc., and to properly segment the background and foreground. If the accuracy and reliability of the semantic segmentation can be improved, this will lead to better style transfer results. Figure 23 shows an experiment in which the number of labels in the semantic masks is increased, and demonstrates the importance of separately labelling all the major components of the face.

Modifying the soft mask weight. There are three parameters in our style transfer model, α_1, α_2 and β which are the weights for the style, content and semantic mask loss terms. Since the effect of α_1 and and α_2 is considered in [29], we focus on studying the effect of β. By default we set the soft mask weight $\beta = 20$. This value can be adjusted to control the importance of semantic compliance. Figure 24 demonstrates the effect of modifying β using the content image in figure 6(c) and style image in figure 3(b), where $\alpha_1 = 10^{-4}$, $\alpha_2 = 20$. When β is too small, the result does not have sufficient semantic control and can produce semantically wrong matches. On the other hand, setting β too large may result in matched patches having poor content/style consistency. According to our experiments, $\beta \in [15, 35]$ achieves best results.

Modifying the content and style weight. Further experiments are carried out, in which α_1 and α_2 are modified while $\beta = 20$ is fixed. Figure 25 demonstrates the effect of modifying α_1 using the content image in figure 6(c) and style image in figure 3(b), where $\alpha_2 = 20$. Using the same images, figure 26 demonstrates the effect of modifying α_2, where $\alpha_1 = 10^{-4}$.

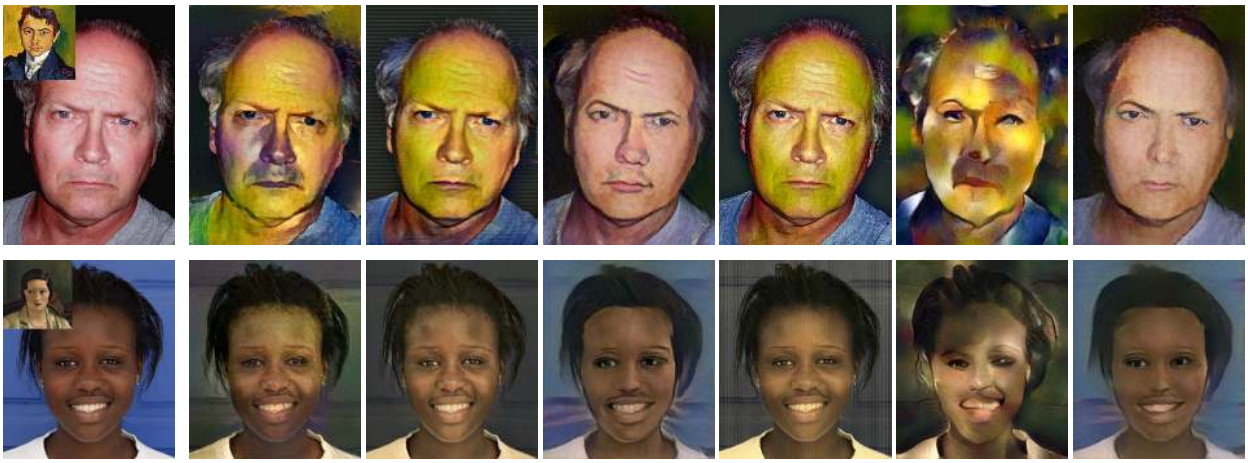
When α_1 or α_2 is too small, the result does not have sufficient content/style information. On the other hand, setting α_1 or α_2 too large may result in matched patches having poor style/content consistency. According to our experiments, $\alpha_1 \in [10^{-3}, 10^{-4}]$ and $\alpha_2 \in [20, 40]$ achieve best results.

User evaluation. In addition to visual inspection, we also performed a quantitative comparison with five existing methods. Since there is no standard automatic style transfer measure or test, we performed a user study in which the users were presented with a style image, a content image, and stylised output images from the following four methods: I [16], II [23], III [29], IV, which is our method, V [42] and VI [30].

The user study is designed using the 2AFC (Two-Alternative Forced Choice) paradigm, widely used in perceptual studies due to its simplicity and reliability. In each trial the user was asked to complete two tasks by answering the following questions:

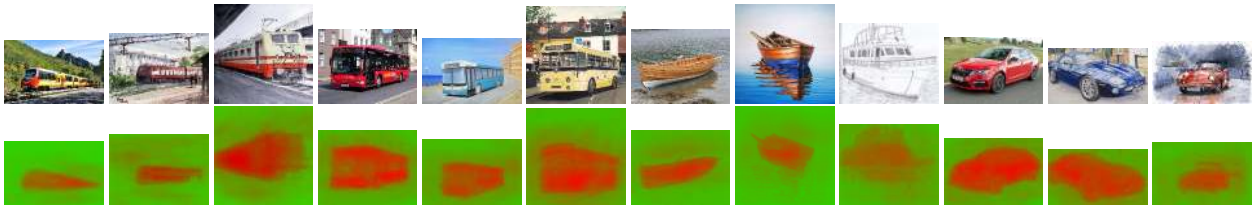
- Task 1: Given the two result images, which image better matches the target style?
- Task 2: Given the two result images, which image do you prefer?

For each question, the user can choose either of the two result images.



(a) input (b) [16] (c) [23] (d) [29] (e) [42] (f) [30] (g) Our

Fig. 9 Portrait style transfer comparison for content images with harsh lighting and dark skin.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 10 Content (a,b) and style (c,d) images and their soft masks.

To make the comparison more meaningful while limiting the user effort to a reasonable level, we used the full set of results that were contained in figures 3, 6, 10, containing $12 \times 6 = 72$ test images (6 sets for human faces and 6 sets without human faces) and stylised results generated by the six methods. For each task, 50 users participated in the user study, with ages ranging from 17 to 54. To avoid bias, we randomised the order of image pairs shown as well as their left/right position. Altogether, results of each method are compared against $12 \times 5 = 60$ results of alternative methods. We recorded the total number of user preferences (clicks) for each method, and treat these as random variables.

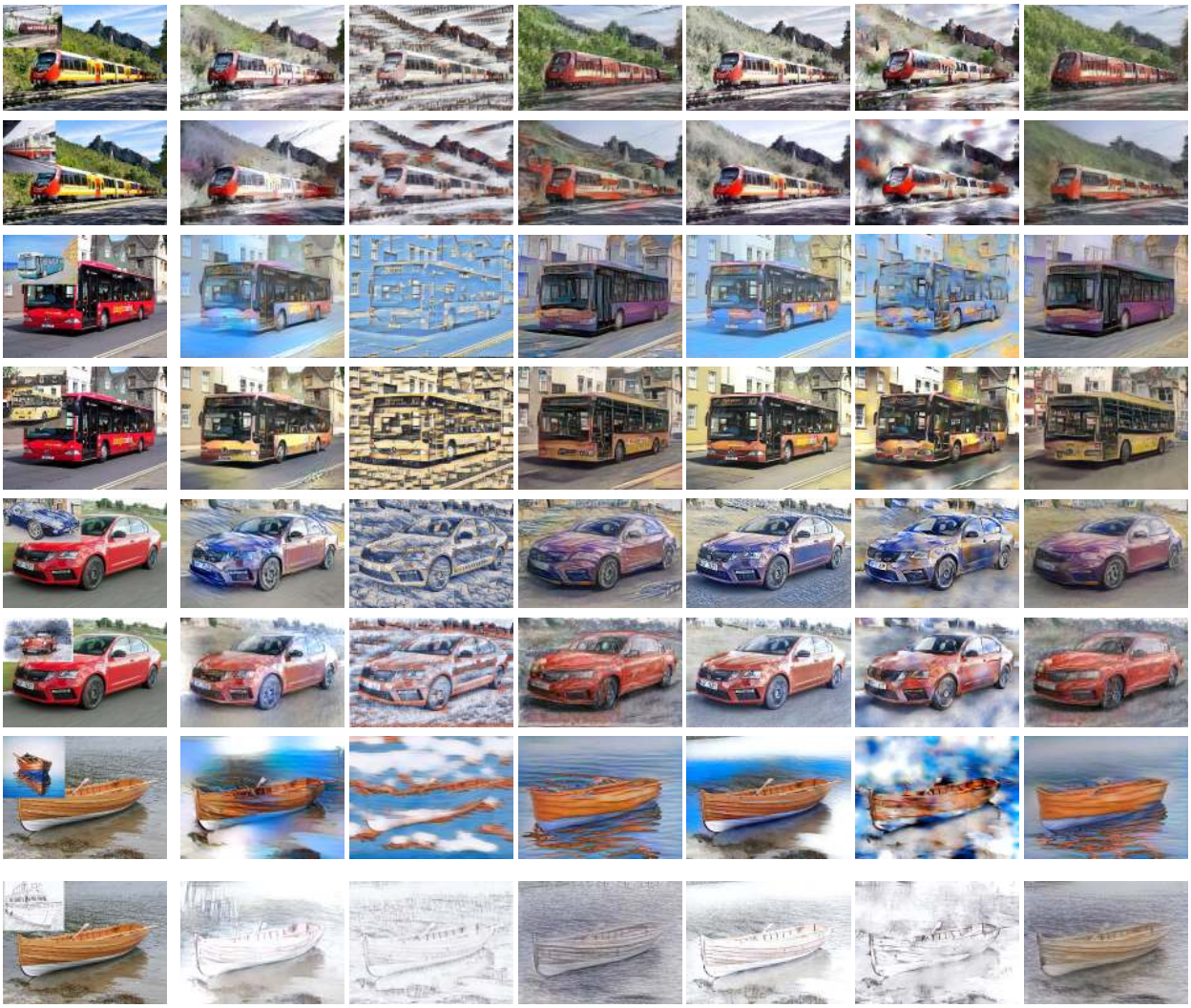
We performed the ANOVA test and results are shown in figures 27 and 28. The p-values comparing our method and alternative methods are shown in table 1. They show that the method proposed in this paper has the highest mean score and is preferred by the majority of the users. The difference between our method and alternative methods is statistically significant (at the level of 0.05).

7 Conclusions

Our paper demonstrates the benefits of automatic semantic mask extraction by combining state-of-the-art methods for both semantic segmentation and facial features. Using soft masks helps mitigate this, but there is certainly scope to improve semantic segmentation, or to develop methods dedicated to generating soft semantic masks. In most cases, soft masks can achieve better results than binary masks, especially in uncertain areas. The probability maps show the likelihood of having specific objects in the image, and can help capture elements of the styles for objects in the style image and preserve the structure of the content image. The artwork can lead to problems with general segmentation methods which are mainly intended for photographs of natural scenes. Therefore we use a different approach to extract facial skin for artworks, as compared to photographs, of people. However, if the artwork is so highly abstracted that automatic segmentation is impossible or unreliable, then a semi-automatic approach to segmentation should be used.

There remain some areas with scope for improvement, which suggests the following future work:

- Fine tuning the weights of the semantic masks can be used to achieve different sylisations. In the future we will carry out more extensive experiments to 1/ determine which semantic weights produce the best style transfer results, and 2/ investigate the relationship between semantic weights and the input images.
- The probability maps show the likelihood of having specific objects in the image, and can help capture elements



(a) input (b) [16] (c) [23] (d) [29] (e) [42] (f) [30] (g) Our

Fig. 11 Objects style transfer comparison.

Table 1 The p-value of the ANOVA test of the proposed method against the other methods for both tasks

method	I [16]	II [23]	III [29]	V [42]	VI [30]
Task 1	8.0319e-05	0.0102	0.0172	0.0041	0.0129
Task 2	8.3080e-07	0.0163	0.0224	0.0304	0.0025

of the styles for objects in the style image and preserve the structure of the content image. Therefore, this approach can be applied to improving special applications for which this is a requirement, such as makeup transfer.

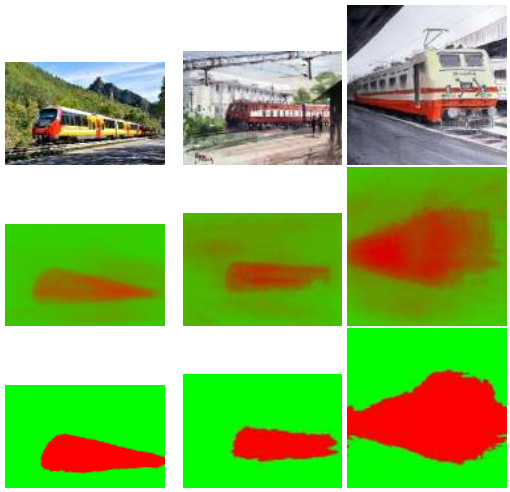
– Although our method is robust to minor segmentation errors, better segmentation would obviously lead to better stylisation results, as more appropriate patches will be matched and chosen, leading to fewer faulty instances of style transfer.

In the future, we will test the effectiveness of using the most recent segmentation methods to obtain semantic maps.

Our current method can only perform 2D image style transfer, not geometric style transfer. It is an interesting future direction which we will investigate as future work.

Acknowledgements

This work was supported by National Natural Science Foundation of China (61503128), Science and Technology Plan Project of Hunan Province (2016TP1020), Scientific Research Fund of Hunan Provincial Education Department (16C0226, 18A333), Hengyang guided science and technology projects and Application-oriented Special Disciplines (Hengkefa [2018]60-



(a) content and content mask (b) style image and style mask (c) style image and style mask

Fig. 12 Images and their masks (second row: soft masks, third row: binary masks).

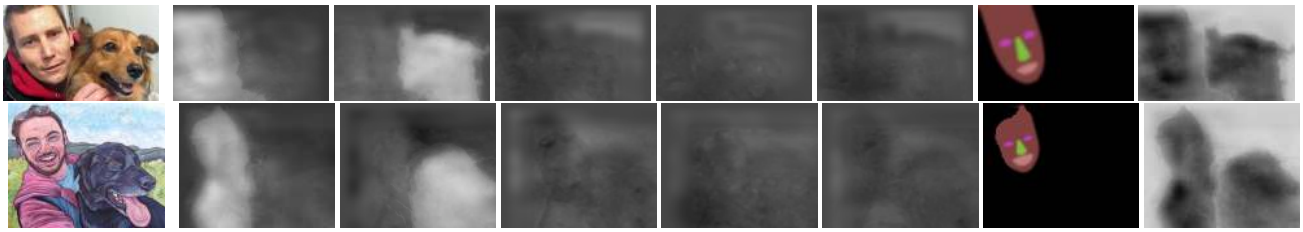
(a) (b) (c) (d)

Fig. 13 Comparison of style transfer results obtained with binary (a,c) and soft (b,d) masks.

31), Double First-Class University Project of Hunan Province (Xiangjiaotong [2018]469), Hunan Province Special Funds of Central Government for Guiding Local Science and Technology Development (2018CT5001) and Subject Group Construction Project of Hengyang Normal University (18XKQ02). We would like to thank NVIDIA for the GPU donation.

References

1. Azadi, S., Fisher, M., Kim, V., Wang, Z., Shechtman, E., Darrell, T.: Multi-content gan for few-shot font style transfer. arXiv preprint arXiv:1712.00516 (2017)
2. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: Winter Conf. on Applications of Computer Vision, pp. 1–10 (2016)
3. Brancati, N., De Pietro, G., Frucci, M., Gallo, L.: Human skin detection through correlation rules between the ycb and ycr subspaces based on dynamic color clustering. Computer Vision and Image Understanding **155**, 33–42 (2017)
4. Champandard, A.J.: Semantic style transfer and turning two-bit doodles into fine artworks. arXiv preprint arXiv:1603.01768 (2016)
5. Chang, H., Lu, J., Yu, F., Finkelstein, A.: Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
6. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing **13**(9), 1200–1212 (2004)
7. Deng, X.: Enhancing image quality via style transfer for single image super-resolution. IEEE Signal Processing Letters (2018)
8. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
9. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp. 341–346. ACM (2001)
10. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proc. Int. Conf. Computer Vision, vol. 2, pp. 1033–1038. IEEE (1999)
11. Face++: Face++. https://www.faceplusplus.com/face-detection/. Accessed April 4, 2015
12. Fišer, J., Jamriška, O., Simons, D., Shechtman, E., Lu, J., Asente, P., Lukáč, M., Sýkora, D.: Example-based synthesis of stylized facial animations. ACM Transactions on Graphics **36**(4) (2017)
13. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. International Journal of Computer Vision **40**(1), 25–47 (2000)
14. Frigo, O., Sabater, N., Delon, J., Hellier, P.: Split and match: example-based adaptive patch sampling for unsupervised style transfer. In: Proc. Conf. Computer Vision and Pattern Recognition, pp. 553–561 (2016)
15. Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 262–270 (2015)
16. 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)
17. Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling perceptual factors in neural style transfer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
18. Gilbert, A., Collomosse, J., Jin, H., Price, B.: Disentangling structure and aesthetics for style-aware image completion. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
19. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. Conf. Computer Vision and Pattern Recognition, pp. 580–587 (2014)
20. Gooch, A., Gooch, B., Shirley, P., Cohen, E.: A non-photorealistic lighting model for automatic technical illustration. In: Conference on Computer Graphics and Interactive Techniques (1998)
21. Hall, P., Cai, H., Wu, Q., Corradi, T.: Cross-depiction problem: Recognition and synthesis of photographs and artwork. Computational Visual Media **1**(2), 91–103 (2015)
22. Isenberg, T.: Visual abstraction and stylisation of maps. Cartographic Journal **50**(1), 8–18 (2013)
23. 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 (2016)
24. Kang, S.B., Kang, S.B., Kang, S.B., Kang, S.B., Kang, S.B.: Visual attribute transfer through deep image analogy. Acm Transactions on Graphics **36**(4), 120 (2017)
25. Kim, B., C. Azevedo, V., Gross, M., Solenthaler, B.: Transport-Based Neural Style Transfer for Smoke Simulations (2019). URL http://arxiv.org/abs/1905.07442
26. Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture optimization for example-based synthesis. ACM Transactions on Graphics (ToG) **24**(3), 795–802 (2005)
27. Kwatra, V., Schödl, A., Essa, I., Turk, G., Bobick, A.: Graphcut textures: image and video synthesis using graph cuts. In: ACM Transactions on Graphics (ToG), vol. 22, pp. 277–286. ACM (2003)



(a) image (b) 1st mask (c) 2nd mask (d) 3rd mask (e) 4th mask (f) 5th mask (g) head masks (h) background

Fig. 14 Object style transfer with automatic probability map selection. (a) content and style images, (b)–(f) the automatically selected top 5 semantic masks, (g) head masks, (h) background mask.

(a) input (b) [16] (c) [23] (d) [29] (e) [42] (f) [30] (g) Our

Fig. 15 Multi-object style transfer.

(a) input (b) [16] (c) [23] (d) [29] (e) [42] (f) [30] (g) Our $\beta = 20$ (h) Our $\beta = 25$

Fig. 16 Multi-object style transfer with different image composition and poses.

(a) (b) (c)

Fig. 17 Content (a) and style (b,c) images and their soft masks.

28. Lerotic, M., Chung, A.J., Mylonas, G., Yang, G.Z.: pq -space Based Non-Photorealistic Rendering for Augmented Reality (2007)
29. Li, C., Wand, M.: Combining markov random fields and convolutional neural networks for image synthesis. In: Proc. Conf. Computer Vision and Pattern Recognition, pp. 2479–2486 (2016)
30. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Advances in Neural Information Processing Systems, pp. 386–396 (2017)
31. Liu, L., Ouyang, W., Wang, X., Fieguth, P.W., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: A survey. CoRR **abs/1809.02165** (2018). URL http://arxiv.org/abs/1809.02165
32. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
33. Luft, T., Kobs, F., Zinser, W., Deussen, O.: Watercolor illustrations of cad data. International Symposium on Computational Aesthetics in Graphics Visualization and Imaging (2008)
34. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proc. Int. Conf. Computer Vision, pp. 1520–1528 (2015)
35. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos. In: German Conference on Pattern Recognition, pp. 26–36. Springer (2016)
36. Selim, A., Elgharib, M., Doyle, L.: Painting style transfer for head portraits using convolutional neural networks. ACM Transactions on Graphics (TOG) **35**(4), 129 (2016)
37. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence **39**(4), 640–651 (2017)
38. Shih, Y., Paris, S., Barnes, C., Freeman, W.T., Durand, F.: Style transfer for headshot portraits. ACM Transactions on Graphics (TOG) **33**(4) (2014)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
40. Thoma, M.: A survey of semantic segmentation. arXiv preprint arXiv:1602.06541 (2016)
41. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.: Texture networks: Feed-forward synthesis of textures and stylized images. In: Int. Conf. on Machine Learning (ICML) (2016)
42. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: Maximizing quality and diversity in feed-forward styliza-

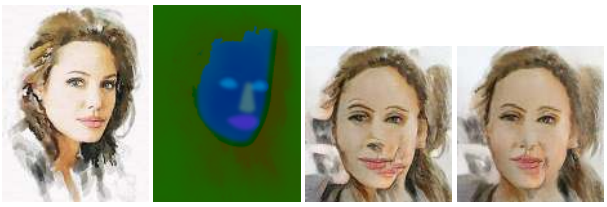
(a) input (b) [16] (c) [23] (d) [29]

(e) input (f) [42] (g) [30] (h) Our

Fig. 18 Multi-object style transfer.

(a) Our result (b) Luminance-only style transfer (c) Colour histogram matching

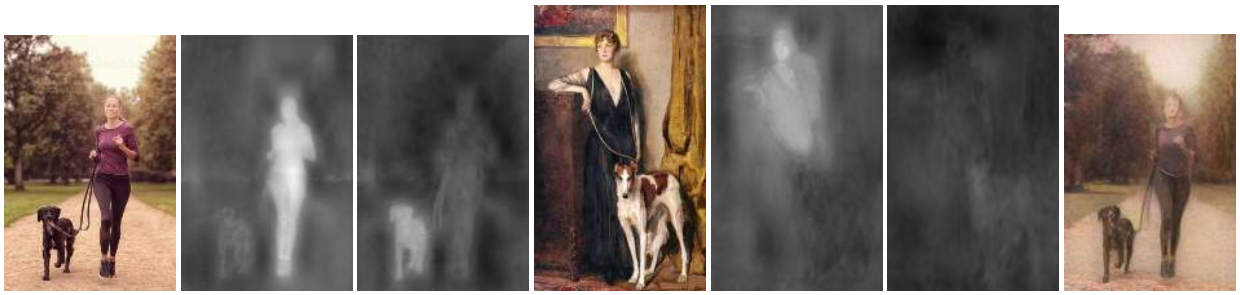
Fig. 19 Colour control results.



(a) style image (b) soft mask (c) [29] (d) Our

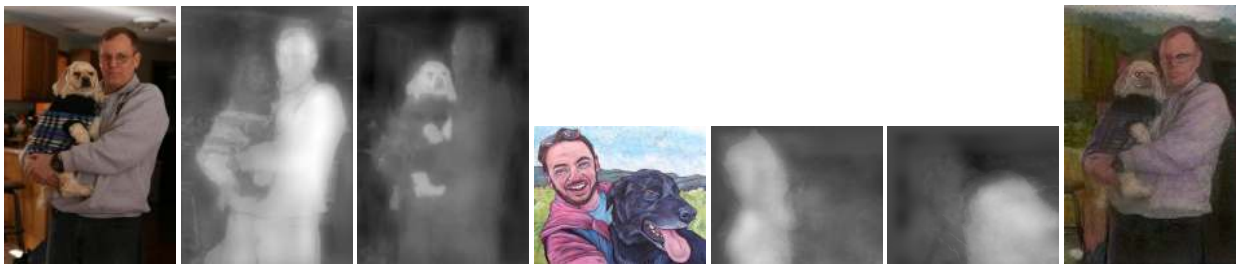
Fig. 20 Result showing the effects on style transfer with faulty semantic segmentation.

tion and texture synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, p. 6 (2017)
43. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: Conf. Computer Vision and Pattern Recognition, pp. 2217–2224. IEEE (2011)
44. Wei, L.Y., Levoy, M.: Fast texture synthesis using tree-structured vector quantization. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pp. 479–488. ACM Press/Addison-Wesley Publishing Co. (2000)
45. Yang, Y., Zhao, H., You, L., Tu, R., Wu, X., Jin, X.: Semantic portrait color transfer with internet images. Multimedia Tools and Applications **76**(1), 523–541 (2017)
46. Zhang, H., Dana, K.: Multi-style generative network for real-time transfer. arXiv preprint arXiv:1703.06953 (2017)
47. Zhang, W., Cao, C., Chen, S., Liu, J., Tang, X.: Style transfer via image component analysis. IEEE Transactions on Multimedia **15**(7), 1594–1601 (2013)
48. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du: Conditional random fields as recurrent neural networks. In: Proc. Int. Conf. Computer Vision, pp. 1529–1537 (2015)



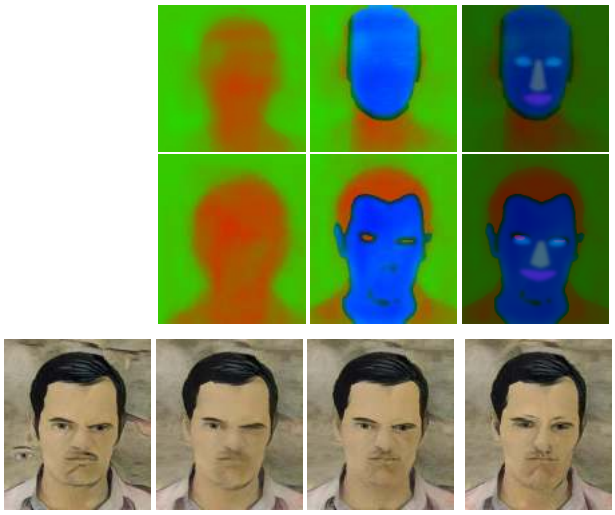
(a) content image (b) woman (c) dog (d) style image (e) woman (f) dog (g) Output

Fig. 21 Result showing the effects of stylisation with a missing style object in semantic segmentation.



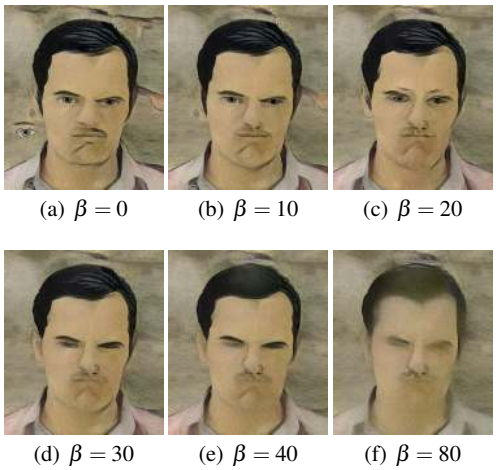
(a) content image (b) man (c) dog (d) style image (e) man (f) dog (g) Output

Fig. 22 Result showing the effects of style transfer with a missing content object in semantic segmentation.



(a) no mask (b) 2 masks (c) 3 masks (d) 6 masks

Fig. 23 Result showing the effects on style transfer with an increasing number of labelled objects in the soft masks.



(a) $\beta = 0$ (b) $\beta = 10$ (c) $\beta = 20$

(d) $\beta = 30$ (e) $\beta = 40$ (f) $\beta = 80$

Fig. 24 Result showing the effects of varying parameter β.

(a) $\alpha_1 = 10^{-1}$ (b) $\alpha_1 = 10^{-2}$ (c) $\alpha_1 = 10^{-3}$

(d) $\alpha_1 = 10^{-4}$ (e) $\alpha_1 = 10^{-5}$ (f) $\alpha_1 = 10^{-6}$

Fig. 25 Result showing the effects of varying parameter α_1.

(a) $\alpha_2 = 10$ (b) $\alpha_2 = 20$ (c) $\alpha_2 = 40$

(d) $\alpha_2 = 60$ (e) $\alpha_2 = 80$ (f) $\alpha_2 = 100$

Fig. 26 Result showing the effects of varying parameter α_2.

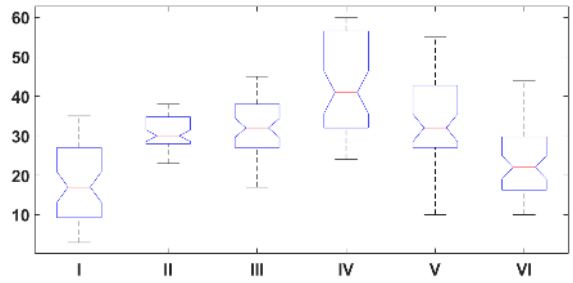


Fig. 27 Boxplots of user preferences for six different style transfer methods in task 1, showing the mean (red lines), quartiles (blues lines), and extremes (black lines) of the distributions.

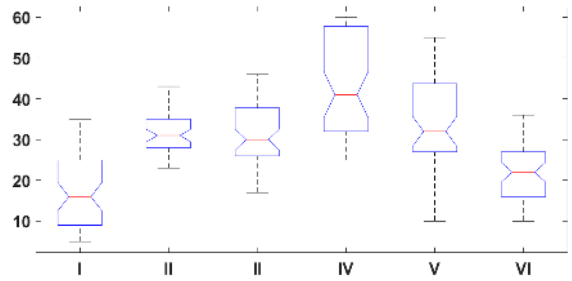


Fig. 28 Boxplots of user preferences for four different style transfer methods in task 2, showing the mean (red lines), quartiles (blues lines), and extremes (black lines) of the distributions.