Semantic Context-Aware Image Style Transfer

Yi-Sheng Liao and Chun-Rong Huang, *Senior Member, IEEE*

***Abstract*— To provide semantic image style transfer results which are consistent with human perception, transferring styles of semantic regions of the style image to their corresponding semantic regions of the content image is necessary. However, when the object categories between the content and style images are not the same, it is difficult to match semantic regions between two images for semantic image style transfer. To solve the semantic matching problem and guide the semantic image style transfer based on matched regions, we propose a novel semantic context-aware image style transfer method by performing semantic context matching followed by a hierarchical local-to-global network architecture. The semantic context matching aims to obtain the corresponding regions between the content and style images by using context correlations of different object categories. Based on the matching results, we retrieve semantic context pairs where each pair is composed of two semantically matched regions from the content and style images. To achieve semantic context-aware style transfer, a hierarchical local-to-global network architecture, which contains two sub-networks including the local context network and the global context network, is proposed. The former focuses on style transfer for each semantic context pair from the style image to the content image, and generates a local style transfer image storing the detailed style feature representations for corresponding semantic regions. The latter aims to derive the stylized image by considering the content, the style, and the intermediate local style transfer images, so that inconsistency between different corresponding semantic regions can be addressed and solved. The experimental results show that the stylized results using our method are more consistent with human perception compared with the state-of-the-art methods.**

***Index Terms*— Semantic image style transfer, image style transfer, semantic context matching, hierarchical local-to-global network, deep learning.**

I. Introduction

IMAGE style transfer aims to change strokes, textures, and colors of a *content* image to those of a *style* image. For high-quality image style transfer, object boundaries and scene structures of the content image should be preserved while the appearances are required to be aligned with the style image. To this end, matching content and style images is essential to image style transfer. Conventional methods for image style transfer apply image- or patch-level deep feature matching between the content and style images, and transfer the learned styles from the latter image to the former one. However, when content regions and style regions are not correctly matched, e.g., matching a building in the content image to a tree in the style image, semantically incoherent transfer typically makes the resulting stylized image inconsistent with human perception.

To address this issue, semantic image style transfer methods [1], [2] transfer each style region to a corresponding content region based on image matching. The matched results guide the semantic image style transfer for generating stylized results which are more consistent with human perception and understanding. These methods typically assume that the content and style images share the same semantic objects. For example, the styles of a face painting image are transferred to a real face of a content image. When semantic categories of the content and style images are not the same, the matching results may lead to unpredictable or sub-optimal performance. In other words, these methods only work when the content and style images contain the same object categories. In addition, how to effectively transfer detailed strokes of the style regions to their corresponding content regions remains a problem.

To solve the problem of unmatched object categories between the content and style images in semantic image style transfer, we propose semantic context matching to identify corresponding semantic regions between these two images. Here, the context [3] represents the co-occurrence relationships among objects and stuff in the environments. For example, buildings and streets often appear jointly, while mountains usually accompany the sky in images. In the cases where the content and style images do not share the same semantic object categories, we suggest transferring styles between object categories of high co-occurrence relationships. The contextual coherence of these categories makes style transfer among them less inconsistent with human perception. To represent the contextual coherence, we propose the contextual co-occurrence which is calculated from [4] to represent the semantic context relationship between different object categories, e.g., mountains and the sky. Semantic context matching is performed based on the contextual co-occurrence to obtain semantic context pairs, i.e., semantic corresponding regions in the content and style images even though no common object categories exist.

Semantic context matching enables style transfer across better aligned local regions, but local style transfer may lead to region-wise inconsistency in the resultant stylized images. To address this issue, we develop a novel hierarchical local-to-global network architecture which can transfer local styles for each corresponding semantic region while maintaining

Manuscript received May 16, 2021; revised November 4, 2021; accepted January 20, 2022. Date of publication February 10, 2022; date of current version February 15, 2022. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST110-2221-E-005-070, Grant MOST110-2634-F-006-022, and Grant MOST110-2327-B-006-006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guo-Jun Qi. *(Corresponding author: Chun-Rong Huang.)*

The authors are with the Department of Computer Science and Engineering, National Chung Hsing University, Taichung 402, Taiwan (e-mail: g107056049@mail.nchu.edu.tw; crhuang@nchu.edu.tw).

Digital Object Identifier 10.1109/TIP.2022.3149237

global consistency of the stylized image. The hierarchical local-to-global network architecture contains a local context network followed by a global context network. The local context network is designed to locally transfer the styles of the semantic context regions of the style image to the corresponding regions of the content image. It is driven by a local content loss and a local style loss to learn the detailed strokes of the styles from each semantic region of the style image. The learned strokes are then transferred to corresponding semantic regions of the content image. After applying this network, a local style transfer image is produced which records the style feature representations for each semantic content region of the content image.

While the local context network enhances style transfer between semantic context pairs, the boundaries among neighboring regions in the resultant stylized image may be incoherent due to diverse local style transfer. To ensure image-wise visual consistency, we develop the global context network to derive the stylized image based on the content, local style transfer, and style images. By jointly considering the three images, the obtained stylized image can better represent the detailed strokes of the style image and be more structurally consistent with the content image.

The contribution of this work is three-fold. First, semantic context matching based on contextual co-occurrence of objects is proposed to effectively match local regions between the content and style images no matter if the two images share common semantic categories or not. Second, a novel hierarchical local-to-global network architecture is developed to enhance the stokes of local semantic context pairs and generate a globally consistent stylized result. Third, our method can transfer arbitrary style images to generate stylized images which are more contextually consistent with human perception compared with the state-of-the-art methods.

The rest of this paper is organized as follows. Section II presents the related work. The proposed method is introduced in Section III. Experimental results are shown in Section IV. Finally, Section V gives the conclusions and future work.

II. Related Work

Conventional image style transfer methods [5]–[7] transfer styles between a pair of content and style images based on hand-crafted texture and color features. Thus, high-level image structures or semantic evidences are not explored in these methods. To address this issue, neural style transfer methods are proposed, which can be further categorized into two groups, including model optimization based and image optimization based methods.

A. Model Optimization Based Methods

Model optimization based methods typically train feed-forward convolutional neural networks (CNNs) offline for one or more style images. They produce a stylized image by performing a single forward pass of the trained networks and can efficiently generate the stylized image. Ulyanov *et al.* [8] train a compact feed-forward CNN for a given style image to transfer the artistic style to another image. Johnson *et al.* [9] develop a feed-forward transformation network based on perceptual loss functions provided by a pre-trained loss network for a given style image. However, to transfer the styles of a particular style image, these methods need to train a specific CNN to capture the styles. Thus, it is hard to adapt these methods to work on multiple style images simultaneously.

To solve the aforementioned issue, Chen and Schmidt [10] present a style swap layer to replace the content features with the most similar style features of the pre-trained network in a patch-by-patch manner. However, the swap becomes the computational bottleneck of their method. To address this problem, an adaptive instance normalization layer [11] is proposed to match the mean and variance of the VGG features of the content and style images. The stylized image is generated by using a decoder network. Li *et al.* [12] replace adaptive normalization with signal whitening and coloring transforms (WCTs) between the features of the content and style images. In each layer, the extracted content features are transformed to style features that have exhibited the same statistical characteristics as the content features after using WCTs. The transformed features are then fed forward into decoder layers to obtain the stylized image. To further improve [12] where only the style loss is considered, Lu *et al.* [13] seek optimal style transfer to preserve image structures by considering the content loss. Li *et al.* [14] propose a learnable linear transformation matrix based on arbitrary pairs of content and style images by two light-weighted CNNs. A linear propagation module is included to correct distortions and artifacts in the stylized results. Zhang *et al.* [15] propose using the style encoder and content encoder to extract style and content representations. Then, a mixer fuses both representations for the decoder to generate stylized result. Qiao *et al.* [16] propose a style-corpus constrained learning method by considering style-specific and style-agnostic properties at the same time and further improve photorealism of stylized results.

To preserve the structure of the content image, Gu *et al.* [17] propose feature reshuffle to spatially rearrange locations of deep features. Reshuffling deep features helps match local style patterns and enhance consistency between the content and the style images. Sheng *et al.* [18] present Avatar-Net which is a patch-based style decorator module to maintain the content structure by decorating content features with the characteristics of style patterns. However, it is not able to balance local and global style patterns. Park and Lee [19] design SANet which is similar to Avatar-Net but improves the performance by employing a learnable soft-attention-based network for the style decoration. Chen *et al.* [20] propose a multi-collection style transfer method based on the adversarial gated networks. Although multiple styles can be pre-trained in the discriminative network, semantic region correspondences between the content and pre-trained style images are not considered.

To enable local awareness of the content image, Yao *et al.* [21] develop an attention-aware multi-stroke (AAMS) model for arbitrary style transfer. Their method uses a spatial attention mechanism for matching corresponding regions between the content and style images. Thus, it can better transfer the style to the content image since the

attention map grasps salient characteristics of the content image. Lu *et al.* [22] propose a fast semantic style transfer method by conducting semantic feature fusion of the deep features of the content and style images, reconstructing feature maps within each semantic region, and decoding the feature maps to produce the stylized image. Cheng *et al.* [23] suggest using the depth map based global structure and the image edge based local structure to describe the spatial distribution of components in the content image. Two networks are employed to drive the image style transfer based on the depths [24] of the global structure and the edges of the local structure. However, when images contain multiple objects with similar depths or unclear structures, the performance degrades. Despite efficiency, these methods are hard to perceptually map an artwork style to content images if the semantic categories between the content and style images are not covered in the pre-trained models.

Compared with model optimization based methods, our method leverages semantic context matching to associate regions of the style and content images even when the two images do not share the same semantic object categories. Moreover, it extracts and transfers region-specific styles for each semantic context pair, while enforcing image-wise coherence.

B. Image Optimization Based Methods

Unlike model optimization based methods, image optimization based methods iteratively optimize the transferred image, instead of the model, based on the given content and style images until it has desired CNN representations of the both images. As a result, methods of this category often offer better image style transfer results. For example, Gatys *et al.* [25] propose to use the hierarchy of CNN [26] to separate and recombine the image content and style of a natural image. They show that CNN can learn deep image representations from an arbitrary artwork and blend the extracted style into the content of an input image. However, the pre-image search for matching feature representations of the content and style images ignores semantic content matching. Li and Wand [27] use generative Markov random fields (MRFs) models to take semantic regions into account in the feature patch level but their method only works for the content image that has similar shaped elements with the style image.

To solve the aforementioned issues, Gatys *et al.* [28] deal with image style transfer by manually controlling spatial, color and scale information to improve the quality of the stylized image. However, their method is hard to achieve meaningful parametric control over the stylization. For instance, straight edges may have spatial distortions. To prevent edge distortion, Luan *et al.* [29] constrain image style transfer only in the color space by using matting Laplacian. Their method may suffer from the loss of the style characteristics of the style image due to the constraints. Kolkin *et al.* [30] propose style transfer by relaxed optimal transport and self-similarity which allow manual control of corresponding regions between the content and style images. By using manual guidance, semantic matching errors by conventional unconstrained style transfer can be reduced.

For better semantic matching between the content and style images, Champandard [31] presents a semantic style transfer algorithm based on patches with semantic segmentation maps and focuses on common object categories appearing in both content and style images. Mechrez *et al.* [32] design a contextual loss to avoid the alignment of the content and the style images during training. However, the content and style images need to have the same semantic regions for the style transfer. Park *et al.* [1] utilize word embedding to compute similarity of the semantic regions between the content and style images. Their method applies style transfer to only the matched regions, which may lead edge distortion and discontinuity between neighboring regions. Furthermore, the assumption of sharing common semantic categories in content and style images reduces the applicability of these methods. Kim *et al.* [2] propose deformable style transfer based on the assumption that the content and style images have approximate alignment. However, when poor matching occurs, the styles are hard to be correctly transferred to the content image. Thus, providing good corresponding matches between the content and style images can significantly affect the quality of the stylized results. For detailed reviews, please refer to [33].

Compared with recent image optimization based methods, our method can effectively match semantic context regions between the content and style images when their semantic object categories do not overlap. Moreover, the local-to-global style transfer helps deliver the specific styles of each corresponding semantic region and achieves globally consistent results, which can reduce artifacts and provide visual consistent results with respect to human perception.

III. Proposed Method

This section describes the proposed method for semantic context-aware image style transfer. Fig. 1 gives an overview of our method, which is composed of three components, including semantic context matching (blue-shaded area), the local context network (green-shaded area), and the global context network (yellow-shaded area). Semantic context matching identifies corresponding regions across the content and style images no matter if the same semantic object categories are shared between the two images or not. The local context network is developed to locally transfer the style between the corresponding regions. The global context network is designed to ensure global image coherence of the stylized image after region-specific transfer. The three components are elaborated as follows.

A. Semantic Context Matching

Given a content image I_c and a style image I_s, image style transfer aims to derive an output stylized image, whose content is similar to I_c while its style is similar to I_s. Some advanced methods, e.g. [21], [28], leverage manually defined semantic information to transfer styles across regions of the same object categories, considerably improving the quality of the stylized image. Nevertheless, automatically obtaining the semantic object categories of the content image is desired. In our approach, the PSPNet [34], a semantic segmentation

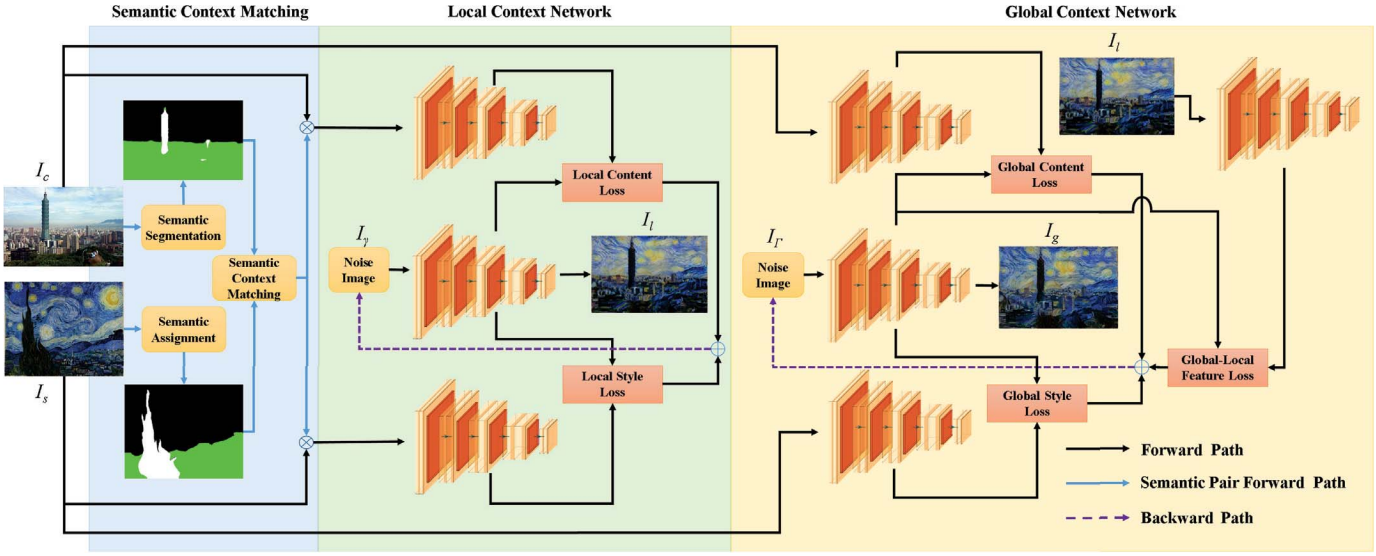

Fig. 1. The overview of our method. Semantic segmentation results of the content image serve as the context information. By applying semantic context matching between the content and style images, we build semantic context pairs which are used to transfer detailed styles of the semantic regions of the style image to the corresponding regions of the content image. For each semantic context pair, our local context network transfers the local strokes of style regions to the corresponding content regions by considering a local content loss and a local style loss and generates the local style transfer image. With the content, style and local style transfer images, our global context network optimizes the stylized image based on a global content loss, a global style loss, and a global-local feature loss.

method, is used to automatically estimate the region-wise semantic object categories of the content and style images. With the extra category information, image style transfer can be region-wise carried out, namely transfer across regions of the same categories as shown in [1]. However, a critical problem arises when the semantic object categories of the content and style images are not exactly matched. Moreover, the style transfer with respect to corresponding semantic object categories may lead to the visual incoherent problem between the stylized results of different object categories.

To address the first problem, we propose to use semantic context relaxation, with which transfer between contextually similar categories is enabled when identical categories are not cross-image available. In this work, we consider the semantic context [3], [32] and model spatial relationships between semantic object categories in natural images. To establish the semantic context of object categories, we use the annotated images in the training set of the ADE20K dataset [4] to calculate the contextual co-occurrence $p_o(o_i, o_j)$ between object categories o_i and o_j as follows:

$$p_o(o_i, o_j) = \frac{1}{Q_j} \sum_{k=1}^{K} |\{p | p \in I_k : p \in o_i \text{ and } \mathcal{N}(p) \in o_j\}|, \tag{1}$$

where I_k is the kth annotated image, K is the number of the annotated images, $\mathcal{N}(p)$ indicates the set of the four-connected neighbors of pixel p, Q_j is the total number of pixels of o_j to reduce the effects of the numbers of pixels between different object categories, and the function $|\cdot|$ returns the size of a set. If o_i and o_j are frequently present in neighboring regions, their contextual co-occurrence $p_o(o_i, o_j)$ is high. High contextual co-occurrence between such o_i and o_j typically makes style transfer across them more consistent with human perception since it preserves the consistency of the transferred styles in the neighboring regions.

Given the content image I_c, we apply the PSPNet to obtain a set of semantic regions, i.e., $\mathcal{C} = \{C_1, \ldots, C_r, \ldots, C_{N_c}\}$, where C_r is the rth semantic region of I_c and N_c is the number of semantic regions of I_c. Please note that we assign each pixel to the most plausible object category. Similarly, another semantic region set $\mathcal{S} = \{S_1, \ldots, S_q, \ldots, S_{N_s}\}$ is obtained by manually assigned to the style image I_s, where S_q denotes the qth semantic region of I_s and N_s is the number of semantic regions of I_s. To establish region correspondences between the content and style images, we match each semantic region C_r of the content image I_c to the semantic region $S_{\pi(r)}$ of the style image I_s via

$$S_{\pi(r)} = \underset{S_q \in \mathcal{S}}{\operatorname{argmax}}\, p_o(\mathcal{O}(C_r), \mathcal{O}(S_q)), \tag{2}$$

where the function $\mathcal{O}(\cdot)$ returns the semantic object category of the input region. It is worth mentioning that according to the definition of the contextual co-occurrence in Eq. (1), for each object category o_i, the maximal value of $p_o(o_i, o_j)$ typically presents when object categories o_i and o_j are the same. It follows that the content region C_r and its matched style region $S_{\pi(r)}$ via Eq. (2) belong to the same object category if the object category of $\mathcal{O}(C_r)$ is covered by the style image. Otherwise, the style region $S_{\pi(r)}$ with the maximal contextual co-occurrence is retrieved.

According to the matching process in Eq. (2), the content region C_r and the corresponding style region $S_{\pi(r)}$ are either of the same object category or contextually similar. Thus, transferring styles across the matched regions is preferable. We consider two matched regions as a semantic context pair

$P_r = (C_r, S_{\pi(r)})$ to represent the target for local style transfer. We repeat the matching process of Eq. (2) for each content region, and get the semantic context matching set $\mathcal{P} = \{P_r = (C_r, S_{\pi(r)})\}_{r=1}^{N_c}$, which then serves as the input to the local context network for region-specific style transfer.

B. Local Context Network

Given a semantic context matching set $\mathcal{P}$, we aim to synthesize a local stylized image based on each semantic context pair in $\mathcal{P}$. To individually transfer the style of $S_{\pi(r)}$ to the corresponding content region of C_r, we propose a local context network which locally transfers the styles of the semantic regions of the style image to the corresponding regions of the content image, while reducing the interference of remaining non-corresponding regions of the style image.

The learning of the network is driven by the proposed local context network loss ℓ_l as:

$$\ell_l = \lambda_{lc}\ell_{lc} + \lambda_{ls}\ell_{ls}, \tag{3}$$

where ℓ_{lc} and ℓ_{ls} are the local content loss and local style loss, respectively, λ_{lc} and λ_{ls} are the weights of the two losses. The front one aims to preserve the content structure of C_r and the latter one aims to extract the style representation of $S_{\pi(r)}$ for corresponding C_r. Based on ℓ_l, the local context network is trained from each semantic context pair in $\mathcal{P}$. For the sake of clarity, ℓ_{lc} and ℓ_{ls} are introduced in the following.

Assume that the local context network contains L layers. Each layer l contains N^l feature maps, where the size of N^l is denoted as M^l. A matrix $F^l \in \Re^{N^l \times M^l}$ is used to store the feature maps of layer l, where F^l_{ik} is the response of the ith feature map at the kth position of layer l. Let I_γ be a white noise image to match the feature responses of the content image I_c and the style image I_s. To learn the content information of each semantic region C_r, the local content loss of C_r is defined as the distances of the feature representations between C_r and I_γ as follows:

$$\ell_{lc}(C_r, I_\gamma, l) = \frac{1}{2N^l M^l}\sum_{i,k}\{(F^l_{ik} - P^l_{ik})^2 | k \in C^l_r\}, \tag{4}$$

where F^l_{ik} and P^l_{ik} are the responses of the ith feature maps of I_c and I_γ at the kth position of layer l, respectively, and C^l_r represents the positions of the feature map with respect to C_r of layer l. In this way, the local context network can learn each semantic region of I_c individually without the interference of other semantic regions. The local content loss ℓ_{lc} is defined as the summation of all losses of C_r as:

$$\ell_{lc} = \sum_{r=1}^{N_c}\ell_{lc}(C_r, I_\gamma, l), \tag{5}$$

where l is layer *conv4_2*. Compared with [25], they consider the content loss of the whole content image, which may miss the local details of the content image during learning. In contrast, our method considers the local content loss for each semantic region of I_c, which is more sensitive to the local content of I_c with respect to layer l for better preserving the content structure.

To obtain the feature representations of I_γ, we build the Gram matrix G^l for each layer l. The Gram matrix contains the correlations between different feature maps of layer l. Let G^l_{ij} be the inner product between the ith and the jth feature maps of layer l as:

$$G^l_{ij} = \sum_k\{F^l_{ik}F^l_{jk}\}. \tag{6}$$

The Gram matrix of each layer l provides style feature space to represent multi-scale texture information of I_γ. Similarly, the Gram matrix A^l_{ij} of each layer l of I_s can also be defined. Based on these style feature spaces, we can then construct an image that matches the style representation of I_s to achieve the style transfer between I_γ and I_s. This process can be done by minimizing the distances between the entries of the Gram matrices of I_γ and I_s.

To individually transfer the detailed strokes of $S_{\pi(r)}$, we propose the local style loss ℓ_{ls} as:

$$\ell_{ls} = \ell_{lws} + \ell_{lps} + \ell_{lns}, \tag{7}$$

where ℓ_{lws}, ℓ_{lps}, and ℓ_{lns} are the local whole style loss, the local positive style loss, and the local negative style loss, respectively and described in the following.

The local whole style loss ℓ_{lws} aims to learn the styles from the whole style image and is computed based on the Gram matrices of I_γ and I_s as follows:

$$\ell_{lws} = \frac{1}{(2N^l M^l)^2}\sum_l\sum_{i,j}(G^l_{ij} - A^l_{ij})^2. \tag{8}$$

Although it provides the fundamental style transfer for I_γ based on I_s, the detailed strokes of each matched semantic region of the style image may not be correctly transferred.

To specifically transfer the style of $S_{\pi(r)}$ to corresponding C_r, we propose a novel local positive style loss ℓ_{lps} based on the semantic context pairs. For each C_r, we generate its Gram matrix $G^l_{ij}(C_r)$ of layer l as follows:

$$G^l_{ij}(C_r) = \sum_k\{F^l_{ik}F^l_{jk} | k \in C^l_r\}. \tag{9}$$

Please note that $G^l_{ij}(C_r)$ is computed based on the feature maps of C_r so that the Gram matrix can represent C_r during the style transfer. To represent the style of $S_{\pi(r)}$, we also generate the Gram matrix of $S_{\pi(r)}$ as:

$$A^l_{ij}(S_{\pi(r)}) = \sum_k\{A^l_{ik}A^l_{jk} | k \in S^l_{\pi(r)}\}, \tag{10}$$

where $S^l_{\pi(r)}$ represents the positions of the feature maps with respect to $S_{\pi(r)}$ in layer l. The loss function $\ell_{lps}(C_r, S_{\pi(r)})$ aims to enhance the local style transfer of the semantic context pairs P_r based on the object context correlations by using the Gram matrices of C_r and $S_{\pi(r)}$ as:

$$\ell_{lps}(C_r, S_{\pi(r)}) = \sum_{i,j}(G^l_{ij}(C_r) - A^l_{ij}(S_{\pi(r)}))^2. \tag{11}$$

By minimizing $\ell_{lps}(P_r)$, the detailed strokes of $S_{\pi(r)}$ can be directly transferred to C_r in I_γ. By accumulating the losses computed based on the semantic context matching set $\mathcal{P}$ and

the network layers, the local positive style loss ℓ_{lps} is defined as:

$$\ell_{lps} = \frac{1}{(2N^l M^l)^2} \sum_{l \in L_s} \sum_{\mathcal{P}} \ell_{lps}(C_r, S_{\pi(r)}), \tag{12}$$

where L_s is the set of layers *conv3_1*, *conv4_1* and *conv5_1* of the local context network for high level feature representations of the style image. By considering ℓ_{lps}, the styles of $S_{\pi(r)}$ can be strongly emphasized and transferred to the semantic region C_r in I_γ.

When ℓ_{lws} aims to minimize the style differences between I_γ and I_s based on the Gram matrices, it implies that the styles of non-corresponding semantic regions of C_r will also be transferred to C_r in I_γ. To reduce the interference of the non-corresponding semantic regions of C_r in $\mathcal{P}$, we propose a local negative style loss ℓ_{lns}. Let $\overline{S}_{\pi(r)} = \{\mathcal{S} \setminus S_{\pi(r)}\}$ be the set of semantic regions that are not matched to C_r. The Gram matrix of $\overline{S}_{\pi(r)}$ is defined as follows:

$$A^l_{ij}(\overline{S}_{\pi(r)}) = \sum_k \{A^l_{ik} A^l_{jk} | k \in \overline{S}_{\pi(r)}\}. \tag{13}$$

The local negative style loss $\ell_{lns}(C_r, \overline{S}_{\pi(r)})$ for each C_r in layer l is defined as:

$$\ell_{lns}(C_r, \overline{S}_{\pi(r)}) = -\sum_{i,j} (G^l_{ij}(C_r) - A^l_{ij}(\overline{S}_{\pi(r)}))^2, \tag{14}$$

where the minus of the loss represents the dissimilarity between the Gram matrices of C_r and $\overline{S}_{\pi(r)}$ to reduce the effects of the style transfer from $\overline{S}_{\pi(r)}$ to C_r in I_γ. The loss ℓ_{lns} is defined as:

$$\ell_{lns} = \frac{1}{(2N^l M^l)^2} \sum_{l \in L_s} \sum_{\mathcal{P}} \ell_{lns}(C_r, \overline{S}_{\pi(r)}). \tag{15}$$

Similar to ℓ_{lps}, we also consider high level feature representations of the style image for computing the loss. By considering ℓ_{lns}, the interference of the styles of the non-corresponding regions will be reduced. The derivative of the loss functions can be computed as shown in [25]. Please note that the derivative is used to iteratively update I_γ until the features of the semantic regions of I_γ can be simultaneously matched to the content features of I_c and style features of I_s. The parameters of the networks used to extract features of I_c, I_s, and I_γ are fixed. Different from previous methods, our local style transfer ensures that detailed strokes of corresponding regions of I_s can be reserved for better style representation based on each semantic context pair. After learning, it generates a local style transfer image I_l as the reference for the global context network.

C. Global Context Network

Based on the local context network, the local style transfer image I_l contains the detailed strokes based on the semantic context matching set $\mathcal{P}$. However, given two neighbor semantic regions of the content image, their neighbor corresponding regions of the style image may not be spatially connected. If the styles of neighbor corresponding regions are significantly different, the strokes of the transferred styles for neighbor regions will become incoherent. As a result, the local style transfer image usually contains visual incoherent edges between the neighbor regions of different object categories.

To solve the visual incoherent problem, we propose a global context network to learn visually consistent transfer results and maintain detailed strokes of styles based on the content image, the style image and the local style transfer image. The global context network is driven by the proposed global loss ℓ_g as follows:

$$\ell_g = \lambda_{gc}\ell_{gc} + \lambda_{gs}\ell_{gs} + \lambda_{gl}\ell_{gl}, \tag{16}$$

where λ_{gc}, λ_{gs} and λ_{gl} are the weights of the global content loss ℓ_{gc}, the global style loss ℓ_{gs}, and the global-local feature loss ℓ_{gl}, respectively. In the global context network, the global content loss and the global style loss give the global transfer constraints to compromise visual incoherent edges between the neighbor regions of different object categories in the local style transfer image which stores the detailed strokes of each semantic context pair. By simultaneously minimizing all of the three losses, the generated stylized image can be globally similar to the content image and contain detailed strokes learning from the style image. For the sake of clarity, we will introduce each loss in the following.

Let I_Γ be a white noise image to match the feature responses of the content image I_c, the style image I_s, and the local style transfer image I_l. The global content loss ℓ_{gc} aims to preserve the global content structure of the stylized image based on the whole content image. We use layer *conv2_2*, which can better represent the boundary structure of the objects in the content image, to compute ℓ_{gc} as:

$$\ell_{gc} = \frac{1}{2N^l M^l} \sum_{i,k} (F^l_{ik} - P^l_{ik})^2, \tag{17}$$

where F^l_{ik} and P^l_{ik} are the responses of the ith feature maps of I_c and I_Γ at the kth position in layer *conv2_2* of the global context network.

The global style loss function aims to extract the styles of the whole style image and transfer the styles to I_Γ. To ensure that all of the styles in different layers can be transferred to the I_Γ, we use the Gram matrix of all of the layers for the global style transfer as follows:

$$\ell_{gs} = \frac{1}{(2N^l M^l)^2} \sum_l \sum_{i,j} (G^l_{ij} - A^l_{ij})^2, \tag{18}$$

where G^l_{ij} and A^l_{ij} are the Gram matrices built from the ith and the jth feature maps in layer l of I_Γ and I_s, respectively.

Based on ℓ_{gc} and ℓ_{gs}, the global content structure and styles can be learned for I_Γ. However, these two losses do not consider the transfer of detailed strokes with respect to semantic context pairs. Because the detailed styles of each semantic context pair are learned in the local style transfer image, the final style transfer results should also be similar to the local style transfer image, which correctly represents the style transfer results of the semantic context pairs. In other word, the feature representations of the final stylized image should also be similar to the those of the local style transfer

image. To achieve the goal, we propose a novel global-local feature loss ℓ_{gl} based on I_l as:

$$\ell_{gl} = \frac{1}{2N^l M^l} \sum_{i,k} (O^l_{ik} - P^l_{ik})^2. \quad (19)$$

where l is layer *conv4_2* and O^l_{ik} is the ith feature responses of position k of I_l. Among the layers, layer *conv4_2* preserves more semantic and structure details of the local style transfer image. Thus, we select layer *conv4_2* as the feature representations of the local style transfer image during computing ℓ_{gl}. By using ℓ_{gc}, ℓ_{gs}, and ℓ_{gl}, the visual incoherent problem can be solved and the detailed strokes of the style regions can be successfully transferred.

Please note that the derivative is used to iteratively update l_Γ until the features of the semantic regions of l_Γ can be simultaneously matched to the content features of I_c, style features of I_s, and the content features of I_l. The parameters of the networks used to extract features of I_c, I_s, I_l and I_Γ are fixed. By using the global context network, the styles of corresponding semantic regions between the content and style images can be successfully transferred. As a result, the problems of matching semantic regions between the content and style images and the visual incoherence between neighbor regions can be solved by the proposed method.

IV. Experimental Results

A. Experimental Settings

In the experiments, the content images were from MS-COCO dataset [35], [2] and the collected photos. The style images were provided by [25] and [2] for evaluation. To perform semantic segmentation on the content images, we trained PSPNet [34] on the ADE20K dataset [4], which contains 150 object categories.

The deep models in the local context network and the global context network were the pre-trained VGG-19 models [36]. In the local context network, we used the features of layer *conv4_2* to compute the local content loss ℓ_{lc}, while we used the features of layers *conv3_1*, *conv4_1* and *conv5_1* to compute the local style loss ℓ_{ls}. In the global context network, we used the features of layer *conv2_2* to compute the global content loss ℓ_{gc}, the features of layers *conv1_1*, *conv2_1*, *conv3_1*, *conv4_1* and *conv5_1* to compute the global style loss ℓ_{gs}, and the features of layer *conv4_2* to compute the global-local feature loss ℓ_{gl}. The weights λ_{lc} and λ_{ls} of the local network were set to 0.002 and 1 which learned more detailed styles of semantic regions of I_s. The weights λ_{gc}, λ_{gs} and λ_{gl} were set to 0.2, 1, and 0.2 which generated visual coherent style transfer results. We used Adam [37] to optimize the stylized images in the local context network and global context network.

Our method was implemented on an Inter I7 CPU computer with a GTX1060 GPU. Given a pair of the content and style images, the average learning time was around 9 minutes to obtain the final stylized image. Please note that the average learning time is similar to [25] under the same environment.

B. Ablation Study

In the ablation study, we show the effects of the local context network, and the loss functions of the global context network to evaluate the proposed schemes. Fig. 2(a) shows the content and style images. The semantic masks of the content and style images are shown in Fig. 2(b). In the local context network, each semantic context pair is used to individually transfer the style of the semantic region of I_s to the corresponding content region of I_c. Although the styles are transferred, visual incoherent boundaries will appear between neighbor regions due to the over-emphasized strokes as shown in Fig. 2(c). Thus, the local style images serve as stroke feature representations of the semantic context pairs for the global context network in our framework. However, if we directly add ℓ_{gs} and ℓ_{gc} to the local context network, the transferred strokes of each style region and the learned content structure of each corresponding content region of the local style transfer image will be affected by the remaining semantic regions in I_s and I_c. Thus, the detailed strokes of each semantic context pair may not be well learned due to the interference of adding ℓ_{gs} and ℓ_{gc}. The stylized results of the local context network with ℓ_{gs} and ℓ_{gc} are shown in Fig. 2(d), where unexpected noise appears and leads to worse visual quality. To solve this problem, our method considers the hierarchical local-to-global network structure to help generate high quality style transfer results.

Three loss functions are employed in the global context network including the global content loss, the global style loss, and the global-local feature loss. Fig. 2(e) shows the stylized results without the global content loss. As shown in the row 1 of Fig. 2(e), the pupil of the left eye becomes very unclear due to the lack of the global content loss which aims to maintain the structure of the content image. Similar situations can also be observed in the building and the tree images (e.g. rows 2 and 3). The stylized results without the global style loss are shown in Fig. 2(f). Although the content structure is correctly preserved, the styles cannot be transferred to the content images. Moreover, the stylized results still maintain partial color tones of the content image.

By only considering the global content loss and the global style loss, the detailed strokes of the style image are not correctly learned as shown in Fig. 2(f) due to the lack of semantic matching information between the content and style images. For example, the style of the sea in the style image is transferred to the sky of the building image in row 2 of Fig. 2(g). Similar non-semantic transferred results can also be observed in the tree image in row 3 of Fig. 2(g). The strokes of the object in the style image are transferred to the backgrounds of the tree image. In contrast, by simultaneously considering all of the three losses, not only the detailed strokes can be properly transferred, but also the original content structure can be well preserved as shown in Fig. 2(h). The styles of the houses in the style image are properly transferred to the building in the content image in row 2 of Fig. 2(h). As shown in row 3 of Fig. 2(h), the styles of the foreground objects of the style images are transferred to the foreground trees of the content image, while the styles of the backgrounds of the style

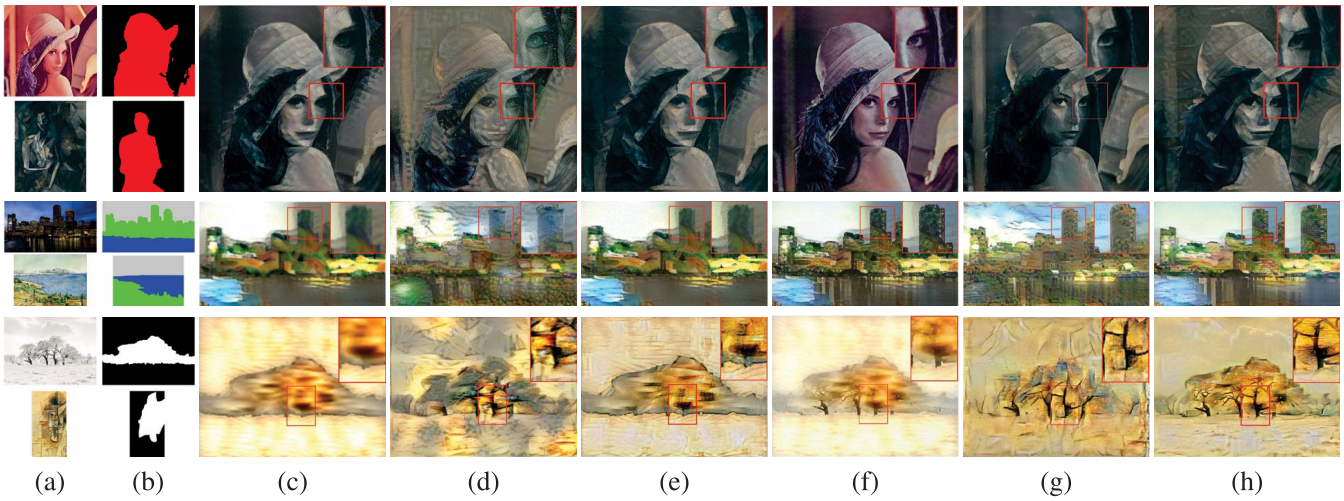

Fig. 2. The stylized results of the ablation study. (a) Content and style images, (b) Semantic masks of content and style images, (c) The local style transfer image, (d) The stylized results of the local context network with the global content loss and global style loss, (e) The stylized results without the global content loss, (f) The stylized results without the global style loss, (g) The stylized results without the global-local feature loss, and (h) The stylized results of the proposed method.

images are transferred to the backgrounds of the content image. As a result, the effectiveness of the proposed schemes is proven.

C. Qualitative Results

We compared the proposed method with four state-of-the-art image optimization based methods [2], [25], [28], [29]. In [25], the image style transfer is achieved by the image optimization based on the whole content and whole style images. To enhance local strokes of the style image, Gatys *et al.* [28] improves [25] by considering spatial region matching between the content and style images. Luan *et al.* [29] constrain image style transfer based on matting Laplacian for real photo style transfer. In [2], Kim *et al.* a propose a geometry-aware stylization method via deformation based on keypoint matching results of the content and style images. To show the advantages of image optimization based methods, we also compared the proposed method with two state-of-the-art model optimization based methods [12], [21]. Please note that the stylized results of the competing methods were obtained by using the codes provided by their authors.

Fig. 3 shows the image style transfer results of the competing methods and the proposed method. As shown in Fig. 3(a), the content and style images contain similar object categories. For example, the rows 1 and 2 show the style transfer results between two portraits. The rows 3, 4 and 5 show the style transfer results between sky, buildings and water. The rows 6, 7, 8 and 9 show the style transfer results between the same objects, i.e. cars, horses, and owls. The corresponding semantic regions are shown in Fig. 3(b). The same colors of the semantic regions represent the matched semantic context pairs between the content and style images. For fair comparison, the semantic regions are also applied for the competing methods if these methods can also apply the information.

Fig. 3(c), (d), (e), (f), (g), (h) and (i) show the stylized results of [25], [28], [29], [12], [21], [2] and our method, respectively. Because the image optimization based method [25] only considers the mapping of the whole content image with respect to the whole style image, the styles of the portraits are incorrectly transferred to the backgrounds of the content images (e.g. rows 1, 2, 6 and 7) as shown in Fig. 3(c). Moreover, the styles of the water are transferred to the sky of the content image (e.g. rows 3 and 5) as shown in Fig. 3(c). Without semantic matching regions, [25] cannot achieve transfer results which are consistent with human perception. To improve [25], [28] considers transferring styles for each segmented rectangular region of the content image. Since it still extracts the style representations of the whole style image, different objects in the content images are still stylized by the same style representation. Thus, the transferred results of the background (e.g. rows 1, 2, 6 and 7), and the water (e.g. rows 3 and 5) contain unnatural noise as shown in Fig. 3(d).

In [29], although the photorealism regularization term helps constrain the structure of the content image, it fails to transfer the styles to the stylized images even under the guidance of the semantic segmentation results. In addition, the styles of the artworks are not correctly transferred to the content images as shown in Fig. 3(e). Although the WCTs in [12] are fast to compute, the trained models are hard to properly represent the semantic correlations between the content and style images. As a result, unnatural noise can be observed in the backgrounds (e.g. rows 1, 2, 5, 6 and 7) as shown in Fig. 3(f). To solve the semantic correspondence problem, attention maps of objects are introduced in [21]. However, when the content regions incorrectly match to the style regions due to the attention maps, the strokes of the style regions are incorrectly synthesized to the content images as shown in Fig. 3(g).

To address the matching issue, geometry-aware deformation between the content and style images is considered in [2] as shown in Fig. 3(h). Although the content and style images contain similar objects, the stylized results (e.g. rows 2, 3, 4,

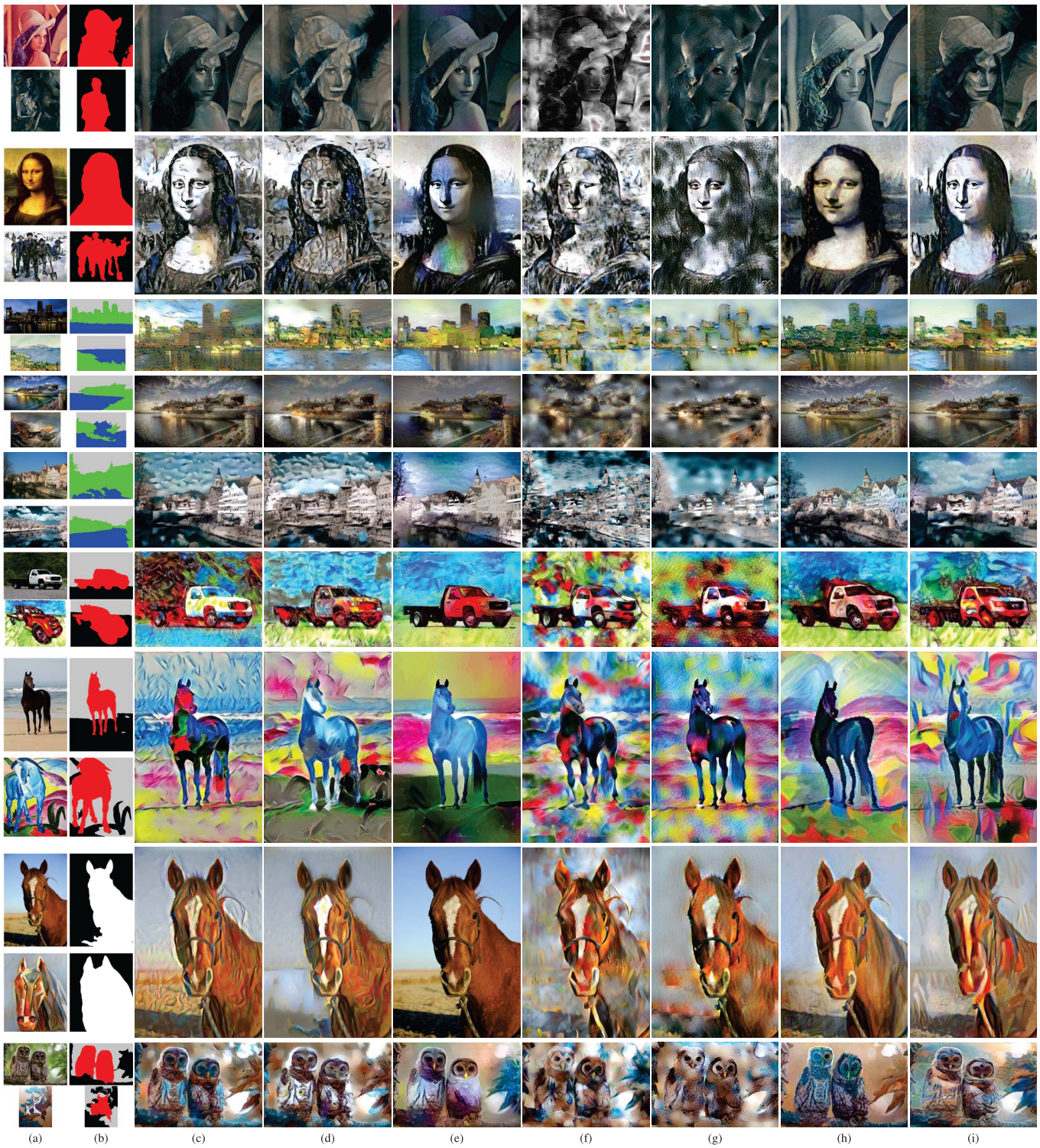

Fig. 3. Qualitative comparisons between our method and competing methods when the content and style images contain similar object categories. (a) Content and style images, (b) Semantic masks of content and style images, (c) Results of [25], (d) Results of [28], (e) Results of [29], (f) Results of [12] (g) Results of [21], (h) Results of [2] and (i) Results of our method.

6 and 7) may still contain deformation which implies that the content structure is not well preserved. Moreover, the detailed strokes of the sky are not properly transferred to the content image (e.g. row 5) due to no matched keypoints between the sky of the content and style images. Compared with the competing methods, our method achieves better image style transfer quality due to the correct matching of semantic context regions as shown in Fig. 3(i). Moreover, the detailed strokes are preserved for each semantic region while the boundaries between different semantic regions are smoothed. Thus, our stylized results are consistent with human perception.

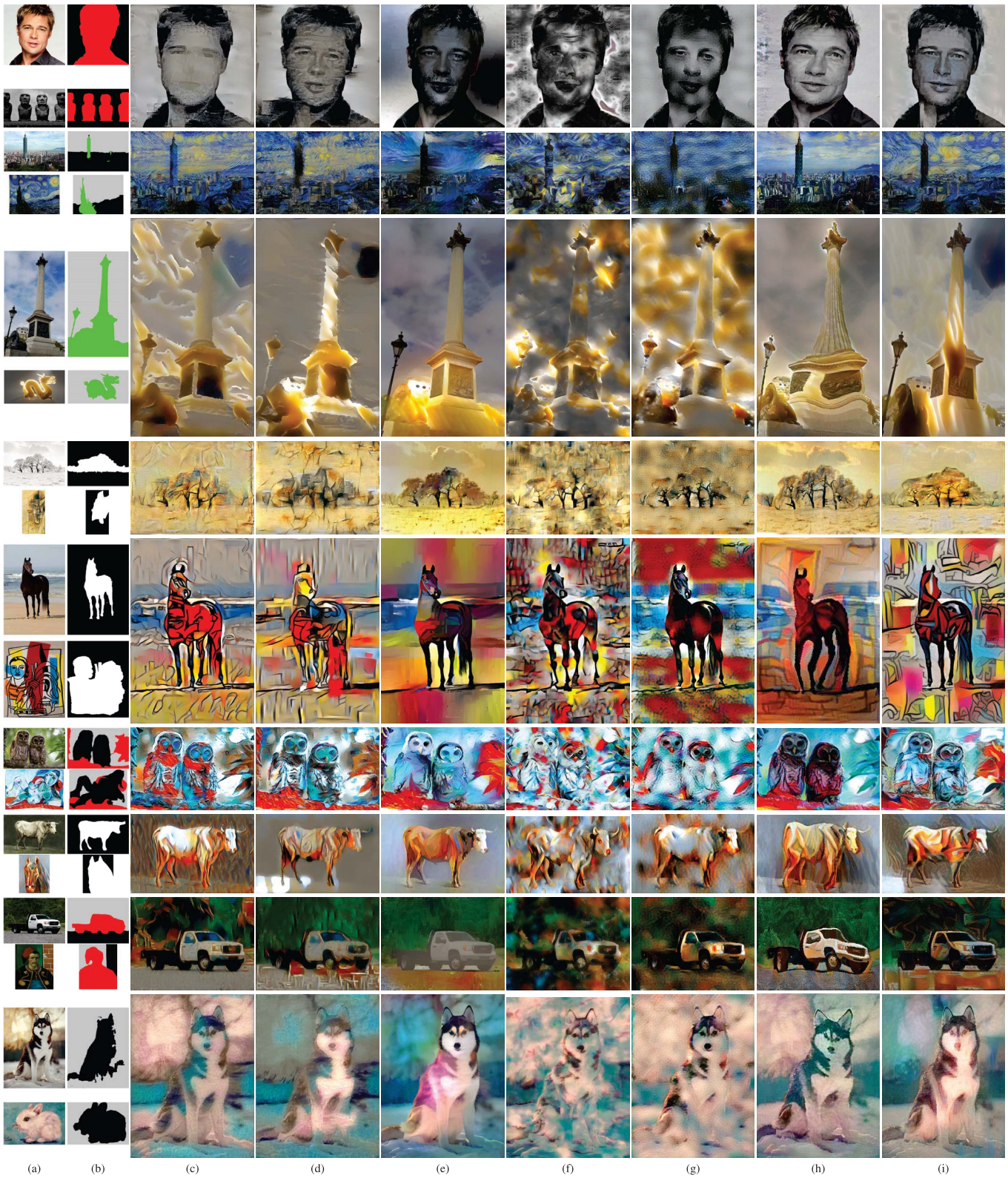

Fig. 4. Qualitative comparisons between our method and competing methods when the appearing objects are different between the content and style images. (a) Content and style images, (b) Semantic masks of content and style images, (c) Results of [25], (d) Results of [28], (e) Results of [29], (f) Results of [12] (g) Results of [21], (h) Results of [2] and (i) Results of our method.

Fig. 4 shows the qualitative comparisons when the appearing object categories are different between the content and style images. For example, the content image contains a portrait while the style image contains statues in row 1 of Fig. 4. Although both images contain different object categories, our semantic context matching can still retrieve

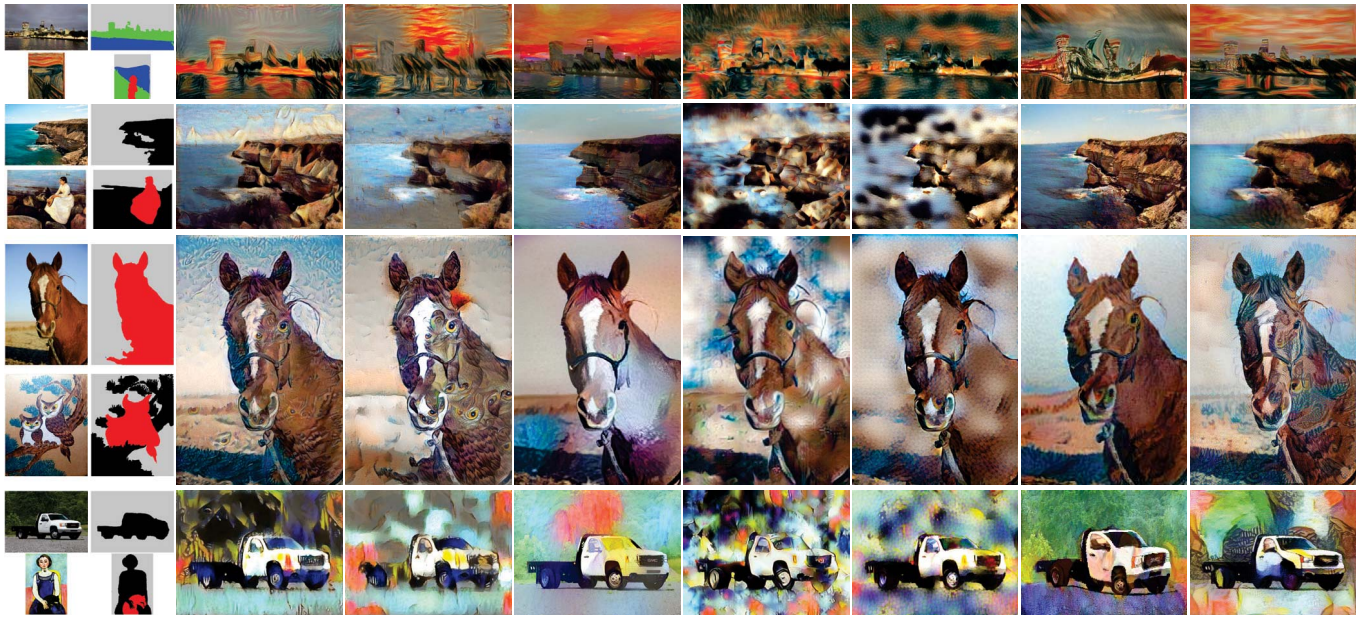

Fig. 5. Qualitative comparisons between our method and competing methods when the numbers of categories between the content and style images are different. (a) Content and style images, (b) Semantic masks of content and style images, (c) Results of [25], (d) Results of [28], (e) Results of [29], (f) Results of [12] (g) Results of [21], (h) Results of [2] and (i) Results of our method.

the corresponding semantic regions based on the contextual co-occurrence of objects. The styles cannot be successfully transferred by the image optimization based methods as shown in Fig. 4(c), (d), (e) and (h), because these methods cannot correctly compile the semantic corresponding regions. As shown in Fig. 4(f) and (g), the stylized results of the backgrounds (e.g. rows 1, 4, 6, 7, 8 and 9), and the sky (e.g. rows 2, 3 and 5) are corrupted by the pre-trained models. Thus, model optimization based methods are hard to achieve visually coherent results with the style images. In contrast, as shown in Fig. 4(i), by considering semantic context matching, the styles of the foreground objects of the style images can be successfully transferred to the foreground objects of the content images. Such results show the proposed hierarchical local-to-global network architecture can properly transfer the local styles of the corresponding semantic regions between the content and style images. Thus, our stylized results can still be visually consistent with human perception even when the content and styles images do not contain the same object categories.

Finally, Fig. 5 shows qualitative comparisons when the numbers of object categories between the content and style images are different. As shown in Fig. 5(b), the first two style images contain humans while the content images do not. The tree of the third style image and the cat of the fourth style image also do not exist in the content images. Without the semantic context matching, the styles of the humans, tree and cat of the style images will be transferred to the content images. Thus, incoherent and unnatural strokes of the stylized results can be observed in the sky, sea or background regions as shown in Fig. 5(c), (d), (e), (f), (g) and (h) of the competing methods. By simultaneously considering the semantic context matching and the hierarchical local-to-global network architecture, the proposed method can correctly transfer styles between the content and style images, and achieves significantly better stylized results as shown in Fig. 5(i). Moreover, it can also successfully maintain the content structure to avoid the deformation of the stylized results.

D. User Study

We performed the user study to evaluate the visual performance of the proposed method and the competing methods. We randomly showed the stylized results of Fig. 3, Fig. 4, and Fig. 5 with the semantic regions of the content and style images to each subject. Three visually quality related questions were asked for each subject. The first one is which stylized image achieves the best quality of the style transfer. The second one is which stylized image is the best style transfer results based on the corresponding semantic regions. The third one is which stylized image is the most visually coherent with the style image. The voting results of 42 subjects are shown in Fig. 6. In the user study, image optimization based methods [2], [25], [28], [29] have better visual transfer quality compared with model optimization based methods [12], [21], because they can perform more specific stylized results based on the content and style images.

Compared with the competing methods, our method is significantly favored in all of three questions. Based on the local context network, the detailed strokes of the style image can be transferred to the corresponding content regions. Then, the final stylized results optimized by the global context network achieve region-wise consistency. As a result, our method can provide visually coherence of the style image and well represent the structure of the content image. In addition, our semantic context matching can obtain the corresponding regions between the content and style images based on the

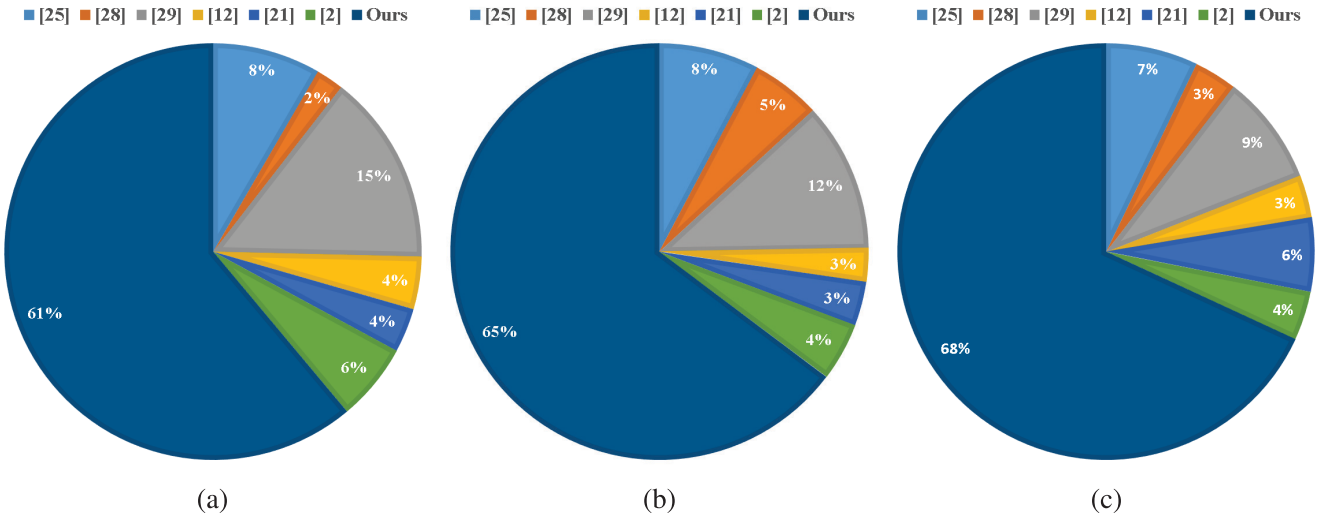

Fig. 6. The user study of stylized images with respect to (a) The best quality of the style transfer, (b) Style transfer results based on the corresponding semantic regions, and (c) Visually coherent with the style image.

context. Thus, even when no common object categories exist between the content and style images, the semantic context pairs can still help generate semantic style transfer results which are consistent with human perception.

V. Conclusion

In this paper, we propose a novel semantic context matching method to automatically obtain the semantic context pairs between the content and style images. Moreover, a hierarchical local-to-global network architecture, which contains the local context network and global context network, is proposed to optimize the transfer results based on the semantic context pairs. By considering semantic context pairs and learning by using hierarchical local-to-global network architecture, detailed strokes of the styles can be successfully transferred to the content image with a visually consistent manner of the human perception. Experimental results show that the proposed method is better than the state-of-the-art methods. In the future, we will further accelerate the computation speed of the proposed method.

References

[1] J. H. Park, S. Park, and H. Shim, "Semantic-aware neural style transfer," *Image Vis. Comput.*, vol. 87, pp. 13–23, Jul. 2019.

[2] S. S. Y. Kim, N. Kolkin, J. Salavon, and G. Shakhnarovich, "Deformable style transfer," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 246–261.

[3] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends Cognit. Sci.*, vol. 11, no. 12, pp. 520–527, 2007.

[4] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5122–5130.

[5] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *Proc. Ann. Conf. Comput. Graph. Interact. Techn.*, 2001, pp. 327–340.

[6] J. E. Kyprianidis, J. Collomosse, T. Wang, and T. Isenberg, "State of the 'art': A taxonomy of artistic stylization techniques for images and video" *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 5, pp. 866–885, May 2013.

[7] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand, "Style transfer for headshot portraits," *ACM TOG*, vol. 33, no. 4, p. 148, 2014.

[8] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1349–1357.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[10] T. Qi Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," 2016, *arXiv:1612.04337*.

[11] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[12] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 386–396.

[13] M. Lu, H. Zhao, A. Yao, Y. Chen, F. Xu, and L. Zhang, "A closed-form solution to universal style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5951–5960.

[14] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast image and video style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3809–3817.

[15] Y. Zhang, Y. Zhang, and W. Cai, "A unified framework for generalizable style transfer: Style and content separation," *IEEE Trans. Image Process.*, vol. 29, pp. 4085–4098, 2020.

[16] Y. Qiao, J. Cui, F. Huang, H. Liu, C. Bao, and X. Li, "Efficient style-corpus constrained learning for photorealistic style transfer," *IEEE Trans. Image Process.*, vol. 30, pp. 3154–3166, 2021.

[17] S. Gu, C. Chen, J. Liao, and L. Yuan, "Arbitrary style transfer with deep feature reshuffle," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8222–8231.

[18] L. Sheng, Z. Lin, J. Shao, and X. Wang, "Avatar-Net: Multi-scale zero-shot style transfer by feature decoration," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8242–8250.

[19] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5873–5881.

[20] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao, "Gated-GAN: Adversarial gated networks for multi-collection style transfer," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 546–560, Feb. 2019.

[21] Y. Yao, J. Ren, X. Xie, W. Liu, Y.-J. Liu, and J. Wang, "Attention-aware multi-stroke style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1467–1475.

[22] 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. (ICCV)*, Oct. 2017, pp. 2488–2496.

[23] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 909–920, 2020.

[24] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single-image depth perception in the wild," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2016, pp. 730–738.

[25] 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.
[26] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 262–270.
[27] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2479–2486.
[28] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3730–3738.
[29] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6997–7005.
[30] N. Kolkin, J. Salavon, and G. Shakhnarovich, "Style transfer by relaxed optimal transport and self-similarity," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10043–10052.
[31] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," *CoRR*, vol. abs/1603.01768, pp. 1–7, Mar. 2016.
[32] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 768–783.
[33] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 11, pp. 3365–3385, Nov. 2020.
[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[35] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, May 2014, pp. 740–755.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–14.
[37] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 1–15.

Yi-Sheng Liao received the B.S. and M.S. degrees from the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, in 2018 and 2020, respectively.

Chun-Rong Huang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 1999 and 2005, respectively. In 2005, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, as a Postdoctoral Fellow. He joined the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, in 2010, where he became a Full Professor in 2019. His research interests include computer vision, computer graphics, multimedia signal processing, image processing, and medical image processing. He is a member of the IEEE Circuits and Systems Society, the IEEE Signal Processing Society, the IEEE Computational Intelligence Society, and the Phi Tau Phi Honor Society.