IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 6, 2020, accepted January 4, 2021, date of publication January 6, 2021, date of current version January 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049637

Pseudo-Supervised Learning for Semantic Multi-Style Transfer

SAEHUN KIM, JEONGHYEOK DO, AND MUNCHURL KIM, (Senior Member, IEEE)
School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

Corresponding author: Munchurl Kim (mkimee@kaist.ac.kr)

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) Grant funded by the Korean Government (MSIT), Intelligent High Realistic Visual Processing for Smart Broadcasting Media, under Grant 2017-0-00419.

ABSTRACT Numerous methods for style transfer have been developed using unsupervised learning and gained impressive results. However, optimal style transfer cannot be conducted from a global fashion in certain style domains, mainly when a single target-style domain contains semantic objects that have their own distinct and unique styles, e.g., those objects in the anime-style domain. Previous methods are incongruent because the unsupervised learning can not provide the semantic mappings between the multi-style objects according to their unique styles. Thus, in this paper, we propose a pseudo-supervised learning framework for the semantic multi-style transfer (SMST), which consists of (i) a pseudo ground truth (pGT) generation phase and (ii) a SMST learning phase. In the pGT generation phase, multiple semantic objects of the photo images are separately transferred to the target-domain object styles in an object-oriented fashion. Then the transferred objects are composed back to an image, which is the pGT. In the SMST learning phase, a SMST network (SMSTnet) is trained with the pairs of the photo images and its respective pGT in a supervised manner. From this, our framework can provide the semantic mappings of multi-style objects. Moreover, to embrace the multi-styles of various objects into a single generator, we design the SMSTnet with channel attentions in conjunction with a discriminator dedicated to our pseudo-supervised learning. Our method has been applied and intensively tested for anime-style transfer learning. The experimental results demonstrate the effectiveness of our method and show its superiority compared to the state-of-the-art methods.

INDEX TERMS Style transfer, image-to-image translation, generative adversarial networks.

I. INTRODUCTION

For the task of stylizing images from one domain to another, unsupervised learning is often used to train the networks toward generating the images of target-domain styles. Previous unsupervised learning methods [1], [3]–[6] mostly using Generative Adversarial Networks (GAN) [7], achieved successful performances in style transfer problems where the target style domains often have common global features. For example with photo to art painting transfer, unsupervised learning networks are trained to learn the common global features (brush stroke texture) of the target style domain (art painting) from a large image set (e.g., Vincent van Gogh's paintings). Since objects in paintings all share a common brush stroke texture, the unsupervised learning framework is able to learn the global brush stroke texture successfully. However, it often fails when the target style domain contains semantic objects that have their own distinct and unique styles, e.g., the anime style domain. We refer to the style transfer problem where the target style domain contains multiple style objects as the 'object-to-object multi-style transfer' problem. The object-to-object multi-style transfer problem has not been considered in the previous style transfer studies where the objects in the target style domain are often assumed to have coherent style characteristics. To the best of our knowledge, our work is the first to define the 'object-to-object multi-style transfer problem' where various objects in a target style domain have their respective unique style characteristics. Fig. 1 shows the results of style transfer learning for various methods. As shown in Fig. 1, the previous unsupervised learning methods [1], [2] generate poorly style-transferred results for the anime style domain. This is due to the fact that unsupervised learning can not provide the semantic mappings for the multi-style objects from real

The associate editor coordinating the review of this manuscript and approving it for publication was Mingjun Dai.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

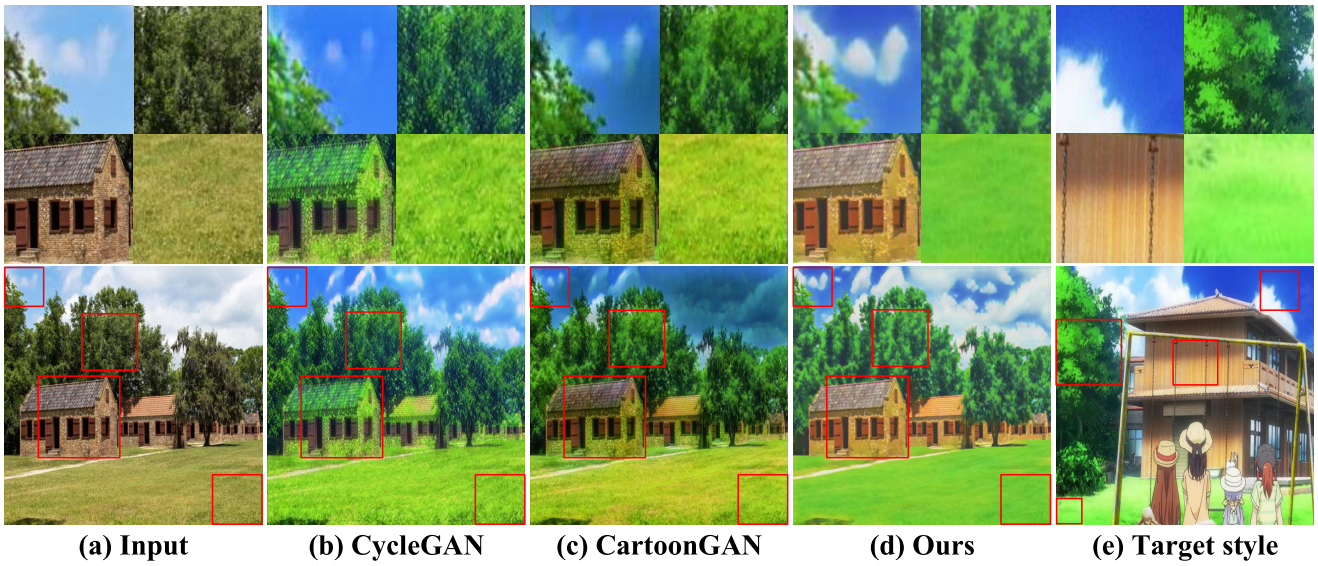

FIGURE 1. (a) Real-world input photo and patches of semantic objects. Results of (b) CycleGAN [1], (c) CartoonGAN [2], and (d) our method. (e) A sample image from the anime "Non Non Biyori the movie: Vacation", which is the target anime style. As shown in the four patches, various objects in anime have their own unique style.

photo domain to anime-style domain. The rough and global learning of the unsupervised learning framework mixes up the different object styles in the feature space. Also, the architectures of their generators are insufficient in embracing the multiple object styles within their single networks. In order to overcome these limitations, we propose a new learning framework, called pseudo-supervised learning, that provides the semantic connections between the objects of the two style domain in the image space. Also, we propose a new generator network and a new discriminator learning method that can embrace the multiple object styles within a single network. It is noted in Fig. 1-(d) that our semantic multi-style transfer network can faithfully generate stylized objects accordingly for anime-style domain. The contributions of our work are summarized as follows:

- Our proposed pseudo-supervised learning framework is the *first* work that can elaborately handle the object-to-object multi-style transfer problem. A novel idea of utilizing pseudo ground truths is used to provide the semantic mappings of the multiple styles that are contained in one target-style domain. Also, three loss functions are carefully balanced to achieve optimal stylized results.
- We propose a generator network called semantic multi-style transfer network (SMSTnet), which can embrace various styles of distinct characteristics for different objects in the target style domain. The SMSTnet is a U-Net [8] based generator incorporating three effective processing blocks: densely-connected channel attention block (DCCAB), down-scaling channel attention block (DSCAB), and up-scaling channel attention block (USCAB). The DCCAB is placed in each spatial resolution level of the U-Net, while the DSCAB and USCAB substitute the max-pooling and the transpose convolution layers. These channel attentions benefit in generating fine details of various object styles.
- Inspired by [9], we propose a new training method for the discriminator in our proposed pseudo-supervised learning framework. When training the discriminator, we utilize not only the pairs of the whole pseudo ground truths and real photo images but also the images with mixed real photo object regions and the pGT object regions. This helps the discriminator locally discern stylized and non-stylized image regions, leading the SMSTnet to generate more locally faithful target object styles.

II. RELATED WORK

A. STYLE TRANSFER

Style transfer methods can be divided into two groups, neural style transfer, and GAN-based style transfer. Neural style transfer methods [10]–[16] require a single reference style image where the network transfers the content image similar to the reference image while preserving the original context. GAN-based style transfer methods [1], [2], [17]–[21] learn the overall domain characteristics and transfer the input image to that domain without a particular reference style image. Since it is time-consuming to find a reference style image for every input image, GAN-based style transfer methods are often preferred when there is a specific target-style domain that the model wants to learn. Since there is no paired data for this problem, the previous works all use an unsupervised learning framework.

B. SEMANTIC MULTI-STYLE TRANSFER

An example domain of the object-to-object multi-style transfer is the anime style domain. Objects in anime have their own unique styles, as shown in Fig. 1-(e). Chen *et al.* [2]

first proposed a GAN-based anime-style transfer model with a dedicated loss function for anime. However, despite the dedicated loss function, Fig. 1-(c) shows that the unsupervised learning framework and the simple network architecture are insufficient in learning the multiple styles of anime objects.

Works that focused on semantic-wise style transfers [22]–[28] aim to stylize multiple image regions to completely different target style domains, that is, the number of target style domains are more than one. Moreover, these works require a segmentation mask in inference time, where they simply transfer different styles to different regions by masking them with the segmentation mask. However, generating precise segmentation maps for every inference image is very time consuming and may lead to unnatural seams, limiting the performance of the models. On the other hand, our proposed pseudo-supervised learning framework does not require segmentation maps in inference time, and the generator can successfully transfer the real-world objects to the corresponding semantic anime object styles in an end-to-end manner. That is, our network internally learns the semantic-wise style transfers without any external information.

C. NETWORK ARCHITECTURE

Most style transfer networks adopt the encoder-decoder architecture of the U-Net [8] as their generator architecture. The generator of the CartoonGAN [2] adopts the U-Net [8] architecture except for skip connections and adds 8 residual blocks between the encoder part and the decoder part. The discriminator of the CartoonGAN adopts the basic idea of the PatchGAN [29]. Reference [30] proposed RCA-U-Net (Residual Channel Attention U-Net), which is a U-Net where a single residual channel attention (RCA) block is placed between the spatial resolution transfer stages. For the object-to-object multi-style transfer problem, channel attention can give attention to the specific channels that are responsible for the stylization of the corresponding object style, benefiting the stylization results.

D. LEARNING FRAMEWORK

Note that the motivation of our proposed pseudo-supervised learning is conceptually different from the weakly supervised learning methods [31]–[35]. Weakly supervised learning methods are mostly used in object detection where manually annotated labels are heavily required and costly. These works substitute one of the manually annotated labels (object class labeling or segmentation mask) with a module-generated pseudo label to reduce the cost of full annotation. On the other hand, in style transfer, the problem is that the training datasets are unorganized so that labels do not exist. This leads to unsatisfying results where unique style characteristics are mixed and are thus lost. To provide semantic mappings to the entangled feature space with no label, we propose pseudo-supervised learning.

III. PSEUDO-SUPERVISED LEARNING FRAMEWORK

The object-to-object multi-style transfer problem aims to stylize various semantic objects of their own distinct styles. However, unsupervised learning is only capable of learning the global features of the target style domain. Thus, the unique styles of various semantic objects in the target style domain are entangled in the feature space.

The main idea of our approach is to incorporate the multiple object styles in a mutually exclusive manner. That is, we aim to disentangle the multiple object styles and guide the network to properly map the corresponding semantic objects between two different style domains. For this, our proposed pseudo-supervised learning framework utilizes the pseudo ground truth to provide the pixel-wise mapping between the semantic objects in the image space. Thus, the generator internally learns the semantic style information without any external data, such as a segmentation map, in inference time. Moreover, the generator architectures of the previous GAN models [2], [29] are insufficient in embracing the details of the multiple object styles. Our proposed pseudo-supervised learning framework incorporates three loss terms to provide the semantic mapping between the objects of different style domain with a new generator network (SMSTnet) and a new training method for the discriminator so that the SMSTnet can embrace the multiple styles of the target style domain objects.

A. FRAMEWORK

Fig. 2 shows the overall pipeline of our proposed pseudo-supervised learning framework. It consists of pseudo ground truth (pGT) generation phase and semantic multi-style transfer learning phase.

In the pGT generation phase, pseudo ground truths (pGTs) are generated to constitute the pairs of ground truths and real photo images. The pGT for a real photo image is a stylized image whose objects are separately stylized by their corresponding unsupervised single-style transfer module (CycleGAN [1] etc.). The number of single-styled transfer modules is the same as the number of semantics. For our experiment, we divided the semantics into 5 classes according to distinct style characteristics as an example. The 5 classes are sky, tree, grass, water, and the remaining miscellaneous anime objects (MAO). The photo and animation images are segmented manually according to the 5 semantic classes. We used a total of 5 networks where each network corresponds to a dedicated semantic. It should be noted that such five objects are known to be representative objects that have unique style in anime. So, we considered the five object semantics as an example to show the effectiveness of our proposed object-to-object multi-style transfer framework. Note that the segmentation maps are only required in the pGT generation phase, where each style transfer module is delicately trained with its corresponding objects for the target style learning. The style of the anime sky segments and the anime grass segments are stylized via the CycleGAN [1] architecture. We experimentally found that for the anime tree segments

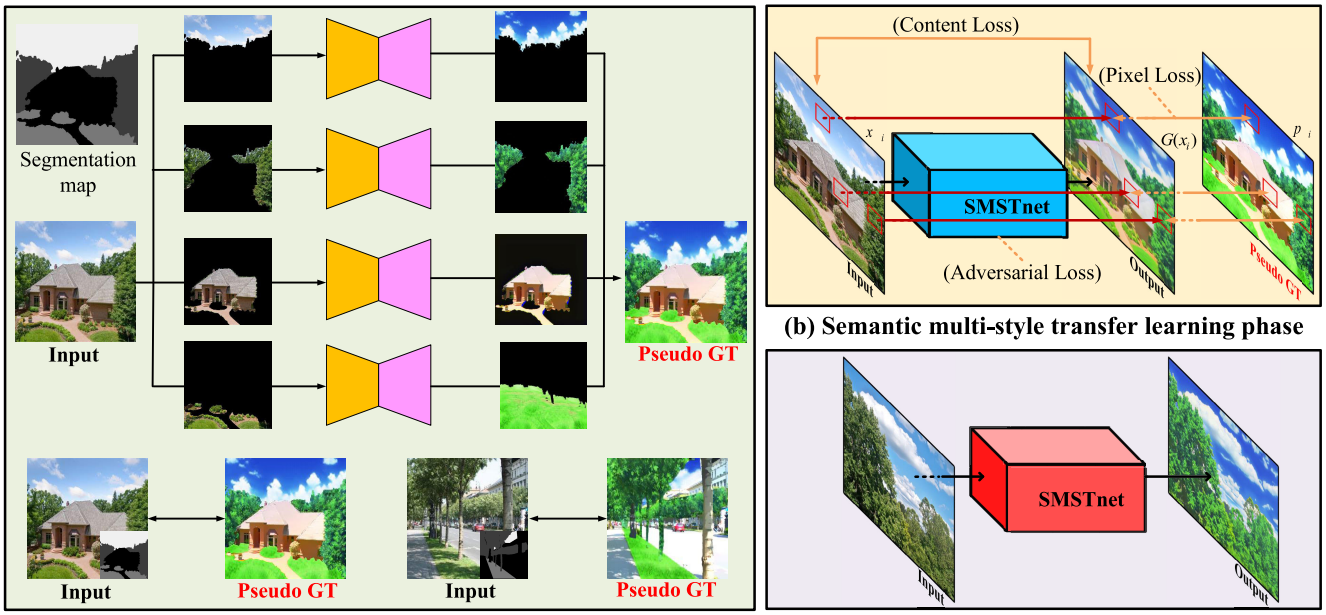

FIGURE 2. The conceptual pipeline of our proposed pseudo-supervised learning. (a) pGT generation phase, where the pseudo ground truths are generated. (b) SMST learning phase, where the SMSTnet is trained with the pseudo ground truths. (c) In inference time, the SMSTnet stylizes the input image in an end-to-end manner. No segmentation maps are required.

and the water segments, the CartoonGAN [2] architecture produces better stylized results. The MAO class includes all the remaining objects except the sky, tree, grass, and water. These objects share common characteristics, which are often (i) smooth texture, (ii) bright color, and (iii) highly saturated color. We model these three characteristics by the MAO module that consists of the three sequent processors: (i) an image smoothing processor, (ii) a brightness scaler, and (iii) a saturation booster of the smoothed image. That is, the MAO module is not a trainable network but a predefined processing module. Note that the single style transfer modules can be implemented using any advanced style transfer network. The designs chosen in this paper is an example to verify our object-to-object multi-style transfer framework for a single target style domain. The style-transferred objects for a real photo image are composed into a pGT in a divide-and-conquer manner. By doing so, the pGTs contain very well stylized objects since the unsupervised style transfer method performs very well in style transfer for a single dedicated style, as mentioned in Section I. However, due to the divide-and-conquer strategy which incorporates segmentation maps, there are inevitable artifacts like unnatural seams along the object boundaries and minor object deletion in the pseudo ground truth images. This is why the generated pairs are called 'pseudo' ground truths. Moreover, the divide-and-conquer strategy is very naive and time-consuming. Our objective of the semantic multi-style transfer (SMST) learning phase is to distill the knowledge of individual small networks for different style objects by designing an end-to-end trainable generator, which is the SMSTnet.

In the semantic multi-style transfer (SMST) learning phase, the SMSTnet is trained in a supervised manner with the pseudo ground truths generated in the pGT generation phase. Note that this is different from the Pix2pix [29] algorithm, where they train a conditional GAN to regress a ground truth target. Since 'pseudo' ground truths contain inevitable artifacts, we incorporate three simple but effective loss functions: an adversarial loss, a pixel loss, and a content loss.

Let S_x be the set of the real-world photos and S_d be the set of images that are used to train the discriminator D. The specific training method for the discriminator will be explained in Section III-C. The pseudo ground truth p_i is generated for the corresponding input photo x_i. $G(x_i)$ refers to the output of our SMSTnet G. d_i refers to an image from the set of images that are used to train the discriminator D. Firstly, the adversarial loss $\mathcal{L}_{adv}$ is defined as:

$$\mathcal{L}_{adv}(G, D) = \mathbb{E}_{d_i \sim S_d}[logD(d_i)] + \mathbb{E}_{x_i \sim S_x}[log(1 - D(G(x_i)))] \quad (1)$$

The adversarial loss trains the generator to learn the global features of the target style domain. To provide the supervision to local details of the multiple object styles, the pixel loss $\mathcal{L}_{pixel}$ is computed between the output $G(x_i)$ and its pseudo ground truth pair p_i and is defined as:

$$\mathcal{L}_{pixel}(G) = \mathbb{E}_{x_i \sim S_x}[(G(x_i) - p_i)^2] \quad (2)$$

As shown in Fig. 2, the red boxes indicate patches from different semantic objects which have their own unique style. That is, by using the pseudo ground truths as pixel-wise

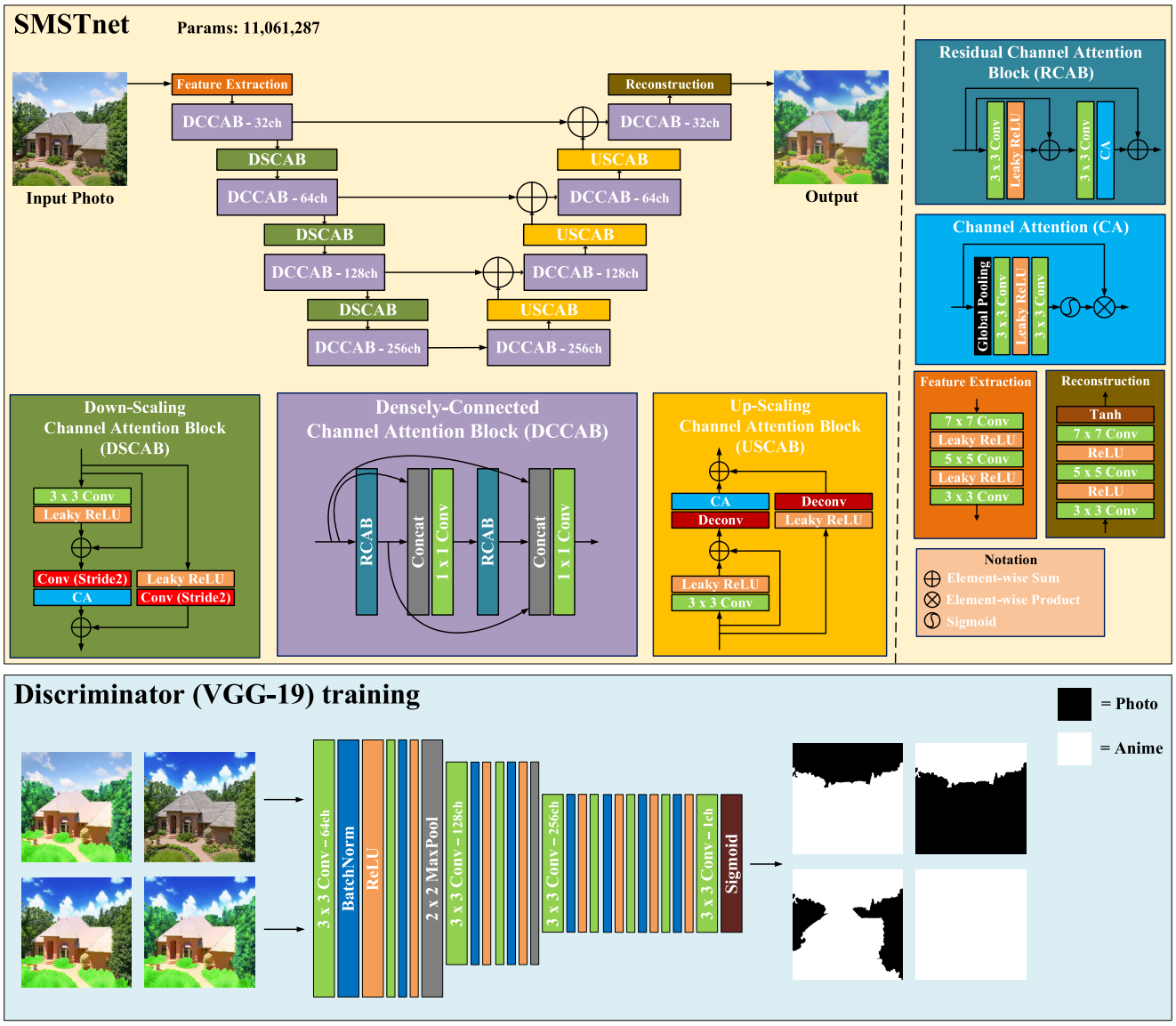

FIGURE 3. Network architectures of our generator and discriminator.

regression targets, our SMSTnet can directly learn the connection between real-world objects and their corresponding anime object styles in the image space. However, as we mentioned above, the pseudo ground truths contain inevitable artifacts due to the recombining of individual segments in the pGT generation phase. The main two artifacts are unnatural seams and minor object deletion. With only the pixel loss, the SMSTnet will learn the artifacts as it is and generates them in the results. Thus, we use the content loss [2] $\mathcal{L}_{con}$ to somewhat restrict the SMSTNet from distorting the original context of the input photo. The content loss is defined as:

$$\mathcal{L}_{con}(G) = \mathbb{E}_{x_i \sim S(x)}[\|VGG_l(G(x_i)) - VGG_l(x_i)\|_1] \quad (3)$$

where VGG_l refers to the pre-trained VGG-19 [39] feature maps in the layer l = (conv4_4). Note that we adopt the L1 sparse regularization of VGG feature maps, which is known to focus on preserving the high-level structures of the original image and relatively ignoring the low-level differences [2].

As a result, the total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda \mathcal{L}_{pixel} + \omega \mathcal{L}_{con} \quad (4)$$

where the hyperparameters are empirically set to $\lambda = 10$ and $\omega = 5$. A good balance leads to an optimal point where the $\mathcal{L}_{con}$ prevents the SMSTnet from generating artifacts from the pseudo ground truths, while the $\mathcal{L}_{adv}$ and $\mathcal{L}_{pixel}$ stylizes the real-world objects to their corresponding target object style where the different object styles are disentangled in the feature space. In summary, the objective of the semantic multi-style transfer learning phase of our proposed pseudo-supervised learning is to train a single generator network that can fully transfer the multiple anime object styles

FIGURE 4. (a) Input photo. Results of (b) CycleGAN [1], (c) DiscoGAN [21], (d) DualGAN [20], (e) MUNIT [36], (f) CartoonGAN [2], (g) U-GAT-IT [37], (h) GANILLA [38], and (i) our method. (j) Real anime sample from the target anime for visual comparison.

with precise detail, while avoiding the inevitable artifacts of the pseudo ground truths.

B. SMSTnet ARCHITECTURE

We propose a new generator network called semantic multi-style transfer network (SMSTnet), as shown in Fig. 3. The SMSTnet is based on the RCA-U-Net [30] architecture, where we implement our proposed three modules, Densely-Connected Channel Attention Block (DCCAB), Down-Scaling Channel Attention Block (DSCAB), and Up-Scaling Channel Attention Block (USCAB). The convolution kernels of the generator are trained to learn the mappings between each patch feature and its corresponding target style. Since every convolution kernel slides through the entire input image, the different semantic features of different objects will be stored in the channels of the feature maps. Therefore, the channel attentions in the SMSTnet benefit in generating fine details of various object styles.

The DCCAB is a module where two Residual Channel Attention Blocks (RCAB) [30] are densely connected. The DCCAB is placed on every level of the SMSTnet. The DSCAB and the USCAB substitute the max pooling and the transpose convolution of the RCA-U-Net architecture, respectively. That is, the two blocks perform channel attention when reducing and enlarging the spatial resolution of the features. From this, the dense features of each level can be adequately processed during the feature resolution change and sent to the next DCCAB block.

C. DISCRIMINATOR TRAINING

Inspired by [9], we propose a new training method for the discriminator, which is dedicated to our proposed

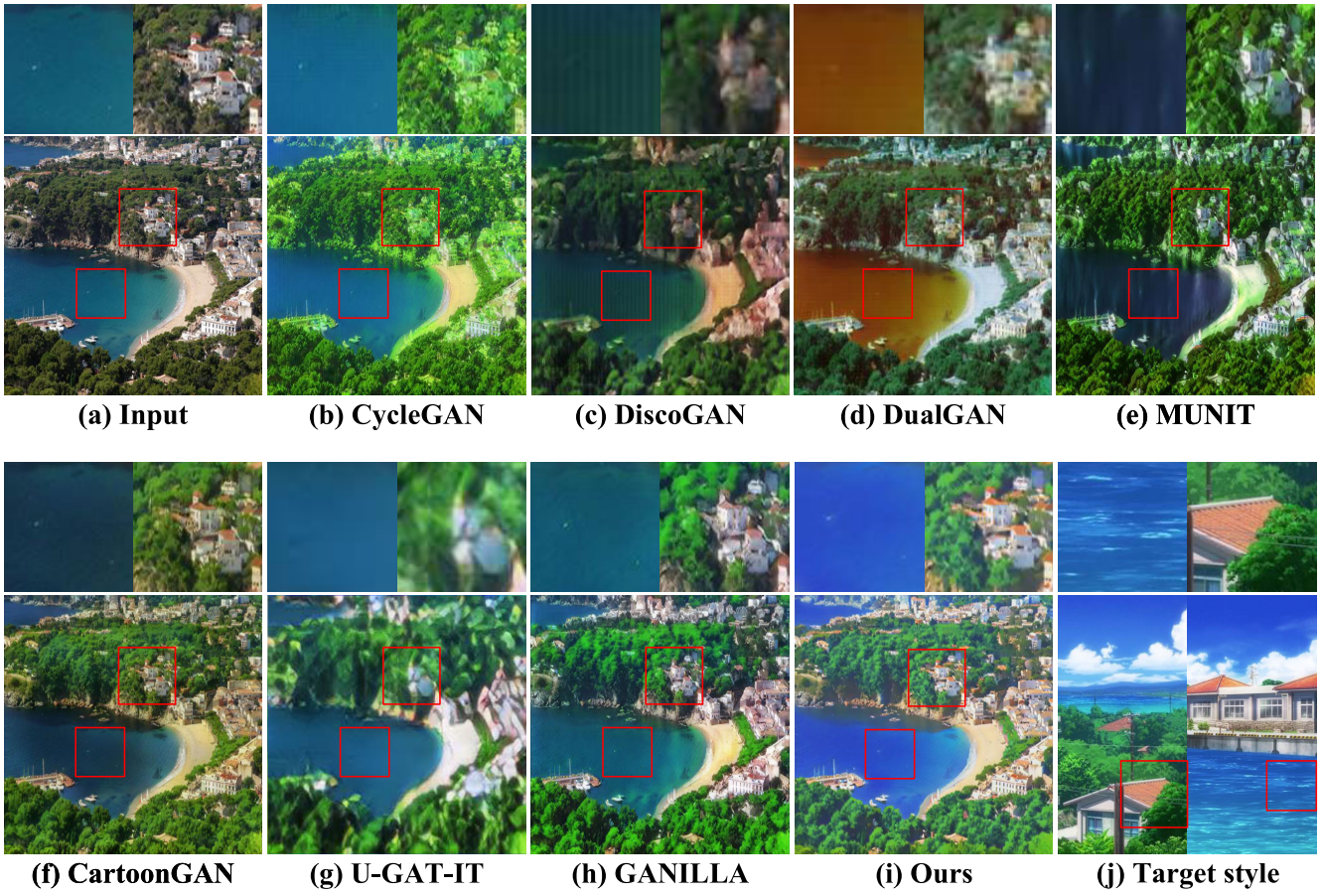

FIGURE 5. (a) Input photo. Results of (b) CycleGAN [1], (c) DiscoGAN [21], (d) DualGAN [20], (e) MUNIT [36], (f) CartoonGAN [2], (g) U-GAT-IT [37], (h) GANILLA [38], and (i) our method. (j) Real anime sample from the target anime for visual comparison.

pseudo-supervised learning framework by utilizing the pseudo ground truth pairs created in the pGT generation phase. The discriminator is trained with three types of images. The pseudo ground truths, the output of the generator, and images that are created by randomly substituting some of the segments in the real-world photo to the corresponding segments taken from the pseudo ground truth pair, e.g., an image containing real-world photo sky and anime-stylized trees (Fig. 3). For the mixed-segment images, a segmentation mask that indicates the photo regions and anime regions are given as supervision. Thus, the discriminator performs semantic segmentation in which it discriminates regions that are properly stylized enough as real and the regions that are not yet stylized enough as fake, instead of only discriminating the entire input image to real or fake. From this, the SMST-net is adversarially trained to be aware of local false-styled segments and generate locally precise anime-styled images. We adopt the first 10 layers of the VGG-19 [39] architecture. Note that the discriminator is not responsible for preventing inevitable artifacts, whereas the content loss handles this point. It is known that the discriminator helps the generator generate more detailed low-level features of the target style (mostly color) [29]. Thus, the discriminator does not have to see a natural boundary.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

To perform style transfer for a specific target anime style, a set of real-world photos and a set of anime images drawn by a certain artist is required. We used the anime object styles in the animation series "Non Non Biyori" drawn by the artist "Atto" as the target style and manually selected 1,434 screen-shot images. We augmented the data by cropping into four sub-images, resulting in 7,170 anime images for our anime dataset. We used a total of 9,646 images from the ADE20K dataset [40], as the real-world photo dataset, where the segmentation map is provided. The animation images are also segmented manually according to the same semantic classes of the photo images. The previous models [1], [2], [20], [21], [36]–[38] were retrained with these datasets for fair comparisons. All images were resized to 256×256 for training.

The time taken in training is as follows. For our method, the GANs in the pGT generation phase each take 12000 sec (= 200 epochs × 60 sec/epoch), totaling 48,000 sec (= 12000 sec × 4 GAN). SMST learning phase takes 7,400 sec (= 5 epochs × 1480 sec/epoch). So, the total training time is 55,400 sec. The CartoonGAN [2] takes 188,000 sec (= 200 epochs × 940 sec/epoch). The number

FIGURE 6. (a) Input photo. Results of (b) CycleGAN [1], (c) DiscoGAN [21], (d) DualGAN [20], (e) MUNIT [36], (f) CartoonGAN [2], (g) U-GAT-IT [37], (h) GANILLA [38], and (i) our method. (j) Real anime sample from the target anime for visual comparison.

of parameters of the SMSTnet (11M) is close to the number of parameters of the CartoonGAN (11M). As a consequence, our SMSTnet takes less training time than the CartoonGAN that takes about 3 times longer time.

B. QUALITATIVE RESULTS

Fig. 4 through Fig. 7 compares the qualitative results of our method with the CycleGAN [1], DiscoGAN [21], DualGAN [20], MUNIT [36], CartoonGAN [2], U-GAT-IT [37], and GANILLA [38]. It can be seen by comparing the results object-by-object that our method produces the best results with the fine details of the target anime object styles. For example, as shown in Fig. 4, the specific texture of the tree in the target style is finely generated in the stylized result, while the previous methods failed to transfer the texture of the tree in the original photo to the target domain style. Also, our method faithfully generates the colors of the sky and grass in the stylized images that highly resemble those in the target style domain. However, the sky and grass in the stylized images of the previous methods lack in the color saturation of the sky and grass of the target style domain and often contain dirty color in various objects. One of the main reasons for this poor performance is that the unsupervised learning framework entangles the distinct styles of multiple objects in the target style domain. Since their generators are not capable of transferring various semantic objects according to their own distinct styles, the various features of the semantic objects are mixed in the feature space, resulting in dirty colors. Also, because the generators of the previous methods fail to learn the dedicated mapping between the tree in the real photo images and that in the target style images,

FIGURE 7. (a) Input photo. Results of (b) CycleGAN [1], (c) DiscoGAN [21], (d) DualGAN [20], (e) MUNIT [36], (f) CartoonGAN [2], (g) U-GAT-IT [37], (h) GANILLA [38], and (i) our method. (j) Real anime sample from the target anime for visual comparison.

the unique texture of the tree in the target style domain is lost in the resulting outputs. The qualitative results show that our proposed pseudo-supervised learning framework can allow for learning the dedicated mapping through which the style details of various semantic objects can be effectively generated.

C. QUANTITATIVE EVALUATION

We perform user studies with 8 anime experts (employees who work in the anime industry) and 126 ordinary viewers. We give the experimenters five groups of images. Each group consists of a real-world photo and the anime-styled images generated by the CycleGAN [1], CartoonGAN [2], and our method. The test is based on the DSIS (Double-Stimulus Impairment Scale) test, which is one of the standard subjective evaluation methods recommended by ITU-R BT.500 [41]. Under the DSIS test, the subjects are supposed to rate the scores between 1 and 5 (1 and 10 in our case) on the impaired image (an anime-stylized image in our case) in comparison with a presented original image (a real anime scene in our case). The highest score (10 in our case) means 'no visual difference' between them. In our case, the subjects are asked to rate how much similar the anime-stylized images are to the real anime drawn by the artist "Atto".

We also compute the user preference score by counting the number of times the model got the highest quality score for a group (including joint first place). The quality scores and the user preference scores are averaged and shown in Table 1. We can see that our method outperforms the previous methods in quality score and is highly preferred for both experts and non-experts.

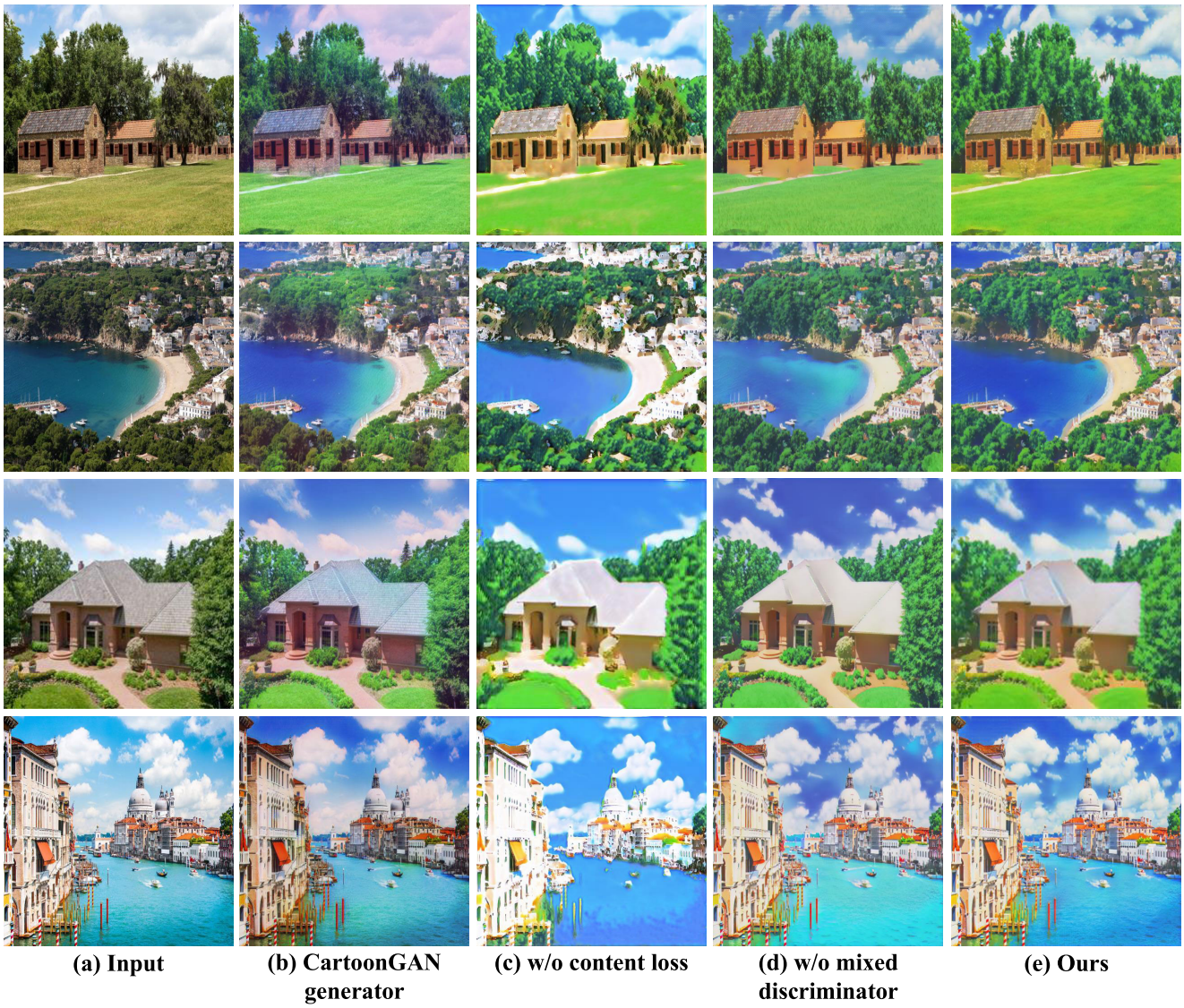

FIGURE 8. Ablation studies. (a) Real-world input photo. (b) Ablation study for the generator. Results using the generator architecture of the CartoonGAN [2]. (c) Ablation study for the content loss. Results without the content loss. (d) Ablation study for the discriminator. Results without using the mixed-component images when training the discriminator. (e) Results of our method.

TABLE 1. User studies for different methods by anime experts and non-experts. The quality score is ranged from 1 to 10. The preference score is computed by counting the number of times the model got the highest quality score for a group.

	Expert		Non-expert	
	Quality	Preference	Quality	Preference
CycleGAN [1]	4.50	0.38	5.69	0.29
CartoonGAN [2]	3.75	0.20	5.23	0.20
Ours	**5.50**	**0.62**	**7.59**	**0.77**

TABLE 2. Computational complexity comparison for generator networks from different methods. The time taken in inference is measured as the amount of time taken when the generator processes one image of size 512 × 512.

	# of parameters	Time taken in inference (sec)
CycleGAN [1]	11.4M	0.0031
CartoonGAN [2]	11.1M	0.0026
GANILLA [38]	7.2M	0.0044
U-GAT-IT [37]	10.6M	0.0088
Ours	11.1M	0.0097

We compare the quantitative computation complexities of different generator networks in Table 2. It can be seen in Table 2 that our network achieves significant performance improvements with a similar number of parameters compared to other methods. The time taken when processing one image of size 512×512 is slightly larger than U-GAT-IT [37] which also utilizes attention modules. However, at the expense of the slightly longer inference time, the novel channel attention blocks of our method have brought a significant performance superiority in subjective visual comparison over the attention module in U-GAT-IT [37].

FIGURE 9. Generator ablation studies. (a) A sample image from the target anime for visual comparison. (b) Input photo. (c) With DSCAB, USCAB, DCCAB. This refers to the SMSTnet. (d) The generator with only DCCAB, where (USCAB) DSCAB is replaced by simple (transposed) convolution layers of stride 2. Artifacts are generated on the trees. (e) The generator where the dense connections between RCABs are removed. Details are slightly lost and the saturation decreased. (f) The generator without both modules. Fails to generate the abundant anime tree texture. Poor saturation.

D. ABLATION STUDY

Fig. 8 shows the results of the ablation studies. We perform ablation studies for the generator and the new training method of the discriminator and the necessity of the content loss. Fig. 8-(b) shows the results generated via the generator network architecture of the CartoonGAN [2] on our proposed pseudo-supervised learning framework, instead of our SMSTnet. Since our proposed framework provides the semantic mapping, the dirty colors of the stylized objects were diminished compared to Fig. 5-(f). However, the result lack degree of stylization. Thus, it can be seen that the generator network architecture of the CartoonGAN fails to embrace the multiple styles of the target style domain.

Fig. 8-(d) shows the results performed without our proposed discriminator training method. That is, the mixed-component images are not used when training the discriminator. Only images of the pseudo ground truths and the output of the generator were used, which is identical to the discriminator of the previous methods. The main difference between this result and the results of our method is the color. This shows that, due to the mixed-component images, the network of our method can be more sensitive to local low-level features.

Finally, Fig. 8-(c) shows the results performed without the content loss. As expected, the inevitable artifacts (unnatural seams and minor object deletion) are shown in the results as well. By comparing these results to the results of our method, we can see that the content loss prevents the generator from generating the artifacts and only filter the details of the object styles from the pseudo ground truths.

Fig. 9 shows additional ablation studies for the SMSTnet. Fig. 9-(c) is the result of the SMSTnet with all three blocks. The SMSTnet generates the abundant anime tree textures, which resemble the tree textures in the real anime sample image. However, as shown in Fig. 9-(d), without the DSCAB and USCAB modules, the generator produces rough artifacts which are mostly in the tree regions. On the other hand, without the DCCAB module, the result slightly loses the texture detail of the tree. Also, the saturation of the tree and sky decreases. Without both of the modules, the saturation decreases even more and fails to learn the texture of anime trees. Thus, we can conclude that the DCCAB has the ability to generate fine details of the anime objects by the dense connection of channel attention blocks, but incorporates artifacts as well. The DSCAB and the USCAB effectively process the dense feature maps of the DCCAB while reducing and enlarging them, alleviating the artifacts.

V. CONCLUSION

In this paper, we first propose a pseudo-supervised learning framework for semantic-wise object-to-object multi-style transfer. Previous methods rely on unsupervised learning frameworks, so they can not guide the networks to learn the semantic mappings for various objects with their distinct styles. This causes the unique features of the semantic object styles to be mixed in the feature space without any distinction, thus leading to unpleasant stylized results. Also, the architectures of the previous networks are insufficient in learning the multiple object styles of the target domain. Our pseudo-supervised learning framework solves these limitations by utilizing pseudo ground truths to learn the semantic mappings. Moreover, we propose a new generator network called SMSTnet, incorporating channel attentions to embrace the multiple object styles. Finally, for the discriminator, we utilize the images with mixed real photo object regions and the pGT object regions. Thus, the discriminator can locally discern true and false image regions, leading the SMSTnet to generate more locally faithful target object styles. Our method is shown to be very effective qualitatively and quantitatively from the intensive experiments, outperforming the state-of-the-art methods, while our framework is a very genetic scheme by which any style transfer learning for the objects of distinct characteristics can be easily realized with high fidelity of style transformation.

REFERENCES

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[2] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9465–9474.

[3] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 35–51.

[4] W. Cho, S. Choi, D. K. Park, I. Shin, and J. Choo, "Image-to-image translation via group-wise deep whitening-and-coloring transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10639–10647.

[5] M. Amodio and S. Krishnaswamy, "TraVeLGAN: Image-to-image translation by transformation vector learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8983–8992.

[6] N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, "Multistage attention network for image inpainting," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107448.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[9] E. Schonfeld, B. Schiele, and A. Khoreva, "A U-Net based discriminator for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8207–8216.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

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

[12] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," in *Proc. Workshop Constructive Mach. Learn. (NIPS)*, 2016, pp. 1–10.

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

[14] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang, "Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5239–5247.

[15] Y. Jing, Y. Liu, Y. Yang, Z. Feng, Y. Yu, D. Tao, and M. Song, "Stroke controllable fast style transfer with adaptive receptive fields," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 238–254.

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

[17] A. Sanakoyeu, D. Kotovenko, S. Lang, and B. Ommer, "A style-aware content loss for real-time HD style transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 698–714.

[18] D. Kotovenko, A. Sanakoyeu, P. Ma, S. Lang, and B. Ommer, "A content transformation block for image style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10032–10041.

[19] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[20] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.

[21] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1857–1865.

[22] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 3693–3703.

[23] 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. 2469–2477.

[24] H.-H. Zhao, P. L. Rosin, Y.-K. Lai, and Y.-N. Wang, "Automatic semantic style transfer using deep convolutional neural networks and soft masks," *Vis. Comput.*, vol. 36, pp. 1–18, Jul. 2019.

[25] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–10.

[26] Z. Huang, J. Zhang, and J. Liao, "Style mixer: Semantic-aware multi-style transfer network," *Comput. Graph. Forum*, vol. 38, pp. 469–480, 2019.

[27] C. Castillo, S. De, X. Han, B. Singh, A. K. Yadav, and T. Goldstein, "Son of Zorn's lemma: Targeted style transfer using instance-aware semantic segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1348–1352.

[28] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," *Comput. Graph. Forum*, vol. 35, pp. 93–102, 2016.

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[30] Z. Fang, Y. Chen, D. Nie, W. Lin, and D. Shen, "RCA-U-Net: Residual channel attention U-net for fast tissue quantification in magnetic resonance fingerprinting," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 101–109.

[31] K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent, "Cap2Det: Learning to amplify weak caption supervision for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9686–9695.

[32] D. Zhang, J. Han, L. Zhao, and T. Zhao, "From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5549–5560, Dec. 2020.

[33] D. Zhang, J. Han, L. Yang, and D. Xu, "SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 475–489, Feb. 2020.

[34] J. Han, Y. Yang, D. Zhang, D. Huang, D. Xu, and F. De La Torre, "Weakly-supervised learning of category-specific 3D object shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 25, 2019, doi: 10.1109/TPAMI.2019.2949562.

[35] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1755–1769, Jul. 2020.

[36] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–189.

[37] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–19.

[38] S. Hicsonmez, N. Samet, E. Akbas, and P. Duygulu, "GANILLA: Generative adversarial networks for image to illustration translation," *Image Vis. Comput.*, vol. 95, Mar. 2020, Art. no. 103886.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, pp. 1–14, Sep. 2014.

[40] 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. 633–641.

[41] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Recommendation BT.500-11, International Telecommunication Union-Radiocommunication Sector, 2002.

SAEHUN KIM received the B.E. and M.E. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2018 and 2020, respectively. His research interests include computer vision and image processing, such as detection, counting, style transfer, image-to-image translation, and deep learning.

JEONGHYEOK DO received the B.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2019, where he is currently pursuing the M.S. degree. His research interests include image dehazing, low-complexity networks, and imbalanced learning.

MUNCHURL KIM (Senior Member, IEEE) received the B.E. degree in electronics from Kyungpook National University, Daegu, South Korea, in 1989, and the M.E. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 1992 and 1996, respectively. He joined the Electronics and Telecommunications Research Institute, Daejeon, South Korea, as a Senior Research Staff Member, where he led the Realistic Broadcasting Media Research Team. In 2001, he joined the School of Engineering, Information, and Communications University (ICU), Daejeon, as an Assistant Professor. Since 2009, he has been with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, where he is currently a Full Professor. He was involved in scalable video coding and high-efficiency video coding in JCT-VC standardization activities of ITU-T VCEG and ISO/IEC MPEG. His current research interests include deep learning for image restoration and visual quality enhancement, deep video compression, perceptual video coding, visual quality assessments, computational photography, machine learning, and pattern recognition.