IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 August 2022, accepted 12 September 2022, date of publication 16 September 2022,
date of current version 26 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3207288

RESEARCH ARTICLE

Cross Modal Facial Image Synthesis Using a Collaborative Bidirectional Style Transfer Network

NIZAM UD DIN[1,2], SEHO BAE[2], KAMRAN JAVED[3], HYUNKYU PARK[2], AND JUNEHO YI[2]
[1]Saudi Scientific Society for Cybersecurity, Riyadh 11442, Saudi Arabia
[2]Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon 16419, South Korea
[3]National Centre of Artificial Intelligence (NCAI), Saudi Data and Artificial Intelligence Authority (SDAIA), Riyadh 11543, Saudi Arabia

Corresponding author: Juneho Yi (jhyi@skku.edu)

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant 2020R1F1A1048438, and in part by the High-performance Computing (HPC) Support Project funded by the Ministry of Science and ICT and National IT Industry Promotion Agency (NIPA).

ABSTRACT In this paper, we present a novel collaborative bidirectional style transfer network based on generative adversarial network (GAN) for cross modal facial image synthesis, possibly with large modality gap. We think that representation decomposed into content and style can be effectively exploited for cross modal facial image synthesis. However, we have observed that unidirectional application of decomposed representation based style transfer in case of large modality gap does not work well for this purpose. Unlike existing image synthesis methods that typically formulate image synthesis as an unidirectional feed forward mapping, our network utilizes mutual interaction between two opposite mappings in a collaborative way to address complex image synthesis problem with large modality gap. The proposed bidirectional network aligns shape content from two modalities and exchanges their appearance styles using feature maps of the layers in the encoder space. This allows us to effectively retain the shape content and transfer style details for synthesizing each modality. Focusing on facial images, we consider facial photo, sketch, and color-coded semantic segmentation as different modalities. The bidirectional synthesis results for the pairs of these modalities show the effectiveness of the proposed approach. We further apply our network to style-content manipulation to generate multiple photo images with various appearance styles for a same content shape. The proposed method can be adopted for solving other cross modal image synthesis tasks. The dataset and source code are available at https://github.com/kamranjaved/Bidirectional-style-transfer-network.

INDEX TERMS Generative adversarial network, image synthesis, unidirectional style transfer network, bidirectional style transfer network, collaborative learning.

I. INTRODUCTION

The goal of this research is to synthesize realistic cross modal face images while retaining the input face identity. We interpret facial images of a person from different modalities as facial images with the same shape content and different appearance styles. We have also observed that decomposed representation into content and style can bring great advantage to cross modal image synthesis [2]. On the other hand, as can be seen in Fig. 1, directly employing style transfer as unidirectional feed forward mapping for cross modal image synthesis does not work well in case of large modality gap.

Based on our interpretation and observation, we aim to develop a novel bidirectional synthesis network that effectively employs style transfer schemes to achieve our goal. We could effectively align the shape content from the two modalities and exchange their appearance styles by exploiting mutual interaction between two opposite mappings. In this work, we consider facial photo, sketch, and color-coded semantic segmentation as different modalities.

Generative adversarial networks (GANs) [3] have achieved significantly advanced image synthesis performance with

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

phenomenal quality and realism. Conditional GAN (cGAN) [4] took GAN into another direction by providing control over the generation of desired outputs. Most notably, Pix2Pix [5] adopted supervised learning as a general-purpose solution to translate a source image into a target image. Zhu *et al.* [6] presented an unsupervised approach to translate an image from one domain to another. Recently, StyleGAN [7], [8] proposed a novel style-based generator architecture which provides the generator with more control and representational capabilities to create images with high visual quality and realism. Numerous image synthesis works [9], [10], [11] fine-tuned this work to attain meaningful performances. Park *et al.* [12] proposed spatially-adaptive normalization layer for synthesizing photo realistic images from color-coded segmentation map. Most of the GAN-based photo-sketch synthesis models [13], [14], [15] formulated the mapping as unidirectional feed forward mapping. Thus, utilization of mutual interaction between two opposite mappings is found lacking. To better utilize the mutual information between opposite mappings, Col-cGAN [16] proposed a bi-directional cGAN based framework in which an intermediate image domain between photo and sketch is learned to enhance the synthesis performance. However, their method yields blurred effects and great deformation over various facial components without decomposing representation into content and style.

On the other hand, style transfer networks [1], [17], [18], [19] factored image representation into content and style components. These methods produced impressive artistic style results by transferring content from one image and style from others. However, as mentioned earlier, utilizing style transfer schemes as unidirectional feed forward mapping for image synthesis with large modality gap does not yield satisfactory performance. Chen *et al.* [20] synthesized sketches in a cascade manner by first generating the shape of face and then adding style details. However, they failed to preserve finer appearance style such as pencil lines and shading.

Unlike most image synthesis methods that typically formulate synthesis mapping as an unidirectional feed forward mapping, we propose a simple yet effective bidirectional style transfer network to exploit a mutual interaction between two opposite mappings for better coping with large modality gap. In the layers of the encoder space, we align the shape content from the two modalities and exchange their appearance styles by employing an AdaIN [1] based style-transfer unit called Bidirectional Style Transfer Module (BSTM) between their feature maps. This allows us to successfully synthesize visually plausible cross modal facial images with large domain gap.

We even view facial sketch and color-coded semantic segmentation as a facial modality and present bidirectional image synthesis between them although their modality gap is large. We further demonstrate our network for content-style manipulated synthesis. In this task, we generate multiple photo images from a single segmentation map via conditioning on photos with different styles. The bidirectional synthesis results for the pairs of facial photo, sketch, and color-coded semantic segmentation shows that the proposed methodology can be adapted for solving other cross modal image synthesis tasks.

The main contributions of this work are as follows.

- We have presented a style-transfer based bidirectional synthesis network to effectively exploit mutual interaction between two opposite mapping to address cross modal image synthesis with large modality gap.
- We demonstrate on challenging bidirectional synthesis from face sketch to semantic segmentation and semantic segmentation to face sketch.
- Our network is capable of generating multiple photo images with various appearance styles from a single segmentation map by conditioning on photo images with different styles.

II. RELATED WORK

A. IMAGE-TO-IMAGE TRANSLATION

Image-to-image (I2I) translation techniques aim to transfer images from a source domain to a corresponding images of a target domain. Pix2Pix [5] first uses a conditional GAN model to translate an image from one domain to another. Since then, their work has been extended for many scenarios: text-to-image synthesis [21], high-resolution synthesis [22], object removal [23], multi-style image synthesis [24] and face de-occlusion [25]. Despite promising performances, they have not utilized mutual interaction between two opposite mappings. In contrast, the proposed network effectively takes advantage of the mutual content information of cross modalities through a bidirectional synthesis framework.

Many studies have investigated face photo-to-sketch and face sketch-to-photo synthesis tasks as an image-to-image translation problem using GANs in their models [13], [14]. However, their methods are unable to effectively deal with the large domain gap between photo and sketch. For the last few years, great progress has been made in developing methods specifically designed for photo-sketch synthesis tasks. Yu *et al.* [26] incorporate facial composition information into their GAN based face photo-sketch synthesis. PS^2-MAN [15] takes an approach of gradually learning low-resolution to high-resolution images using multi-adversarial networks. Although these methods formulate photo-sketch transformation through end-to-end mapping, they do not utilize the mutual interaction between two modalities. To effectively reduce the modality gap for photo-sketch synthesis task, Col-cGAN [16] learns an intermediate modality between photo and sketch by utilizing the mutual interaction of the two opposite mapping. CUT [27] maximize the mutual interaction between different modalities based on contrastive learning of corresponding patches. StarGAN v2 [28] learns mapping between multiple modalities by utilizing a style encoder and mapping network. These approaches produce plausible results when the domain gap is small but struggles in cases where the domain gap is large. On the other hand,

Bae *et al.* [29] exploited a bidirectional synthesis network for face photo-sketch recognition.

B. USING DECOMPOSED IMAGE REPRESENTATION

The separation of an image into content and style components has widely been studied for artistic style transfer [1], [17], [30], [31]. Image synthesis can be achieved through image style transfer. Gatys *et al.* [17] showed that the feature statistics of a convolutional neural network could effectively capture the style information of an image. In particular, AdaIN [1] demonstrated impressive stylized outputs by simply aligning the channel-wise mean and variance of content input features to those of style input features. StyleGAN [7] used AdaIN operation at each convolution layer in their generative network to adjust the style of the image. Richardson *et al.* [10] introduced an encoder architecture built upon a pre-trained StyleGAN network. It directly generates a series of style vectors to solve image-to-image translation tasks, yielding impressive results. MUNIT [2] decomposed image representation into content and style codes. They recombined content code with random style code sampled from the style space of the target domain to produce cross domain outputs. SEAN [32] manipulated the style of an image via given style images and semantic masks. Chen *et al.* [20] subdivided the test photo into non-overlapping patches and tried to find the best matching photo from data samples to estimate the target style for photo-sketch synthesis task. Peng *et al.* [33] translated photo image into the style of the entire training sketch collection when training photos are unavailable. Although these methods give plausible results, they are unable to well preserve the structure of the transferred samples and often produce stylized results with messy texture.

III. PROPOSED METHOD

A. OVERVIEW

The overall architecture of our method is illustrated in Fig. 2. Our network consists of encoders E_A, E_B, BSTM (Bidirectional Style Transfer Module) units, two generators $G_{A\to B}$, $G_{B\to A}$ and two discriminators $D_{A\to B}$, $D_{B\to A}$. A, B denote two different modalities and $A \to B$, $B \to A$ represent the transformation from A to B and from B to A, respectively. The encoders consist of two main blocks, where each block consists of multiple layers. The encoders in each block first extract the individual features, F_A and F_B. The BSTM unit then decomposes each feature F_A, F_B into content and style components, denoted as C_A, C_B and S_A, S_B, respectively as shown in Fig. 2 (b). The cross style transferred features $F_{A\to B}$, $F_{B\to A}$ are obtained by exchanging the style components using AdaIN layer [1]. These transferred features $F_{A\to B}, F_{B\to A}$ are fed into the next block of the encoder and the same process is repeated. The two generators $G_{A\to B}$, $G_{B\to A}$ then alternatively map the original features F_A, F_B and the style-transferred feature $F_{A\to B}$, $F_{B\to A}$ into the desired output image space $I_{A\to B}$, and $I_{B\to A}$, respectively. Two discriminators $D_{A\to B}$, $D_{B\to A}$ are used to distinguish generated images from real sample by imposing the adversarial loss [3] on both modalities.

B. BIDIRECTIONAL STYLE TRANSFER MODULE (BSTM)

As stated earlier, a synthesis method that decomposes representation into content and style can bring great advantages to cross modal image synthesis [2]. In BSTM, the network learns individual domain characteristics and adopts the cross domain style by incorporating the transferred style factor into the content factor.

As shown in Fig. 2 (b), we first extract features F_A, F_B for images I_A, I_B in the first block of the encoders E_A and E_B, respectively.

$$F_A = E_1(I_A), F_B = E_2(I_B). \tag{1}$$

These features $F_A, F_B \in \mathbb{R}^{C\times H\times W}$, where W and H indicates spatial dimensions, and C the number of channels, are fed into a BSTM unit and are decomposed into content and style components. Channel-wise mean and standard deviation represent image style while normalized feature map represents content or shape in an image. We obtain style and content components as follows:

$$\mu\left(F_A\right) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F_{A_{hw}}, \tag{2}$$

$$\sigma\left(F_A\right) = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(F_{A_{hw}} - \mu(F_A))^2 + \epsilon}, \tag{3}$$

$$S_A = <\mu\left(F_A\right), \sigma\left(F_A\right)>,\ C_A = \frac{F_A - \mu\left(F_A\right)}{\sigma\left(F_A\right)}. \tag{4}$$

For simplicity, we show here only the style and content component computation for modality A. The style and content representations, S_B, C_B for modality B is computed in the same manner. This decomposed representation is then used to transfer the style components across modalities by simply scaling and shifting the content component of one modality with channel-wise mean (μ) and standard deviation σ, of the other modality. This produces the feature maps, $F_{A\to B}$ and $F_{B\to A}$ that contain the shape content of one modality with the appearance style of the other modality as follows:

$$F_{A\to B} = \mu\left(F_B\right)\cdot C_A + \sigma\left(F_B\right), \tag{5}$$

$$F_{B\to A} = \mu\left(F_A\right)\cdot C_B + \sigma\left(F_A\right). \tag{6}$$

Along with style transfer, we also align the shape contents, C_A and C_B from the two modalities by computing the l_1 distance between them. This process is repeated in the next block of the encoder.

C. ARCHITECTURE DETAILS

1) ENCODERS

The architecture of the proposed encoders is shown in Fig. 3 (a). The encoders consists of two main blocks. The first blocks consist of two convolution layers while the second

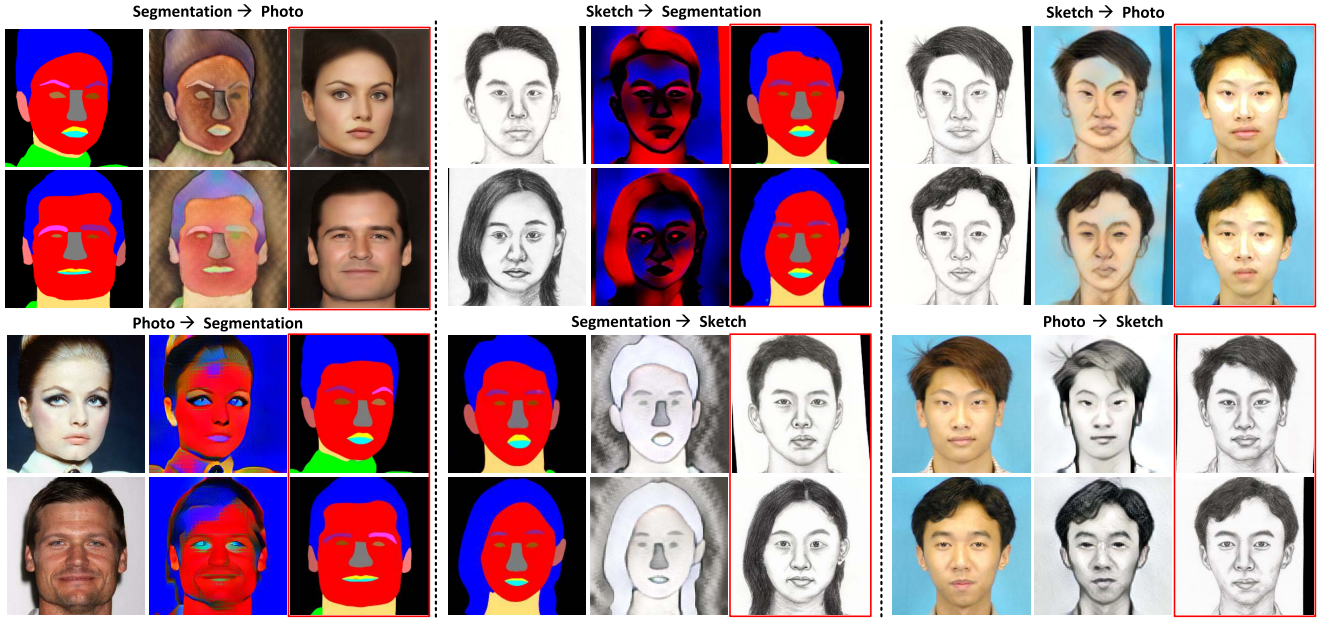
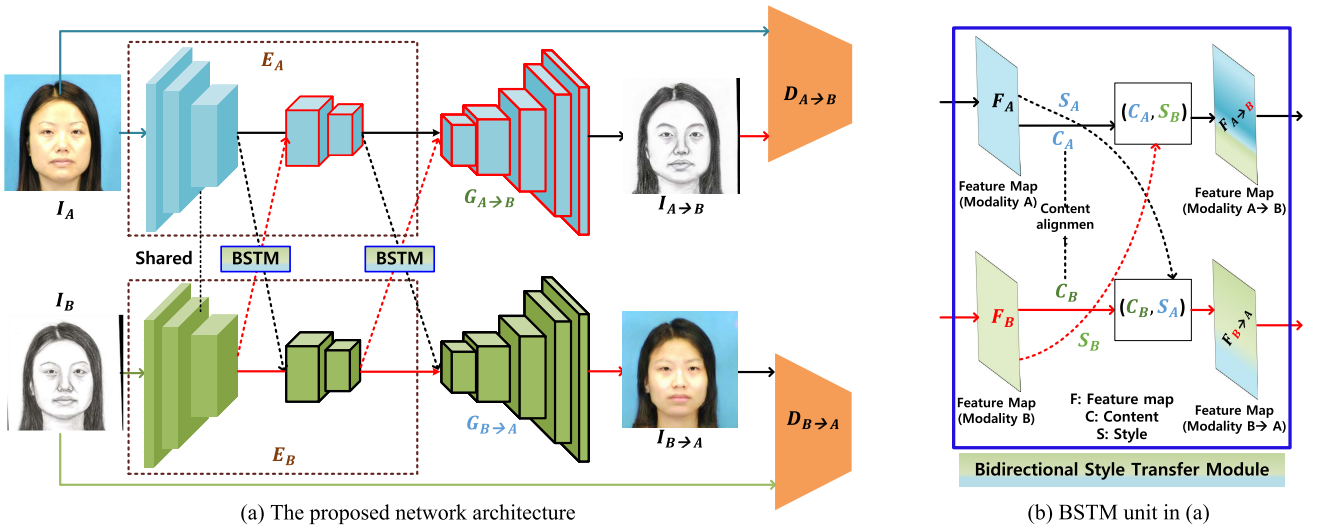


FIGURE 1. Unidirectional style transfer by AdaIN [1] vs. collaborative bidirectional style transfer proposed for cross modal facial image synthesis. We have observed that unidirectional application of style transfer in case of large modality gap does not work well. In order to warrant better synthesis quality, our network utilizes mutual interaction between two opposite mappings in a collaborative way. The three columns for each transformation problem shows the input (first column), the result for the unidirectional style transfer (second column), and the result for the proposed collaborative bidirectional transfer (third column).



(a) The proposed network architecture

(b) BSTM unit in (a)

FIGURE 2. An overview of the proposed network. (a) The proposed network involves two encoders E_A, E_B, BSTM (Bidirectional Style Transfer Module) units, and two generators $G_{A\to B}$, $G_{B\to A}$. The images I_A, I_B from the modality A, modality B are the input and the cross modal synthesized images $I_{A\to B}$, $I_{B\to A}$ are the outputs. (b) Detailed picture of the proposed BSTM in (a).

blocks are composed of one convolution layer and residual blocks [34]. The first blocks of the encoders share their weights. We apply Batch Instance Normalization (BIN) [35] to all the layers in the encoders.

2) GENERATORS

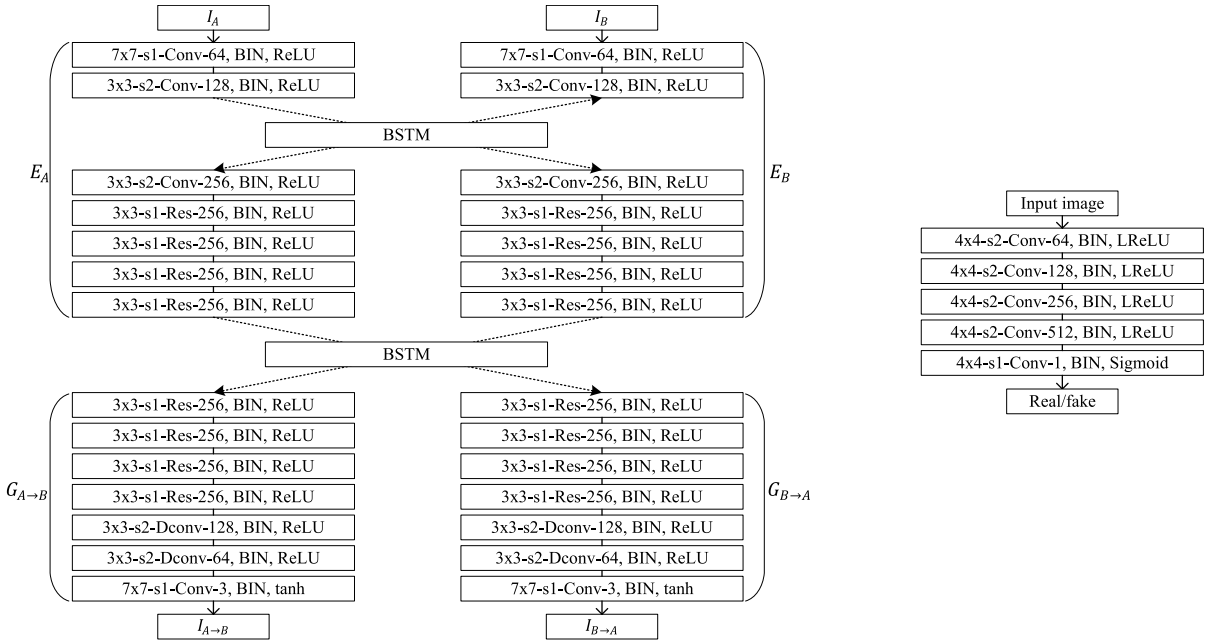
The architecture of the generator is a mirror copy of the encoders except that convolution is replaced by deconvolution layers as shown in Fig. 3 (a). The last layer of the generators uses *tanh* activation function.

3) DISCRIMINATORS

The architecture of both discriminators follows the one used in the pix2pix [5]. We use a patch-level discriminator that discriminates the image structure at the patch scale of 70 × 70. The details of the discriminator architecture is given in Fig. 3 (b).

D. TRAINING LOSS

We train our bidirectional network using the joint loss function in Eq. 7 which is a weighted combination of multiple



(a) Generator architecture (b) Discriminator architecture

FIGURE 3. The proposed model architecture. (a) The encoders and generators consist of a series of convolution layers and residual blocks. For example, 7 × 7-s1-Conv-64 denotes a 7-by-7 convolution layer of stride one with convolution filters 64. BIN indicates the batch instance normalization. Res and DConv denote Residual block and Deconvolution layer, respectively. (b) Our discriminator architecture largely follows the pix2pix [5] discriminator architecture. It takes the concatenation of the generated image and real image as input image and classifies it as real or fake at the patch level of 70 × 70.

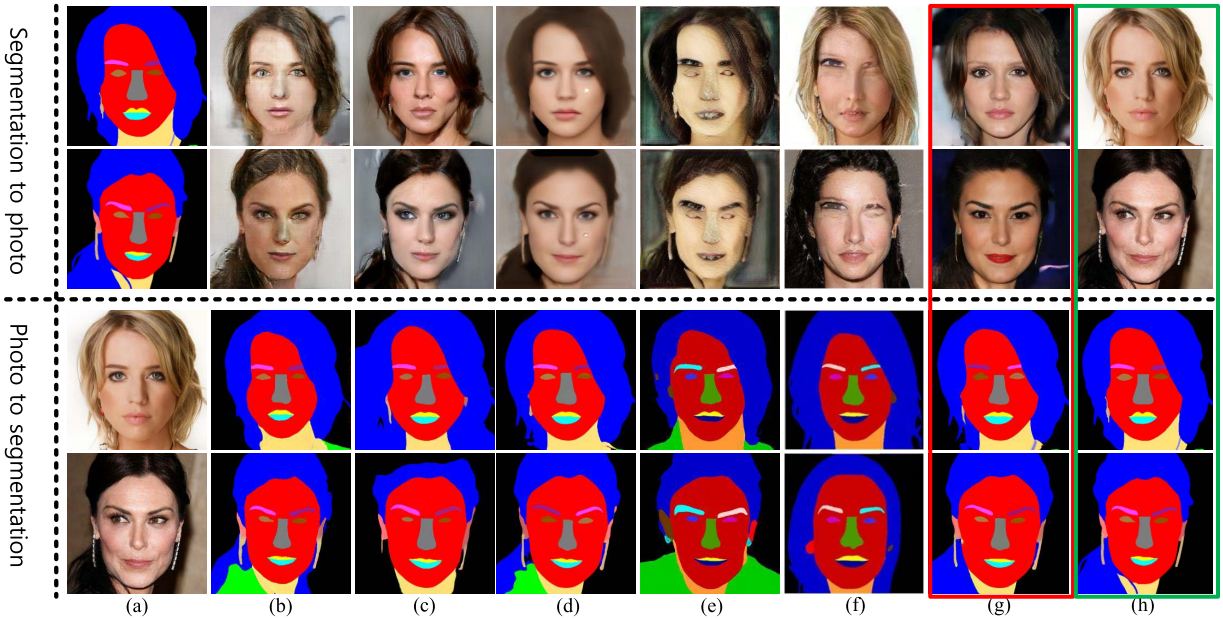


FIGURE 4. Comparison of photo ⇄ segmentation synthesis results on the CelebA-HQ dataset. Top two rows show results for segmentation-to-photo synthesis while bottom two rows present photo-to-segmentation synthesis results. From left to right: (a) Input, (b) Pix2Pix [5], (c) SPADE [12], (d) Col-cGAN [16], (e) CUT [27], (f) StarGAN v2 [28], (g) Ours, and (h) ground truth.

objectives.

$$L = \lambda_{GAN} L_{GAN} + \lambda_s L_s + \lambda_c L_c. \tag{7}$$

To generate real and natural looking synthetic outputs, we trained the bidirectional network using GAN loss function, L_{GAN} [3], along with the similarity loss, L_s. The similarity loss, L_s, measures pixel-wise l_1 distance and structural similarity (*SSIM*) between synthetic and real images. This similarity loss for both modalities is as follows:

$$\begin{aligned} L_s(A) &= L_{l_1}(A) + L_{ssim}(A), \\ L_s(B) &= L_{l_1}(B) + L_{ssim}(B). \end{aligned} \tag{8}$$

TABLE 1. Quantitative comparison of our method to the other state-of-the-art representative methods for photo ⇄ segmentation synthesis task. The best result are boldfaced.

Method	Segmentation → **Photo**		Photo → **Segmentation**
	SSIM	PSNR	mIoU
Pix2Pix [5]	0.229	9.291	0.686
SPADE [12]	**0.558**	13.514	0.500
Col-cGAN [16]	0.460	12.714	0.711
CUT [27]	0.351	10.801	0.142
StarGAN v2 [28]	0.303	9.688	0.169
Ours	0.519	**13.726**	**0.734**

TABLE 2. Quantitative comparison of our method to the other state-of-the-art representative methods for photo ⇄ sketch synthesis task. The best result are boldfaced.

Method	Sketch → **Photo**		Photo → **Sketch**	
	SSIM	PSNR	SSIM	PSNR
Pix2Pix [5]	0.512	10.824	0.451	9.762
PS^2-MAN [15]	0.558	9.624	0.575	9.7205
Col-cGAN [16]	0.692	18.533	0.583	14.085
CUT [27]	0.616	10.533	0.533	9.084
StarGAN v2 [28]	0.536	10.620	0.443	9.494
Ours	**0.694**	**18.689**	**0.642**	**17.771**

TABLE 3. Quantitative comparison of our method to the other state-of-the-art representative methods for sketch ⇄ segmentation synthesis task. The best result are boldfaced.

Method	Segmentation → **Sketch**		Sketch → **Segmentation**
	SSIM	PSNR	mIoU
Pix2Pix [5]	0.429	9.750	0.469
SPADE [12]	0.558	12.899	0.434
Col-cGAN [16]	0.579	14.479	**0.470**
CUT [27]	**0.597**	12.417	0.372
StarGAN v2 [28]	0.467	9.645	0.257
Ours	**0.597**	**15.263**	0.469

L_{l_1} loss is the pixel difference between the generated image and the ground truth as:

$$L_{l_1}(A) = \mathbb{E}_{a,b}\left[\|I_{A\to B} - I_B\|\right],$$
$$L_{l_1}(B) = \mathbb{E}_{a,b}\left[\|I_{B\to A} - I_A\|\right]. \quad (9)$$

SSIM measures the structural similarity between the generated and real samples and its corresponding loss function is written as:

$$L_{ssim}(A) = 1 - SSIM(I_{A\to B} - I_B),$$
$$L_{ssim}(B) = 1 - SSIM(I_{B\to A} - I_A). \quad (10)$$

We also introduce a collaborative loss, L_c, that minimizes l_1 distance between C_A and C_B of the same identity. This helps enforcing and regularizing the same content distribution for modality A, and modality B, in the content feature space.

λ_{GAN}, λ_s, and λ_c in Eq. (7) are the weight coefficients used to control the relative importance of each loss function. We have empirically found that $\lambda_{GAN} = 1$, $\lambda_s = 10$, and $\lambda_c = 0.25$ produce best results in our experiments.

E. IMPLEMENTATION AND TRAINING DETAILS

For the task of segmentation ⇄ photo synthesis in Sec. IV-A, we use the CelebAMask-HQ dataset [36] that has the total of 30,000 pairs of face photo and corresponding segmentation mask. Out of these, we use 25,000 paired samples for training and the rest of the samples for inference. We use the photo/sketch paired CUFS dataset [37] for photo ⇄ sketch synthesis task in Sec. IV-B. This dataset contains 168 samples for training and 142 for test. For the sketch ⇄ segmentation synthesis task in Sec. IV-C, we have constructed our own dataset as there are no currently available public datasets for colored segmentation map with corresponding sketches. More details about this dataset is described in Sec. IV-C. For all experiments, we use images of size 272 × 272, which are randomly cropped to 256 × 256 for training. We train our model for 5,000 epochs for photo ⇄ sketch in Sec. IV-B and sketch ⇄ segmentation synthesis tasks in Sec. IV-C, and for 200 epochs for photo ⇄ segmentation synthesis task in Sec. IV-A.

We train our model in three steps. For one third of the iterations, we first train the part of the network for one directional synthesis with the synthesis in the opposite direction fixed. We then train the network for another one third of the iterations for the synthesis in the opposite direction with the already trained part fixed. For the remaining iterations, we train the network for the bidirectional synthesis with the BSTM units on. Our model alternatively uses BSTM units. For example, in one epoch we train our network using BSTM, while in the next epoch we do not use BSTM. However, we apply shape content alignment throughout the training epochs. This training scheme helps our model overcoming the problem of directly utilizing style transfer technique for image synthesis and producing results with correct structure and stylized results with smooth texture. In inference time, we do not use the BSTM module for our results except content-style manipulated image synthesis.

IV. APPLICATION AND EXPERIMENTS

We give the performance evaluation of our method of bidirectional cross modal facial image synthesis for photo ⇄ segmentation in Sec. IV-A, photo ⇄ sketch in Sec. IV-B and sketch ⇄ segmentation in Sec. IV-C, respectively. We train all the methods to be compared, except Col-cGAN [16], in two opposite directions separately as they do not support bidirectional synthesis.

A. PHOTO ⇄ SEGMENTATION SYNTHESIS

For this task, the collaborative loss weightage is kept small, $\lambda_c = 0.1$. We do this because photos in the training data contain background information while no background information is available in the segmentation images. Otherwise, a large value of λ_c sometimes produces photo images with messy background.

Results: We compare the performance of our method with that of Pix2Pix [5], SPADE [12], Col-cGAN [16], CUT [27], and StarGAN v2 [28] on the CelebAMask-HQ dataset [36]. Top two rows in Fig. 4 compare the results for synthesized segmentation map from photo images and bottom two rows for synthesized photo images from segmentation map, respectively. As can be seen in the first two rows of Fig. 4, synthesized photos produced by Pix2pix and SPADE contain deformation for complex face semantics. Moreover, Pix2pix also yields noise and messy face texture. Col-cGAN gives

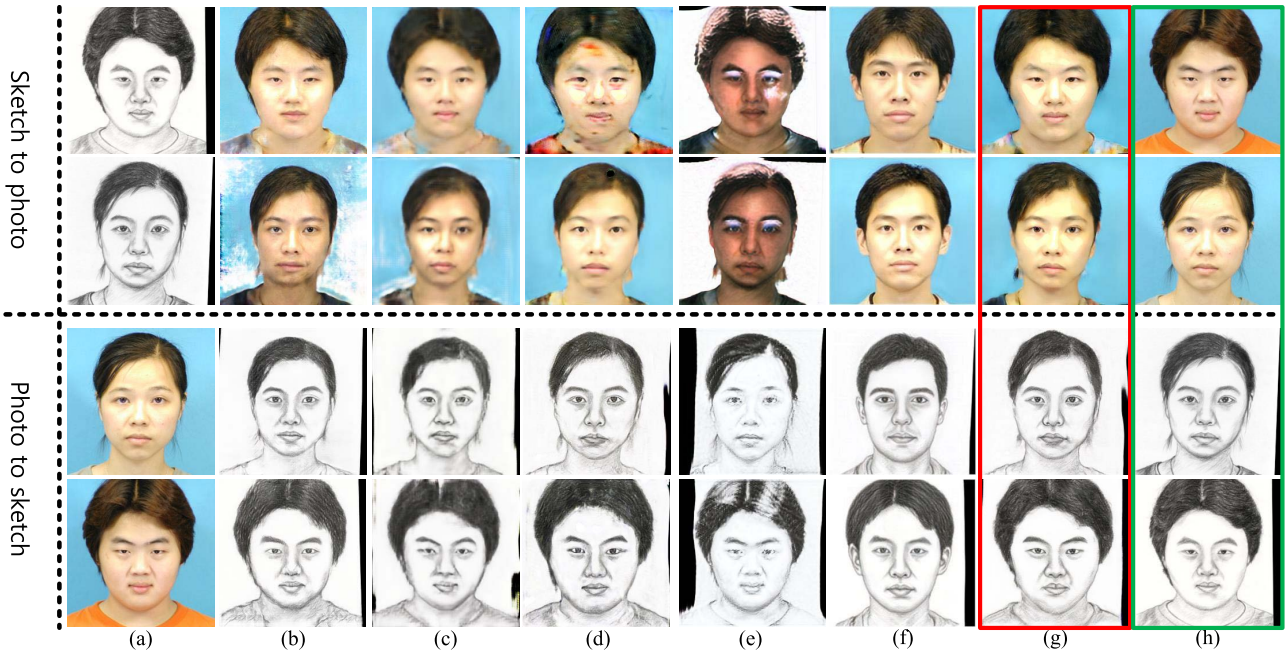


FIGURE 5. Comparison of photo ⇄ sketch synthesis results on the CUFS dataset. Top two rows show results for sketch-to-photo synthesis while bottom two rows present photo-to-sketch synthesis results. From left to right: (a) Input, (b) Pix2Pix [5], (c) PS^2-MAN [15], (d) Col-cGAN [16], (e) CUT [27], (f) StarGAN v2 [28], (g) Ours, and (h) ground truth.

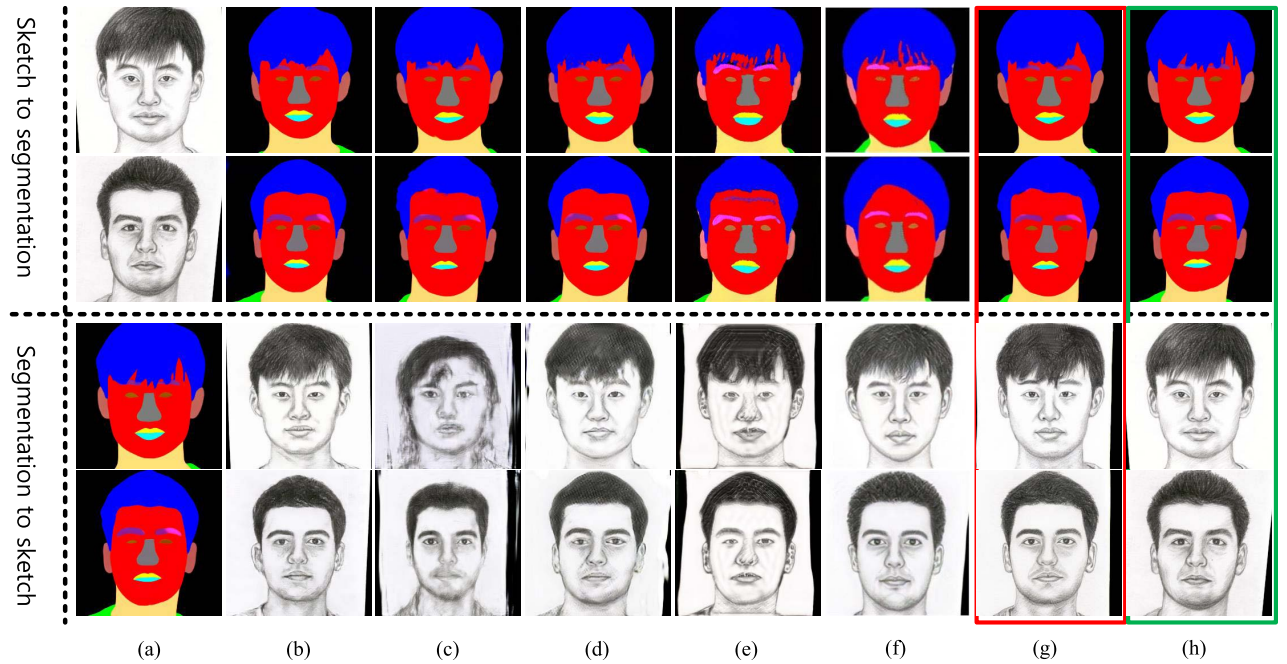


FIGURE 6. Comparison of sketch ⇄ segmentation synthesis results. Top two rows show results for sketch-to-segmentation synthesis while bottom two rows presents segmentation-to-sketch synthesis results. From left to right: (a) Input, (b) Pix2Pix [5], (c) SPADE [12], (d) Col-cGAN [16], (e) CUT [27], (f) StarGAN v2 [28], (g) Ours, and (h) ground truth.

better results compared to Pix2Pix and SPADE, but still produces blurred effects and dotted artifacts. CUT and StarGAN v2 fail to produce complex region of the face, e. g., the eye region is severely distorted. In contrast, our method generates sharp photo images with finer details. For synthesizing segmentation map from photo, SPADE and CUT do not provide plausible output. StarGAN v2 not only changes the identity, but also fails to produce the correct segmentation map for the hair region. Pix2Pix and Col-cGAN generate plausible results, however, they still cannot preserve the finer details, e. g., the earrings and the strap on the neck in the third row of Fig. 4.

We also provide quantitative comparisons in Table 1. We use Structural SIMilarity (SSIM) and Peak Signal to Noise Ratio (PSNR) for segmentation→photo and mean Intersection-over-Union (mIoU) for photo→segmentation. Table 1 indicates that our method outperforms the other methods in terms of PSNR and mIoU, but SPADE gives the best SSIM score.

We have additionally experimented on the FFHQ-Aging dataset [38] for photo ⇄ segmentation synthesis. The FFHQ-Aging dataset is built based on the FFHQ face photo dataset [7] to be used for face aging related tasks. The segmentation maps in the FFHQ-Aging dataset are generated through a pre-trained Deeplab v3 network [39]. However, they contain many inaccurate and mislabeled segmentation maps which prevent proper training of the network for photo ⇄ segmentation synthesis. We think that the performance evaluation on this dataset is not informative. For reference, we have included the experimental results on this dataset in the supplementary material.

B. PHOTO ⇄ SKETCH SYNTHESIS

Photo ⇄ sketch synthesis is a challenging task due to large modality gap between the two modality and lack of sufficient paired training data. We compare the performance of our method with those of Pix2Pix [5], PS^2-MAN [15], Col-cGAN [16], CUT [27], and StarGAN v2 [28] on the CUFS database [37].

Results: Fig. 5 shows qualitative comparison for synthesized photos and sketches. Top two rows of Fig. 5 show results for sketch-to-photo synthesis while bottom two rows present photo-to-sketch synthesis results. In top two rows, we can see that the Pix2Pix, PS^2-MAN and Col-cGAN not only yield blurred effects but also contain prominent dotted artifacts. Unsupervised approaches such as CUT and StarGAN v2 do not yield plausible photos from sketch. CUT generates photos with unnatural skin color while StarGAN v2 fails to maintain the identity of the input sketch. Also, sketches generated by those methods are unable to well preserve the artistic appearance such as sketch-line texture. For example, PS^2-MAN and CUT are not capable of producing those pencil lines while Pix2Pix and Col-cGAN blend those pencil line shadows. In contrast, our method not only retains the face identity but also produces sharp and realistic sketches, i. e., sketch-like texture on hair region and pencil line shadows.

Table 2 shows quantitative comparisons using SSIM and PSNR. Our method achieves the best performance for both tasks.

C. SKETCH ⇄ SEGMENTATION SYNTHESIS

Synthesizing sketch images from color-coded segmentation map is a very challenging task. To the best of our knowledge, there are no research works that presented results on this task. Although a color coded semantic segmentation map provides enough information about face semantics, it contains no information about artistic appearance of face. Sketches add more complexity as the artistic appearance are very minute, e. g., pencil lines on the face in CUFS database [37]. However, we think that some state-of-the-art image synthesis frameworks such as Pix2Pix [5], SPADE [12], Col-cGAN [16], CUT [27], and StarGAN v2 [28] can be used to synthesize sketches from segmentation and segmentation from sketch samples. For this, we have trained all those methods with our constructed dataset.

TABLE 4. A user preference study. The numbers indicate user preference percentage for the proposed method over those of the compared methods. The best result are boldfaced. Note: Seg. denotes color coded segmentation map.

Method	Photo ⇄ Sketch		Photo ⇄ Seg.		Sketch ⇄ Seg.	
	Photo	Sketch	Photo	Seg.	Sketch	Seg.
Pix2Pix [5]	13.23	17.67	4.42	8.85	8.35	24.50
PS^2-MAN [15]	18.62	8.35	NA	NA	NA	NA
Col-cGAN [16]	11.77	24.50	7.42	13.22	21.15	18.50
SPADE [12]	NA	NA	3.95	2.45	2.45	17.75
CUT [27]	6.40	11.27	0.50	0.98	0.50	5.5
StarGAN v2 [28]	0.00	0.00	0.00	0.50	2.00	0.75
Ours	**49.98**	**38.20**	**83.71**	**74.00**	**65.55**	**33.00**

1) DATASET CONSTRUCTION

Currently, there are no publicly available datasets to train sketch ⇄ segmentation synthesis task in a supervised manner. For this, we have created a dataset for color coded segmentation map and their corresponding sketches using the publicly available photo/sketch paired dataset (CUFS) [37]. To achieve this, we use the model trained for the photo ⇄ segmentation synthesis task. We translate all photos from the CUFS dataset into segmentation map and use those synthesized segmentation maps along with the corresponding sketches as paired segmentation/sketch samples. Fig. 7 shows examples of pairs we have created for this task.

2) RESULTS

Results for sketch ⇄ segmentation synthesis are illustrated in Fig. 6. Pix2Pix, SPADE, Col-cGAN, CUT, and StarGAN v2 obtain almost equivalent results for segmentation outputs from a given sketch. However, they are unable to produce plausible sketches from a segmentation map. As can be seen in the last two rows of Fig. 6, SPADE and CUT fail to produce plausible sketches from segmentation map. Col-cGAN outputs are blurred and totally ignore sketch-like appearance styles in hair region and pencil line shadows. Also, they show artifacts on hair texture. StarGAN v2 produces plausible results, but fails to synthesize hair region with finer details. Pix2Pix blends the pencil line shadows and does not give plausible face semantics, e. g., ears in the third row of Fig. 6. In contrast, our method not only produces visually pleasing results, but also obtains more diverse outputs that better retain finer details, especially in segmentation-to-sketch synthesis.

We also provide quantitative comparisons in Table 3 using SSIM and PSNR for segmentation→sketch and mIoU for sketch→segmentation. Our method achieves the best SSIM and PSNR scores for segmentation→sketch. For sketch→segmentation, Pix2Pix, Col-cGAN, and our method yield equivalent performance.

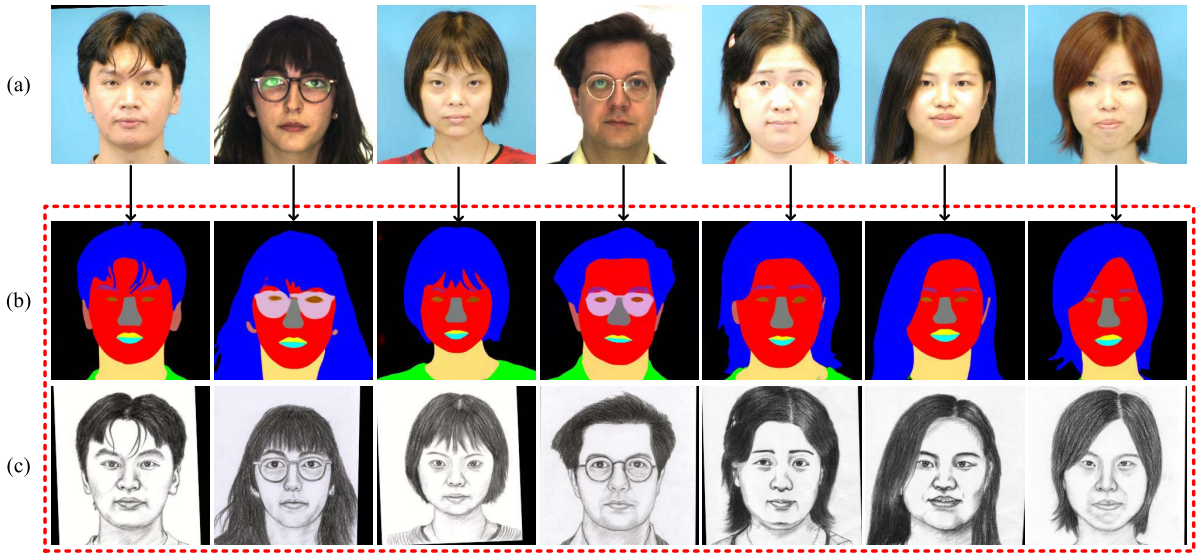


FIGURE 7. Examples of paired data (a) and (c) from our constructed dataset for sketch ⇄ segmentation synthesis task. (a) The input photo image from the CUFS dataset [37]. (b) Synthesized segmentation map of (a) generated from photo ⇄ segmentation synthesis task by our method. (c) The corresponding sketches for the input photo image in the CUFS dataset in (a). The red dotted box shows the constructed segmentation/sketch paired dataset we used for sketch ⇄ segmentation synthesis task.

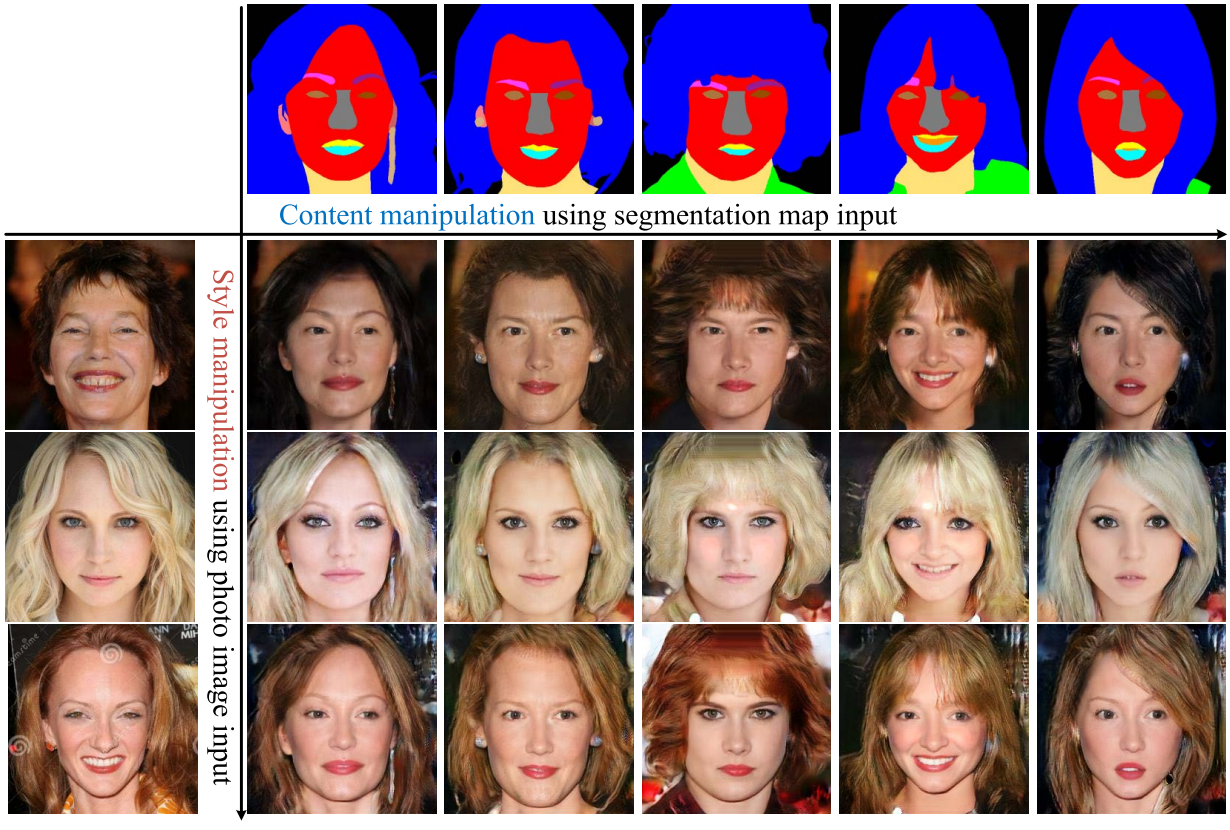


FIGURE 8. Style/content manipulated image synthesis results from segmentation maps. The results display the capability of our network that synthesizes outputs with diverse appearance style for the same content shape.

D. STYLE/CONTENT MANIPULATED IMAGE SYNTHESIS

Style/content manipulated image synthesis aims at generating multiple photo-realistic images for the same shape content with various appearance styles. To achieve this, we use the model trained for photo ⇄ segmentation synthesis in Sec. IV-A. Unlike the other synthesis tasks in Sec. IV-A ∼ IV-C, we exploit BSTM units in inference time. For the same input segmentation map, we use different photo

image inputs from the test dataset to generate photo images with the appearance style of the selected photo. Our model extracts the shape content information from the segmentation map and style information from photo images using the BSTM units. This results in outputs containing the shape content similar to that of the input segmentation map and appearance style similar to that of the input photo image. Fig. 8 demonstrates that our model is capable of generating multiple high-quality photo realistic images for a same identity. More results are included in the supplementary material.

V. USER STUDY

We have additionally performed a pilot user study to evaluate our results using perceptual assessment of people. We have asked fifty two participants to select which output looks more realistic and natural. Each participant is given the total of twenty four questions, four questions for each synthesis task. For every test sample, participants are shown input image along with six images synthesized by different methods for the given input. Table 4 shows that our method significantly outperforms the other representative methods in all three bidirectional synthesis tasks.

We think that for performance comparison, a user study like ours can give better performance evaluation because except for segmentation, there is no perfect quantitative evaluation metric that quantifies the quality of generated image.

VI. CONCLUSION

This research features a novel collaborative bidirectional style transfer network for cross modal image synthesis. In our method, we effectively exploit mutual interaction between two opposite mappings to align the content from two modalities and exchange their appearance styles for cross modal facial image synthesis. Extensive evaluation demonstrates the effectiveness of our model for bidirectional synthesis, between segmentation and photo, between photo and sketch, and between sketch and segmentation. Moreover, the proposed methodology can be adapted for solving other cross modal image synthesis tasks. We also think that our method can be applied to generative methods for cross modal image matching because better synthesis results are very likely to lead to better matching accuracy.

REFERENCES

[1] 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. 1501–1510.
[2] 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.
[3] 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.
[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[5] 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.*, Jul. 2017, pp. 1125–1134.
[6] 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.*, Oct. 2017, pp. 2223–2232.
[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.
[8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.
[9] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9243–9252.
[10] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.
[11] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "StyleRig: Rigging StyleGAN for 3D control over portrait images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6142–6151.
[12] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.
[13] J. Yu, X. Xu, F. Gao, S. Shi, M. Wang, D. Tao, and Q. Huang, "Towards realistic face photo-sketch synthesis via composition-aided GANs," 2017, *arXiv:1712.00899*.
[14] N. Wang, W. Zha, J. Li, and X. Gao, "Back projection: An effective postprocessing method for GAN-based face sketch synthesis," *Pattern Recognit. Lett.*, vol. 107, pp. 59–65, May 2018.
[15] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 83–90.
[16] M. Zhu, J. Li, N. Wang, and X. Gao, "A deep collaborative framework for face photo–sketch synthesis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3096–3108, Oct. 2019.
[17] 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.
[18] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 694–711.
[19] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2479–2486.
[20] C. Chen, X. Tan, and K.-Y.-K. Wong, "Face sketch synthesis with style transfer using pyramid column feature," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 485–493.
[21] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.
[22] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[23] N. U. Din, K. Javed, S. Bae, and J. Yi, "Effective removal of user-selected foreground object from facial images using a novel GAN-based network," *IEEE Access*, vol. 8, pp. 109648–109661, 2020.
[24] S. Bae, N. Ud Din, K. Javed, and J. Yi, "Efficient generation of multiple sketch styles using a single network," *IEEE Access*, vol. 7, pp. 100666–100674, 2019.
[25] N. Ud Din, K. Javed, S. Bae, and J. Yi, "A novel GAN-based network for unmasking of masked face," *IEEE Access*, vol. 8, pp. 44276–44287, 2020.
[26] J. Yu, X. Xu, F. Gao, S. Shi, M. Wang, D. Tao, and Q. Huang, "Toward realistic face photo–sketch synthesis via composition-aided GANs," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4350–4362, Sep. 2020.
[27] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 319–345.
[28] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8188–8197.

[29] S. Bae, N. Ud Din, H. Park, and J. Yi, "Face photo-sketch recognition using bidirectional collaborative synthesis network," 2021, *arXiv:2108.09898*.
[30] Z. Zou, T. Shi, S. Qiu, Y. Yuan, and Z. Shi, "Stylized neural painting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15689–15698.
[31] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCV)*, Sep. 2018, pp. 1–16.
[32] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "SEAN: Image synthesis with semantic region-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5104–5113.
[33] C. Peng, N. Wang, J. Li, and X. Gao, "Universal face photo-sketch style transfer via multiview domain translation," *IEEE Trans. Image Process.*, vol. 29, pp. 8519–8534, 2020.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[35] H. Nam and H.-E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," 2018, *arXiv:1805.07925*.
[36] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5549–5558.
[37] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2008.
[38] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman, "Lifespan age transformation synthesis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 739–755.
[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

NIZAM UD DIN received the Bachelor of Science (B.Sc.) degree in electrical (computer) engineering from the COMSATS Institute of Information Technology, Pakistan, in 2013, the Master of Science (M.Sc.) degree in computer engineering from the University of Engineering and Technology Taxila, Taxila, Pakistan, in 2016, and the Ph.D. degree in computer engineering from Sungkyunkwan University, South Korea, in 2021.

He was a Visiting Faculty Member at Quaid-i-Azam University, Islamabad, Pakistan, from 2016 to 2017. He is currently working as an AI Research Scientist at Saudi Federation for Cybersecurity, Programming and Drones, Riyadh, Saudi Arabia. His research interests include developing artificial intelligence systems specifically, deep learning-based systems to solve computer vision problems. His awards and honors include a University Scholarship for B.Sc. degree, in 2009, and a Higher Education Commission HRDI—Faculty Development of UESTPSUETS Scholarship for Ph.D. degree from HEC, Pakistan, in 2017.

SEHO BAE received the B.S. degree from the Department of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering. He has been a member of the Computer Vision Laboratory, Sungkyunkwan University, since 2015. His research interests include cross-modal image matching and cross-modal image synthesis using generative adversarial networks.

KAMRAN JAVED received the B.Sc. degree (Hons.) in electronic engineering and the M.Sc. degree in computer engineering from the University of Engineering and Technology (UET) Taxila, Taxila, Pakistan, in 2012 and 2014, respectively, and the Ph.D. degree in electronic and computer engineering from Sungkyunkwan University, South Korea, in 2020.

He was a Lecturer with the Electronic Engineering Department, UET Taxila, from 2013 to 2016, and an Assistant Professor with the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan, from 2020 to 2021. He is currently working as an AI Research Scientist with the National Center of Artificial Intelligence (NCAI), Creativity Research Department, Riyadh, Saudi Arabia. His research interests include generative adversarial networks and its application to computer vision for image unmosaicing and object removal. His awards and honors include the Award of Honors for B.Sc. degree, in 2012, University Scholarship for M.Sc. degree, in 2012, and Higher Education Commission Scholarship for Ph.D. degree (abroad) from HEC, Pakistan, in 2016.

HYUNKYU PARK received the B.S. degree from the Department of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea, in 2019. He is currently pursuing the Ph.D. degree in electrical and computer engineering. He has been a member of the Computer Vision Laboratory, Sungkyunkwan University, since 2019. His research interests include image harmonization using generative adversarial networks and segmentation.

JUNEHO YI received the B.S. degree from Seoul National University, South Korea, in 1985, the M.S. degree from Pennsylvania State University, University Park, PA, USA, in 1987, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1994, all in electrical engineering. In 1989, he was a Research Scientist with the Samsung Advanced Institute of Technology. From 1994 to 1995, he was a Research Scientist at the University of California at Riverside, Riverside. From 1995 to 1996, he was a Senior Research Scientist at the Korea Institute of Science and Technology, Seoul, South Korea. Since 1997, he has been with Sungkyunkwan University, South Korea, where he is currently a Professor with the School of Electronic and Electrical Engineering. His pioneering works include masked fake face detection and depth filtering using parametrized structured light imaging. His research interests include the areas of computer vision and statistical pattern recognition.