Biphasic Face Photo-Sketch Synthesis via Semantic-Driven Generative Adversarial Network With Graph Representation Learning

Xinggun Qi*, Muyi Sun*, Zijian Wang, Jiaming Liu, Member, IEEE, Qi Li, Member, IEEE, Fang Zhao†, Shanghang Zhang*, and Caifeng Shan*, Senior Member, IEEE

Abstract—Biphasic face photo-sketch synthesis has significant practical value in wide-ranging fields such as digital entertainment and law enforcement. Previous approaches directly generate the photo-sketch in a global view, they always suffer from the low quality of sketches and complex photograph variations, leading to unnatural and low-fidelity results. In this article, we propose a novel semantic-driven generative adversarial network to address the above issues, cooperating with graph representation learning. Considering that human faces have distinct spatial structures, we first inject class-wise semantic layouts into the generator to provide style-based spatial information for synthesized face photographs and sketches. In addition, to enhance the authenticity of details in generated faces, we construct two types of representation graphs via semantic parsing maps upon input faces, dubbed the intraclass semantic graph (IASG) and the interclass structure graph (IRSG). Specifically, the IASG effectively models the intraclass semantic correlations of each facial semantic component, thus producing realistic facial details. To preserve the generated faces being more structure-coordinated, the IRSG models interclass structural relations among every facial component by graph representation learning. To further enhance the perceptual quality of synthesized images, we present a biphasic interactive cycle training strategy by fully taking advantage of the multilevel feature consistency between the photograph and sketch. Extensive experiments demonstrate that our method outperforms the state-of-the-art competitors on the CUHK Face Sketch (CUFS) and CUHK Face Sketch FERET (CUFSF) datasets.

Index Terms—Face photo-sketch synthesis, generative adversarial network, graph representation learning, intraclass and interclass, iterative cycle training (ICT).

I. INTRODUCTION

BIPHASIC face photo-sketch synthesis refers to generating sketches from face photographs and, conversely, generating photographs from face sketches. The wide-ranging application fields of the biphasic face photo-sketch synthesis include digital entertainment, law enforcement, and criminal case judgment. Specifically, face sketch is one of the most popular and fundamental portrait painting styles in the scope of digital entertainment [1]. In law enforcement and criminal case judgment, police commonly just hold the sketches of suspects drawn from the description of the witnesses. Face photographs synthesized from these sketches with clear identities and manifest features can provide a feasible way to promote the efficiency of justice criminal cases [2]. However, it requires a vast time and effort to create distinct face sketches by professional artists. Due to the vital practical value, it is especially essential to automatically synthesize face photographs and sketches with realistic effects and consistent identity preservation.

Numerous approaches have been proposed to address this task, which can be roughly grouped into exemplar-based approaches [3], linear-regression-based approaches [4], and generative-model-based ones [5]. Exemplar-based approaches [1], [3], [6], [7], [8] focus on bridging the mapping from photograph to sketch via limited exemplar-paired patches, leading to over-smoothed and low personal identity results. Besides, linear-regression-based approaches [9], [10], [11], [12] always directly construct the linear mapping between photographs and sketches. Recently developed generative-based methods [14], [15], [16], [17], [18], [19], [20] show better performance compared with the other two
types. However, although these methods could generate sketches or face photographs automatically, they adopted a holistic manner for generation and overlooked the local specificity of different facial regions. Therefore, the face directly generated by these methods might yield distortion and noises in detailed parts. Meanwhile, the fidelity of the image and the consistency of the face identity also need to be improved.

Motivated by the above studies, we propose a novel semantic-driven generative adversarial network with graph representation learning for biphasic face photo-sketch synthesis. Our model is built upon the key insight that the human face has obvious spatial structures that determine the fidelity and identity consistency of face images. Therefore, the structural information should be enhanced by incorporating external prior or learning by designing specific objectives during the training process. In this article, we design three methods to incorporate and learn structural information: saliency detection-based guidance, face semantic injection (SI), and two graph-based training objectives.

The saliency detection guidance utilizes the face saliency map as prior input to our network. Thanks to the recently advanced face saliency detection technique [21], we first leverage a pretrained saliency detector to obtain the saliency maps, which incorporate overall facial structural information by identifying the most conspicuous or prominent parts of face images. As displayed in Fig. 1, it is worth noting that the saliency map is different from the corresponding target sketch since the sketch contains more personal styles and textures of the painter, e.g., the shadows and outlines.

Then, considering the generated results should be structured consistent with the input human face, we perform face SI in the network. Specifically, we transform the class-wise semantic layouts into two modulation parameters and inject them into the network decoder to provide style-based spatial information. In this fashion, the facial structure is well-preserved as well and the person’s identity is effectively guaranteed during the generation. Here, we utilize a superior face parser to extract accurate semantic layouts for both photograph and sketch, as depicted in Fig. 1.

To ensure the details of generated results are authentic, we propose two training objectives, which employ two novel graph representations for measuring facial structured information dubbed the intraclass semantic graph (IASG) and the interclass structure graph (IRSG). In particular, IASG leverages the human face semantic layouts as guidance to model the intraclass correlation of each facial component represented by a graph node. Here, we calculate the graph node as the mean center and variance center of the corresponding facial components. Considering that different facial components always contain large-range pixel numbers, we present an adaptive reweighting algorithm to balance the contribution of detailed facial parts. Then, intraclass correlation can be effectively expressed by the similarity between the mean center (or variance center) and each pixel in the corresponding facial component. By enforcing the generated fake IASG to be consistent with the target ones, the details in the synthesized images achieve high-fidelity effects.

As a complement, the IRSG aims to keep the generated face photograph and sketch more structure-coordinated with targets, globally. IRSG models the interclass relations among every facial component. Concretely, the interclass relations are built by computing the affinity scores between every two different mean centers (and variance centers). Similar to IASG, we utilize the target IRSG to supervise the synthesized fake ones, thus producing the natural human faces.

Moreover, based on the observation that the paired photo-sketch shared many personal characters, we designed a novel biphasic iterative cycle training (ICT) strategy to improve the perceptual quality of synthesized images. Here, we first train two models of photograph generation and sketch generation, respectively. Then, we exploit the pretrained sketch generation model as a personal knowledge extractor to boost photograph generation. Iteratively, the sketch generation model is improved with the help of a pretrained photograph personal knowledge extractor. In this fashion, our framework enables high-quality biphasic face photo-sketch synthesis through multistage iterations. Extensive experiments demonstrate that our method significantly outperforms various counterparts, displaying natural and realistic facial details.

The main contributions of this study are summarized as follows.

1) We propose a novel semantic-driven generative adversarial network with graph representation learning for generating realistic photographs and distinct sketches.
2) We construct two types of representational graphs and design corresponding constraints to facilitate the
preservation of the details and coordinate structures in generated face photographs and sketches.

3) We propose a novel biphasic ICT to improve the perceptual quality of the synthesized images by effectively taking advantage of the multilevel feature consistency between the photograph and sketch.

4) Extensive comparison experiments are conducted on CUF S and CUFSF datasets, showing our method obtains state-of-the-art performance.

Compared to our preliminary work in [22], the improvements and extensions are consulted threefold: 1) we propose a novel IRSG that facilitates the preservation of the details and personal identities in generated face photographs and sketches; 2) a novel biphasic ICT strategy is proposed to improve the perceptual quality of synthesized images; and 3) we conduct the extension experiments on both sketch synthesis and face photograph synthesis tasks. Driven by the aforementioned improvements, our framework achieves superior performance on biphasic photo-sketch synthesis in a unified manner. This significantly facilitates the application community on digital entertainment and law enforcement.

II. RELATED WORK

A. Biphasic Face Photo-Sketch Synthesis

Biphasic face photo-sketch synthesis has developed rapidly in the last few decades. Massive works have been proposed to solve this problem, which include two closely related subtasks: face sketch synthesis and face photograph synthesis. Therefore, researchers often analyze and discuss these two subtasks in a unified manner.

Deep neural network-based approaches are the mainstream routine of biphasic photo-sketch synthesis in recent years, which have gradually emerged with the boost of generative adversarial networks. Zhang et al. [16] employed multidomain adversarial methods to construct a mapping from photo-domain to sketch-domain. Zhu et al. [19] borrowed knowledge from transfer learning and proposed a lightweight network supervised by a high-performance larger network. Zhang et al. [23] embedded the photograph parsing priors and designed a parametric sigmoid activation function in a GAN-based framework to facilitate robust sketch synthesis. Recently, Yu et al. [20] decomposed the face parsing layouts into multiple compositions and encoded them into conditional GAN (cGAN) for biphasic face photo-sketch synthesis which achieved state-of-the-art performance. Moreover, Duan et al. [18] introduced the gradient-based self-attention mechanism to combine the global residual connection and local residual connection in the proposed network which achieved better results. Besides, Zhang et al. [24] designed a dual transfer strategy to promote biphasic face photo-sketch synthesis. Lin et al. [25] proposed the feature injection module to preserve the identity of synthesized sketches and photographs. However, the sketches and photographs synthesized by these methods are short of realistic detailed depictions.

Inspired by previous works, we inject the semantic information into the generator of our proposed network. However, our injection module is different from previous works which embedded the multilevel identity feature into the network. In contrast, we aim to provide class-wise style-based spatial supervision for synthesized face photographs and sketches. Furthermore, the previous work [25] did not follow the common experiment setting which ignored the influence of background on the generated results.

B. Paired Image-to-Image Style Transfer

Paired image-to-image style transfer tasks can be regarded as a combination of image-to-image translation tasks and style transfer tasks. The image-to-image translation is often formulated as pixel-wise image generation tasks applied with paired images like biphasic face photo-sketch synthesis. Isola et al. [26] proposed a cGAN architecture to solve the image-to-image translation task with paired input and output named Pix2Pix. Due to the eminent performance of Pix2Pix on the paired dataset, researchers have made numerous improvements based on Pix2Pix and applied them to a wide range of research fields [27], [28], [66], [67]. By combining Pix2Pix and residual blocks, Wang et al. [29] proposed a novel network architecture to generate high-resolution images named pix2pixHD. Moreover, Park et al. [30] introduced the semantic layouts as spatial supervision injected in the pix2pixHD for synthesizing photorealistic images. Motivated by previous research, we exploit cGAN like Pix2Pix as our backbone network.

The biphasic face photo-sketch synthesis task can be treated as an image style transfer task between realistic photographs and vivid sketch portraits. Gatys et al. [31] successfully applied pretrained CNNs to the image style transfer task. Furthermore, Ulyanov et al. [32], [33] optimized the style transfer process by manipulating the batch normalization (BN) layers and instance normalization (IN) layers. Dumoulin et al. [34] utilized a group of parameters to realize the transfer of various image styles. Consecutively, Huang and Belongie [35] proposed the adaptive IN (AdaIN) layers which could perform arbitrary style transfer without training repeatedly. Recently, Wang et al. [13] put forward the spatially-adaptive normalization (SPADE) layers that inject the image style from the semantic layouts to obtain photorealistic images. Motivated by the previous research, we inject the style-based statistic information into the network to generate images with more distinct characteristics.

C. Graph Representation Learning

Graph representation learning plays a significant role in computer vision tasks which could encode each node represented by a low-dimensional dense embedding in the graph structure [36]. Perozzi et al. [37] effectively embedded information of nodes by randomly sampling the graph structure. Furthermore, Defferrard et al. [38] designed a novel pooling strategy to rearrange the nodes to preserve more useful information about the graph. Li et al. [39] constructed the structural graph of actions to capture the high-order dependencies between successive skeletons in the proposed actional-structural graph convolution network for skeleton-based action recognition. Besides, Ren et al. [40] adopted the graph generator to build the connections among...
spatial parts and construct the feature graph of these node’s representation for biometrics. Yang et al. [41] utilized a novel graph representation specifically designed for sketches by bridging the structural hierarchical relationship of sketches. This work has significantly improved the recognition accuracy of sketches. Wu et al. [64] proposed an adaptive graph representation learning scheme for video person Re-ID, which enables the contextual interactions between relevant regional features. Jin et al. [65] introduced a self-supervised approach to learning graph node representations by enhancing Siamese self-distillation with multiscale graph representation learning. As for the biphasic face photo-sketch synthesis task, Zhu et al. [7] combined graphical exemplar-based features with deep neural networks to synthesize high-quality sketches. This work has significantly improved the recognition accuracy of sketches. Wu et al. [64] proposed an adaptive graph representation learning scheme for video person Re-ID, which enables the contextual interactions between relevant regional features. Jin et al. [65] introduced a self-supervised approach to learning graph node representations by enhancing Siamese self-distillation with multiscale graph representation learning. As for the biphasic face photo-sketch synthesis task, Zhu et al. [7] combined graphical exemplar-based features with deep neural networks to synthesize high-quality sketches. The results synthesized by [7] are robust against lighting variations and clutter backgrounds. However, they still suffer from over-smoothing and lacking specific-identify characters in the synthesized photographs and sketches. On the contrary, our graph representation learning algorithms can restrain intra-class features and interclass features to help keep the realistic personal details of synthesized photographs and sketches.

III. Method

In this section, details about our proposed semantic-driven generative adversarial network with graph representation learning are presented. First, we introduce the preliminaries and problem formulation of our method. Afterward, the network architecture is described. Subsequently, we elaborate on the graph representation learning algorithms and biphasic ICT strategy. Finally, the objective functions of our model are introduced.

A. Preliminaries

Our semantic-driven network aims to construct a biphasic mapping of paired photo-sketch by utilizing class-wise semantic layouts as guidance. Previous researchers directly synthesize sketches or photographs in a holistic manner, leading to unnatural and low-fidelity details in results. Our key insight is based on the observation that the human face has a distinct spatial structure. Here, we leverage the class-wise semantic layouts to model this structure information via graph representation learning. Specifically, the IASG and the IRSG are customized and designed. Moreover, we present a novel biphasic ICT strategy to enhance the perceptual quality of the synthesized images. Besides, we conduct facial saliency detection on the input images to provide overall prior information on facial structure. For brevity, we take the sketch synthesis task as the prototype to introduce the details.

Given paired photo-sketch training samples \( \{(x_i, y_i) \mid x_i \in X, y_i \in Y \}_{i=1}^{N} \), where \( x_i \) represents photograph and \( y_i \) represents sketch. The purpose of face sketch synthesis is to construct a mapping from the source photograph domain \( X \) to the target sketch domain \( Y \). As illustrated in Fig. 1, we find that there are complex variations in the source domain, resulting in severe impacts on the identity and fidelity of the generated sketches. Conversely, in the photograph generation task, the low-quality sketches of the source domain would affect the clarity of the generated image. Advanced face saliency detectors effectively capture the global structure information of the input photographs or sketches. Regarding this, we first utilize face saliency detection results as prior information to provide the global facial structure. We concatenate the saliency map \( M \) and the face photograph as the input to the generator. Besides, we also employ the pretrained face parser to acquire semantic layouts \( S \) as guidance. Then, we inject this semantic information into our network to produce the final synthesized result. Therefore, the overall mapping can be formulated as \( \{X, M, S\} \rightarrow Y \).

B. Semantic Driven Network Architecture

As shown in Fig. 2(c), the SI module takes two inputs: the forward activation features after the BN layer and semantic layouts obtained by pretrained BiSeNet [42]. In order to prevent semantic ambiguity, we divide the human face into 12 spatial components: two eyes, two eyebrows, two ears, glasses, upper and lower lips, inner mouth, hair, nose, skin, neck, cloth, and background. Therefore, we have \( S = \{s^{(1)}, \ldots, s^{(C)}\} \in \mathbb{R}^{h \times w \times C} \), where \( C \in \{1, 2, \ldots, 12\} \), \( s^{(c)} \in [0, 1] \), \( h \) and \( w \) denote the height and width of the semantic maps. Here, to inject the spatial structure information into the SI module, we perform the convolutional operation on semantic layouts. Thus, two modulation parameters \( \gamma \) and \( \beta \) are produced to normalize the final output. These two parameters encode sufficient spatial structure information. Then, we conduct multiplication and addition between these two modulation parameters and the normalized activation maps in an element-wise pattern as shown in Fig. 2(c). In this fashion, the facial structure of the synthesized sketch is well-preserved with the input photo. Meanwhile, since the semantic layouts are robust to the complex variations of background, the generated sketches achieve the high-fidelity effect. Finally, we adopt a patch-wise discriminator to enforce the generated sketches keeping realism.

C. Graph Representation Learning

Previously, researchers always imposed global supervision on the entire generated sketches, resulting in defective performances in facial details. Thus, we construct two representational graphs named the IASG and the IRSG to facilitate the preservation of the details in generated sketches as depicted in Fig. 3. Since different face components contain widely varying amounts of pixels, details in small components such as the two eyes might be easily overlooked. Thus, to balance
the contributions of different components, we treat each component as a graph node extracted in an adaptive reweighting pattern.

1) Intraclass Semantic Graph: As illustrated in Fig. 3, the synthesized sketch is represented as $F \in \mathbb{R}^{h_f \times w_f \times c_f}$, where $h_f$, $w_f$, and $c_f$ denote the height, width, and channel of the sketch, respectively. Here, we divide the synthesized sketch into 12 facial components guided by the aforementioned semantic layouts. Thus, each node of IASG is formulated as

$$\mu(c) = \frac{1}{|S(\cdot, \cdot, c)|} \sum_{i=1}^{h_f} \sum_{j=1}^{w_f} S(i, j, c) F(i, j)$$  \hspace{1cm} (1)$$

where $|S(\cdot, \cdot, c)|$ represents the summation of pixel numbers in each facial component with the same semantic class $c$. Obviously, this strategy leverages the pixel number summation to normalize the contribution of each node, adaptive to different face components. Moreover, $\mu(c)$ is considered as the mean center of all pixels in the $c$th semantic category. Furthermore, the modulation variance $\nu(c)$ can faithfully react to the semantic variation of the intraclass feature distribution. Note that both $\mu$ and $\nu$ are tensors in practice. As depicted in Fig. 3, each circle in the graph nodes represents different semantic classes by different colors. To model the intraclass correlation in each facial component, we calculate the cosine similarity between the synthesized sketch and each mean center (and variance center). Formally, the computation is listed as follows:

$$C_1 = \frac{F \cdot \mu}{\|F\|_2 \cdot \|\mu\|_2} \hspace{1cm} C_2 = \frac{F \cdot \nu}{\|F\|_2 \cdot \|\nu\|_2}$$ \hspace{1cm} (3)$$

In this way, we construct the IASG for both the synthesized sketch and the target ones.

2) Interclass Structure Graph: Subsequently, to ensure the generated sketch is structure-coordinated with the target, we model the interclass structure relations among every two facial components in IRSG. Formally, the IRSG is expressed as

$$G_1 = (\mu, E_1)$$

where $\mu$ and $\nu$ are the mean and variance centers of all pixels in each semantic category, respectively. As shown in Fig. 3, each edge in the IRSG represents different semantic classes by different colors. To model the interclass correlation, we calculate the cosine similarity between each node and each mean center (and variance center). Formally, the computation is listed as follows:

$$C_1 = \frac{F \cdot \mu}{\|F\|_2 \cdot \|\mu\|_2} \hspace{1cm} C_2 = \frac{F \cdot \nu}{\|F\|_2 \cdot \|\nu\|_2}$$ \hspace{1cm} (3)$$

In this way, we construct the IASG for both the synthesized sketch and the target ones.
where the $\mu$ and $v$ represent nodes in the structure graph. Then, we utilize the edge $E$ between every two nodes to represent the structure relation of different facial components. Concretely, the graph edges are computed as the Euclidean distance

$$
E(c_1, c_2) = \mathcal{E}(\mu(c_1), \mu(c_2))
$$

$$
E(c_1, c_2) = \mathcal{E}(v(c_1), v(c_2))
$$

where the $c \in \{1, 2, \ldots, 12\}$ denotes different facial components and $E$ represents the Euclidean distance. Once we obtain the ITSG of both generated sketches and target ones, we exploit the target ITSG to constrain the synthesized sketches.

### D. Iterative Training Strategy

Inspired by the observation that paired photo-sketch shared personal characters, we propose a biphasic ICT strategy to improve the perceptual quality of the synthesized results as illustrated in Fig. 4. There are multistage in our iterative training strategy as reported in Algorithm 1. We name the generator and discriminator as $G^i_k$, $D^i_k$, $G^0_o$, and $D^0_o$, where $i \in \{0, 1, \ldots, n\}$ denote the iterative stages; $k$ and $o$ represent sketch synthesis and photograph synthesis, respectively.

First, we train the generator and discriminator for the sketch synthesis task. Then, the photograph generator and discriminator are trained similarly. Afterward, to improve the perceptual quality of the generated sketch, we take the pretrained photograph generator as a knowledge extractor to boost the sketch generator and discriminator. Concretely, we take the generated fake sketch concatenated with saliency detection maps (produced from real sketch) and feed it into the pretrained photograph generator to acquire the fake photo. Then, we exploit the corresponding real sketch concatenated with saliency map feed into the photograph generator to acquire the reconstructed photo. Then, we leverage the reconstructed photograph and corresponding feature maps to provide multiscale supervision of the generated fake ones, thus significantly improving the sketch generator and discriminator. Once we obtain the updated sketch generation, we utilize it to enhance the performance of the photograph generator and discriminator, iteratively. Note that this ICT strategy could be conducted on multistage until we obtain the optimal models.

### E. Objective Function

The overall objective of our model includes following functions: $\mathcal{L}_{GAN}$, $\mathcal{L}_{content}$, $\mathcal{L}_{IASG}$, $\mathcal{L}_{perceptual}$, $\mathcal{L}_{BCE}$, $\mathcal{L}_{IRSG}$, and $\mathcal{L}_{ICT}$. Here, we adopt the sketch synthesis task as the prototype to elaborate the objective functions.

1) Adversarial Loss: The adversarial loss is leveraged to correctly distinguish the real sketches or generated sketches. Similar to [32], the adversarial loss is formulated as

$$
\mathcal{L}_{GAN} = E_{X,M,Y} [\log D(X, M, Y)] + E_{X,M} [\log(1 - D(X, M, G(X, M)))]
$$

where $X$, $Y$, and $M$ denote the source photographs, target sketches, and saliency detection maps, respectively.

2) Content Loss: In addition, we utilize the normalized $L_1$ distance to represent content loss

$$
\mathcal{L}_{content}(G) = E_{X,M,Y} [\|Y - G(X, M)\|_1].
$$

3) Perceptual Loss: In order to ensure the generated sketch and the target sketch maintain similar specificity, we employ the pretrained VGG-19 net [48] as a feature extractor to obtain high-level feature representations. We compare the features after the pool1 and pool2 layers

$$
\mathcal{L}_{perceptual} = \sum_{l=1}^{2} \|o^l(Y) - o^l(G(X, M))\|_2^2
$$

where $o^l(\cdot)$ represents the output feature maps and $l$ denotes the selected pool1 and pool2 layers.

---

**Algorithm 1 The Iterative Training Strategy of Sketch Synthesis**

**Input:**

- Input photo: $X$; saliency detection map: $S_o$; parsing mask: $S_i$;
- Target sketch: $Y$; saliency detection map: $M_k$;
- Target sketch: $M_o$;
- The number of iteration: $i$; max iteration: $T = 4$;

**Output:**

- The optimal models $G^i_k$ and $D^i_k$;

1) **Step1:** Initialize training $G^0_k$, $D^0_k$, $G^0_o$, and $D^0_o$;

2) **Step2:** Iterative training;

3) for $i = 0$ to $T$

4) Feed the $X$, $M_o$ and $S_o$ to train the $G^{i+1}_k$ and $D^{i+1}_k$

5) which obtain fake sketch $\hat{Y}$;

6) Feed the $\hat{Y}$, $M_k$ and $S_k$ to the pre-trained $G^i_o$ to obtain corresponding $\hat{X}$; Extract multi-level feature maps from $G^i_o$ as $\hat{F}_o$;

7) Feed the $Y$, $M_k$ and $S_k$ to the pre-trained $G^i_o$ to obtain corresponding $\hat{X}$; Extract multi-level feature maps from $G^i_o$ as $\hat{F}_o$;

8) Compute the loss between $\hat{X}$ and $\hat{X}$, $\hat{F}_o$ and $\hat{F}_o$;

9) Back-propagate the gradients;

10) end
4) Binary Cross-Entropy (BCE) Parsing Loss: Moreover, we introduce the BCE loss to further refine the synthesized sketch at the semantic level. We contrast the semantic mask of the generated sketch and the target sketch obtained by the pretrained parsing network [42]

\[ \mathcal{L}_{\text{BCE}} = (P(Y), P(G(X, M))) \]  

where \( P \) denotes the inference process of parsing the network.

5) IASG Loss: In practice, we extract the IASGs from the target sketch and synthesized sketch, respectively. Then, we reinforce the supervision of these IASGs to restrain the generated sketch from matching the feature distribution of the target domain. The IASG loss is formulated as

\[ \mathcal{L}_{\text{IASG}}(C_{\text{target}}, C_{\text{fake}}) = \frac{2}{r=1} \sum_{c=1}^{12} \left\| C_{\text{target}}(c) - C_{\text{fake}}(c) \right\|_2^2 \]  

where the \( r \in [1,2] \) denotes two types of nodes of IASG represented by mean center and variance center.

6) IRSG Loss: Based on our preliminary work [22], we employ the IRSGs to facilitate the coordinate structure preservation of synthesized sketches. Consequently, we apply the constraints between the IRSGs of synthesized sketches and target sketches named IRSG loss formulated as

\[ \mathcal{L}_{\text{IRSG}}(G_{\text{target}}, G_{\text{fake}}) = \frac{2}{r=1} \sum_{c=1}^{12} \left\| G_{\text{target}}(c) - G_{\text{fake}}(c) \right\|_2^2 \]  

where the \( r \in [1,2] \) denotes two types of IRSGs with mean center and variance center.

7) ICT Loss: Finally, we design the ICT loss for the biphasic iterative training strategy. Four iterations are conducted in the experiments to reach the optimal results

\[ \mathcal{L}_{\text{ICT}} = \frac{5}{i=1} \left\| G_{\text{i}}(Y) - G_{\text{i}}(G_{\text{i}}(X, M)) \right\|_1 \]  

where \( G_{\text{i}}(\cdot) \) denotes the photograph generator and \( i \) represents the times of iteration. We select the feature maps to extract multilevel identity-specific information represented by \( i \).

8) Full Objective: Eventually, we combine all loss functions to achieve overall supervision

\[ \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \alpha \mathcal{L}_{\text{content}} + \lambda \mathcal{L}_{\text{perceptual}} + \delta \mathcal{L}_{\text{BCE}} + \eta \mathcal{L}_{\text{IASG}} + \tau \mathcal{L}_{\text{IRSG}} + \xi \mathcal{L}_{\text{ICT}} \]  

where the \( \alpha, \lambda, \delta, \eta, \tau, \) and \( \xi \) are weighting factors. Furthermore, the generator \( G \) and the discriminator \( D \) could be optimized by the following formulation:

\[ \min_G \max_D \mathcal{L}_{\text{total}}. \]

IV. EXPERIMENTS

In this section, we first introduce the implementation details of our method. Then, we describe the datasets and evaluation criteria. Next, the experimental results are presented from both quantitative and qualitative perspectives, showing the effectiveness of our proposed method. Finally, we conducted the ablation study to verify each module.

A. Implementation Details

Both the generator and discriminator are implemented on the platform Pytorch [44] with a single NVIDIA GeForce Titan X GPU. We leverage the Adam optimizer with \( \beta_1 = 0.5 \) and \( \beta_2 = 0.999 \). The total training epochs are 200, then the initial learning rate is set to 0.0002 for the first 100 epochs and decays linearly in the last 100 epochs. In addition, we utilize the IN [33], and set the batchsize = 1. Meanwhile, the weighting factors are set as \( \alpha = 100, \lambda = 10, \delta = 15, \eta = 100, \tau = 100, \) and \( \xi = 5 \), respectively.

B. Datasets and Evaluation Criteria

We conduct extensive experiments on the CUHK Face Sketch Dataset (CUFS) [1] and the CUHK Face Sketch FERET Dataset (CUFSF) [45]. In the CUFS dataset, there are 606 faces, of which 188 faces are from the Chinese University of Hong Kong (CUHK) student database [3], 123 faces from the AR database [46], and 295 faces from the XM2VTS database [47]. For each sample, there are paired face photo-sketch drawn by the artist in natural lighting conditions. The CUFSF dataset contains 1194 face photographs with paired sketches. However, all the photographs in the CUFSF dataset are acquired under complex illumination variations as illustrated in Fig. 1(d). For both datasets, we exploit the geometrical alignment strategy between the photographs and sketches, based on the points of two eye centers and the mouth centers. Then, the aligned photo-sketch pair is cropped to \( 200 \times 250 \) following [48]. Meanwhile, we adopt the reshaping and padding conventions in [20] to expand the input image size to \( 256 \times 256 \).

The experimental performance is measured by multiple metrics. We employ the feature similarity index metric (FSIM) [49] to evaluate the feature quality of synthesized sketches. FSIM is commonly utilized to measure the low-level similarity between the paired images, which extracts the phase congruence (PC) and the image gradient magnitude (GM) as features to index the quality. Consequently, blurring and noise of the generated images are evaluated by FSIM, ordinarily.

In addition, we apply the structural similarity index metric (SSIM) [50] to demonstrate the perceptual similarity between synthesized results and ground-truth images, which follow the visibility of humans. However, some works point out that SSIM tends to favor over-smoothed images and ignores the texture of the results, which is not completely consistent with human perception [51]. Therefore, we introduce the learned perceptual image patch similarity (LPIPS) [52] combined with SSIM to measure the perceptual visibility of synthesized results. LPIPS calculates the distance of embedding features between the generated images and target images. In this article, LPIPS is exploited by three classification networks which are SqueezeNet [53], AlexNet [54], and VGGNet [43].

Besides, we adopt the Fréchet inception distance (FID) [55] to compute the Earth-mover distance (EMD) of distributions between the target domain and the synthesized image domain. Specifically, a pretrained Inception-v3 network [56] is raised to measure the 2048-dimension features between the two
contrast domains. FID is widely used in biphasic photo-sketch synthesis tasks and presents high confidence in image realism.

To evaluate the preservation of personal identity characteristics of the generated human faces, we introduce the face verification rate (FVR) which is implemented by Face++ API [57]. Previously, other researchers often used null-space linear discriminant analysis (NLDA) to measure identity-specific. However, NLDA is found to be seriously affected by image texture and deformations so it might not make an accurate assessment of identity-specific characteristics [19]. We utilize the Face++ APIs which the threshold is set as $73.975@\text{FAR} = 1e^{-5}$ in our identity preservation experiments.

To demonstrate the superiority of our method, we adopt several benchmark approaches (e.g., CycleGAN [58], Pix2Pix [26], MDAL [16], Col-cGAN [60], KD+ [19], MSG-SARL [18], SCAGAN [20], DP-GAN [61], and DIR-MFP [62]) for comparison. In addition, to further demonstrate the effectiveness of the proposed method, we also conduct comparison experiments between our preliminary work SDGAN [22] and our current model.

### C. Results on Face Sketch Synthesis Task

We verify the performance of our method on the face sketch synthesis on the CUFS and the CUFSF datasets. The experimental results are reported in Table I. The proposed method decreases the previous best FID from 34.2 to 33.256. In addition, our method realizes the best performance on all indicators in the CUFSF dataset except FID. Obviously, the LPIPS (alex, squeeze, and vgg) is significantly decreased with the large margin in both datasets which implies the higher fidelity and realism of the sketches generated by our method. In the sketch synthesis experiments, the MSG-SARL, DIR-MFP, and SCAGAN reached better results in limited metrics on the FSIM, SSIM, and FID. We observe that in Fig. 5, the generated images of these methods are more smoothed. Since metrics such as SSIM are more kind to the smoothed image texture, this leads to better results than ours. However, we exploit the saliency map to provide global structure information. This results in better performance on the LPIPS and a more structural-coordinated human face.

We visualize the comparison of generated sketches among our method and various state-of-the-art counterparts as depicted in Fig. 5. Since the KT [5] is the preliminary conference version of KD+ [4], we only display the better version in the experiments. We observe that CycleGAN and MDAL produce low-fidelity sketches compared with ground truth. The results of Col-cGAN, MSG-SARL, and SCAGAN contain blur and noisy effects. Conversely, our method achieves the best performance, especially in the detailed facial components such as the eyes, ears, and hair contours (e.g., in the red boxes of the last column).

### D. Results on Face Photograph Synthesis Task

Compared with face sketch synthesis, there are fewer previous methods that focus on the photograph synthesis task. Our method obtains the best performance with LPIPS (alex, squeeze, and vgg) and SSIM on both datasets. According to Table II, the SCAGAN [20] achieves better results than our method with FSIM and FID because of its stacking architecture. In contrast, our network is end-to-end and has stronger practical significance on edge devices. It is more challenging to synthesize photographs from the CUFSF dataset. As shown in Fig. 1(d), there are illumination variations in the CUFSF dataset. Meanwhile, all photographs are grayscale and sketches are drawn with deformation.
TABLE II
RESULTS ON FACE PHOTOGRAPH SYNTHESIS IN THE CUFS AND CUFSF DATASETS. ↑ INDICATES THE HIGHER IS BETTER, ↓ INDICATES THE LOWER IS BETTER. OUR METHOD REACHES THE OPTIMAL AND SUBOPTIMAL RESULTS IN THE CUFS DATASET AND THE CUFSF DATASET. NOTE THAT THE RESULTS OF THE PS2MAN MODEL ARE DIRECTLY BORROWED FROM [19].

| Model          | CycleGAN | Pix2Pix | KT | KD+ | MSG-SARL | SCAGAN | PS2MAN | SDGAN | ours |
|----------------|----------|---------|----|-----|----------|--------|--------|-------|------|
| **Cufs**       |          |         |    |     |          |        |        |       |      |
| LPPS(alex)     | 0.2898   | 0.1687  | 0.1919 | 0.1717 | -        | -      | 0.2464 | 0.1674 | 0.1497 |
| LPPS(squeeze)  | 0.2509   | 0.1453  | 0.1747 | 0.1474 | -        | -      | 0.2158 | 0.1370 | 0.1225 |
| LPPS(vgg-16)   | 0.4383   | 0.3031  | 0.3208 | 0.2806 | -        | -      | 0.3254 | 0.2640 | 0.2367 |
| FSIM           | 0.7270   | 0.7723  | 0.7851 | 0.7819 | 0.7866   | 0.795  | 0.7819 | 0.7845 | 0.8001 |
| SSIM           | 0.4461   | 0.6086  | -   | -   | 0.6242   | -      | -      | 0.6543 | 0.6822 |
| FID            | 124.540  | 86.996  | -   | -   | -        | 66.17  | 40.3   | 63.937 | 49.925 |
| **CufsF**      |          |         |    |     |          |        |        |       |      |
| LPPS(alex)     | 0.2271   | 0.2115  | 0.2440 | 0.2322 | -        | -      | 0.3145 | 0.2011 | 0.1998 |
| LPPS(squeeze)  | 0.1725   | 0.1669  | 0.2023 | 0.1791 | -        | -      | 0.2853 | 0.1581 | 0.1556 |
| LPPS(vgg-16)   | 0.3690   | 0.3579  | 0.3758 | 0.3565 | -        | -      | 0.4237 | 0.3422 | 0.3376 |
| FSIM           | 0.7544   | 0.7855  | 0.7931 | 0.7789 | 0.7734   | 0.845  | 0.7812 | 0.7902 | 0.7935 |
| SSIM           | 0.5994   | 0.6194  | -   | -   | 0.6114   | -      | -      | 0.6305 | 0.6441 |
| FID            | 29.584   | 60.286  | -   | -   | 59.61    | 20.6   | -      | 38.776 | 38.372 |

Fig. 6. Visualization of our generated photographs against various state-of-the-art methods, e.g., CycleGAN [58], Pix2Pix [26], KD+ [19], MSG-SARL [18], SCAGAN [20], and PS2MAN [63]. From top to bottom, samples of the first three rows are from the CUFS dataset, and the others are from the CUFSF dataset. Best viewed on screen.

We display the visual comparison in Fig. 6. Pix2Pix, MSG-SAR, and SCAGAN synthesize the natural photographs compared with the ground truth faces. CycleGAN, KD+, and PS2MAN generate better results while their faces contain less stylized textures of the painter. On the contrary, our method achieves vivid and realistic photograph synthesis. As highlighted in the red boxes, the facial structures of our results are much more natural than other competitors.

E. Comparison With Preliminary Work

To evaluate the improvement between our current model and preliminary work SDGAN, we conduct the performance comparison as reported in Table III.

Since SDGAN supervises the synthesis of sketches and photographs in the intraclass semantic space, the generated human faces always have abundant detailed features. However, SDGAN does not effectively extract the structure knowledge in the facial images, which is modeled by the IRSG in the current framework. Therefore, we take advantage of both intraclass semantic and IRSG representations that allow the synthesized human faces to be more clearly structured and contain distinctive features, as illustrated in columns Fig. 8(g)–(i). Besides, the SDGAN is designed for unidirectional photo-sketch synthesis that overlooks there are personal-identity consistency between the paired photo-sketch. On the contrary, with the help of the biphasic ICT strategy, the identity-preserving ability of our current model is significantly enhanced. This is essential in the image-to-image translation task with geometrical alignment datasets. As reported in Table IV and Table V, the FVR significantly increased after applying the biphasic ICT strategy. Besides, as illustrated in Table III, our current model performance has been significantly raised compared with preliminary SDGAN, especially in the photograph synthesis task. From a more intuitive point of view, the human faces synthesized by the current model give the expression of more fidelity and consistency than SDGAN.

TABLE III
COMPARISON BETWEEN THE PRELIMINARY WORK SDGAN AND OUR CURRENT NETWORK NAMED OURS IN THE TABLE. ALL THE PERFORMANCES OF OUR CURRENT MODEL ARE BETTER THAN PRELIMINARY WORK. ↑ INDICATES THE HIGHER IS BETTER, ↓ INDICATES THE LOWER IS BETTER.

| Task          | Sketch Synthesis | Photo Synthesis |
|---------------|------------------|----------------|
| Model         | SDGAN | Ours | Gap | SDGAN | Ours | Gap |
| **Cufs**      |      |      |    |      |      |    |
| LPPS(alex)    | 0.1444 | 0.1432 | 0.0012 | 0.1674 | 0.1497 | 0.0177 |
| LPPS(squeeze) | 0.1017 | 0.0986 | 0.0031 | 0.1370 | 0.1225 | 0.0145 |
| LPPS(vgg-16)  | 0.2767 | 0.2646 | 0.0021 | 0.2640 | 0.2367 | 0.0273 |
| FSIM          | 0.7446 | 0.7494 | 0.0048 | 0.7845 | 0.8001 | 0.0156 |
| SSIM          | 0.5360 | 0.5493 | 0.0133 | 0.6543 | 0.6822 | 0.0279 |
| FID           | 33.408 | 33.256 | 0.152 | 63.937 | 49.925 | 14.012 |
| FVR           | 86.179 | 87.542 | 1.363 | 79.040 | 81.780 | 2.74 |
| **CufsF**     |      |      |    |      |      |    |
| LPPS(alex)    | 0.1906 | 0.1867 | 0.0039 | 0.2011 | 0.1998 | 0.0013 |
| LPPS(squeeze) | 0.1370 | 0.1341 | 0.0029 | 0.1581 | 0.1556 | 0.0025 |
| LPPS(vgg-16)  | 0.3358 | 0.3347 | 0.0017 | 0.3422 | 0.3376 | 0.0046 |
| FSIM          | 0.7328 | 0.7332 | 0.0004 | 0.7902 | 0.7955 | 0.0053 |
| SSIM          | 0.4339 | 0.4407 | 0.0068 | 0.6305 | 0.6441 | 0.0136 |
| FID           | 30.594 | 24.577 | 6.017 | 38.776 | 38.372 | 0.404 |
| FVR           | 86.758 | 87.109 | 0.351 | 62.763 | 63.344 | 0.581 |
In this section, we conduct extensive ablation experiments to verify the effectiveness of each module and the loss function we proposed.

1) Saliency Detection Map $M$: As reported in Table IV, we concatenate the saliency detection map $M$ with input images to provide overall structure prior knowledge. We observe that the SSIM is affected after the utilization of the saliency map since SSIM is more sensitive to texture information. However, as a whole, the performance of the model is improved, especially the FSIM indicators. As displayed in Fig. 1, due to the low quality and lighting variations of the source images, we find that the detected saliency maps

F. Ablation Study

In this section, we conduct extensive ablation experiments to verify the effectiveness of each module and the loss function we proposed.

1) Saliency Detection Map $M$: As reported in Table IV, we concatenate the saliency detection map $M$ with input images to provide overall structure prior knowledge. We observe that the SSIM is affected after the utilization of the saliency map since SSIM is more sensitive to texture information. However, as a whole, the performance of the model is improved, especially the FSIM indicators. As displayed in Fig. 1, due to the low quality and lighting variations of the source images, we find that the detected saliency maps
have noises and distortions such as the hairs, and faces on the CUFSF dataset. Considering that, we drop the saliency maps and still achieve the best performance on the biphasic photo-sketch synthesis results on the CUFSF dataset. Besides, under normal light conditions, the saliency detector [21] expresses superior performance on both face photograph and sketch, thus leading to the positive impact as shown in Table IV and Fig. 7. Here, as shown in Table V, we discard the saliency maps in the following experiments after the injection of semantic layouts.

2) Parsing Layouts Injection: Furthermore, we adopt the semantic parsing layouts as two modulation parameters with spatial dimensions injected into the decoder of our network. The semantic layouts aim to provide a kind of spatial supervision of synthesized images. As depicted in Figs. 7(e) and 8(d), the details of the synthesized images are more realistic and vivid in the class region. The experiment results also show the effectiveness of parsing layouts injection as shown in Tables IV and V.

3) Perceptual Loss and BCE Loss: In addition, we apply the perceptual loss to improve the high-frequency quality of synthesized images. As displayed in Tables IV and V, the perceptual loss could significantly reduce the value of FID while increasing FSIM in the face sketch synthesis task. Besides, the BCE loss is implemented by BiSeNet [42] to refine the structure layouts of synthesized images. As illustrated in Figs. 7(g) and 8(f), the synthesized human faces retain more distinct facial contours.

4) IASG Loss: Moreover, we propose a novel IASG loss to restrain the generated images with ground truth. The IASG loss forces the synthesized images to hold more semantic intra-class knowledge. The details of the sketches and photographs we generated are more similar to ground truth, such as the texture of the hair, the contour of the ears, the position of the eyebrows, and the eyes are more consistent as represented in Figs. 7(h) and 8(g). More than that, the IASG Loss raises multiple indicators considerably such as LPIPS (alex, squeeze, and vgg) as expressed in Tables IV and V.

5) IRSG Loss: Combined with the IASG loss we proposed, we bring forward a novel IRSG loss that remarkably enhances the performance of our network. As reflected in Tables IV and V, the SSIM is increased in whole experiments, especially on the sketch synthesis task on the CUFS dataset. The IRSG loss allows the synthesized sketches to become more structure-coordinated. It can be observed that in Figs. 7(i) and 8(h), the sketches are more vivid and distinct.

6) ICT Loss: Eventually, we designed a novel biphasic ICT loss to boost the training of sketch and photograph synthesis. There are multistage iterations in this training strategy. To examine the comprehensive training procedure, we express the results in Table VI. After conducting three times iterations, the network tends to converge and reach its optimum at the third iteration. As depicted in Figs. 7(j) and 8(i), the synthesized sketches and photographs contain more personal characteristics. The details of the face images are more meticulous without less blurring and noise such as the eyelashes, hairstyles, and teeth.

G. Heterogeneous FVR

Meanwhile, we exploit the FVR, deployed by the Face++ API, to report the identity preservation ability of our multivariation models, as highlighted in Tables IV and V. Our method synthesizes face photographs and sketches from heterogeneous domains. Comprehensively, in the CUFS dataset, the FVR increases distinctly from the backbone to the final model by 3.332 (84.252→87.542) for sketch synthesis and 8.253 (73.527→81.780) for face photograph synthesis.
Similarly, the proposed method improves the FVR considerably despite the deformation between the paired sketch photograph in the CUFSF dataset. To be more specific, the FVR increases significantly after we exploit the intraclass and interclass representational graph loss functions, especially in the CUFSF dataset. Although there are deformations between the photographs and sketches in the CUFSF dataset, the proposed graph loss functions improve the FVR by 0.511 (86.512 → 87.023) for sketch synthesis and 1.026 (62.026 → 63.052) for photograph synthesis. This highlights the robustness of our proposed graph representation algorithms.

**H. User Study**

We conducted a user study to further analyze the visual quality of the generated biphasic photo-sketch from different methods. In particular, we recruited 20 volunteers and asked them to give feedback scores ranging from 0 to 5 (the higher and the better) of the generated faces. The evaluation perceptions consist of naturalness and authenticity. The statistical results are reported in Fig. 9. Our method (i.e., the red column) achieves the best performance on both sketch and photograph synthesis tasks. This strongly proves our key insight on building the two types of representational graphs to model the spatial structure information of human faces.

**V. Conclusion**

In this article, we propose a semantic-driven generative adversarial network with graph representation learning for biphasic face photo-sketch synthesis by utilizing saliency detection and face parsing layouts as prior information. In particular, parsing layouts are employed to construct two types of representational graphs that restrain the intraclass and interclass features of the synthesized images. These graphs could enforce our network to generate the considerable face structure and details. In addition, based on the observation that the paired photo-sketch shared many personal characters, we designed a novel biphasic ICT strategy to refine the high-frequency quality of the synthesized images. Extension experiments are conducted to verify the effectiveness of proposed each module.

Although the proposed method achieves state-of-the-art performance, there are still limitations in our work. The photographs generated in the CUFSF dataset are not vivid and realistic leading to the low FVR of the synthesized photographs. These issues are caused by the deformation between photographs and sketches in the dataset. In addition, in terms of a few indicators, our method is still not optimal implying there are many improvements that can be made. In the future, we will pay more attention to the biphasic face photo-sketch synthesis when facing the large deformations between the paired human photograph and sketch.

**REFERENCES**

[1] 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.

[2] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, “A comprehensive survey to face hallucination,” *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, Jan. 2014.

[3] X. Tang and X. Wang, “Face sketch synthesis and recognition,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2003, pp. 687–694.

[4] M. Zhu and N. Wang, “A simple and fast method for face sketch synthesis,” in *Proc. 16th ACM Int. Conf. Int. Multimedia. Comput. Serv. (ICMCS)*, Aug. 2016, pp. 168–171.

[5] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[6] M. Zhang, N. Wang, Y. Li, and X. Gao, “Deep latent low-rank representation for face sketch synthesis,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3109–3123, Oct. 2019.

[7] M. Zhu, N. Wang, X. Gao, and J. Li, “Deep graphical feature learning for face sketch synthesis,” in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, Aug. 2017, pp. 3574–3580.

[8] 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.

[9] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, “Transductive face sketch-photo synthesis,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1364–1376, Sep. 2013.

[10] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, “A nonlinear approach for face sketch synthesis and recognition,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 1005–1010.

[11] J. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[12] H. Zhou, Z. Kuang, and K. K. Wong, “Markov weight fields for face sketch synthesis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1091–1097.

[13] N. Wang, X. Gao, L. Sun, and J. Li, “Bayesian face sketch synthesis,” *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1264–1274, Mar. 2017.

[14] I. Goodfellow et al., “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[15] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, “End-to-end photo-sketch generation via fully convolutional representation learning,” in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 627–634.

[16] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, “Face sketch synthesis by multidomain adversarial learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1419–1428, May 2019.

[17] H. Bi, N. Li, H. Guan, D. Lu, and L. Yang, “A multi-scale conditional generative adversarial network for face sketch synthesis,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3876–3880.

[18] S. Duan, Z. Chen, Q. M. J. Wu, L. Cai, and D. Lu, “Multi-scale gradients self-attention residual learning for face photo-sketch transformation,” *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1218–1230, 2021.

[19] M. Zhu, J. Li, N. Wang, and X. Gao, “Knowledge distillation for face photo-sketch synthesis,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 893–906, Feb. 2022.
Xingguo Qi received the B.E. and M.S. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in individualized interdisciplinary (artificial intelligence) with the Academy of Interdisciplinary Studies Department, The Hong Kong University of Science and Technology, Hong Kong, China.

His research interests include computer vision, pattern recognition, and image processing.

Qi Li (Member, IEEE) received the B.E. degree in automation from the China University of Petroleum, Qingdao, China, in 2011, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2016. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, CASIA. His research interests include face recognition, computer vision, and machine learning.

Muyi Sun received the B.E. degree in automation and the Ph.D. degree in control science and engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015 and 2020, respectively.

He was also a Post-Doctoral Researcher/Assistant Research Fellow at the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing. Currently, he is an Associate Research Fellow with the School of Artificial Intelligence, BUPT. He has published research papers in reputable journals and conferences, such as IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, and CVPR. His research interests include biometric image generation and medical image analysis.

Fang Zhao received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. From October 2021 to June 2023, he worked as a Senior Researcher with the Tencent AI Laboratory, Shenzhen, China. He is currently an Associate Professor at the School of Intelligence Science and Technology, Nanjing University, Suzhou, China. His research interests include computer vision and machine learning.

Shanghang Zhang received the B.Eng. degree from Southeast University, Nanjing, China, in 2010, and the M.Eng. degree from Peking University, Beijing, China, in 2013.

She has been a Post-Doctoral Research Fellow at the Berkeley AI Research Laboratory (BAIR), UC Berkeley, Berkeley, CA, USA. She is currently a Tenure Track Assistant Professor at the School of Computer Science, Peking University. Her research interests include OOD generalization which enables machine learning systems to generalize to new domains using limited labels.

Dr. Zhang received the AAAI 2021 Best Paper Award and has been selected to “2018 Rising Stars in EECS, USA. She has been the Chief Organizer of several workshops on ICML/NeurIPS, and the Special Issue on “ICMR.”

Zijian Wang received the B.E. and M.S. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2020 and 2023, respectively. He is currently pursuing the Ph.D. degree with The University of Sydney, Sydney, NSW, Australia.

His research interests include computer vision and multimodal learning.

Qi Li received the B.E. and M.S. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with Peking University, Beijing.

His research interests include out-of-distribution, multimodal large language models, and autonomous driving.

Dr. Shan has served as Associate Editor for journals including IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.