S2FGAN: Semantically Aware Interactive Sketch-to-Face Translation

Yan Yang[1,2]   Md Zakir Hossain[1,2]   Tom Gedeon[3]   Shafin Rahman[4]

[1] BDSI, Australian National University, Australia   [2]A&F, CSIRO, Australia

[3] EECMS, Curtin University, Australia   [4] ECE, North South University, Bangladesh

{Yan.Yang, zakir.hossain}@anu.edu.au   Tom.Gedeon@curtin.edu.au   shafin.rahman@northsouth.edu

Abstract

Interactive facial image manipulation attempts to edit single and multiple face attributes using a photo-realistic face and/or semantic mask as input. In the absence of the photo-realistic image (only sketch/mask available), previous methods only retrieve the original face but ignore the potential of aiding model controllability and diversity in the translation process. This paper proposes a sketch-to-image generation framework called S2FGAN, aiming to improve users' ability to interpret and flexibility of face attribute editing from a simple sketch. First, to restore a vivid face from a sketch, we propose semantic level perceptual loss to increase the translation quality. Second, we dedicate the theoretic analysis of attribute editing and build attribute mapping networks with latent semantic loss to modify latent space semantics of Generative Adversarial Networks (GANs). The users can command the model to retouch the generated images by involving the semantic information in the generation process. In this way, our method can manipulate single or multiple face attributes by only specifying attributes to be changed. Extensive experimental results on the CelebAMask-HQ dataset empirically show our superior performance and effectiveness on this task. Our method successfully outperforms state-of-the-art sketch-to-image generation and attribute manipulation methods by exploiting greater control of attribute intensity.

1. Introduction

Generative Adversarial Networks (GANs) [7] is one of the emerging techniques for image synthesis and achieves tremendous success in image-to-image (I2I) translation. Now, generating photo-realistic faces from sketches becomes possible by learning pixel-wise correspondence using conditional GANs [24, 23, 33, 31, 12]. Although such techniques remain successful in enabling a novice artist to restore a face from a sketch, they often fail to control specific facial attributes' intensity. For example, a user may wish to add a specific attribute, such as a smile, after generating a face from a drawing and/or s/he may wish to control intensities of facial attributes. For non-artists, modifying the sketches is very hard for describing facial features such as happiness, age, or chubby. However, it is easy to specify
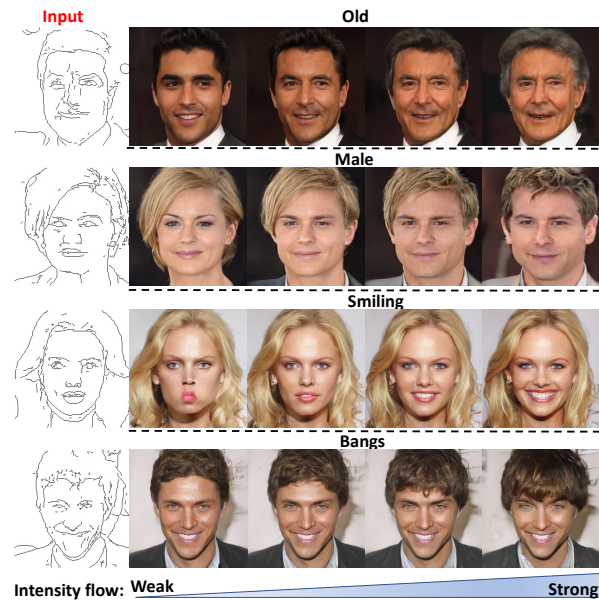

Figure 1: Our proposed *S2FGAN* framework generates high resolution and photo-realistic face from a face sketch image as input by using semantic conceptual loss and manipulates single or multiple attributes (age, gender, nose, smiles, lips, bangs, and so on) of faces with greater control by using latent semantic loss.

those abstract appearances (attribute/semantic) as text input. This paper investigates how a GAN model can increase such users' interpretability and flexibility of attributes editing from a simple sketch.

We identify several limitations in this line of investigation. *(a)* Existing methods perform attribute editing when the input is a photo-realistic face [4, 9, 22, 18, 29]. These methods cannot work in the absence of a photo-realistic image as input. *(b)* Although methods can successfully add or remove attributes (like age, gender, beard) from photo-realistic faces, they are not able to provide adequate measures to control intensities of attributes. Recent approaches have somewhat addressed this issue, but they mostly fail to preserve subjects identify at higher intensity boundary of attributes. Moreover, many methods ignore state-like attributes, such as *chubby*, which are expected to have diverse intensities. [4, 22, 9, 18, 29] *(c)* For attribute editing, users need to specify both the attributes intended to be preserved and changed, increasing the necessity of manual annotations [9, 18, 4]. Recent approaches, InterFaceGAN [29] and

STGAN [22] allow users to edit face attributes by specifying attributes to be changed only. However, none of those methods consider sketch inputs.

This paper presents a novel sketch-to-face GAN framework, *S2FGAN*, to aid the controllability of image attributes in the sketch-to-image generation process. We encapsulate two tasks in a single framework: sketch-to-face generation and face attribute editing. *For the sketch-to-face generation*, approaches widely use encoder-decoder GAN structures [31, 20, 3]. Since a sketch describes a face's layout but fails to indicate the low-level facial features, the latent code of GAN (semantically encoded low dimensional vectors after discarding redundant information [29]) struggles to describe low-level facial features, which is essential to restore a photo-realistic face. This problem gets intensified with the increase of output image resolution. To address this problem, we propose a semantic level perceptual loss. It encourages to compose latent code of sketch and ground-truth face indistinguishable. In this way, similar to the latent code of ground-truth face, the resultant latent code of sketch also describes low-level facial features in addition to face layout. *For attribute editing on the generated face*, we perform conditional manipulation of desired facial attributes without affecting the rest attributes of interests. We present two attribute mapping networks with latent semantic loss to modify semantics in the latent space. Moreover, we provide the theoretical underpinning to establish that proposed attribute mapping networks focus on preserving the semantic and intensity of non-edited attributes. It also helps the GAN decoder correctly constructing the rare attribute combinations (e.g., female and beard). In this way, we preserve identity and edit multiple attributes simultaneously based on only specifying attributes to be changed and manipulate the attribute intensity with greater control. Figure 1 shows sample outputs of our method. Comparing with the state-of-the-art methods in Figure 6, our model provides smooth attributes intensity control. Furthermore, we achieve superior attribute editing and diversity control performance, especially when manipulating multiple attributes, as shown in Figure 5. Overall, our contributions are summarised below:

- We propose the *S2FGAN* framework for sketch-to-image translation with the facility of face reconstruction, attribute editing, and interactive manipulation of attribute intensity. Our model can work on both single and multiple attribute editing and manipulation problems. Further, users can control intensities of attributes by only specifying/changing target attributes (semantics) and preserving face identity.
- We present a semantic level perceptual loss to increase the sketch-to-image translation quality. Our attribute editing models use latent semantic loss, which calibrates face attributes with broader control, diversity, and smoothness.

- We compare with state-of-the-art (DeepFaceDrawing [3], DeepFacePencil [20], Deep Plastic Surgery [33], Pix2PixHD [31]) sketch-to-image translation model and attribute editing model (AttGAN [9], STGAN [22] and alternative baselines). Our model is capable of translating the human badly drawn sketch with detecting the desired facial structures (See Figure 3), and then perform attribute editing and intensity control.

2. Related Work

Sketch-to-Image Generation. Generative Adversarial Networks (GANs) have shown great potential in computer vision tasks such as high-resolution image synthesis [14, 15, 16, 31], image completion [13, 33, 26], image translation [23, 33, 26, 29, 12, 4, 9, 18, 21, 11, 17, 1, 28, 30, 25], and conditional image synthesis [24, 23, 13, 26]. Among these applications, face-to-face translation is one of the most widely studied tasks, because faces carry influential social cues essential for human communication. This enables non-artists to simulate diverse images by sketching their abstract ideas. The goal is to map the sketch to their corresponding ground truth images [33, 23, 31, 12, 5]. Previous work such as Deep Plastic Surgery [33] and ContextualGAN [23] had a target of increasing the model robustness by adapting poorly drawn sketches. However, they did not consider improving the generation process by including semantic descriptions from users. As a result, they could not control the diversity of image attributes. This paper aims to help the sketch-to-image generation processes by providing the opportunities to control intensities of attributes.

Attribute Editing. In Face-to-Face translation, attribute editing attempts to change certain features (e.g., big lips/big smile) of a given face, preserving identity information. Some methods edit attributes of photo-realistic faces by using pencil sketch as input [13, 33]. Others edit the face attributes by keyword descriptions (e.g., Male and Young) [4, 9, 18, 22]. Because it requires less manual interaction, this paper explores the latter approach. A known issue about keyword-based editing is that it requires manually specifying both attributes to be edited and preserved [9, 4, 18]. Recently, STGAN [22] addressed this issue by improving the generation process of AttGAN [9]. However, they did not consider assisting the generation process of diverse image-to-image translation problems by involving semantic attributes. In another work, InterfaceGAN [29] successfully included a control on diverse image generation by editing the latent code proposed in StyleGAN [15] or PGGAN [14]. However, they have weak GAN inversion results and lack analysis and effect on preserving multiple attributes that do not want to change. All existing works of attribute editing operate on the domain of photo-realistic images, which lacks generalization ability. This paper fo-
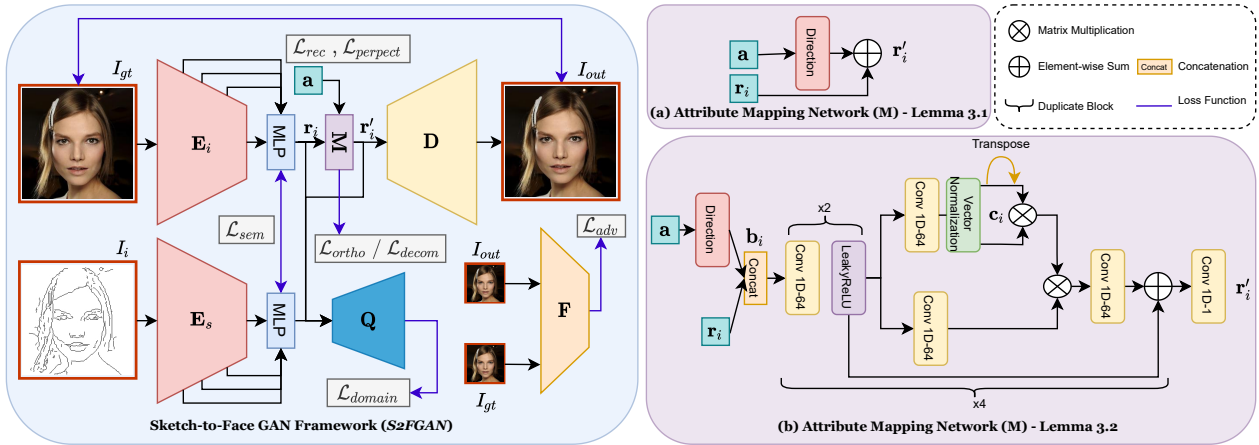
Figure 2: The network architecture of *S2FGAN* framework. The ground-truth face, I_{gt} is projected to the latent code via Image Encoder, $\mathbf{E}_i$. The Sketch Encoder, $\mathbf{E}_s$ learns to map latent code, $\mathbf{E}_i(I_{gt})$ via minimizing the semantic perceptual loss. A domain discriminator, $\mathbf{Q}$ encourages $\mathbf{E}_s(I_i)$ to be indistinguishable from ground-truth image latent code. Then, Attribute Mapping Networks enables interactive semantic editing from $\mathbf{a}$. The discriminator structure provides necessary supervision to support the generator. There are two versions of our Attribute Mapping Network. **(a)** After disentangling attributes semantics by using $\mathcal{L}_{ortho}$, we build the attribute editing by using Lemma 3.1. **(b)** By modeling Lemma 3.2 with $\mathcal{L}_{decom}$, we can decompose latent code semantic with a goal of attribute editing.

cuses on attribute editing in the absence of photo-realistic images (e.g., sketch or edges) that can help users express their abstract ideas by compensating for the limited representation power of sketches, masks, and so on.

3. Method

Problem Formulation. Suppose, $\mathbf{A}$ represents the set of valid attributes (i.e. smile, old and so on) of faces. Given a sketch image $I_i \in \mathbb{R}^{H \times W}$ and desired attributes shifting vector $\mathbf{a} = [a_1, a_2 \dots a_{|\mathbf{A}|}] \in \mathbb{R}^{|\mathbf{A}|}$, our goal is to find a parameterized generator $\mathbf{G}$ that learns a mapping, $\mathbf{G}(I_i, \mathbf{a}) \to I_{out}$ from the sketch, I_i to photo-realistic image, $I_{out} \in \mathbb{R}^{H \times W \times 3}$ described by the attributes shifting vector $\mathbf{a}$. By manipulating $\mathbf{a}$, users can control intensity of attributes in a generated face I_{out}. When $\mathbf{a} = \mathbf{0} \in \mathbb{R}^{|\mathbf{A}|}$, $\mathbf{G}(I_i, \mathbf{0})$ reconstructs the ground truth photo-realistic image I_{gt}. Otherwise, $\mathbf{G}(I_i, \mathbf{a})$ generates manipulated photo-realistic face according to the value of $\mathbf{a}$. Shifting the value a_j manipulates the jth attribute of face. Users have the flexibility to manipulate single or multiple attributes a_j simultaneously where $j = 1, 2 \dots |\mathbf{A}|$.

3.1. Attribute Editing on Latent Code

We further decompose the generator, $\mathbf{G}(\cdot)$ into encoder, $\mathbf{E}(\cdot)$, attribute manipulation network, $\mathbf{M}(\cdot)$ and decoder, $\mathbf{D}(\cdot)$. The relations are defined below:

$$\mathbf{r}_i = \mathbf{E}(I_i), \quad \mathbf{r}'_i = \mathbf{M}(\mathbf{r}_i, \mathbf{a}), \quad I_{out} = \mathbf{D}(\mathbf{r}'_i)$$

Where $\mathbf{r}_i$ and $\mathbf{r}'_i \in \mathbb{R}^d$ and d is the dimensionality of the latent code. We assume, $S^+ \cup S^- = \mathbf{A}$, where, S^+ contains the set of attributes to be edited, and S^- are the rest of attributes in $\mathbf{A}$.

Definition 3.1. $[\mathbf{w}_1 \dots \mathbf{w}_{|\mathbf{A}|}]$ defines a list of hyperplanes that classifies $|\mathbf{A}|$ attributes. Each of the hyperplane is a unit

vector such that $\forall_{i \in [1 \dots |\mathbf{A}|]} \|\mathbf{w}_i\|^2 = 1$. The scalar product of $\mathbf{r}_i$ and $\mathbf{w}_i$ defines the intensity of i_{th} attribute. $\mathbf{w}_i^T \mathbf{r}_i \geq 0$ means that i_{th} attribute is present in input image and otherwise absent.

Remark. Face identify information is a summary of different facial attributes. Attribute editing with preserving the identity could be a paradox. For example, shifting attributes of a "female" to a "male" face loses the feminine identity of the input. Thus, we assume that the attribute editing operation preserves identity information if and only if the operation is invertible and effective when a sufficiently large dataset is available. Mathematically,

$$\mathbf{r_i} = \mathbf{M}(\mathbf{M}(\mathbf{r}_i, \mathbf{a}), -\mathbf{a}), \ \mathbf{r}_i\mathbf{w}_j + \mathbf{a} = \mathbf{M}(\mathbf{r}_i, \mathbf{a})\mathbf{w}_j \ \forall j \in [1 \dots |\mathbf{A}|]$$

We propose two different ways to accomplish attribute editing. Firstly, we learn a disentangled latent space that forces all the attributes in set $\mathbf{A}$ to be orthogonal with each other. Editing a subset of attributes in $\mathbf{A}$ by simple additions would not affect the rest of the attributes. Lemma 3.1 confirms the claim.

Lemma 3.1. *Given the attributes shifting vector* $\mathbf{a} = [a_1, a_2 \dots a_{|\mathbf{A}|}]$. *If* $\forall j, k \in |\mathbf{A}| \times |\mathbf{A}|$ *S.T.* $\mathbf{w}_i$ *is orthogonal with* $\mathbf{w}_j$ *and all hyperplanes correctly classify the input latent code* $\mathbf{r}_i$, *then we can have a disentangled latent space. Specifically, we can editing the attributes without affecting the rest attributes by* $\mathbf{r}_i + a_j\mathbf{w}_j$.

Proof. We proof this claim by contradiction. Assume, editing the set of attributes S^+ will affect the set of attributes S^-. Then,

$$(\mathbf{r_i} + \sum_{j \in S^+} a_j\mathbf{w}_j)\mathbf{w}_k \neq \mathbf{r_i}\mathbf{w}_k \quad \forall k \in S^- \tag{1}$$

$$\mathbf{r_i}\mathbf{w}_k + \sum_{j \in S^+} a_j\mathbf{w}_j\mathbf{w}_k \neq \mathbf{r_i}\mathbf{w}_k \quad \forall k \in S^- \tag{2}$$

$$\mathbf{r_i}\mathbf{w}_k \neq \mathbf{r_i}\mathbf{w}_k \quad \forall k \in S^- \tag{3}$$
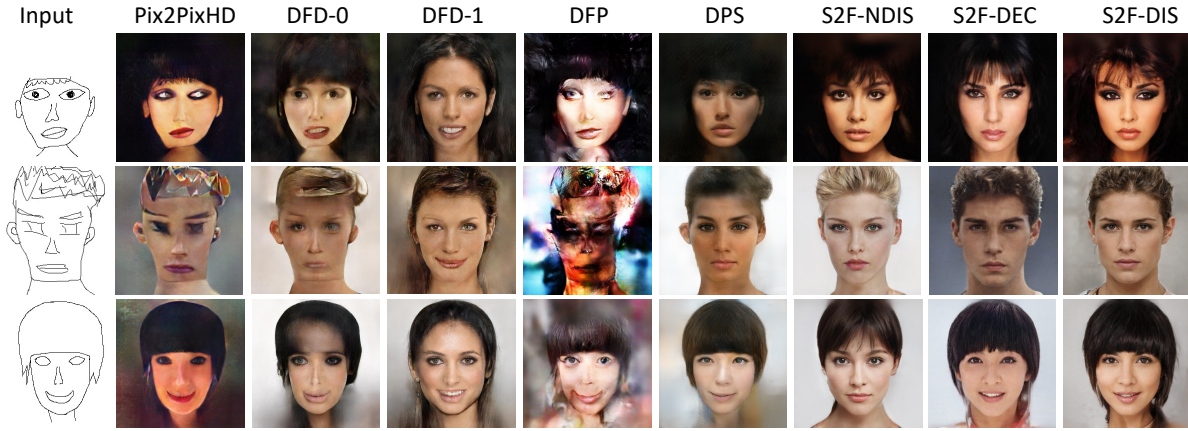
1271

Figure 3: Comparison of translating human-drawn sketches [33] with Pix2PixHD [31], DeepFaceDrawing (DFD) [3], DeepFacePencil (DFP) [20] and Deep Plastic Surgery (DPS) [33]. DFD-0 and DFD-1 represent the medium refinement and fully refinement of DFD.

Eq. 3 can be inferred from Eq. 2 because all hyperplanes are orthogonal with each other. Eq. 3 derives the contradiction. ∎

Secondly, attribute editing can be accomplished via semantic decomposition. We can edit the attributes in set S^+ by preserving the intensity of attributes in set S^-. However, the semantic of attributes in S^- may change. Lemma 3.2 describes detailed proves.

Lemma 3.2. *Given the attributes shifting vector* $\mathbf{a} = [a_1, a_2 \ldots a_{|\mathbf{A}|}]$. *There exists a linear decomposition* $\Gamma = [\Gamma_0(\cdot), \Gamma_1(\cdot) \ldots]$ *S.T.* $\mathbf{r}_i = \sum_{k \in |\Gamma|} \Gamma_k(\mathbf{r}_i)$ *and* $\forall j \in |\mathbf{A}|$, $\mathbf{w}_k = \sum_{k \in |\Gamma|} \Gamma_k(\mathbf{w}_j)$, *where* $\Gamma_k(\cdot)$ *is a index function. Denoting* Γ^+ *be the set of component need to be varied when editing attributes* S^+. *Then we can formulate the attribute editing below,*

$$\Big(\sum_{j \in S^+} \sum_{k \in |\Gamma^+|} \eta_k \Gamma_k(\mathbf{w}_j) \Big) \mathbf{w}_l = 0 \qquad \forall l \in S^- \quad (4)$$

$$\Big(\sum_{j \in S^+} \sum_{k \in |\Gamma^+|} \eta_k \Gamma_k(\mathbf{w}_j) \Big) \mathbf{w}_l = a_l \qquad \forall l \in S^+ \quad (5)$$

η_k *is a scalar variable that varies the magnitude of sub-semantic* $\Gamma_k(\mathbf{w}_j)$. *Eq. 4 constrains that intensity scores of* S^- *will be preserved during attribute editing. Eq. 5 constrains that the intensity score of* S^+ *should be changed with expectation. The Eq. 4 and Eq. 5 have feasible solutions if and only if* $d >= |\mathbf{A}|$. *It means we can decompose the attribute semantic into different sub-semantics. Then, we can manipulate the attributes of interests* S^+ *by preserving the intensity of other attributes* S^- *and modifying the magnitude of sub-semantics.*

3.2. *S2FGAN* Framework

We introduce our proposed framework, *S2FGAN*, in Figure 2. It has four components: a Encoder (Image Latent Encoder, $\mathbf{E}_i$, Sketch Latent Encoder, $\mathbf{E}_s$), Attribute Mapping Network, $\mathbf{M}$, Style Aware Decoder, $\mathbf{D}$ and a Discriminator,

$\mathbf{F}$. For clarification, we use $\mathbf{G}_i$ and $\mathbf{G}_s$ to denote the image reconstruction branch $\mathbf{D}(\mathbf{M}(\mathbf{E}_i(I_{gt}), \mathbf{a}))$ and sketch-to-image generation branch $\mathbf{D}(\mathbf{M}(\mathbf{E}_s(I_i), \mathbf{a}))$ of our generation. Our sketch-to-image translation with attribute editing relies on encoding and manipulating the sketch's latent code. Our decoder and discriminator adopt StyleGAN [16] as backbone. Here, we describe the architecture of the encoder and attribute mapping network.

Encoder. Regardless of any input sketch, I_i, learning a direct mapping to photo-realistic image, I_{out} is difficult. With the increase of the input's spatial resolution, the network should learn more fine-grained facial features to construct a photo-realistic look. However, the sketches often fail to describe such facial texture. To solve this problem, we design an Image Latent Encoder, $\mathbf{E}_i$ and a Sketch Latent Encoder, $\mathbf{E}_s$, that treat image construction as an auxiliary task to help the sketch-to-image generation. Our encoder is a simple $(\lfloor \sqrt{\log(HW)} \rfloor - 2)$ layers ResNet [8]. Considering different facial attribute desires different sizes of convolution feature maps, we pool the feature maps and pass the summarised feature to a multilayer perceptron (MLP) before each down-sampling operations. We then sum the processed feature from different encoder hierarchies and refine these features through another MLP. These operations enable the encoder to learn the facial feature spatially. For example, a pale skin attribute would more reasonably present in early convolution features of the encoder.

Attribute Mapping Network. We describe two different versions of Attribute Mapping Network using Lemma 3.1 and Lemma 3.2. The objectives for training are presented in the next subsection.

According to Lemma 3.1, after disentangling the attribute semantics of interest by ensuring them to be orthogonal to each other, the attribute editing task becomes a simple addition operation. We perform the attribute editing by Eq. 6 in Figure 2 (a).

$$\mathbf{r}'_i = \mathbf{r}_i + \sum_{j \in [1 \ldots |\mathbf{A}|]} a_j \mathbf{w}_j \qquad (6)$$
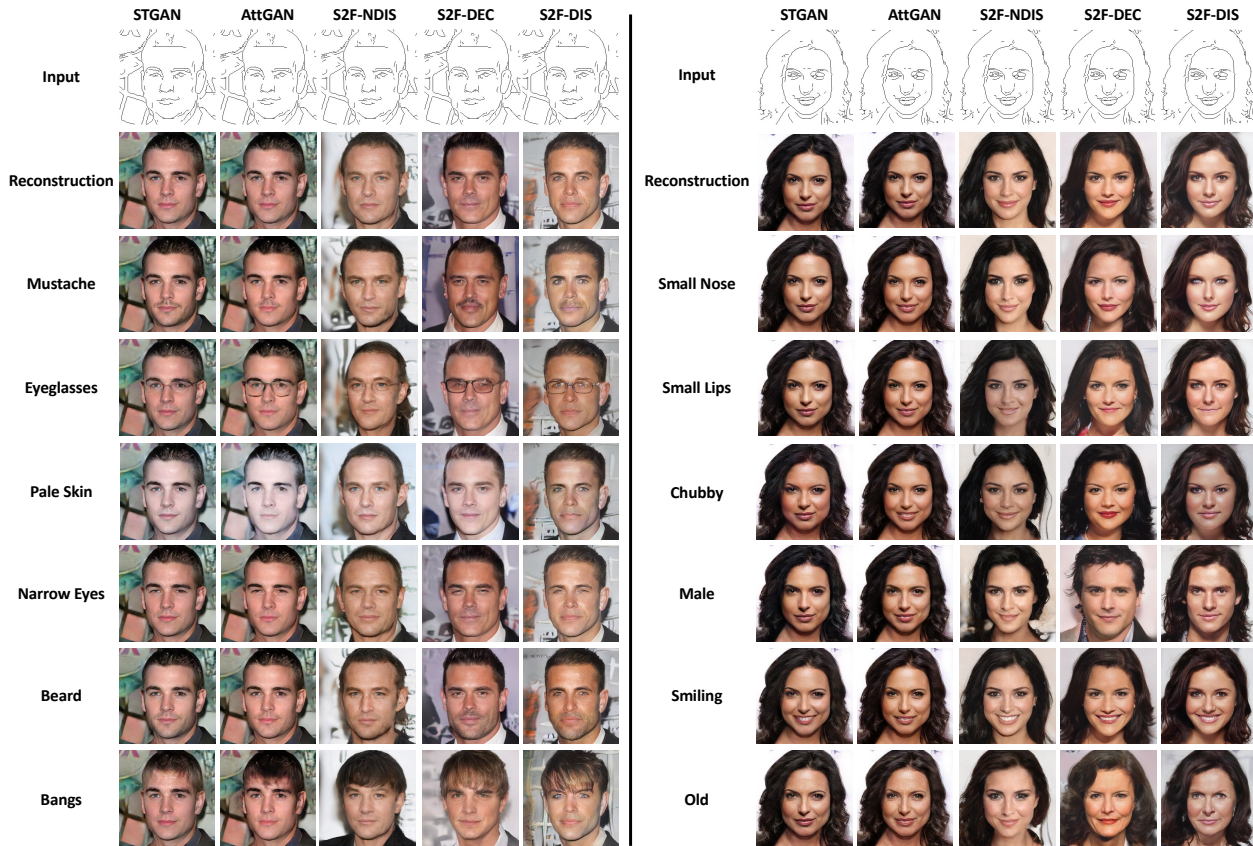
Figure 4: Single attribute editing results for STGAN [22], AttGAN [9], S2F-NDIS, S2F-DEC and S2F-DIS.

According to Lemma 3.2, we model Eq. 4 and Eq. 5 inside the subnetwork Attribute Mapping Network (Figure 2 (b)) which maintains a trainable embedding layer $\mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^{|\mathbf{A}| \times d}$. This embedding layer (named Direction in Figure 2) could also be a replica of attributes hyperplane $[\mathbf{w}_1 \ldots \mathbf{w}_{|\mathbf{A}|}]$. Let f_1, f_2, f_3 and f_4 denote four convolution 1D layer. And δ denote a multi-layer convolution block, which models semantic decomposition. The key operations of attribute mapping network are:

$$\mathbf{b}_i = \delta(concat(\mathbf{e} \cdot \mathbf{a}, \mathbf{r}_i)), \quad \mathbf{c}_i = \mathbf{Norm}(f_1(\mathbf{b}_i)),$$
$$\mathbf{r}'_i = f_2(\mathbf{c}_i \mathbf{c}_i^T f_3(\mathbf{b}_i)) + f_4(concat(\mathbf{e} \cdot \mathbf{a}, \mathbf{r}_i))$$

We repeat operations above four times. *concat* and $\mathbf{Norm}$ represent channel-wise concatenation and vector normalization along channel dimension, respectively. $\mathbf{b_i}$ and $\mathbf{c_i}$ are intermediate variables. First, δ learns to decompose the current latent code and desired attribute shifting latent code into sub-semantics. Second, we calculate the cosine similarity between each of the sub-semantics, which scores each unit semantic. It aims to incorporate the magnitude changes for different sub-semantics for preserving the intensity of non-modified attributes. For attributes with non-zero attribute shifting value, they may correlate with each other, and the sub-semantics among them also need to be aligned. Meanwhile, the best-edited representation is searched with the awareness of the current latent code $\mathbf{r}_i$. This idea

(Lemma 3.2) is especially helpful when multiple attribute editing is involved in a latent space, and a pretrained generator (encoder and decoder) is available. In this way, we can adapt to arbitrary complex attribute editing while preserving the appearance of the sketch I_i.

3.3. Multi-Objective Learning

There are two extra objective functions in our compared with conventional GAN framework [15, 16]. Let $\mathbf{u}_i$ denote attributes of ground truth images I_{gt}.

Semantic Level Perceptual Loss. If the attribute shifting vector, $\mathbf{a} = \mathbf{0} \in \mathbb{R}^{|\mathbf{A}|}$, the generator, $\mathbf{G}_s$ will reconstruct the ground-truth image, I_{gt} from a sketch input, I_i. Instead of directly imposing the L1 loss [12, 11, 9, 4, 22], Perceptual Loss [31, 33], and Feature Matching Loss [31, 33, 3, 20] between the synthesis I_{out} and ground-truth I_{gt} image, we propose to match the semantics of synthesis and ground-truth image on the latent space. With the increase of desired synthesis resolution, the size of model output increases exponentially. In that case, optimization using L1, perceptual, and feature matching loss become more challenging. In contrast, our GAN's latent space calculates a discriminative summarization of low-dimensional image features. It maintains domain-specific perceptual context and ensures faster optimization. We define the Semantic Level Percep-
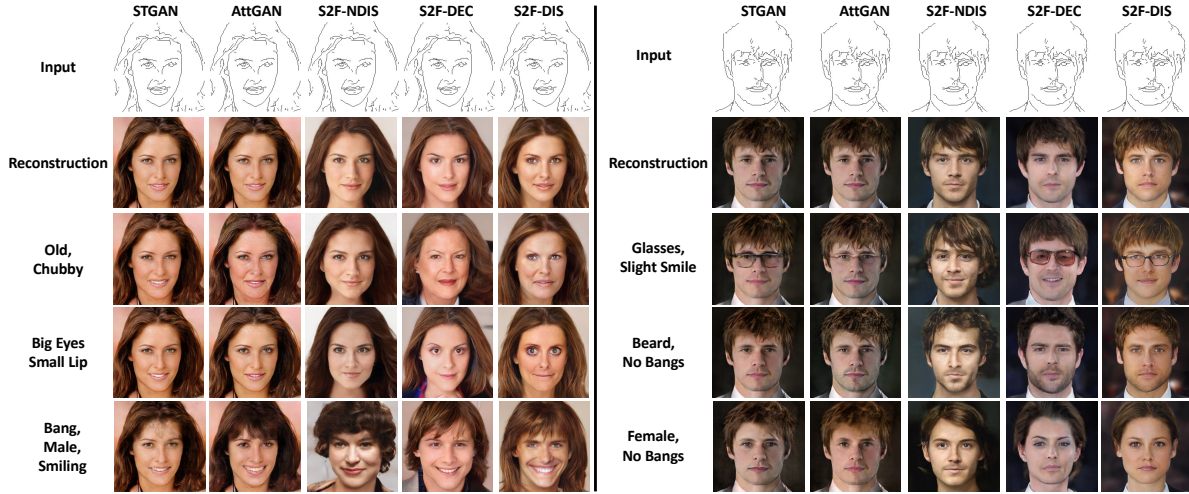
Figure 5: Multi-attribute editing results for AttGAN [9], STGAN [22], S2F-NDIS, S2F-DEC and S2F-DIS.

tual Loss as,

$$\mathcal{L}_{sem} = \mathbb{E}\big[\|\mathbf{E}_i^*(I_{gt}) - \mathbf{E}_s(I_i)\|_2\big] \tag{7}$$

where '*' indicates the component removed from the computation graph during backpropagation. A domain discriminator $\mathbf{Q}$ regularizes the sketch encoder E_s. It encourages sketch latent code, $\mathbf{E}_s(I_i)$ to stick around the latent space of $E_i(I_{gt})$. This is helpful when translating the bad drawn sketches. (See Figure 3). To avoid two-stage adversarial training, GRL layer [6] is used to reversal the gradient before updating the sketch encoder $\mathbf{E}_s$ during training.

$$\mathcal{L}_{domain} = \log\Big(1 - \mathbf{Q}\big(\mathbf{E}_s(I_i)\big)\Big) + \log\big(1 - \mathbf{Q}\big(\mathbf{M}(\mathbf{E}_s(I_i), \mathbf{a})\big)\big)$$
$$+ \log\mathbf{Q}\big(\mathbf{E}_i^*(I_{gt})\big) + \log\mathbf{Q}\big(\mathbf{M}(\mathbf{E}_i^*(I_{gt})), \mathbf{a}\big) \tag{8}$$

Meanwhile, to discover the latent code of ground-truth image I_{gt}, we use image reconstruction as an auxiliary task. Another intuition is that image reconstruction is simpler than sketch-to-image translation. We use the L1-Loss $\mathcal{L}_{rec}$ and perceptual loss $\mathcal{L}_{percept}$ for image reconstruction.

Latent Semantic Loss. As classifying the attributes from the sketch is noisy, we leverage the side effect of auxiliary image reconstruction tasks. Here, we describe the attribute reconstruction loss corresponding to Lemma 3.1 and Lemma 3.2, separately.

For Lemma 3.1, we focus on learning a disentangled latent space, where all attributes in $\mathbf{A}$ are orthogonal to each other. The loss is:

$$\mathcal{L}_{ortho} = \mathbb{E}\big[\|\mathbf{E}_i(I_{gt})\mathbf{W}^T - \mathbf{u}_i\|_2^2\big] + \|\mathbf{W}\mathbf{W}^T - \mathbb{1}\|_2 \tag{9}$$

We denote $\mathbf{W} = [\mathbf{w}_1 \ldots \mathbf{w}_{|\mathbf{A}|}] \in \mathbb{R}^{|\mathbf{A}| \times d}$ as trainable parameters and $\mathbb{1}$ as a $|\mathbf{A}| \times |\mathbf{A}|$ identity matrix.

Remark. We have to maintain a proper dimensionality d for the latent code to prevent the co-adaption of $\mathcal{L}_{rec}$ and $\mathcal{L}_{ortho}$. If d is extremely large, they ($\mathcal{L}_{rec}$ and $\mathcal{L}_{ortho}$) will depend on different latent code chunks, leading to the failure of further editing of attributes. In our experiment, we set d to 512 during our training.

For Lemma 3.2, we aim to learn a decomposition that preserves attributes' intensity when it is not edited. The associated loss is:

$$\mathcal{L}_{decom} = \quad \mathbb{E}\big[\|\mathbf{E}_i^*(I_{gt})\mathbf{W}^T - \mathbf{u}_i\|_2^2\big]$$
$$+ \mathbb{E}\big[\|\mathbf{M}(\mathbf{E}_i^*(I_{gt}), \mathbf{a}) - \mathbf{E}_i^*(I_{gt})(\mathbf{W}^*)^T - \mathbf{a}\|_2^2\big]$$
$$+ \mathbb{E}\big[\|\mathbf{M}(\mathbf{M}(\mathbf{E}_i^*(I_{gt}), \mathbf{a}), -\mathbf{a}) - \mathbf{E}_i^*(I_{gt})\|_2\big] \tag{10}$$

The first component learns to classify the attributes from the latent code. The second component enforces the attribute mapping networks manipulating the attributes with the desired intensity. The third component ensures the identity-invariant of attribute editing, i.e., the latent code should be reconstructed with two inverse editing behaviors. There is a significant difference between $\mathcal{L}_{ortho}$ and $\mathcal{L}_{decom}$. $\mathcal{L}_{ortho}$ forces the encoder to decorrelate the attributes $\mathcal{A}$ with gradient support. $\mathcal{L}_{decom}$ learns to classify the attributes from latent code while removing the latent code from the computation graph. During our experiments, the $\mathcal{L}_{ortho}$ is stronger at attribute editing when the image quality of $\mathcal{L}_{decom}$ is better. $\mathcal{L}_{ortho}$ decorrelates the attribute semantics that contradicts with data distribution. In contrast, $\mathcal{L}_{decom}$ allows correlating attributes. Additionally, we use the conventional adversarial loss $\mathcal{L}_{adv}$ with gradient penalty $\mathcal{L}_{R1}$ from [15, 16] to train our model.

Final Objectives. The Lemma 3.2 and 3.1 lead two distinct models. They use the same semantic level perceptual loss, adversarial loss, and gradient penalty but corresponding with different latent semantic loss as indicated above. See **supplementary material** for details.

4. Experiments

Dataset. We use a high-quality face dataset CelebAMask-HQ [19]. For this work, we resize all images to resolution 256×256. We choose every 20th image as the testing set ($1,500$ images) from the whole dataset, and the remaining images are in the training set ($28,500$ images).
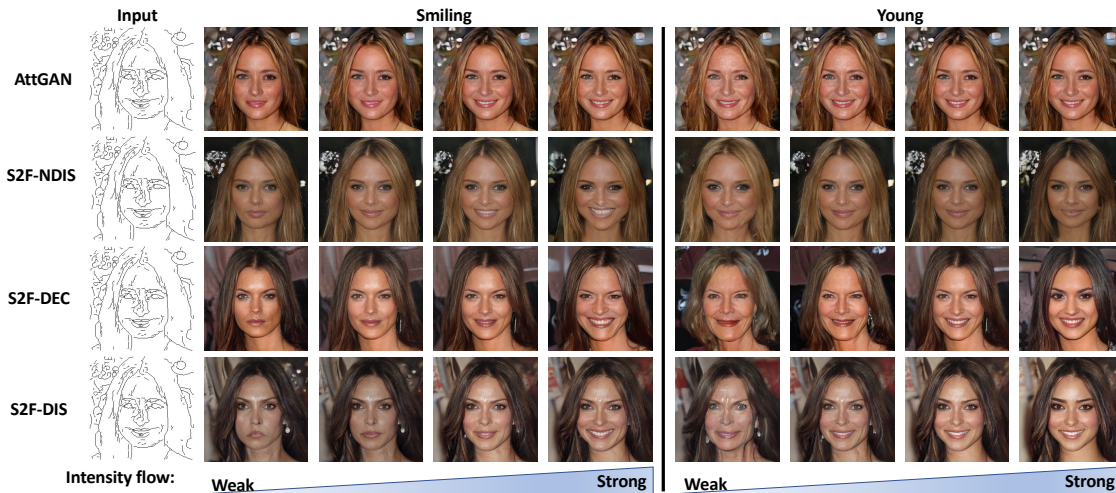
Figure 6: Intensity control for AttGAN [9], S2F-NDIS, S2F-DEC and S2F-DIS on two state-like attributes (Smiling and Young).

We experiment with twelve state-like attributes (Smiling, Male, No_Beard, Eyeglasses, Young, Bangs, Narrow_Eyes, Pale_Skin, Big_Lips, Big_Nose, Mustache, and Chubby). The statistic of each attribute in the training and testing set are summarised in supplementary materials.To obtain face sketches, we use HED edge detector [32], and post-process with the steps employed by Isola *et al.* [12]. We use synthesized real drawings from edge-based sketches [33, 12, 13, 26] to avoid laborious costs and difficulty in collecting hand drawings and images pair.

Method for Comparison. *For sketch-to-image translation*, we compare our model with Pix2PixHD [31], Deep plastic Surgery (DPS) [33], Deep Face Drawing (DFD) [3] and Deep Face Pencil (DFP). *For attribute edining*, we compare our work with the most recent model-based state-of-the-art attribute generation methods, AttGAN , and STGAN. These methods are originally designed to manipulate attributes of photo-realistic images. For this, we train a Pix2PixHD model [31] to translate sketches to images before applying attribute manipulation. The local enhancer of Pix2PixHD [31] are removed to compare in resolution 256×256. Thus, we train all models from scratch with their published recommended settings if original authors do not provide checkpoints in the CelebAMask-HQ dataset. We denote *our work* as S2F-DIS and S2F-DEC motivated from Lemma 3.1 and Lemma 3.2, respectively. DIS and DEC are short for disentangle and decompose, respectively. They are our core concepts for the two lemmas. We also compare the final recommendation of our approach with several plausible alternatives. To illustrate the effectiveness of the proposed image reconstruction auxiliary task and semantic perceptual loss, we remove the image encoder $\mathbf{E}_i$ of S2F-DIS while denoting this ablation with S2F-NDIS.

Implementation Details.[1] We train our model from scratch under Adam optimizer with learning rate 0.002. The decays

of Adam optimizer are set to 0.0 and 0.99. Our model is trained on two Nvidia Tesla Volta V100-SXM2-32GB with Intel Xeon Cascade Lake Platinum 8268 (2.90GHz) CPUs. We use batch size 24. During training, images and corresponding sketches are randomly cropped and resized with ratio 0.8 and probability 0.5. We randomly used horizontal flip with probability of 0.5 also. When trains S2F-DIS, a always set to zero. See the **supplementary material** for detailed network architecture and parameters.

4.1. Sketch-to-Image Translation

Translating Human Drawn Sketches. We compare with the state-of-the-art models (with their provided checkpoint) by using the human drawn sketches [33] in Figure 3. Although [33, 3, 20, 31] used different sketch extraction methods, the ultimate goal is to adapt the human-drawn sketches to retrieve photo-realistic images. The Pix2PixHD, DFD, and DFP fail in generating high-quality and sensitive images. Though DPS and S2F-NDIS can capture the outline of input sketches, the image quality remains questionable. The layouts and shape variations between input sketch and output synthesis are expected because we subjectively defined the refinement level. Depends on users' confidence in their drawn sketches, they have the flexibility of controlling the fitness between input sketches and output synthesises (See Figure 4-7 in the **supplementary material** for synthesis without any refinement). Our method combines the truncation trick [2] (used in StyleGAN [15, 16]) and k-nearest neighborhood for adapting badly drawn sketches (used in DFP [3]). Because of the superior semantic level perceptual loss (with domain regularization), our sketch encoder $\mathbf{E}_s$ searches the sketches' latent code within the scope of ground truth image's I_{gt} latent space. It helps the model to refine the badly drawn human sketches. See **supplementary material** for algorithm details, more comparison and style transfer in refining drawn sketches and comparison for translating machine extracted sketches.

---

[1] Code and models are available at: `https://github.com/Yan98/S2FGAN`

Table 1: Quantitative evaluation: facial attributes editing accuracy (%).

Model	Smiling	Male	No Beard	Eye glasses	Young	Bangs	Narrow Eyes	Pale Skin	Big Lips	Big Nose	Mustache	Chubby
AttGAN [9]	56.0	24.0	44.0	82.0	46.0	56.0	80.0	66.0	66.0	54.0	10.0	12.0
STGAN [22]	72.0	36.0	64.0	92.0	56.0	84.0	81.0	86.0	80.0	72.0	45.0	32.0
S2F-NIDS	**100**	**100**	97.8	38.4	96.5	99.5	**99.5**	91.0	68.2	84.5	**66.4**	**78.4**
S2F-DEC	**100**	99.6	57.3	99.3	98.8	**99.8**	99.4	**96.2**	81.3	92.1	63.9	75.5
S2F-DIS	**100**	99.9	**98.9**	**99.7**	**99.5**	98.0	91.4	72.5	**90.0**	**98.9**	66.1	58.3
Ground-truth	93.0	97.6	95.8	99.5	87.9	94.0	91.0	95.7	68.7	80.3	95.9	94.4

Table 2: Quantitative evaluation of synthesis quality and diversity.

Metric / Model	IS	FID
DFP [3]	2.61 ± 0.02	339.2 ± 0.85
DFD [20]	$\mathbf{3.09} \pm 0.03$	313.8 ± 0.53
DPS [33]	2.65 ± 0.03	325.7 ± 0.88
Pix2PixHD [31]	2.99 ± 0.04	38.71 ± 0.69
S2F-NDIS	2.58 ± 0.04	34.18 ± 0.37
S2F-DEC	2.73 ± 0.03	28.21 ± 0.76
S2F-DIS	3.04 ± 0.04	$\mathbf{26.14 \pm 0.43}$

Attribute Editing. We present the comparison of S2F-DIS, S2F-DEC, and baseline methods for single and multiple attribute editing. For the single attribute case (see Figure 4), AttGAN [9] and STGAN [22] fail in the majority of attribute editing cases and have weak editing effect, for example, for the "Male" case. We also compare our proposed architecture with AttGAN [9] and S2F-NIDS for attribute intensity control in Figure 6. Here, we present the easiest editing examples, smiling and young. Because of learning the orthogonal semantic vectors, the S2F-DIS leads to a better editing performance. Meanwhile, this learning process needs to de-correlate the data, making the image quality slightly worse than the S2F-DEC. For multi-attribute editing cases (see Figure 5), [22, 9] sometimes ignores part of the attributes that need to be changed, such as "Old", and "Chubby", "Beard" and "No Bang" editing cases. Moreover, those methods usually fail to retouch the existing attributes such as to change "Lip Size". AttGAN [9] and STGAN [22] control the attribute editing based on concatenation of down-sampled inputs and scaled labels that lack continuity and variation. S2F-DEC aims to find a potentially attribute editing under Eq. 4 and 5. It only constrains the intensity score of attributes, where attributes' semantic may change as we suggested in Lemma 3.2. Thus, after attribute editing, there might be a different visual representation for all/some attributes but only guarantees the equivalence of intensity score between before- and after-editing. For S2F-DIS, semantic and intensity can be preserved simultaneously, as Lemma 3.1 indicates. However, both do not provide any performance guarantees for the attributes that are not in interest. S2F-NIDS lacks semantic level perceptual loss during training, making the learning noisy and leading to poor editing results. Our final recommendation (S2F-DEC, S2F-DIS) generates photo-realistic images from a sketch with superior attribute intensity control. We present more qualitative results in the **supplementary material.**

4.2. Quantitative Evaluation.

In Table 1, we use a classifier suggested by He *et.al* [9] to validate the accuracy of attribute editing for our model. The evaluation classifier is trained on our training data of CelebAMask-HQ dataset [19] for resolution 128×128. We bilinearly downsample the generated images to resolution 128×128 before feeding them into the evaluation classi-

fier. Note, there are around $\frac{1}{7}$ of training data where the target resolution is higher than their original implementation (compared with the original AttGAN [9] and STGAN [22]). Our proposed framework S2F-DEC and S2F-DIS outperform the rest of the methods in the majority of cases. In the last row (GT) of Table 1, we also present the classifier's performance on real testing data that serves the upper bound of performance. Our model can control attribute intensity, which can construct obvious attributes for the evaluation. It also can enable the evaluation classifier to identify the edited attributes easily. Moreover, we use Frechet Inception Distance (FID) [10] and Inception Score (IS) [27] to measure diversity and quality of synthesized images (1,500 testing images) in Table 2. We do not set any refinement level for all methods during the comparison because it is a subjective metric. Similar to [20], we calculate the FID and IS based on the simulated human-drawn sketches by randomly dilating, deforming, and removing some edges from machine extracted sketches. DFD [3] and DPS [33] heavily rely on defining appropriate refinement levels and hence not performing well. Though S2F-DEC has worse IS than Pix2PixHD, S2F-DEC can better translate human-drawn sketches and provide advanced rendering facial textures (Figure 3). The S2F-DIS earns the best and second-best performance in FID and IS, respectively. To better understand the performance of our model, we present a user study in **supplementary material** for comparing with state-of-the-art methods.

5. Conclusion

This paper proposes two photo-realistic face generation models, S2F-DEC and S2F-DIS, given a sketch image as input. They can ascribe attributes on the generated face and include smooth manipulation over intensities of attributes. Further, our generated face preserves subject identity considering both single and multi-attribute editing cases. By adopting proposed semantic level perceptual loss and latent semantic loss, we can construct photo-realistic faces with the flexibility of shifting desired face attributes. Experiments on large face collection datasets demonstrate that S2F-DEC and S2F-DIS can accurately edit face attributes with more excellent controllability, even with non-photo-realistic inputs.

References

[1] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 599–608. PMLR, 2018.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[3] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: deep generation of face images from sketches. *ACM Trans. Graph.*, 39(4):72, 2020.

[4] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020, 2017.

[5] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T. Freeman. Sparse, smart contours to represent and edit images. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3511–3520. IEEE Computer Society, 2018.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[9] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Trans. Image Process.*, 28(11):5464–5478, 2019.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017.

[11] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, pages 179–196. Springer, 2018.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[13] Youngjoo Jo and Jongyoul Park. SC-FEGAN: face editing generative adversarial network with user's sketch and color. *CoRR*, abs/1902.06838, 2019.

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019.

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. IEEE, 2020.

[17] Samuli Laine. Feature-based metrics for exploring the latent space of generative models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.

[18] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5967–5976, 2017.

[19] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5548–5557. IEEE, 2020.

[20] Yuhang Li, Xuejin Chen, Binxin Yang, Zihan Chen, Zhihua Cheng, and Zheng-Jun Zha. Deepfacepencil: Creating face images from freehand sketches. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20:*

The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, pages 991–999. ACM, 2020.

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 700–708, 2017.

[22] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. STGAN: A unified selective transfer network for arbitrary image attribute editing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3673–3682. Computer Vision Foundation / IEEE, 2019.

[23] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual GAN. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 213–228. Springer, 2018.

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[25] Stanislav Pidhorskyi, Donald A. Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14092–14101. IEEE, 2020.

[26] Tiziano Portenier, Qiyang Hu, Attila Szabó, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: deep sketch-based face image editing. *ACM Trans. Graph.*, 37(4):99:1–99:13, 2018.

[27] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016.

[28] Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The riemannian geometry of deep generative models. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 315–323. IEEE Computer Society, 2018.

[29] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9240–9249. IEEE, 2020.

[30] Paul Upchurch, Jacob R. Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Q. Weinberger. Deep feature interpolation for image content changes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6090–6099. IEEE Computer Society, 2017.

[31] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8798–8807. IEEE Computer Society, 2018.

[32] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1395–1403. IEEE Computer Society, 2015.

[33] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. *CoRR*, abs/2001.02890, 2020.