Human-Inspired Facial Sketch Synthesis with Dynamic Adaptation

Fei Gao[1], Yifan Zhu[2], Chang Jiang[2], Nannan Wang[3*]
[1]Hangzhou Institute of Technology, Xidian University [2]Hangzhou Dianzi University [3]Xidian University
fgao@xidian.edu.cn, 2961695289@qq.com, jc233@hdu.edu.cn, nnwang@xidian.edu.cn

Abstract

Facial sketch synthesis (FSS) aims to generate a vivid sketch portrait from a given facial photo. Existing FSS methods merely rely on 2D representations of facial semantic or appearance. However, professional human artists usually use outlines or shadings to covey 3D geometry. Thus facial 3D geometry (e.g. depth map) is extremely important for FSS. Besides, different artists may use diverse drawing techniques and create multiple styles of sketches; but the style is globally consistent in a sketch. Inspired by such observations, in this paper, we propose a novel Human-Inspired Dynamic Adaptation (HIDA) method. Specially, we propose to dynamically modulate neuron activations based on a joint consideration of both facial 3D geometry and 2D appearance, as well as globally consistent style control. Besides, we use deformable convolutions at coarse-scales to align deep features, for generating abstract and distinct outlines. Experiments show that HIDA can generate high-quality sketches in multiple styles, and significantly outperforms previous methods, over a large range of challenging faces. Besides, HIDA allows precise style control of the synthesized sketch, and generalizes well to natural scenes and other artistic styles. Our code and results have been released online at: https://github.com/AiArt-HDU/HIDA.

1. Introduction

Making computers create arts like human beings, is a longstanding and challenging topic, in the artificial intelligence (AI) area [2]. To this end, researchers have made great efforts and proposed numerous methods, such as neural style transfer (NST) [17] and image-to-image translation (I2IT) [18, 46]. These methods mainly tackle cluttered image styles, such as oil paintings [17]. In this paper, we are interested in creating artistic sketches from facial photos, which is referred to as face sketch synthesis (FSS) [34].

For now, there has been significant progress in FSS inspired by the excellent success of Generative Adversarial Networks (GANs) [18]. Specially, researchers have proposed various techniques, including embedding image prior [43], semi-supervised learning [5], self-attention/transformer based methods [11, 47, 9], hierarchical GANs [28, 40, 8], composition assistance [37], and semantic adaptive normalization [21], to boost the quality of synthesized sketches. However, all these methods merely use 2D appearance or semantic representations of the input photo. They may fail to handle serious variations in appearance, such as the pose, lighting, expression, and skin color.

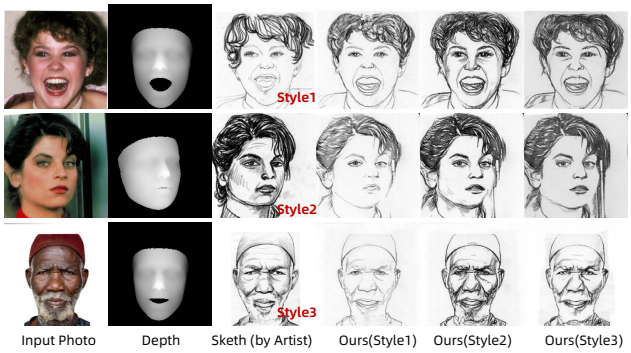


Figure 1: Illustration of facial photos, depth maps, multi-style facial sketches drawn by human artists [8], and the corresponding results synthesized by our method.

To tackle this challenge, we propose a novel method, inspired by how human artists draw a sketch. We observe that facial 3D geometry plays a significant role in human artists' drawing process. Besides, a professional human artist considers comprehensive information, including facial 3D geometry, 2D appearance, and the artistic style, to execute a sketch portrait. We summarize the drawing methodologies of human artists [22] into the following four folds:

- **Local 3D geometry conveyor**: First, artists typically use abstract and deformable outlines to characterize major geometry, and use different shading methodologies, e.g. hatching, blending, and stippling, to convey local 3D structures [3].

* Corresponding Author

- **Local 2D appearance representation**: Second, artists may use different shading or tonal techniques to represent local 2D facial appearance, so as to depict variations in lighting, color, texture, etc.
- **Sketches in diverse styles**: Third, different artists may use diverse drawing methods and create multiple styles of sketches. In other words, they may use divergent textures to represent the same facial area. Fig. 1 shows three styles of sketches drawn by artists [8]. Obviously, Style1 is extremely abstract and mainly contains sketchy outlines. In contrast, Style3 depicts facial 3D geometry with a lot of shading textures.
- **Globally consistent style**: Finally, the style of pencil-drawing is usually consistent in a single sketch. As shown in Fig. 1, although there are distinct inter-style divergences, the style of pencil-drawing is globally consistent across different regions inside each sketch.

Inspired by these observations, we seek to guide the synthesis of sketch portraits by using comprehensive information, including facial 3D geometry and 2D appearance, as well as global style control. In the implementation, given a facial photo, we use the depth map to represent its 3D geometry, and use the encoding features to represent its 2D appearance. Afterwards, we combine them with a style map to dynamically modulate deep features for generating a sketch. Inspired by the success of SPADE [27] in style control [41] and the local flexibility of dynamic neural networks [14], we propose to dynamically modulate neuron activations, based on a joint consideration of all these information. Such modulation is conducted though both dynamic normalization and activation. Specially, we propose a novel dynamic activation function, termed Informative ACON (InfoACON), and a dynamic normalization module, termed DySPADE. In addition, we use deformable convolutions [7] to align deep features [16] at coarse scales for generating abstract and distinct sketchy outlines. Initially, the dynamic adaptation and deformation simulate the flexibility and abstract process of human artists during drawing.

Based on the above mentioned contributions, we build a Human-Inspired Dynamic Adaptation (HIDA) method for FSS. We conduct experiments on several challenging datasets, including the FS2K [8], the FFHQ [20], and a collection of faces in-the-wild. Our method outperforms state-of-the-art (SOTA) methods both qualitatively and quantitatively. Besides, our method allows precise style control and can produce high-quality sketches in multiple styles. Even for faces with serious variations, the synthesized sketches present realistic textures and preserve facial geometric details. In addition, extensive ablation studies demonstrate the effectiveness of the proposed dynamic and adaptive modulation techniques. Finally, our model, although trained for faces, can generate high-quality sketches for natural scenes.

2. Related Works

Our work is related to GANs-based FSS methods. Besides, our method is highly inspired by semantic adaptive normalization [27] and dynamic activation [25].

GANs-based FSS. The latest FSS methods are typically based on GANs [18, 12], where the mapping from a facial photo to a sketch is modeled as an image-to-image translation task [18]. Some latest methods use 2D semantic information to guide the generation process. For example, Yu et al. [37] propose a stacked composition-aided GANs to boost quality of details. Inspired by the great success of spatially adaptive (de)normalization (SPADE) [27] in semantic image generation, Wang et al. [47] and Qi et al. [29] spatially modulate decoding features according to facial parsing masks. Li et al. [21] propose an enhanced SPADE (eSPADE) by using both facial parsing masks and encoding features for feature modulation.

Recently, researchers seek to solve the challenge of unconstrained faces by constructing large datasets. Fan et al. [8] release a challenging FS2K dataset, which consists of multi-style sketches for faces with diverse variations. Nie et al. [26] propose a novel WildSketch dataset and a Perception-Adaptive Network (PANet). In PANet, deformable feature alignment (DFA) and patch-level adaptive convolution are used. Different from [26], we analyze the effects of DFA, and only use DFA over the coarse scales. Besides, we propose to dynamically modulate neuron activations based on facial depth and artistic style.

Semantic Adaptive Normalization. Recently, Park et al. [27] propose to modulate deep features based on semantic layouts for semantic image synthesis. In SPADE, deep features are modulated based on semantic layouts. Afterwards, Zhu et al. [48] propose Semantic Region-Adaptive Normalization (SEAN) to control the style of each semantic region individually. To boost the efficiency of SPADE, Tan et al. [32] propose a Class-Adaptive (DE)Normalization (CLADE) layer by replacing the modulation networks with class-level modulating parameters. All these adaptive normalization layers use 2D semantic maps and show amazing performance in generating photo-realistic images [24] and face sketches [37]. In this paper, we use pix-wise dynamic activation in the normalization block, so that the modulating parameters would flexibly adapt to local information. Experimental results show that the dynamic normalization are essential for detailed synthesis of facial sketches.

Dynamic Activations. Recently, Chen et al. [6] propose a Dynamic ReLU (DY-ReLU) function, where parameters in Leaky ReLU are learned from all input elements. Ma et al. [25] propose a costumed activation function, termed ACON, which automatically decides whether a neuron is active or not. ACON has several variants, among which the pixel-wise version of metaACON shows remarkable performance. Give a neuron activation x, the output of metaA-

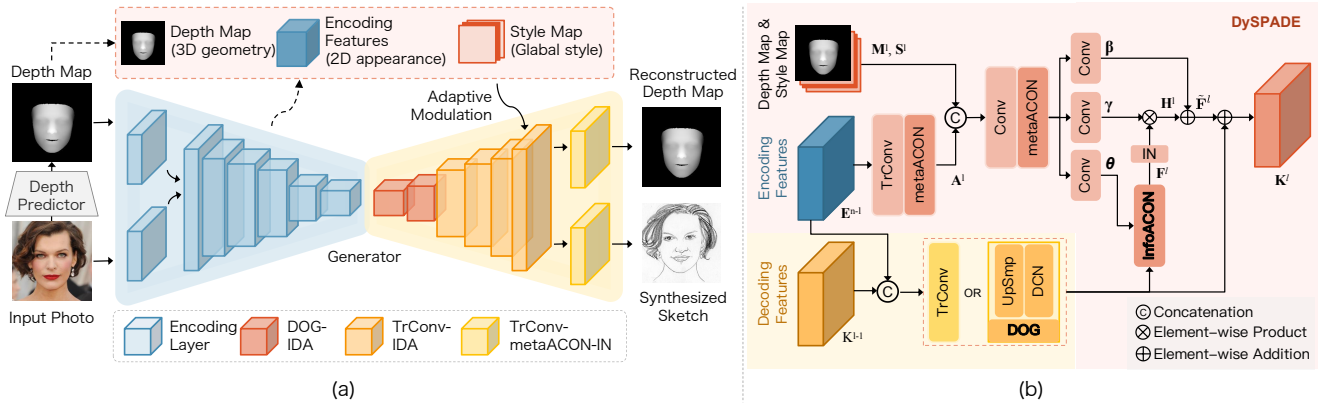


Figure 2: Pipeline of the proposed *Human-Inspired Dynamic Adaptation* (HIDA) method for facial sketch synthesis. (a) The overall generator architecture, (b) an decoding layer with DySPADE, InfoACON, and DOG.

CON is formulated as:

$$y = (p_1 - p_2) \cdot \sigma(\theta(p_1 - p_2)x) + p_2 x, \tag{1}$$

where $\theta = \sigma(x)$, σ is a Sigmoid function, p_1, and p_2 are learnable parameters. In this paper, we use metaACON, instead of ReLU or Leaky ReLU, in part of our networks. Besides, we propose to learn spatially-adaptive parameter θ according to the 3D geometry, 2D appearance, and global style control. Our activation function proves boosting the performance and allowing precise style control.

3. The Proposed

We aim to translate a facial photo $\mathbf{X}$ to a sketch $\mathbf{Y}_s$, in style s, drawn by an artist. Here $s = 1, 2, ..., S$ is a style label, S is the total number of styles. In this work, we seek to guide the synthesis of sketch portraits by using comprehensive facial information, including both 3D geometry and 2D appearance, as well as the global style control. Given a facial photo, we use the corresponding depth map $\mathbf{D}$ to represent its 3D geometry. Afterwards, we combine them with a global style map $\mathbf{S}$ to decode a facial sketch. In this way, our goal is formulated as learning a mapping from $\{\mathbf{X}, \mathbf{D}, \mathbf{S}\}$ to Y_s, i.e. $G : \{\mathbf{X}, \mathbf{D}, \mathbf{S}\} \mapsto \mathbf{Y}_s$.

To supervise our model, it is necessary to obtain depth maps for input facial photos. However, it is usually impossible to obtain ground truth depth information in practical applications. Therefore we use state-of-the-art (SOTA) depth prediction methods to estimate the depth map of an input facial photo. In practice, we use 3DDFA [13] as the depth predictor, because it has been widely used and shown excellent performance in various 3D face reconstruction tasks.

The overall pipeline of our model is as shown in Fig. 2. It contains an off-line facial depth predictor P, a generator G, and a patch-wise discriminator D. In addition to a facial sketch, we enforce G to reconstruct the input depth map $\mathbf{D}$ from features representing the sketch. In this way, the generated sketch $\hat{\mathbf{Y}}_s$ would convey the 3D geometry of $\mathbf{X}$. Besides, we boost the capacity of generator by using a dynamic normalization module and a dynamic activation function. Finally, to formulate the abstraction methodology of human artists in drawing sketchy outlines, we propose using deformable convolutions to align features at coarse scales. Details will be introduced bellow.

3.1. Informative and Dynamic Adaptation (IDA)

To simulate the drawing methodology of human artist, we first propose a novel *Informative and Dynamic Adaptation (IDA)* module, to modulate deep features based on a combination of the facial depth map $\mathbf{D}$, the style map $\mathbf{S}$, and the appearance representations $\mathbf{A}$, i.e. $\{\mathbf{D}, \mathbf{S}, \mathbf{A}\}$. Specially, we propose a novel dynamic activation function, termed Informative ACON (InfoACON), and a dynamic normalization module, termed DySPADE.

Informative ACON (InfoACON). The original metaACON function automatically allows whether a neuron is active or not, based on its value, as previously presented in Eq. 1. During the drawing process, a human artist typically decides whether to draw a stroke or not based on the 3D geometry, 2D appearance, and style type. Inspired by this observation, we propose to learn the parameter θ in Eq. 1 from $\{\mathbf{D}, \mathbf{S}, \mathbf{A}\}$, i.e.

$$\theta = \sigma(\phi_\theta(\mathrm{Cat}(\mathbf{D}, \mathbf{S}, \mathbf{A})), \tag{2}$$

where ϕ_θ is a two-layer Convolutional network (Fig. 3). We refer to the modified metaACON function as *Informative ACON* (InfoACON). In our networks, we apply this InfoACON function in all the decoding layers. In this way, the decoder would pixel-wisely decides whether to depict a stroke, or the type of a stroke, in a generated sketch.

Dynamic Normalization (DySPADE). Following [27], we additionally transform neuron activations by shifting

the mean values and scaling the standard deviations, in the instance-wise and channel-wise manner [33]. Different from the original SPADE, we use dynamic activation here to introduce more flexibility on the learned modulating parameters. Let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denote the input features of the current DySPADE module. H, W, and C are the height, width and the number of channels. The activation value at site (c, h, w) is modulated as:

$$\tilde{f}_{c,h,w} = \gamma_{c,h,w}(\mathbf{D}, \mathbf{S}, \mathbf{A}) \frac{f_{c,h,w} - \mu_c}{\sigma_c} + \beta_{c,h,w}(\mathbf{D}, \mathbf{S}, \mathbf{A}), \tag{3}$$

where $f_{c,h,w}$ and $\tilde{f}_{c,h,w}$ are the input and modulated activation at site (c, h, w), respectively. μ_c and σ_c are the mean and standard deviation of $f_{c,h,w}$ in the c-th channel. $\gamma_{c,h,w}(\mathbf{D}, \mathbf{S}, \mathbf{A})$ and $\beta_{c,h,w}(\mathbf{D}, \mathbf{S}, \mathbf{A})$ are learned scale and bias parameters at site (c, h, w).

As shown in Fig. 2, we use a two-layer and three-branched Convolutional network to predict the parameter θ in InfoACON, and the modulating parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in DySPADE. To improve the flexibility of the adaptation block, we use metaACON (Eq. 1) [25] instead of ReLU, after the first Convolutional layer. In this way, the modulating factors would pixel-wisely adapt to an integration of the facial 3D geometry, 2D appearance, and global style.

In IDA, the activation at each position is modulated according to a joint consideration of local facial 3D geometry, appearance, and artistic style. This mechanism is consistent with the drawing methodology of human artists. To execute a facial sketch, an artist usually uses diverse textures to represent 3D geometry or illustration variations. Besides, the style of all pencil strokes are consistent inside a single sketch. As a result, IDA is promising to produce realistic sketchy textures in globally consistent style.

3.2. Deformable Outline Generation (DOG)

Human artists usually draw abstract lines to capture facial geometric structures, such as the boundaries of facial organs, and facial mood. To this end, the resulting outlines typically convey such structures abstractly, instead of pixel-wisely tracing them. In other words, there are geometric deformations between the input photo and the sketches drawn by artists. To simulate such an abstraction drawing methodology, we propose to align decoding features at coarse scales. In this way, the generated sketches would present abstract and distinct outlines, instead of scattered outlines with a lot of subtle variations.

In practice, we use deformable convolution (DCN) [7] instead of standard Transposed Convolution over the first and second decoding layers. As will be presented in the ablation study (Section 4.5), this deformable outline generation (DOG) module significantly boosts the clarity of generated outlines. Besides, DOG enables the network produce abstract sketches (e.g. Style1 in the FS2K dataset), which contains a sparse set of sketchy line drawings.

3.3. Overall Generator Architecture

Our generator follows the U-Net architecture [18] in whole. In the encoder, the facial photo $\mathbf{X}$ and the depth map $\mathbf{D}$ are first fed into a Convolutional layer, separately. Afterwards, the corresponding feature maps are concatenated and fed into the following encoding layers. Each encoding layer follows a Conv-metaReLU-IN architecture, and down-samples the size of feature maps by 1/2. The encoding features are adopted as appearance representations, $\mathbf{A}$.

In the decoder, we expand an DySPADE block to every decoding layer, except the last one. Fig. 2 illustrates the pipeline of a decoding layer with DySPADE. Over the l-th decoding layer, let $\mathbf{D}^l$ be the corresponding depth map, $\mathbf{S}^l$ the style map, and $\mathbf{A}^l$ the appearance features. We down-sample the original depth map $\mathbf{D}$ to $\mathbf{D}^l$ by building a Gaussian Pyramid, and expand the one-hot style vector $\mathbf{s}$ to $\mathbf{S}^l$. Besides, we obtain $\mathbf{A}^l$ by upsampling $\mathbf{E}^{l-1}$ through a Transposed-Convlotional (TrConv) layer, followed by a metaACON activation layer. We finally apply the residual connection to obtain the output of the l-th decoding layer:

$$\mathbf{K}^l = \mathbf{H}^l \oplus \tilde{\mathbf{F}}^l, \text{ with } \tilde{\mathbf{F}}^l = \text{DySPADE}(\mathbf{F}^l), \tag{4}$$

where $\oplus$ denotes element-wise addition. $\mathbf{H}^l$ is the initial upsampled feature map, output by a DOG layer (over the 1^{st} and 2^{nd} decoding layer) or a TrConv layer (over the rest layers). $\mathbf{K}^l$ is fed into subsequent layers for generating final predictions.

3.4. Loss Functions

To train our model, we use the following loss functions.

Geometric loss. First, we use a geometric constraint to supervise depth reconstructions from features of sketches. The geometric loss is the L2 distance between the input depth map and the reconstructed one:

$$\mathcal{L}_{geo} = \|\hat{\mathbf{D}} - \mathbf{D}\|_2^2. \tag{5}$$

Textural loss. The synthesized sketch $\hat{\mathbf{Y}}_s$ should present similar textures as that drawn by an artist $\mathbf{Y}_s$. In this work, we constrain $\hat{\mathbf{Y}}_s$ and $\mathbf{Y}_s$ to have similar pixel-wise adjacent correlations [21]. To this end, we calculate their gradients by using the Sobel operator, and calculate the average Cosine distance between them. Let $\mathbf{g}_{i,j} = [g^x_{i,j}, g^y_{i,j}]^T$ denote the x-directional and y-directional gradients of $\mathbf{Y}_s$ at site (i, j); and $\mathbf{f}_{i,j} = [f^x_{i,j}, f^y_{i,j}]^T$ the corresponding gradients in $\hat{\mathbf{Y}}_s$. The textural loss is formulated as:

$$\mathcal{L}_{tex} = \frac{1}{MN} \sum_{i,j} \frac{\mathbf{g}^T_{i,j}\mathbf{f}_{i,j}}{\|\mathbf{g}_{i,j}\| \cdot \|\mathbf{f}_{i,j}\|}, \tag{6}$$

where $\|\cdot\|$ denotes the magnitude of a vector, M and N are the width and height of the sketch.

Pixel loss. In addition, we use the pixel-wise reconstruction loss between the synthesized sketch $\hat{\mathbf{Y}}_s$ and the target sketch $\mathbf{Y}_s$, i.e.

$$\mathcal{L}_{pix} = \|\hat{\mathbf{Y}}_s - \mathbf{Y}_s\|_1. \quad (7)$$

Adversarial loss. Finally, we use adversarial loss to measure whether a pair of depth map and synthesized sketch is real or fake. Here, we use the Cross Entropy loss, i.e.

$$\mathcal{L}_{adv} = -\log D(\mathbf{D}, \mathbf{Y}_s) - \log(1 - D(\hat{\mathbf{D}}, \hat{\mathbf{Y}}_s)). \quad (8)$$

Full objective. We use a combination of all the aforementioned losses as our full objective:

$$\mathcal{L}_{all} = \mathcal{L}_{adv} + \lambda_1 \mathcal{L}_{pix} + \lambda_2 \mathcal{L}_{tex} + \lambda_3 \mathcal{L}_{geo}, \quad (9)$$

where λ_1, λ_2, and λ_3 are weighting factors. We train the generator G and the discriminator D in an alternative manner, to minimize $\mathcal{L}_{all}$.

4. Experiments

We present a thorough experimental comparison on the challenging FS2K dataset [8]. Besides, we conduct a series of ablation study to analyse impacts of the proposed DySPADE, InfoACON, and DOG modules.

4.1. Experimental Settings

Data. We conduct experiments on the challenging FS2K dataset. The FS2K dataset is the largest publicly released FSS dataset, consisting of 2,104 photo-sketch pairs from a wide range of image backgrounds, skin colors, sketch styles, and lighting conditions. These sketches are mainly in three styles. Following standard settings [8], we have 1,058 photo-sketch pairs for training, and 1,046 pairs for testing. For each style, we have 357/351/350 training pairs, and 619/381/46 testing pairs, from Style1 to Style3, respectively. All the images are aligned and resized to 250×250. In the inference stage, we use the same style of sketch as the ground truth in default. In addition, we collect a number of challenging Faces in-the-wild from FFHQ [20] and Web. We align and resize these images in the same way as those in the FS2K dataset.

Comparison Methods In this section, we compare our method with various state-of-the-art (SOTA) ones, including FSGAN [8], GENRE [21], SCA-GAN [37], and MDAL [44]. Besides, we compare with several advanced GANs, including CycleGAN [46], Pix2Pix [18], and Pix2PixHD [35]. We use results and codes of these methods released by the corresponding authors [8]. All these methods and ours follow the same experimental settings.

Criteria In this work, we choose four performance indices as the criteria, i.e. the *Fréchet Inception distance* (FID) [15], *Learned Perceptual Image Patch Similarity* (LPIPS) metric [42], *Structure Co-Occurrence Texture* (SCOOT) metric [10], and *Feature Similarity Measure* (FSIM) [39]. Lower values of FID and LPIPS indicate higher realism of synthesized sketches. In contrast, greater values of SCOOT and FSIM generally indicate higher similarity between a synthesized sketch and the corresponding sketch drawn by an artist. We here report the average LPIPS, SCOOT, and FSIM values across all the test samples, respectively. In the following sections, $\downarrow$ indicates that lower value is better, while $\uparrow$ higher is better.

Implementation Details We implemented our model in PyTorch. All experiments are performed on a computer with a Titan 3090 GPU. We use a batch size of 4, a learning rate of $1e-4$. We use the Adam Optimizer, and train the model for 800 epochs on the training set. Our code will be released after peer review.

4.2. Qualitative Comparison with SOTAs

We further qualitatively compare with SOTA FSS methods. Fig. 3 illustrates synthesized sketches on the FS2K dataset. Although our method can generate multiple styles of sketches, here we only show the synthesized sketch in the same style as the ground truth. For the face in constrained condition (the first row), most methods successfully generate a quality sketch. For the face with extreme lighting condition (the second row) or pose variation (the third row), most synthesized sketches present unpleasant geometric deformations and fail to precisely reproduce the style. Although sketches generated by CycleGAN seems acceptable, the textures aren't like pencil-drawings. The sketches generated by FSGAN show the same styles as the ground truths, since FSGAN contains a style control module. However, these sketches show unpleasant structural distortions. This might be caused by the geometric deformations between facial photos and free-hand sketches drawn by artists, in the training data. GENRE successfully produces quality sketches, but they are all almost in the same style, since no style information is considered in GENRE.

In contrast, our HIDA generates high-quality sketches in all three styles. Specially, our synthesized sketches preserve the geometries of input faces. This implies that, HIDA doesn't overfit to the training samples and combats geometric deformations. We achieve such success mainly due to the informatively adaptive normalization module, i.e. DySPADE, and the constraint of reconstructing the input depth. Besides, our synthesized sketches present the same style of strokes as the corresponding ground truths. The drawing textures are consistent inside each sketch. The style consistency demonstrates the effectiveness of our global style control mechanism through DySPADE and InfoACON. Based

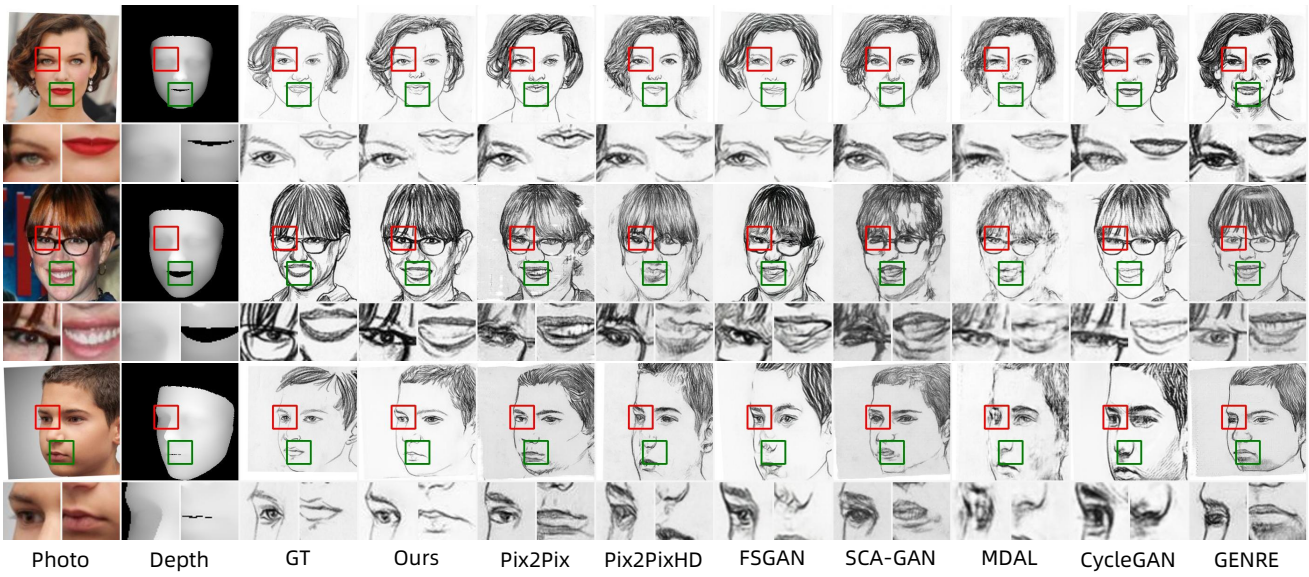


Figure 3: Comparison with SOTAs on the FS2K dataset.

Table 1: Comparison with SOTAs on the FS2K dataset.

	FID↓	LPIPS↓	SCOOT↑	FSIM↑
Pix2Pix [18]	<u>18.34</u>	0.304	0.493	0.541
Pix2PixHD [35]	32.03	0.468	0.374	0.531
CycleGAN [46]	26.49	0.505	0.348	0.501
MDAL [44]	50.18	0.492	0.355	0.530
SCA-GAN [37]	39.63	0.305	**0.600**	**0.782**
FSGAN [8]	34.88	0.483	0.405	<u>0.610</u>
GENRE [27]	20.67	<u>0.302</u>	0.483	0.534
HIDA (Ours)	**15.06**	**0.263**	<u>0.575</u>	0.551

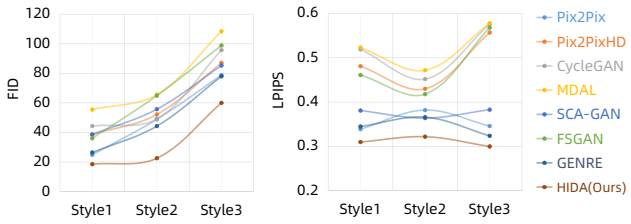


Figure 4: FID and LPIPS values w.r.t. each style in FS2K.

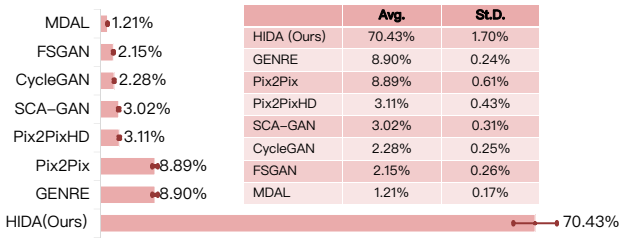
on all these observations, we conclude that our HIDA model can generate high-quality and style-consistent sketches.

4.3. Quantitative Comparison with SOTAs

Overall Performance. Table 1 shows the quantitative performance criteria of each method on the whole FS2K testing dataset. Obviously, our method achieves the lowest FID and LPIPS values. In contrast to previous benchmark method, FSGAN, our HIDA dramatically decrease both FID and LPIPS by about 20 and 0.22, respectively. Besides, compared to SOTA 2D-semantic driven methods, i.e. SCA-GAN and GENRE, HIDA decreases FID by about 24 and 5, respectively. HIDA also decreases LPIPS by about 0.04, i.e. 10% relatively. Such dramatic decreases of both FID and LPIPS mean that our method produces the most realistic sketches in terms of style and stroke.

In addition, HIDA achieves the second best value of SCOOT, which is significantly better than FSGAN and GENRE, but slightly lower than SCA-GAN. Such a high value of SCOOT means that the sketches produced by our method are similar to those drawn by artists in terms of structure and textures. Finally, HIDA achieves the third best FSIM value. Recall that there are geometric deformations between facial photos and sketches drawn by artists. Thus an excessively high value of FSIM might indicate the potential that: a FSS model overfits to the training data, and cannot precisely preserve facial structures in the translation process. Correspondingly, as shown in Fig. 3, both SCA-GAN and FSGAN produce deformable sketches. In contrast, HIDA preserves the structure of input faces.

Performance on Each Style. We further analyse the performance of FSS methods on each style subset. Since both FID and LPIPS measure the realism of synthesized sketches in terms of style and textures, we report them in Fig. 4. Obviously, our HIDA model consistently achieves the lowest FID and LPIPS values, across all the styles. Especially, our method significantly outperforms previous SOTA method, FSGAN, according to both criteria. Such distinct superiority over existing methods demonstrates that



	Avg.	St.D.
HIDA (Ours)	70.43%	1.70%
GENRE	8.90%	0.24%
Pix2Pix	8.89%	0.61%
Pix2PixHD	3.11%	0.43%
SCA–GAN	3.02%	0.31%
CycleGAN	2.28%	0.25%
FSGAN	2.15%	0.26%
MDAL	1.21%	0.17%

Figure 5: Average subjective preference percent (*Avg.*) and the standard deviations among different subjects (*St.D.*).

our method effectively learns the style information and allows precise control over the style of synthesized sketches.

4.4. User Study

We further conduct a series of subjective study to evaluate the performance of HIDA, in contrast to existing methods. Specially, we have 10 participators, all of whom are not professional artists. For each participator, we show them 1,000 randomly selected samples from the testing set in FS2K. Each time, we show a facial photo, the corresponding sketch drawn by an artist, and 8 synthesized sketches produced by different methods. Participators are requested to choose the best sketch, according to (1) the similarity between a synthesized sketch and the ground truth, and (2) the quality of a sketch, based on their own preferences. Finally, we collect totally 10,000 preference labels.

Fig. 5 shows the average preference percent about each model, and the standard deviation among different participants. Obviously, our method dramatically outperforms all the other methods. In average, subjective participators think our model generates the best sketch over 70% of facial photos. The subjective comparison result demonstrate that our method significantly outperforms SOTAs in generating high-quality and style-specific facial sketches. In addition, the sketches synthesized by our HIDA model meet the preference of most users.

4.5. Ablation Study

We first conduct a series of ablation study on the FS2K dataset. To this end, we build several model variants, by gradually adding different modules to the base model, i.e. Pix2Pix [18]. The modules we aims to analyse include the use of depth map $\mathbf{D}$ as auxiliary input, the DySPADE transformation, the InfoACON function, and the DOG layer.

Qualitative Analysis. Fig. 6 illustrates sketches produced by these model variants. The second column shows the depth maps predicted by 3DDFA. These maps convey well with the corresponding facial geometry in general. The third column shows sketches generated by the base model (i.e. Model-A). Obviously, these sketches occasionally show chaotic facial structures. Besides, there is

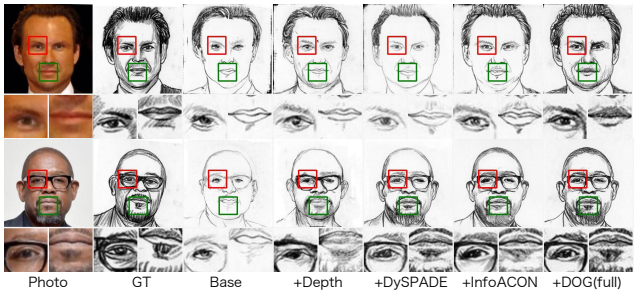


Figure 6: Comparison between different model variants in the ablation study, on the FS2K dataset.

Table 2: Quantitative results of the ablation study on the FS2K dataset.

	FID↓	LPIPS↓	SCOOT↑	FSIM↑
Base (Model-A)	18.34	0.304	0.493	0.541
+ Depth (Model-B)	20.1	0.307	0.489	0.543
+ DySPADE (Model-C)	18.30	0.298	0.479	0.539
+ InfoACON (Model-D)	17.41	0.291	0.493	0.536
+ DOG (full)	**15.06**	**0.263**	**0.575**	**0.551**

no distinct difference between the generated two sketches in terms of style. In contrast, using the DySPADE module (i.e. Model-C) enables the model precisely preserving tiny facial structures. For example, the shapes of eyebrows in both examples become consistent between the synthesized sketches and the input photos.

If we further use the InfoACON function in the decoder (i.e. Model-D), the generator produces more details. For example, the textures precisely present the 3D structure of lips. Besides, the major boundary of eyeglass is generated. The synthesized sketches of these two examples also show different types of strokes over the same semantic regions, e.g. lips. Finally, using DOG (i.e. the full model) enables the model generating abstract and distinct outlines. For example, the result in the top row is consistent with Style1 in terms of line drawings. All the other model variants produce obvious rendering textures to present 3D geometry. Such comparisons demonstrate our motivation of using DOG to simulate the abstraction process of human artists.

Quantitative Analysis. Table 2 lists the performance criteria achieved by these model variants. Using depth alone, although improves the geometrical structures (Fig. 6), doesn't consistently contribute to quantitative performance. As we gradually add the DySPADE, InfoACON, and DOG modules, both FID and LPIPS consistently decrease. At the same time, Model-C and Model-D achieve comparable SCOOT and FSIM values, in contrast to Model-A. This means that both DySPADE and InfoACON improve the realism of the generated sketch portraits, without signifi-

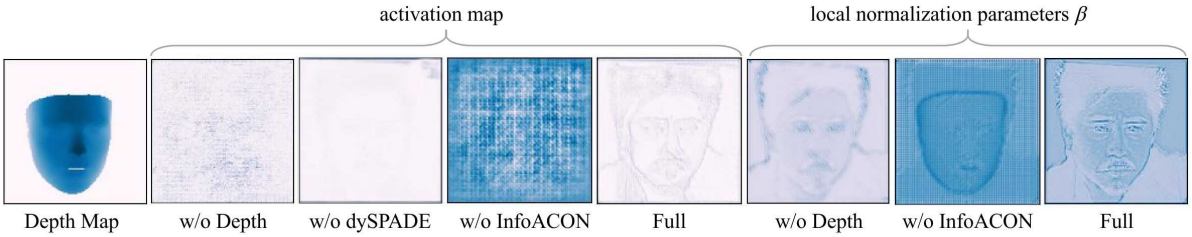


Figure 7: Visualization of activations and adaptation parameters, w.r.t. different model variants of HIDA.

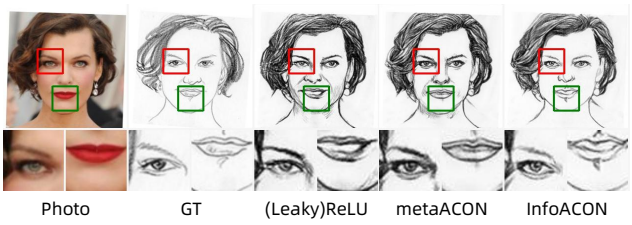


Figure 8: Comparison between activation functions.

cantly changing facial structures. Inspiringly, our full model achieves the best performance in terms of all the quantitative criteria.

Parameter Visualization. We further analyse the impact of each module by removing it from our full model. Fig. 7 visualizes activation maps and adaptation parameters w.r.t. the corresponding model variants. We can see that depth helps learning effective geometric representations (Full *vs*. w/o Depth). Besides, the proposed dynamic adaptation (DySPADE and InfoACON) boosts the representations, and migrates the artefacts introduced by the incomplete depth map. Based on previous analysis, we conclude that our method achieves such inspiring performance, due to a combination of depth and the IDA modules.

Analysis of InfoACON. We further analyze the impacts of dynamic activation functions, including metaACON and the proposed InfoACON. To this end, we build model variants based on Model-B, by (1) using ReLU in the encoder and LeakyReLU in the decoder and discriminator; (2) using metaACON [25] in all layers; and (3) using InfoACON in the decoder and metaACON in the other layers. As shown in Fig. 8, InfoACON makes the generator merely produce distinct sketchy outlines over the mouth region, which is most similar to the ground truth, in terms of style. As shown in Table 3, InfoACON achieves the lowest FID and LPIPS, as well as highly comparable SCOOT and FSIM. Besides, both metaACON and InfoACON outperform ReLU/LeakyReLU. This means that dynamic activation significantly improves the consistency between the synthesized sketches and those drawn by human artists, in terms of textures.

Analysis of DOG. In our framework, we apply deformable convolutions only at coarse-scale layers, i.e. the top 2 layers in the decoder. To verify such motivation, we conduct variants of our final model, by using DOG at top 2 layers (*top*2), middle 3 layers (*mid*3), bottom 2 layers (*btm*2), and all layers (*all*), respectively. In this experiment, HIDA w/o DOG is the base model. Fig. 9 illustrates the corresponding synthesized sketches. Obviously, the sketches synthesized with DOG present distincter geometric outlines than those without DOG. If we apply DOG over the bottom layer, the model fails to generate sketches in Style1. This might due to the fact that human painters usually abstract in large areas rather than small ones. Besides, the sketch synthesized by *top*2 has the most consistent style compared to the ground truth. Using DOG over all decoding layers leads to an integrated effects on the synthesized sketch, e.g. confused styles and distinct boundaries. We therefore merely use DOG over the top 2 decoding layers in our final model.

Table 3: Comparison between different activation functions.

	FID↓	LPIPS↓	SCOOT↑	FSIM↑
(Leaky)ReLU	<u>18.30</u>	0.298	0.479	<u>0.539</u>
metaACON	19.05	<u>0.292</u>	**0.498**	**0.541**
InfoACON	**17.41**	**0.291**	<u>0.493</u>	0.536

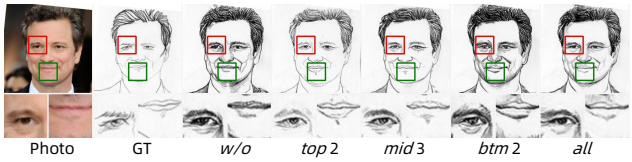


Figure 9: Comparison between different settings of DOG.

4.6. Generalization Ability

To evaluate the generalization ability of our framework, we apply the previously learned HIDA model to challenging faces in-the-wild and natural images. Here, we compare with Pix2Pix, SCA-GAN, and GENRE, because the models of MDAL and FSGAN haven't been released. All models are learned from the training set of the FS2K dataset.

Performance on Faces In-the-wild. Fig. 10 illus-

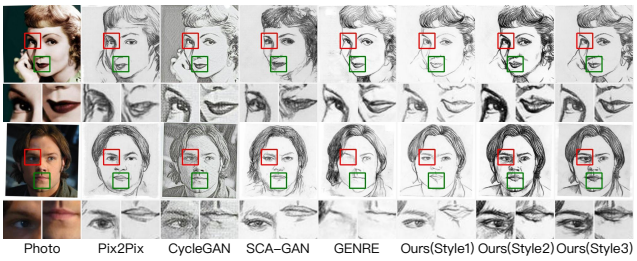


Figure 10: Performance of our method on faces in-the-wild.

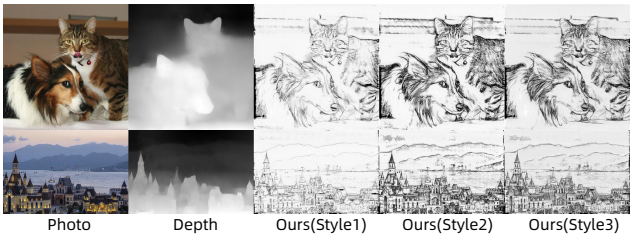


Figure 11: Performance of our model on natural images.

trates synthesized sketches on unconstrained faces. These faces have extreme variations in occlusion, pose, lighting, and tone. Generally speaking, our method produces high-quality sketches, in multiple styles, for both examples. Inspiringly, for the example shown in the bottom row, our HIDA model successfully depicts the eyes in shading areas. In contrast, the other methods fail to generate some quality details, e.g. eyes of both examples. Moreover, they produce geometric deformations over the mouth of the bottom example. Finally, our synthesized sketches vividly characterize the moods shown in the photographic faces.

Extension to Natural Images. We here apply our previously learned model to several natural images, collected from the Web. Here, we use MiDas [31] instead of 3DDFA for depth estimation. Fig. 11 shows that our model still produces high-quality sketches, in multiple styles. The synthesized sketches vividly present the geometry and appearance of natural images.

Extension to other Image-to-Image translation tasks. We additionally apply our method to pen-drawing generation (with paired data on the APDrawing dataset [36]) and exemplar-based image translation (with unpaired data on the MetFace dataset [19]). In the former task, we train and test our full model following standard settings. In the latter task, we use CoCosNet [41] as the baseline, and modify it by (1) using depth, and (2) replacing the standard SPADE modules in CoCosNet by DySPADE and InfoACON. As shown in Table 4, our method outperforms previous SOTA methods, in terms of most performance indices. Fig. 12 shows that our method generates distinct and accurate facial structures, compared to the other methods. Such results

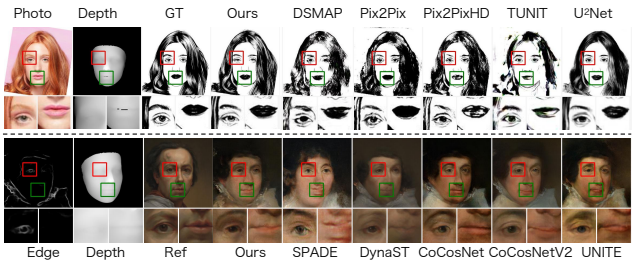


Figure 12: Generated pen-drawings and oil-paintings, on the APDrawing and MetFace datasets.

Table 4: Comparison with SOTAs on the APDrawing and MetFace datasets.

APDrawing	FID↓	LPIPS ↓	**MetaFace**	FID↓	LPIPS↓	Sem.↑
DSMAP [4]	71.38	0.466	GauGAN [27]	76.42	0.391	0.915
TUNIT [1]	91.64	0.458	DyNaST [23]	29.25	0.375	0.917
Pix2Pix [18]	80.11	0.250	Cocosnet [41]	34.14	0.355	0.930
Pix2PixHD [35]	60.55	0.206	CocosnetV2 [45]	27.98	0.296	0.939
U^2-Net [30]	77.19	0.232	UNITE [38]	35.91	0.356	0.930
HIDA (Ours)	**56.58**	**0.194**	HIDA (Ours)	**22.62**	**0.174**	**0.981**

demonstrate that the proposed techniques are robust and applicable to other image translation tasks.

5. Conclusions

In this work, we use comprehensive facial information for synthesizing sketchy portraits. Technically, we propose two informative and dynamic adaptation methods, including a normalization module and an activation function. Extensive experiments show that our method, termed HIDA, can generate high-quality and style-controllable sketches, over a wide range of challenging samples. Our work also implies promising applications of dynamic adaptation, or dynamic networks, in more image generation tasks. Besides, it is promising to boost the performance of FSS models by combining multi-source datasets. We will explore such works in the near future.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61971172, 61971339, 62176230, U22A2096, 62036007; in part by the Proof of Concept Foundation of Xidian University Hangzhou Institute of Technology under Grant No. GNYZ2023YL0301; in part by the Fundamental Research Funds for the Central Universitics under Grant QTZX23042; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; and in part by Open Research Projects of Zhejiang Lab under Grant 2021KG0AB01.

References

[1] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14154–14163, October 2021. 9

[2] Eva Cetinic and James She. Understanding and creating art with ai: Review and outlook. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(2), feb 2022. 1

[3] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7915–7925, 2022. 1

[4] Hsin-Yu Chang, Zhixiang Wang, and Yung-Yu Chuang. Domain-specific mappings for generative adversarial style transfer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 573–589. Springer, 2020. 9

[5] C Chen, W Liu, X Tan, and KKY Wong. Semi-supervised learning for face sketch synthesis in the wild. In *ACCV*, 2018. 1

[6] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic relu. In *European Conference on Computer Vision*, pages 351–367. Springer, 2020. 2

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE ICCV*, pages 764–773, 2017. 2, 4

[8] Fan Deng-Ping, Huang Ziling, Zheng Peng, Liu Hong, Qin Xuebin, and Van Gool Luc. Deep facial synthesis: A new challenge. *Machine Intelligence Research*, 19:257–287, 2022. 1, 2, 5, 6

[9] Shuchao Duan, Zhenxue Chen, QM Jonathan Wu, Lei Cai, and Dan Lu. Multi-scale gradients self-attention residual learning for face photo-sketch transformation. *IEEE Transactions on Information Forensics and Security*, 16:1218–1230, 2020. 1

[10] D. Fan, S. Zhang, Y. Wu, Y. Liu, M. Cheng, B. Ren, Paul L Rosin, and R. Ji. Scoot: A perceptual metric for facial sketches. In *ICCV*, pages 5612–5622, 2019. 5

[11] Fei Gao, Jingjie Zhu, Hanliang Jiang, Zhenxing Niu, Weidong Han, and Jun Yu. Incremental focal loss gans. *Information Processing & Management*, 57(3):102192, 2020. 1

[12] J Gui, Z Sun, Y. Wen, D. Tao, and J. Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–28, 2021. 2

[13] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[14] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7436–7456, 2021. 2

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637. 2017. 5

[16] Shihua Huang, Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 864–873, October 2021. 2

[17] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, oct 2017. 1

[18] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 1, 2, 4, 5, 6, 7, 9

[19] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 9

[20] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, June 2019. 2, 5

[21] Xiang Li, Fei Gao, and Fei Huang. High-quality face sketch synthesis via geometric normalization and regularization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1, 2, 4, 5

[22] Yijun Li, Chen Fang, Aaron Hertzmann, Eli Shechtman, and Ming-Hsuan Yang. Im2pencil: Controllable pencil illustration from photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1525–1534, 2019. 1

[23] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. Dynast: Dynamic sparse transformer for exemplar-guided image generation. In *European Conference on Computer Vision*, pages 72–90. Springer, 2022. 9

[24] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10806–10815, 2021. 2

[25] Ningning Ma, Xiangyu Zhang, Ming Liu, and Jian Sun. Activate or not: Learning customized activation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8042, 2021. 2, 4, 8

[26] Lin Nie, Lingbo Liu, Zhengtao Wu, and Wenxiong Kang. Unconstrained face sketch synthesis via perception-adaptive network and a new benchmark. *arXiv preprint arXiv:2112.01019*, 2021. 2

[27] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, June 2019. 2, 3, 6, 9

[28] C. Peng, N. Wang, J. Li, and X. Gao. Face sketch synthesis in the wild via deep patch representation-based probabilistic graphical model. *IEEE TIFS*, 15:172–183, 2020. 1

[29] Xingqun Qi, Muyi Sun, Qi Li, and Caifeng Shan. Biphasic face photo-sketch synthesis via semantic-driven generative adversarial network with graph representation learning. *arXiv preprint arXiv:2201.01592*, 2022. 2

[30] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 9

[31] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 9

[32] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, 2021. 2

[33] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. pages 4105–4113, 2017. 4

[34] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. Transductive face sketch-photo synthesis. *IEEE TNNLS*, 24(9):1364–1376, 2013. 1

[35] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. pages 8798–8807, 2018. 5, 6, 9

[36] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '19)*, pages 10743–10752, 2019. 9

[37] J. Yu, S. Shi, F. Gao, D. Tao, and Q. Huang. Towards realistic face photo-sketch synthesis via composition-aided GANs. *IEEE TCYB*, 51(9):4350–4362, 2021. 1, 2, 5, 6

[38] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021. 9

[39] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: a feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011. 5

[40] M. Zhang, R. Wang, X. Gao, J. Li, and D. Tao. Dual-transfer face sketch–photo synthesis. *TIP*, 28(2):642–657, Feb 2019. 1

[41] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2, 9

[42] R. Zhang, P. Isola, A. A Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[43] S. Zhang, R. Ji, J. Hu, Y. Gao, and Lin C.-W. Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid. In *IJCAI*, pages 1163–1169, 2018. 1

[44] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li. Face sketch synthesis by multidomain adversarial learning. *IEEE TNNLS*, 30(5):1419–1428, May 2019. 5, 6

[45] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11465–11475, 2021. 9

[46] J. Zhu, T. Park, P. Isola, and A. A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017. 1, 5, 6

[47] Mingrui Zhu, Changcheng Liang, Nannan Wang, Xiaoyu Wang, Zhifeng Li, and Xinbo Gao. A sketch-transformer network for face photo-sketch synthesis. In *International Joint Conference on Artificial Intelligence*, 2021. 1, 2

[48] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 2