AnyFace: Free-style Text-to-Face Synthesis and Manipulation

Jianxin Sun[1,2]*, Qiyao Deng[1,2]*, Qi Li[1,2]†, Muyi Sun[1], Min Ren[1,2], Zhenan Sun[1,2]
[1] Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
{jianxin.sun, dengqiyao, muyi.sun, min.ren}@cripac.ia.ac.cn, {qli, znsun}@nlpr.ia.ac.cn

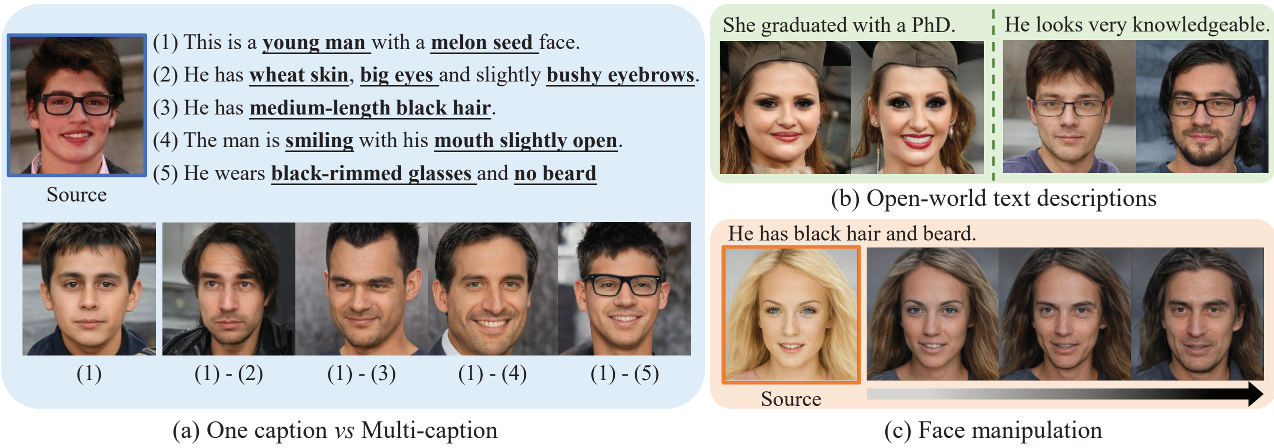

Figure 1. Our AnyFace framework can be used for real-life applications. (a) Face image synthesis with optical captions. The top left is the source face. (b) Open-world face synthesis with out-of-dataset descriptions. (c) Text-guided face manipulation with continuous control. Given source images, AnyFace can manipulate faces with continuous changes. The arrow indicates the increasing relevance to the text.

Abstract

Existing text-to-image synthesis methods generally are only applicable to words in the training dataset. However, human faces are so variable to be described with limited words. So this paper proposes the first free-style text-to-face method namely AnyFace enabling much wider open world applications such as metaverse, social media, cosmetics, forensics, etc. AnyFace has a novel two-stream framework for face image synthesis and manipulation given arbitrary descriptions of the human face. Specifically, one stream performs text-to-face generation and the other conducts face image reconstruction. Facial text and image features are extracted using the CLIP (Contrastive Language-Image Pre-training) encoders. And a collaborative Cross Modal Distillation (CMD) module is designed to align the linguistic and visual features across these two streams. Furthermore, a Diverse Triplet Loss (DT loss) is developed to model fine-grained features and improve facial diversity. Extensive experiments on Multi-modal CelebA-HQ and CelebAText-HQ demonstrate significant advantages of AnyFace over state-of-the-art methods. AnyFace can achieve high-quality, high-resolution, and high-diversity face synthesis and manipulation results without any constraints on the number and content of input captions.

*Equal contribution
†Corresponding author

1. Introduction

A picture is worth a thousand words. Human face, as one of the most important visual signal for social interactions, deserves at least 10,000 words to describe the great diversity in shape, color, hair, expression, and intention, etc. Therefore, it is highly desirable to generate variable face images from arbitrary text descriptions for increasing demands of metaverse, social media, cosmetics, advertisement, etc. This problem is firstly defined as a free-style Text-to-Face task in this paper because current technology only supports one or two facial captions in the training dataset. This paper aims to explore the first plausible solution for free-style Text-to-Face synthesis with successful applications on face image synthesis and manipulation.

Existing face image synthesis [18, 23, 24, 29, 33, 34] and manipulation [3, 14, 40] methods can synthesize impressive results with the powerful unconditional generative model (e.g., StyleGAN [10] and NVAE [27]), but vividly generating faces according to specific requirements is still a challenging problem. Thus, more and more researchers tend to explore text-to-image (T2I) synthesis [1, 5, 13, 15, 16, 28, 31, 39]. In earlier research, text embedding is directly concatenated with encoded image features and then adopted to generat semantically consistent images by learning the relationship between text and image. However, such concatenation operation can only synthesize images from a single caption, limiting the utility of these models in real-life scenarios. In practice, one caption can only describe limited information for target images while multiple captions provide fine details and accurate representation. As shown in Figure 1(a), compared with one caption, the face synthesized from 5-caption is more consistent with the source face.

To handle this issue, several methods [1, 24] attempt to implement multi-caption text-to-image synthesis. They usually use multi-stage architecture and introduce multi-caption embedding to each stage with special modules. Such caption fusion modules extract the features of multiple captions at the cost of external computational resources. And more importantly, since an image-text matching network in these methods is pre-trained on the training set, they are still failed in out-of-dataset text descriptions.

Existing text descriptions for image synthesis and manipulation are only limited to a fixed style (*e.g.* fixed number of sentences, formatted grammar, and existing words in datasets), severely limiting the user's creativity and imagination. In real-world applications, any style of text description is more in line with the user's operating needs. Free-style text descriptions include three perspectives: 1) any number of captions (i.e., one or more captions are allowed to describe an image); 2) any content of captions (allowing users to explore the content or concepts of real-world scenarios); 3) any format of captions (allowing users to describe an image in their own sentences, rather than following a fixed template).

In this paper, free-style text descriptions are explored, leveraging the power of the recently introduced Contrastive Language-Image Pre-training (CLIP) model and a free-style text-to-face method is proposed namely AnyFace for open-world applications. Due to CLIP's ability to learn visual concepts from text-guided natural language, feature extraction of out-of-dataset text descriptions is possible for open-world scenarios (see Figure 1(b)). AnyFace is a two-stream framework that utilizes the interaction between face image synthesis and faces image reconstruction. The face image synthesis stream generates target images consistent with given texts encoded by the Cross Modal Distillation Module (CMD). The face image reconstruction stream reconstructs the face images paired with the given texts. These two streams are trained independently, whose mutual information is interacted through a cross-modal transfer loss. To align the linguistic and visual features, previous attempts [29, 30] often leverage a pairwise loss to strictly narrow the distance between text and image features. We argue that there is no need to restrict the text embedding to the corresponding image features, as the relationship between text and image is a one-to-many problem. The one-to-one constraint may even produce unique results. As a remedy, a new Diverse Triplet Loss is proposed to encourage diverse text embedding along with correct visual semantics.

Overall, the key contributions of this paper are summarized as follows:

- To the best of our knowledge, it is the first definition, solution, and application of the free-style text-to-face problem, which is a breakthrough to remove the constraints in face image synthesis.
- A novel two-stream framework, which consists of a face synthesis stream and a face reconstruction stream, is proposed for accurate and high fidelity text-to-face synthesis and manipulation. Our method provides a good fusion solution of CLIP visual concepts learning and StyleGAN high-quality image generation.
- Extensive experiments are conducted to demonstrate the advantages of our method in synthesizing and manipulating high fidelity face images. Our work will definitely inspire more creative and wider applications of text-to-face technology.

2. Related Works

2.1. Text-to-Image Synthesis.

Currently, multi-stage framework has been widely used for text-to-image synthesis [1, 24, 31, 35, 36, 39]. Specifically, sentence features extracted from a pre-trained Bi-LSTM [22] network are concatenated with noise as input, and multiple generators are applied to synthesize images. However, multiple generators will consume huge computing resources. To address this problem, DFGAN [25] injects noise by affine transformation [8] and uses a single Generator to synthesize images. However, the above method can only generate low-quality images with 256×256 resolution. Recently, StyleGAN-based models [23, 28–30] have been proposed, which advance the traditional methods by a large margin in terms of image quality. These methods usually take StyleGAN [10] or its followups [9, 11] as their backbone and train a mapper to project text descriptions into the latent space. Most of them can synthesize high-quality images with resolution at 1024×1024. Among them, TediGAN-B [30] attempts to achieve face synthesis with open-world texts, but it requires to re-train a separate

Table 1. Comparison of different text-to-image synthesis methods

Methods	AttnGAN [31]	DFGAN [25]	RiFeGAN [1]	SEA-T2F [24]	CIGAN [28]	TediGAN-B [30]	**AnyFace**
Single Model	✓	✓	✓	✓	✓	-	✓
One Generator	-	✓	-	-	✓	✓	✓
Multi-caption	-	-	✓	✓	-	-	✓
High Resolution	-	-	-	-	✓	✓	✓
Manipulation	-	-	-	-	✓	✓	✓
Open-world	-	-	-	-	-	✓	✓

model for each sentence and the results are highly random, which hinders its application. Unlike them, as shown in Table 1, we can synthesize high-resolution images using a single modal and one generator with multiple and open-world text descriptions.

2.2. Text-guided Image Manipulation

Text-guided image manipulation aims at editing images with given text descriptions. Due to the domain gap between image and text, how to map text description to the image domain or learn a cross-modal language-vision representation space is the key to this task. Previous methods [4, 17] take a GAN-based encoder-decoder architecture as the backbone, and the text embedding is directly concatenated with encoded image features. Such coarse image-text concatenation is ineffective in exploiting text context, thus these methods are limited in manipulated performance. ManiGAN [14] proposes a multi-stage network with a novel text-image combination module for high-quality results. Albeit these methods have explored high-quality results, they are still limited to text descriptions on the dataset. Recently, a contrastive language-image pre-training model (CLIP) [20] is proposed to learn visual concepts from natural language. Based on CLIP's generalization performance, several methods [19, 30] achieve face manipulation with out-of-dataset text descriptions. However, these methods require retraining a new model for each given text. Different from all existing methods, we explore a general framework with a single model for arbitrary text descriptions.

3. Method

Figure 2 shows an overview of AnyFace, which mainly consists of two streams: the face image reconstruction stream and the face image synthesis stream. The training stage is comprised of two streams, while we only keep the face image synthesis stream in the inference stage. In the following, we will briefly introduce some notations used in this paper. Given a face image $\mathbf{I}$ and its corresponding caption $\mathbf{T}$, the face image synthesis stream aims to synthesize a face image $\mathbf{I_t}$, whose description is consistent with $\mathbf{T}$. The face image reconstruction stream aims to generate a face image $\hat{\mathbf{I}}$, which can reconstruct the face image $\mathbf{I}$. To overcome the mode collapse, we also introduce another arbitrary caption $\mathbf{T}'$ and propose a diverse triplet learning strategy for face image synthesis.

3.1. Two-stream Text-to-face Synthesis Network

Face Image Synthesis. As shown in Figure 2, this stream is responsible for generating a face image given the input text description $\mathbf{T}$. We first employ CLIP text encoder to extract a 512 dimensional feature vector $f_t \in \mathbb{R}^{1\times512}$ from $\mathbf{T}$. Due to the large gap between linguistic and visual modality, we utilize a transformer based network, namely Cross Modal Distillation (CMD), to embed $f_t \in \mathbb{R}^{1\times512}$ into the latent space of StyleGAN:

$$w_t = \boldsymbol{CMD}(f_t), \tag{1}$$

where $w_t \in \mathbb{R}^{18\times512}$ denotes the latent code of text features.

Previous works [19, 40] have proved that the latent space in StyleGAN has been shown to enable many disentangled and meaningful image manipulations. Instead of generating images by w_t directly, we select the last m dimensions of w_t and denote it as $w_s \in \mathbb{R}^{m\times512}$. Intuitively, w_s provides the high level attribute information of $\mathbf{I_t}$ learned from the given texts. In order to keep the diversity of the generated images, other dimensions of the latent code are harvested from noise by a pre-trained mapping network to provide the random low level topology information. We represent it as $w_c \in \mathbb{R}^{n\times512}$. Unless otherwise specified, we set $m = 14$ and $n = 4$ in this paper. Then w_s and w_c are concatenated together and sent to StyleGAN decoder to generate the final result $\mathbf{I_t}$:

$$\mathbf{I}_t = \boldsymbol{Dec}\left(w_s \oplus w_c\right), \tag{2}$$

where $\boldsymbol{Dec}$ denotes StyleGAN decoder, and $\oplus$ represents the concatenation operation. Note that different with text-to-face synthesis, in the manipulation stage, w_c is generated from the original image using an inversion model to keep the low level topology information of the original image.

Face Image Reconstruction. Text-to-face synthesis is an ill-posed problem. Given one text description, there may be multiple images that are consistent with it. Since the diversity of the generated images is crucial for text-to-face synthesis, we have to carefully design the objective functions

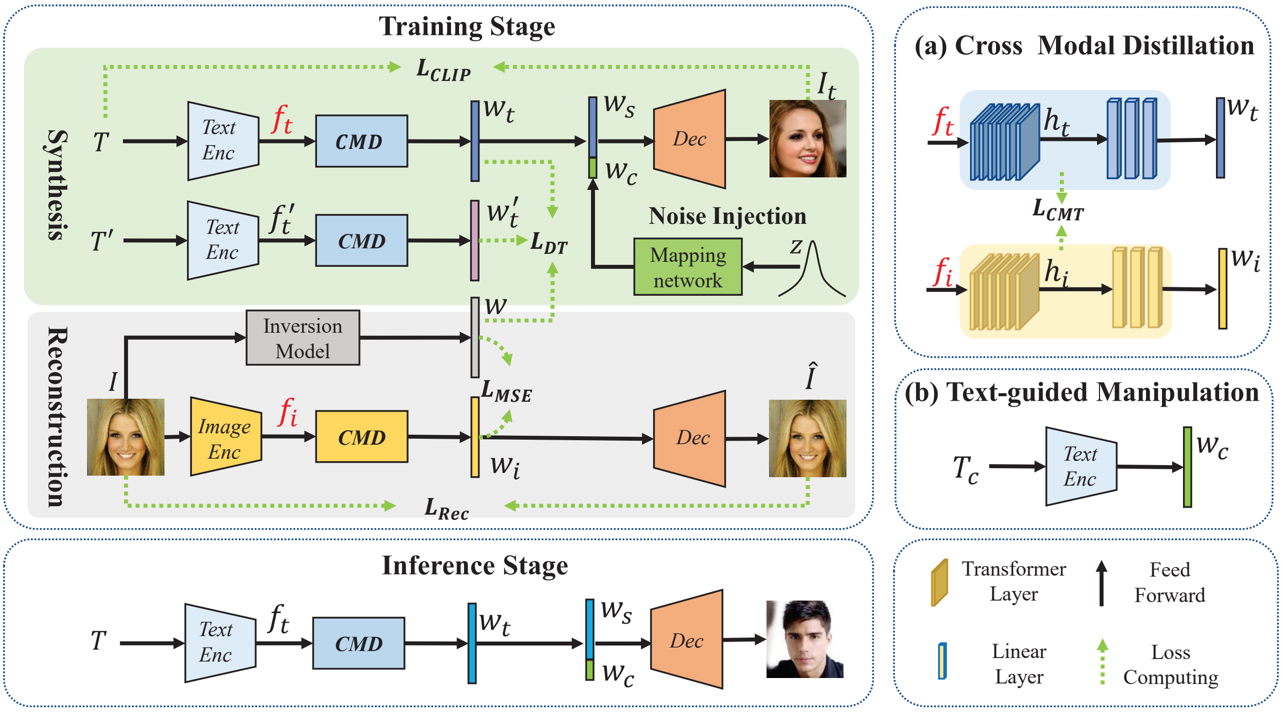

Figure 2. Overview of AnyFace. It consists of a face synthesis network and a face reconstruction network. Text Enc and Image Enc represent CLIP text and image encoder respectively. Dec is the pre-trained StyleGAN decoder. (a) Detailed architecture of Cross Modal Distillation (CMD) module. (b) For text-guided face manipulation, the content code w_c is replaced by the latent code of source image.

to regularize the face image synthesis network so that it can learn more meaningful representations. However, directly computing the pixel-wise loss between the original image $\mathbf{I}$ and the synthesis image $\mathbf{I}_t$ is infeasible. Thus we design a face reconstruction network, which performs face image reconstruction and provides visual information for the face synthesis network.

We first exploit CLIP image encoder to extract a 512 dimensional feature vector $f_i \in \mathbb{R}^{1\times512}$ from $\mathbf{I}$. As shown in Figure 2 (a), the reconstructed image can be formulated as:

$$\hat{\mathbf{I}} = \boldsymbol{Dec}\left(w_i\right), \tag{3}$$

where $w_i = \boldsymbol{CMD}\left(f_i\right) \in \mathbb{R}^{18\times512}$.

Note that both the face synthesis network and the face reconstruction network have Cross Modal Distillation module, which is important for the interaction between w_i and w_t. In the following, we will introduce $\boldsymbol{CMD}$ in detail.

3.2. Cross Modal Distillation

Previous works [7, 32, 38] have revealed that model distillation can make two different networks benefit from each other. Inspired by them, we also exploit model distillation in this paper. In order to align the hidden features of the face synthesis network and the face reconstruction network and realize the interaction of linguistic and visual information, we propose a simple but effective module named Cross Modal Distillation ($\boldsymbol{CMD}$). Recently, the transformer network has achieved impressive performance in natural language processing and computer vision area.

As shown in Figure 2 (a), $\boldsymbol{CMD}$ is also a transformer network, which consists of 6 transformer layers and 3 linear layers. Suppose that the output of the transformer layers can be represented as $h_t \in \mathbb{R}^{1\times512}$ and $h_i \in \mathbb{R}^{1\times512}$, we should align these features before projecting them into the latent space. Specifically, $\boldsymbol{CMD}$ first normalizes the hidden features by $\mathrm{softmax}$, and then learns mutual information by the cross modal transfer loss:

$$\mathcal{L}_{CMT}^{T} = D_{KL}\left(\mathrm{softmax(h_t)}||(\mathrm{softmax(h_i)}\right). \tag{4}$$

where D_{KL} denotes the Kullback Leibler(KL) Divergence. Note that $\mathcal{L}_{CMT}^{T}$ is designed for face synthesis network. We can easily deduce $\mathcal{L}_{CMT}^{I}$ for face reconstruction network.

Benefiting from the ability of transformer networks to grasp long-distance dependencies, our method can also accept multiple captions as the input without any other fancy modules. To be specific, text features extracted by CLIP text encoder are Stacked as $F_t \in \mathbb{R}^{n\times512}$, and then projected into the latent space by $\boldsymbol{CMD}$ module.

3.3. Diverse Triplet Learning

Recall that in subsection 3.1, $\boldsymbol{CMD}$ embeds the text feature f_t into the latent space of StyleGAN: w_t $=$

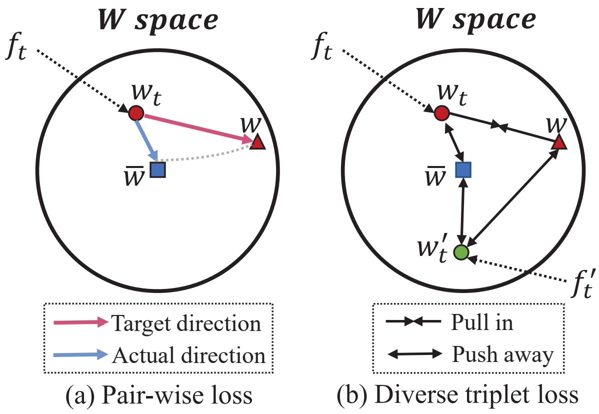

Figure 3. Comparison between (a) pairwise loss and (b) diverse triplet loss. The pairwise loss learns an one-to-one mapping from text embedding w_t to target w, but in fact it often converges to average embedding $\overline{w}$ with an unique result. In contrast, diverse triplet loss follows an one-to-many mapping without strict constraints, which encourages positive embedding w_t to be close to w and negative embedding w_t' to stay away from w. Meanwhile, both w_t and w_t' stay away from $\overline{w}$.

$\boldsymbol{CMD}(f_t)$. Idealy, $\boldsymbol{CMD}$ reduces the large gaps between the text features extracted by CLIP text encoder and the image features extracted by CLIP image encoder. However, there still exists large gaps between the latent space of CLIP and the latent space of StyleGAN. A straightforward way to bridge the gap between them is to minimize the pairwise loss:

$$\mathcal{L}_{pair} = \|w_t - w\|_p \,, \tag{5}$$

where p is a matrix norm, such as the ℓ_0-norm, ℓ_1-norm, w is the corresponding latent code of StyleGAN, which is encoded by a pre-trained inversion model [21, 26]. However, we find that there is a problem with this approach. As shown in Figure 3 (a), $\boldsymbol{CMD}$ will cheat by converging to the latent code of average face $\overline{w}$, which is possibly close to all of the latent codes in StyleGAN.

To improve the diversity and synthesize language-driven high fidelity images, we design a *diverse triplet loss*. First, the latent code of average face $\overline{w}$ is adopted to penalize the model collapse of $\boldsymbol{CMD}$. Besides, arbitrary caption $\mathbf{T}'$ is introduced to form a negative pair in face synthesis network. The constraints on the positive and the negative pairs are also taken into account:

$$\mathcal{L}_{DT} = max\left\{\frac{\langle w_t, w\rangle}{\langle w_t, \overline{w}\rangle} - \frac{\langle w_t', w\rangle}{\langle w_t', \overline{w}\rangle} + m, 0\right\}, \tag{6}$$

where $\langle\cdot,\cdot\rangle$ refers to the cosine similarity, m represents the margin, w, w_t' correspond to positive image embedding and the negative text embedding respectively.

As shown in Figure 3 (b), diverse triplet loss urges positive samples to approach the anchor and negative samples to stay away from the anchor while at the same time encouraging positive and negative samples to diverge and fine-grained features.

3.4. Objective Functions

Face Image Synthesis. The objective function of the face synthesis stream can be formulated as:

$$\mathcal{L}_S = \mathcal{L}_{DT} + \lambda_{CMT}\mathcal{L}^T_{CMT} + \lambda_{CLIP}\mathcal{L}_{CLIP}, \tag{7}$$

where λ_{CLIP} is the coefficient of the CLIP loss which is used to ensure the semantic consistency between the generated image and the input text:

$$\mathcal{L}_{CLIP} = \frac{f_t \cdot f_i^t}{||f_t|| \times ||f_i^t||}, \tag{8}$$

f_t and f_i^t are features of input text $\mathbf{T}$ and target image $\hat{\mathbf{I}}$.

Face Image Reconstruction. The objective function of face image reconstruction stream can be defined as:

$$\mathcal{L}_T = \mathcal{L}_{MSE} + \lambda_{CMT}\mathcal{L}^I_{CMT} + \lambda_{Rec}\mathcal{L}_{Rec}, \tag{9}$$

where $\mathcal{L}_{Rec}$ is employed to to encourage the generated image $\hat{I}$ to be consistent with the input image I, $\mathcal{L}_{MSE}$ is adopted to project the image feature into the latent space of StyleGAN, λ_{Rec} is the coefficient for reconstruction loss, λ_{CMT} is the coefficient for cross modal transfer loss. $\mathcal{L}_{Rec}$ and $\mathcal{L}_{MSE}$ are defined as:

$$\mathcal{L}_{Rec} = ||\hat{\mathbf{I}} - \mathbf{I}||_1, \tag{10}$$

$$\mathcal{L}_{MSE} = ||w_i - w||_2^2. \tag{11}$$

4. Experiments

Dataset. We conduct experiments on Multi-Modal CelebA-HQ [29] and CelebAText-HQ [24] to verify the effectiveness of our method. The former has $30,000$ face images with text descriptions synthesized by attribute labels, and the latter contains $15,010$ face images with manually annotated text descriptions. All of the face images come from CelebA-HQ [12] and each image has ten different text descriptions. We follow the setup of SEA-T2F [24] to split the training and testing set.

Metric. We evaluate the results of text-to-face synthesis from two aspects: image quality and semantic consistency. Image quality includes reality and diversity, which is evaluated by Fréchet Inception Distance (FID) [6] and Learned Perceptual Image Patch Similarity (LPIPS) [37], respectively. As for semantic consistency, R-precision [31] is used

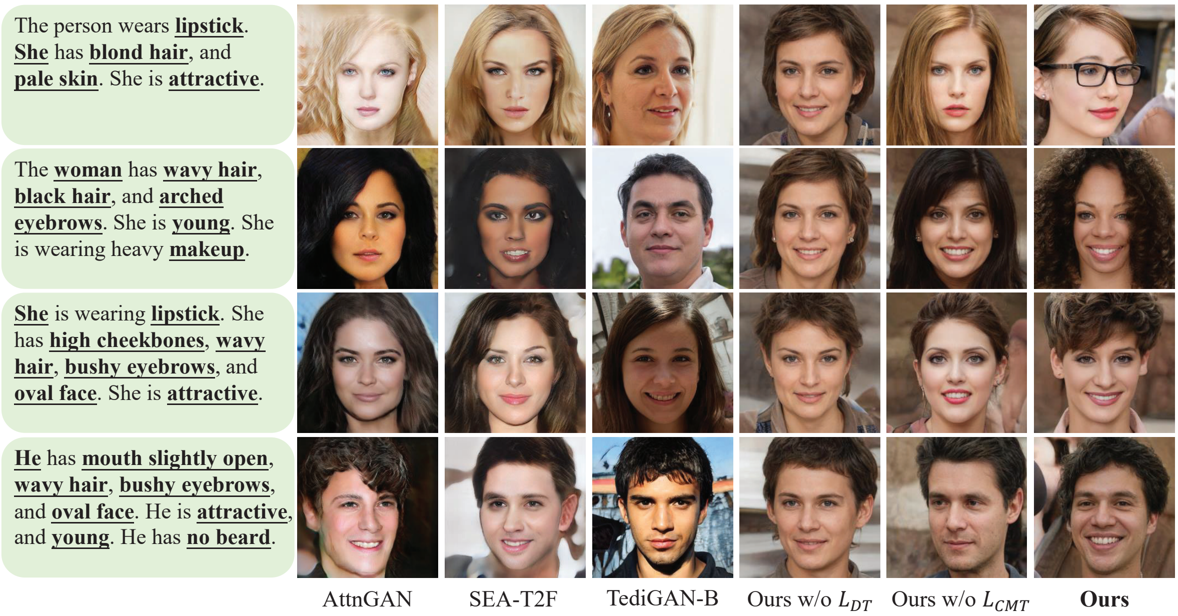

Figure 4. Qualitative comparisons with state-of-the-art methods. The leftmost is the input sentences, columns from left to right represent the results of AttnGAN, SEA-T2F, TediGNA-B, AnyFace without $\mathcal{L}_{DT}$, AnyFace without $\mathcal{L}_{CMT}$ and our AnyFace respectively.

Table 2. Quantitative comparison of different methods on Multi-modal CelebA-HQ dataset. ↓ means the lower the better while ↑ means the opposite.

Methods	FID ↓	LPIPS ↓	RFRR ↑
AttnGAN	54.22	0.548	22.15%
ControlGAN	77.41	0.572	24.5%
SEA-T2F	96.55	0.545	24.3%
AnyFace	**50.56**	**0.446**	**29.05%**

Table 3. Quantitative comparison of different methods on CelebAText-HQ dataset.

Methods	FID ↓	LPIPS ↓	RFRR ↑
AttnGAN	70.47	0.524	26.54%
ControlGAN	91.92	0.512	25.85%
SEA-T2F	140.25	0.502	18.31%
AnyFace	**56.75**	**0.431**	**29.31%**

to evaluate traditional T2I methods. An image-text matching network pre-trained on the training set is utilized to retrieval texts for each target image and then calculate the matching rate. One problem of this evaluation metric is that the performance of the pre-trained model is not accurate due to the limited training samples. Furthermore, there are no pre-trained networks that can match the face and text. Text-to-face synthesis aims to generate the face image based on the given text description, and we expect the synthesized result to be similar to the original one. Thus, we propose to use the Relative Face Recognition Rate (RFRR) to measure the semantic consistency. Specifically, given the same text description, we use ArcFace [2] to extract features from images generated by all methods, and calculate the cosine similarity between these features and the feature of the original image. Then we choose the method with the maximum cosine similarity as the successfully matched one.

4.1. Quantitative Results

We compare our AnyFace with previous start-of-the-art models on CelebAText-HQ and Multi-modal CelebA-HQ dataset. As shown in Table 2, on Multi-Modal CelebA-HQ dataset, AnyFace outperforms the current best method with an improvement of 3.66 FID, 0.099 LPIPS, and 4.55% RFRR points. More impressively, AnyFace reduces the best reported FID on the CelebAText-HQ dataset from 70.49 to 56.75, a 19.5% reduction relatively. The CelebAText-HQ is much more challenging because it is manually manuated, but we can also synthesize high reality, diverse and accurate face images given the text description.

4.2. Qualitative Results

In this subsection, we compare our results with other state-of-the-art methods qualitatively. Moreover, we present the generalization of the proposed method in real-life applications, including open-world, multi-caption and text-guided face manipulation scenarios.

4.2.1 Qualitative Comparisons

Both image quality and semantic consistency should be considered when evaluating the results of different T2I methods qualitatively. As shown in Figure 4, though multi-

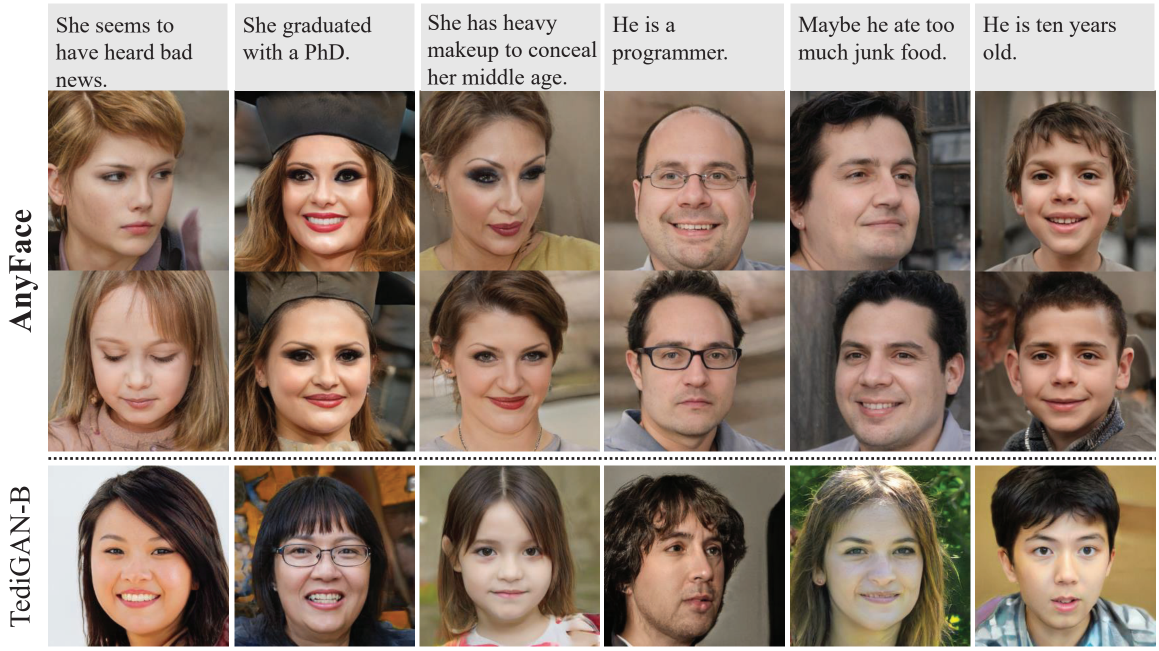

Figure 5. Illustrations of open-world text-to-face synthesis. The first row represents the text descriptions harvested from the Internet, and the rest of column are generated results with given text.

stage method (AttnGAN [31] and SEA-T2F [24]) can synthesize images corresponding to part of the text descriptions, they can only generate low quality images with the resolution at 256×256. While TediGAN-B [30] successfully synthesizes images with high quality. However, the results seem to be text-irrelevant. For example, in the second row, it generates a male image while it should be a female according to the text description. Our method introduces $\mathcal{CMD}$ module and Diverse Triplet loss to help the network to learn some fine-grained features such as beard, smile, makeup, mouth type, hair color and shape, etc., which illustrates that the images synthesized by our method are much better in terms of image quality and semantic consistency.

4.2.2 Real-life Applications

Open-world Scenarios. In this subsection, we compare with TediGAN-B [30] for open-world scenarios. TediGAN-B adopts instance level optimization for open-world texts. Thus it requires retraining the network for each text description, which is tedious and time-consuming. As for our method, a single model is trained on one dataset and can be easily extend to unseen text descriptions.

As shown in Figure 5, TediGAN-B can generate photo-realistic results. However, these results are almost text-irrelevant. While our method can not only synthesize vivid images which are text-relevant, but can generate images with more diversity due to the sophisticated design of the face synthesis network. Besides, benefiting from the powerful representation of CLIP, text descriptions with similar meanings will be encoded into similar latent codes of StyleGAN even though the text descriptions contains abstract information. Overall, AnyFace learns visual concepts from text descriptions. For example, "bad news" produces the expression “sad”, “PhD” corresponds to “mortarboard”, and “programmer” is related to “bald” and “glasses”.

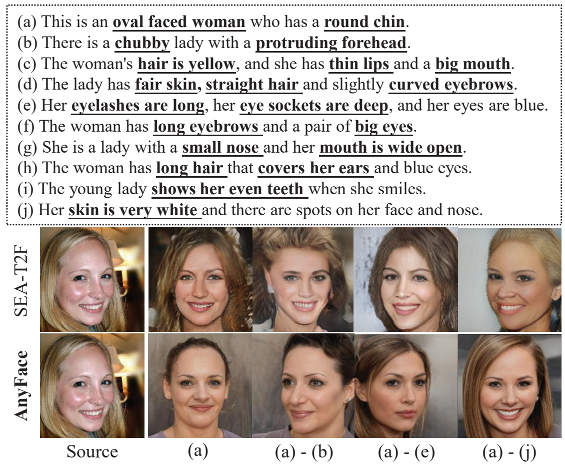

Figure 6. Illustrations of multi-caption synthesis. The text description above corresponds to the source image, synthesized results are conditioned by (a), (a-b), (a-e) and all captions, respectively.

Multi-caption Scenarios. In this subsection, we compare with SEA-T2F [24] for multi-caption scenarios. As

Figure 7. Illustrations of text-guided face manipulation with continuous control. Given source images (1st column), manipulated images show continuous changes according to the text. The arrow indicates the increasing relevance to the text

shown in Figure 6, AnyFace produces natural and smooth appearance without obvious ghosting artifacts and color distortions. As the number of sentences increases, the difficulty of matching all texts correctly also increases. Compared with SEA-T2F, the results of AnyFace are more consistent with the text descriptions (e.g., "long hair" and "curves her ears") . Note that due to the subjectivity of the labeling data samples in [24], even the source images may not perfectly match all the descriptions.

Text-guided Face Manipulation. In this subsection, we show that AnyFace can also be easily adapted to manipulate face images continuously given the text description by changing the size of text embedding (i.e., changing m and n in subsecton 3.1). From Figure 7, we observe that AnyFace can manipulate faces with any text descriptions, including global (e.g., smiling), local (e.g., brown hair) and abstract (e.g., middle-aged) attributes. Meanwhile, as the text information increases, the generated face will be more consistent with the text information.

4.2.3 Ablation Study.

In this part, we conduct ablation studies to evaluate the contributions of diverse triplet loss and cross-modal transfer loss to our framework. As shown in Figure 4 ("w/o $\mathcal{L}_{DT}$ "), we replace the diverse triplet loss with the pairwise loss. Synthesized faces tend to have similar appearance, such as smile, short hair and natural skin. In contrast, AnyFace with $\mathcal{L}_{DT}$ produces diverse and personalized characteristics, e.g., different hairstyles, accessories, skin tones, etc.

We further present the impact of the cross-modal transfer loss on our framework by comparing the qualitative results with or without $\mathcal{L}_{CMT}$ in Figure 4. It is difficult for "AnyFace w/o $\mathcal{L}_{CMT}$ " to learn structured features. For example, faces of the second and third rows fail to synthesize " wavy hair", and the synthesized mouth in the last row is not consistent to part of the text description: "mouth slightly open". The effect of $\mathcal{L}_{CMT}$ is further explored from a quantitative perspective. In Figure 8, we demonstrate that FID changes with or without $\mathcal{L}_{CMT}$ as training steps increases. We observe that the solid line ("w $\mathcal{L}_{CMT}$") improves FID by a large margin and speeds up convergence on all of the datasets. Both quantitative and qualitative assessments demonstrate the effectiveness of proposed losses.

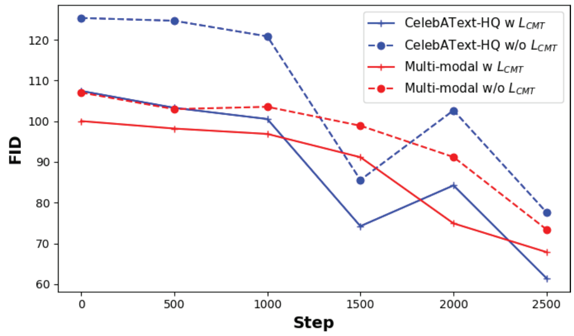

Figure 8. FID of AnyFace with or without $\mathcal{L}_{CMT}$ in different steps.

5. Limitation

Our method has three limitations. 1) Our method cannot infer the identity information from the text descriptions, such as 'Donald Trump'; 2) The results generated by the same text description have a similar style; 3) Sometimes, the attributes that are irrelevant to the text will also be changed during manipulation process.

6. Conclusion

In this paper, we propose a two-stream framework for text-to-face synthesis, which consists of a face synthesis stream and a face reconstruction stream. A Cross Modal Distillation module is further introduced to align the information between the two streams and a diverse triplet loss helps the network to produce images with diverse and fine-grained face components. Furthermore, our method can be applied to real-world scenarios such as open-world or multiple caption text-guided face synthesis and manipulation.

Acknowledgements: This work was supported in part by the National Key Research and Development Program of China under Grant No. 2020AAA0140002, in part by the Natural Science Foundation of China (Grant Nos. U1836217, 62076240).

References

[1] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10911–10920, 2020. 2, 3

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 6

[3] Qiyao Deng, Jie Cao, Yunfan Liu, Zhenhua Chai, Qi Li, and Zhenan Sun. Reference guided face component editing. In *Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 502–508, 2020. 2

[4] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5706–5714, 2017. 3

[5] Fangda Han, Ricardo Guerrero, and Vladimir Pavlovic. CookGAN: Meal image synthesis from ingredients. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1450–1458, 2020. 2

[6] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 5

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2

[9] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020. 2

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 2

[12] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[13] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2065–2075, 2019. 2

[14] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7880–7889, 2020. 2, 3

[15] Jiadong Liang, Wenjie Pei, and Feng Lu. CPGAN: Content-parsing generative adversarial networks for text-to-image synthesis. In *European Conference on Computer Vision (ECCV)*, pages 491–508, 2020. 2

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[17] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 42—-51, 2018. 3

[18] Osaid Rehman Nasir, Shailesh Kumar Jha, Manraj Singh Grover, Yi Yu, Ajit Kumar, and Rajiv Ratn Shah. Text2FaceGAN: Face generation from fine grained textual descriptions. In *IEEE Fifth International Conference on Multimedia Big Data (Big MM)*, pages 58–67. IEEE, 2019. 1

[19] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 3

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3

[21] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in Style: A StyleGAN encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 5

[22] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing (TSP)*, 45(11):2673–2681, 1997. 2

[23] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje ter Hoeve. Conditional image generation and manipulation for user-specified content. *arXiv preprint arXiv:2005.04909*, 2020. 1, 2

[24] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. Multi-caption text-to-face synthesis: Dataset and algorithm. In *ACM International Conference on Multimedia (ACM MM)*, pages 2290–2298, 2021. 1, 2, 3, 5, 7, 8

[25] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Fei Wu, and Xiao-Yuan Jing. DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 2, 3

[26] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 5

[27] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020. 2

[28] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. Cycle-consistent inverse GAN for text-to-image synthesis. In *ACM International Conference on Multimedia (ACM MM)*, pages 630–638, 2021. 2, 3

[29] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. TediGAN: Text-guided diverse face image generation and manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5

[30] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *arXiv preprint arXiv:2104.08910*, 2021. 2, 3, 7

[31] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324, 2018. 2, 3, 5, 7

[32] Yifang Yin, Harsh Shrivastava, Ying Zhang, Zhenguang Liu, Rajiv Ratn Shah, and Roger Zimmermann. Enhanced audio tagging via multi-to single-modal teacher-student mutual learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 10709–10717, 2021. 4

[33] Junchi Yu, Jie Cao, Yi Li, Xiaofei Jia, and Ran He. Pose-preserving cross spectral face hallucination. In *IJCAI*, pages 1018–1024, 2019. 1

[34] Junchi Yu, Tingyang Xu, Yu Rong, Junzhou Huang, and Ran He. Structure-aware conditional variational auto-encoder for constrained molecule optimization. *Pattern Recognition*, 126:108581, 2022. 1

[35] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5907–5915, 2017. 2

[36] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(8):1947–1962, 2018. 2

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[38] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4320–4328, 2018. 4

[39] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5810, 2019. 2

[40] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4834–4844, 2021. 2, 3