TediGAN: Text-Guided Diverse Face Image Generation and Manipulation

Weihao Xia[1] Yujiu Yang[1*] Jing-Hao Xue[2] Baoyuan Wu[3,4†]
weihaox@outlook.com yang.yujiu@sz.tsinghua.edu.cn jinghao.xue@ucl.ac.uk wubaoyuan@cuhk.edu.cn
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2]Department of Statistical Science, University College London, UK
[3]School of Data Science, Chinese University of Hongkong, Shenzhen, China
[4]Secure Computing Lab of Big Data, Shenzhen Research Institute of Big Data, Shenzhen, China

Abstract

In this work, we propose TediGAN, a novel framework for multi-modal image generation and manipulation with textual descriptions. The proposed method consists of three components: StyleGAN inversion module, visual-linguistic similarity learning, and instance-level optimization. The inversion module maps real images to the latent space of a well-trained StyleGAN. The visual-linguistic similarity learns the text-image matching by mapping the image and text into a common embedding space. The instance-level optimization is for identity preservation in manipulation. Our model can produce diverse and high-quality images with an unprecedented resolution at 1024^2. Using a control mechanism based on style-mixing, our TediGAN inherently supports image synthesis with multi-modal inputs, such as sketches or semantic labels, with or without instance guidance. To facilitate text-guided multi-modal synthesis, we propose the Multi-Modal CelebA-HQ, a large-scale dataset consisting of real face images and corresponding semantic segmentation map, sketch, and textual descriptions. Extensive experiments on the introduced dataset demonstrate the superior performance of our proposed method. Code and data are available at `https://github.com/weihaox/TediGAN`.

1. Introduction

How to create or edit an image of the desired content without tedious manual operations is a difficult but meaningful task. To make image generation and manipulation more readily and user-friendly, recent studies have been focusing on the image synthesis conditioned on a variety of guidance, such as sketch [9, 37], semantic label [11, 36], or textual description [26, 39]. Despite the success of its label and sketch counterparts, most state-of-the-art text-guided image generation and manipulation methods are only able to produce low-quality images [28, 8]. Those aiming at generating high-quality images from texts typically design a multi-stage architecture and train their models in a progressive manner. To be more specific, there are usually three stages in the main module, and each stage contains a generator and a discriminator. Three stages are trained at the same time, and progressively generate images of three different scales, *i.e.*, $64^2 \rightarrow 128^2 \rightarrow 256^2$. The initial image with rough shape and color would be refined to a high-resolution one. However, the multi-stage training process is time-consuming and cumbersome, making the afore-

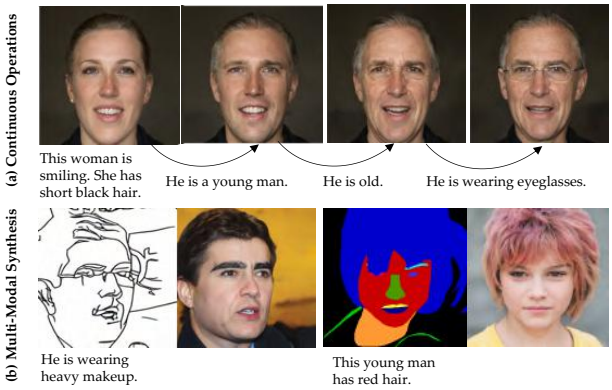


Figure 1. Our TediGAN is the first method that unifies text-guided image generation and manipulation into one same framework, leading to naturally continuous operations from generation to manipulation (a), and inherently supports image synthesis with multi-modal inputs (b), such as sketches or semantic labels with or without instance (texts or real images) guidance.

*Yujiu Yang is the corresponding author. This research was partially supported by the Key Program of National Natural Science Foundation of China under Grant No. U1903213 and the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011387).

†Baoyuan Wu is supported by the Natural Science Foundation of China under Grant No. 62076213, the University Development Fund of the Chinese University of Hong Kong, Shenzhen under Grant No. 01001810, and the Special Project Fund of Shenzhen Research Institute of Big Data under grant No. T00120210003.

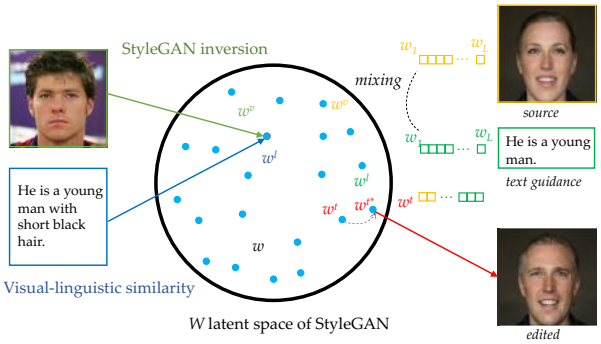


Figure 2. Projecting Multi-Modal Embedding into the $\mathcal{W}$ Space of StyleGAN. Taking visual and linguistic embedding for example, the left illustrates visual-linguistic similarity learning, where the visual embedding $\mathbf{w}^v$ and linguistic embedding $\mathbf{w}^l$ are expected to be close enough. The right demonstrates text-guided image manipulation. Given a source image and a text guidance, we first get their embedding $\mathbf{w}^v$ and $\mathbf{w}^l$ in $\mathcal{W}$ space through corresponding encoders. We then perform style mixing for target layers and get the target latent code $\mathbf{w}^t$. The final $\mathbf{w}^{t*}$ is obtained through instance-level optimization. The edited image can be generated from the StyleGAN generator.

mentioned methods unfeasible for higher resolution. Furthermore, the pretrained text-image matching model they used fails to exploit attribute-level cross-modal information and leads to mismatched attributes when generating images from texts [39, 19, 47, 5], or undesired changes of irrelevant attributes when manipulating images [8, 26, 20, 21].

Recent progress on generative adversarial networks (GANs) has established an entirely different image generation paradigm that achieves phenomenal quality, fidelity, and realism. StyleGAN [16], one of the most notable GAN frameworks, introduces a novel style-based generator architecture and can produce high-resolution images with unmatched photorealism. Some recent work [16] has demonstrated that the intermediate latent space $\mathcal{W}$ of StyleGAN, inducted from a learned piece-wise continuous mapping, yields less entangled representations and offers more feasible manipulation. The superior characteristics of $\mathcal{W}$ space appeal to numerous researchers to develop advanced GAN inversion techniques [38, 2, 1] to invert real images back into the StyleGAN's latent space and perform meaningful manipulation. The most popular way [45, 29] is to train an additional encoder to map real images into the $\mathcal{W}$ space, which leads to not only faithful reconstruction but also semantically meaningful editing. Furthermore, it is easy to introduce the hierarchically semantic property of the $\mathcal{W}$ space to any GAN model by simply learning an extra mapping network before a fixed, pretrained StyleGAN generator. We thoroughly investigated the existing GAN inversion methods, and found all is about how to map images into the latent space of a well-trained GAN model. The other modalities like texts, however, have not received any attention.

In this paper, for the first time, we propose a GAN inversion technique that can map multi-modal information, *e.g.*, texts, sketches, or labels, into a common latent space of a pretrained StyleGAN. Based on that, we propose a very simple yet effective method for Text-guided diverse image generation and manipulation via GAN (abbreviated TediGAN). Our proposed method introduces three novel modules. The first StyleGAN inversion module learns the inversion where an image encoder can map a real image to the $\mathcal{W}$ space, while the second visual-linguistic similarity module learns linguistic representations that are consistent with the visual representations by projecting both into a common $\mathcal{W}$ space, as shown in Figure 2. The third instance-level optimization module is to preserve the identity after editing, which can precisely manipulate the desired attributes consistent with the texts while faithfully reconstructing the unconcerned ones. Our proposed method can generate diverse and high-quality results with a resolution up to 1024^2, and inherently support image synthesis with multi-modal inputs, such as sketches or semantic labels with or without instance (texts or real images) guidance. Due to the utilization of a pretrained StyleGAN model, our method can provide the lowest effect guarantee, *i.e.*, our method can always produce pleasing results no matter how uncommon the given text or image is. Furthermore, to fill the gaps in the text-to-image synthesis dataset for faces, we create the Multi-Modal CelebA-HQ dataset to facilitate the research community. Following the format of the two popular text-to-image synthesis datasets, *i.e.*, CUB [34] for birds and COCO [22] for natural scenes, we create ten unique descriptions for each image in the CelebA-HQ [15]. Besides real faces and textual descriptions, the introduced dataset also contains the label map and sketch for the text-guided generation with multi-modal inputs.

In summary, this work has the following contributions:

- We propose a unified framework that can generate diverse images given the same input text, or text with image for manipulation, allowing the user to edit the appearance of different attributes interactively.
- We propose a GAN inversion technique that can map multi-modal information into a common latent space of a pretrained StyleGAN where the instance-level image-text alignment can be learned.
- We introduce the Multi-Modal CelebA-HQ dataset, consisting of multi-modal face images and corresponding textual descriptions, to facilitate the community.

2. Related Work

Text-to-Image Generation. There are basically two categories of GAN-based text-to-image generation methods.

Figure 3. Diverse High-Resolution Results from Text. Our method can achieve text-guided diverse image generation and manipulation up to an unprecedented resolution at 1024^2.

The first category produces images from texts directly by one generator and one discriminator. For example, Reed *et al.* [28] propose to use conditional GANs to generate plausible images from given text descriptions. Tao *et al.* [32] propose a simplified backbone that generates high-resolution images directly by Wasserstein distance and fuses the text information into visual feature maps to improve the image quality and text-image consistency. Despite the plainness and conciseness, the one-stage models produce dissatisfied results in terms of both photo-realism and text-relevance in some cases. Thus, another thread of research focuses on multi-stage processing. Zhang *et al.* [42] stack two GANs to generate high-resolution images from text descriptions through a sketch-refinement process. They further propose a three-stage architecture [43] that stacks multiple generators and discriminators, where multi-scale images are generated progressively in a course-to-fine manner. Xu *et al.* [39] improve the work of [43] from two aspects. First, they introduce attention mechanisms to explore fine-grained text and image representations. Second, they propose a Deep Attentional Multimodal Similarity Model (DAMSM) to compute the similarity between the generated image and the sentence. The subsequent studies basically follow the framework of [39] and have proposed several variants by introducing different mechanisms like attention [19] or memory writing gate [47]. However, the multi-stage frameworks produce results that look like a simple combination of visual attributes from different image scales.

Text-Guided Image Manipulation. Similar to text-to-image generation, manipulating given images using texts also produces results that contain desired visual attributes. Differently, the modified results should only change certain parts and preserve text-irrelevant contents of the original images. For example, Dong *et al.* [8] propose an encoder-decoder architecture to modify an image according to a given text. Nam *et al.* [27] disentangle different visual attributes by introducing a text-adaptive discriminator, which can provide finer training feedback to the generator. Li *et al.* [20] introduce a multi-stage network with a novel text-image combination module to produce high-quality results. Similar to text-to-image generation, the text-based image manipulation methods with the best performance are basically based on the multi-stage framework. Different from all existing methods, we propose a novel framework that unifies text-guided image generation and manipulation methods and can generate high-resolution and diverse images `directly` without multi-stage processing.

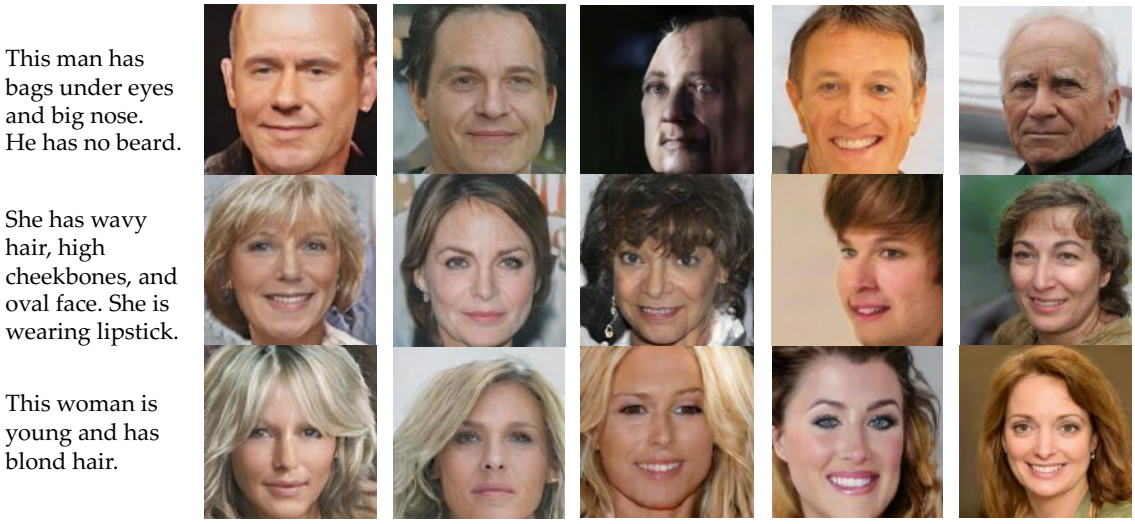
Image-Text Matching. One key of text-guided image generation or manipulation is to match visual attributes with corresponding words. To do this, current methods usually provide explicit word-level training feedback from the elaborately-designed discriminator [20, 21]. There is also a rich line of work proposed to address a related direction named image-text matching, or visual-semantic alignment, aiming at exploiting the matching relationships and making the corresponding alignments between text and image. Most of them can be categorized into two-branch deep architecture according to the granularity of representations for both modalities, *i.e.*, global [17, 24, 23] or local [14, 13, 18] representations. The first category employs deep neural networks to extract the global features of both modalities, based on which their similarities are measured [24]. Another thread of work performs instance-level image-text matching [25, 18, 31], learning the correspondences between words and image regions [13].

3. The TediGAN Framework

We first learn the inversion, *i.e.*, training an image encoder to map the real images to the latent space such that all codes produced by the encoder can be recovered at both the pixel-level and the semantic-level. We then use the hierarchical characteristic of $\mathcal{W}$ space to learn the text-image matching by mapping the image and text into the same joint embedding space. To preserve identity in manipulation, we propose an instance-level optimization, involving the trained encoder as a regularization to better reconstruct the pixel values without affecting the semantic property of the inverted code.

3.1. StyleGAN Inversion Module

The inversion module aims at training an image encoder that can map a real face image to the latent space of a fixed



(a) AttnGAN (b) ControlGAN (c) DM-GAN (d) DF-GAN (e) Ours

Figure 4. Comparison of Text-to-Image Generation on Our Multi-modal CelebA-HQ dataset.

StyleGAN model pretrained on the FFHQ dataset [16]. The reason we invert a trained GAN model instead of training one from scratch is that, in this way, we can go beyond the limitations of a paired text-image dataset. The StyleGAN is trained in an unsupervised setting and covers much higher quality and wider diversity, which makes our method able to produce satisfactory edited results with images in the wild. In order to facilitate subsequent alignment with text attributes, our goal for inversion is not only to reconstruct the input image by pixel values but also to acquire the inverted code that is semantically meaningful and interpretable [30, 40].

Before introducing our method, we first briefly establish problem settings and notations. A GAN model typically consists of a generator $G(\cdot) : \mathcal{Z} \rightarrow \mathcal{X}$ to synthesize fake images and a discriminator $D(\cdot)$ to distinguish real data from the synthesized. In contrast, GAN inversion studies the reverse mapping, which is to find the best latent code $\mathbf{z}^*$ by inverting a given image $\mathbf{x}$ to the latent space of a well-trained GAN. A popular solution is to train an additional encoder $E_v(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$ [46, 3] (subscript v means visual). To be specific, a collection of latent codes $\mathbf{z}^s$ are first randomly sampled from a prior distribution, *e.g.*, normal distribution, and fed into $G(\cdot)$ to get the synthesis $\mathbf{x}^s$ as the training pairs. The introduced encoder $E_v(\cdot)$ takes $\mathbf{x}^s$ and $\mathbf{z}^s$ as inputs and supervisions respectively and is trained with

$$\min_{\Theta_{E_v}} \mathcal{L}_{E_v} = ||\mathbf{z}^s - E_v(G(\mathbf{z}^s))||_2^2, \tag{1}$$

where $||\cdot||_2$ denotes the l_2 distance and Θ_{E_v} represents the parameters of the encoder $E_v(\cdot)$.

Despite of its fast inference, the aforementioned procedure simply learns a deterministic model with no regard to whether the codes produced by the encoder align with the semantic knowledge learned by $G(\cdot)$. The supervision by only reconstructing $\mathbf{z}^s$ is not powerful enough to train $E_v(\cdot)$, and $G(\cdot)$ is actually not fully used to guide the training of $E_v(\cdot)$, leading to the incapability of inverting real images. To solve these problems, we use a totally different strategy to train an encoder for GAN inversion as in [45]. There are two main differences compared with the conventional framework: (a) the encoder is trained with real images rather than with synthesized images, making it more applicable to real applications; (b) the reconstruction is at the image space instead of latent space, which provides semantic knowledge and accurate supervision and allows integration of powerful image generation losses such as perceptual loss [12] and LPIPS [44]. Hence, the training process can be formulated as

$$\begin{aligned}\min_{\Theta_{E_v}} \mathcal{L}_{E_v} = &||\mathbf{x} - G(E_v(\mathbf{x}))||_2^2 + \lambda_1||F(\mathbf{x}) - F(G(E_v(\mathbf{x})))||_2^2 \\ &- \lambda_2 \mathbb{E}[D_v(G(E_v(\mathbf{x})))],\end{aligned} \tag{2}$$

$$\min_{\Theta_{D_v}} \mathcal{L}_{D_v} = \mathbb{E}[D_v(G(E_v(\mathbf{x})))] - \mathbb{E}[D_v(\mathbf{x})] + \frac{\lambda_3}{2}\mathbb{E}[||\nabla_{\mathbf{x}} D_v(\mathbf{x})||_2^2], \tag{3}$$

where Θ_{E_v} and Θ_{D_v} are learnable parameters, λ_1, λ_2, and λ_3 are the hyper-parameters, and $F(\cdot)$ denotes the VGG feature extraction model.

Through the learned image encoder, we can map a real image into the $\mathcal{W}$ space. The obtained code is guaranteed to align with the semantic domain of the StyleGAN generator and can be further utilized to mine cross-modal similarity between the image-text instance pairs.

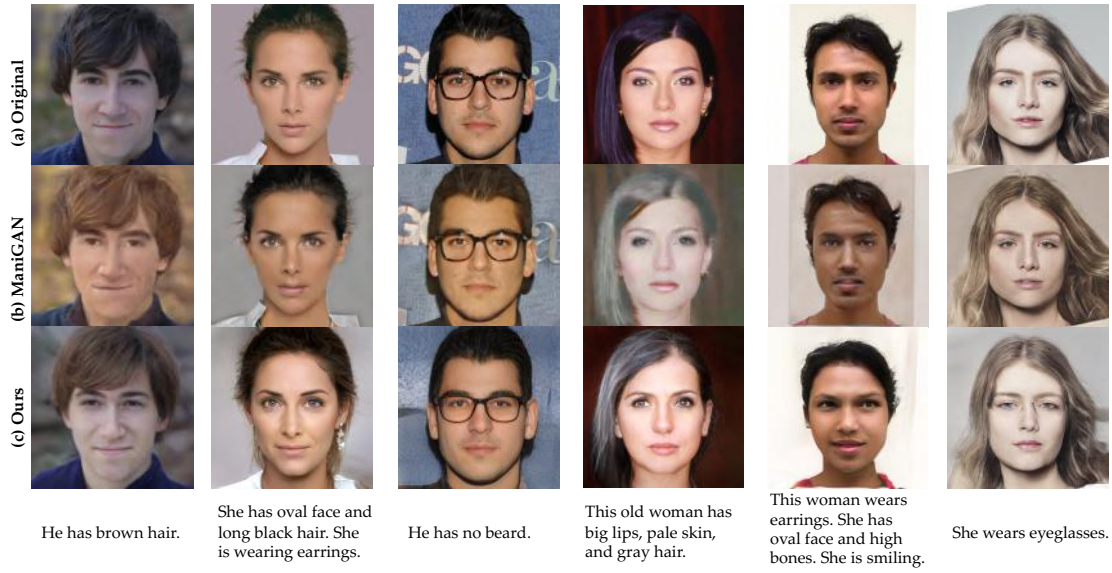


Figure 5. Qualitative Comparison of Image Manipulation using Natural Language Descriptions.

3.2. Visual-Linguistic Similarity Learning

Once the inversion module is trained, given a real image, we can map it into the $\mathcal{W}$ space of StyleGAN. The next problem is how to train a text encoder that learns the associations and alignments between image and text. Instead of training a text encoder in the same way as the image encoder or the aforementioned DAMSM, we propose a visual-linguistic similarity module to project the image and text into a common embedding space, *i.e.*, the $\mathcal{W}$ space, as shown in Figure 2. Given a real image and its descriptions, we encode them into the $\mathcal{W}$ space by using the previously trained image encoder and a text encoder. The obtained latent code is the concatenation of L different C-dimensional $\mathbf{w}$ vectors, one for each input layer of StyleGAN. The multi-modal alignment can be trained with

$$\min_{\Theta_{E_l}} \mathcal{L}_{E_l} = || \sum_{i=1}^{L} p_i(\mathbf{w}_i^v - \mathbf{w}_i^l)||_2^2, \tag{4}$$

where Θ_{E_l} represents the parameters of the text encoder $E_l(\cdot)$ and subscript l means linguistic; $\mathbf{w}^v, \mathbf{w}^l \in \mathcal{W}^{L\times C}$ are the obtained image embedding and text embedding; $\mathbf{w}^v = f(E_v(\mathbf{x}))$ is the projected code of the image embedding $\mathbf{z}$ in the input latent space $\mathcal{Z}$ using a non-linear mapping network $f : \mathcal{Z} \rightarrow \mathcal{W}$; $\mathbf{w}^l$ shares a similar definition; $\mathbf{w}^v$ and $\mathbf{w}^l$ are with the same shape $L \times C$, meaning to have L *layers* and each with a C*-dimensional* latent code; and p_i is the weight of i-th layer in the latent code.

Compared with DAMSM, our proposed module is lightweight and easy to train. More importantly, this module achieves instance-level alignment [35], *i.e.*, learning correspondences between visual and linguistic attributes, by leveraging the disentanglability of StyleGAN. The text encoder is trained with the proposed visual-linguistic similarity loss together with the pairwise ranking loss [17, 8], which is omitted from Equation 4.

3.3. Instance-Level Optimization

One of the main challenges of face manipulation is the identity preservation. Due to the limited representation capability, learning a perfect reverse mapping with an encoder alone is not easy. To preserve identity, some recent methods [33, 29] incorporate a dedicated face recognition loss [7] to measure the cosine similarity between the output image and its source. Different from their methods, for text-guided image manipulation, we implement an instance-level optimization module to precisely manipulate the desired attributes consistent with the descriptions while faithfully reconstructing the unconcerned ones. We use the inverted latent code $\mathbf{z}$ as the initialization, and the image encoder is included as a regularization to preserve the latent code within the semantic domain of the generator. To summarize, the objective function for optimization is

$$\begin{aligned}\mathbf{z}^* = \arg\min_{\mathbf{z}} \; ||\mathbf{x} - G(\mathbf{z})||_2^2 \; &+ \lambda_1' ||F(\mathbf{x}) - F(G(\mathbf{z}))||_2^2 \\ &+ \lambda_2' ||\mathbf{z} - E_v(G(\mathbf{z}))||_2^2,\end{aligned} \tag{5}$$

where $\mathbf{x}$ is the original image to manipulate, λ_1' and λ_2' are the loss weights corresponding to the perceptual loss and the encoder regularization term, respectively.

3.4. Control Mechanism

Attribute-Specific Selection. The two different tasks, *i.e.*, text-to-image generation and text-guide image manipu-

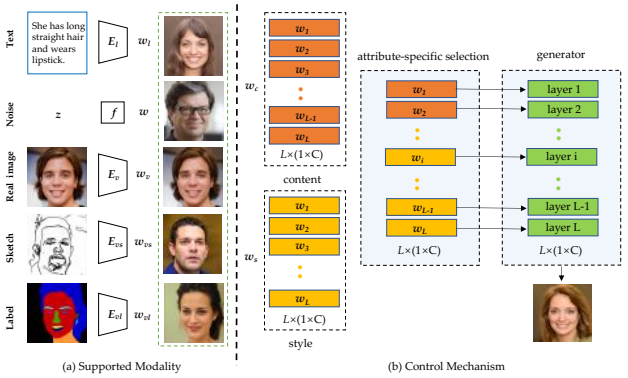


Figure 6. Control Mechanism of Our TediGAN Framework. Different layer in the StyleGAN generator represents different attributes. Changing the value of a certain layer would change the corresponding attributes of the image. Since the texts and images are mapped into the common latent space, we can synthesize images with certain attributes by selecting attribute-specific layers. The control mechanism mixes layers of the style code $\mathbf{w}^s$ by partially replacing corresponding layers of the content $\mathbf{w}^c$. When $\mathbf{w}^s$ is a randomly sampled latent code, it is the text-to-image generation and when $\mathbf{w}^s$ is the image embedding, it performs text-guided image manipulation.

lation, are unified into one framework by our proposed control mechanism. Our mechanism is based on the style mixing of StyleGAN. The layer-wise representation of StyleGAN learns disentanglement of semantic fragments (attributes or objects). In general, different layer $\mathbf{w}_i$ represents different attributes and is fed into the i-th layer of the generator. Changing the value of a certain layer would change the corresponding attributes of the image. As shown in Figure 2, given two codes with the same size $\mathbf{w}^c, \mathbf{w}^s \in \mathcal{W}^{L\times C}$ denoting content code and style code, this control mechanism selects attribute-specific layers and mixes those layers of $\mathbf{w}^s$ by partially replacing corresponding layers of $\mathbf{w}^c$. For text-to-image generation, the produced images should be consistent with the textual description, thus $\mathbf{w}^c$ should be the linguistic code, and randomly sampled latent code with the same size acts as $\mathbf{w}^s$ to provide diversity (results are shown in Figure 7). For text-guided image manipulation, $\mathbf{w}^c$ is the visual embedding while $\mathbf{w}^s$ is the linguistic embedding, the layers for mixing should be relevant to the text, for the purpose of modifying the relevant attributes only and keeping the unrelated ones unchanged.

Supported Modality. The style code $\mathbf{w}^s$ and content code $\mathbf{w}^c$ could be sketch, label, image, and noise, as shown in Figure 6, which makes our TediGAN feasible for multi-modal image synthesis. The control mechanism provides high accessibility, diversity, controllability, and accurateness for image generation and manipulation. Due to the control mechanism, as shown in Figure 1, our method inherently supports continuous operations and multi-modal synthesis for sketches and semantic labels with descriptions. To produce the diverse results, all we need to do is to keep the layers related to the text unchanged and replace the others with the randomly sampled latent code. If we want to generate images from other modality with text guidance, take the sketch as an example, we can train an additional sketch image encoder E_{vs} in the same way as training the real image encoder and leave the other parts unchanged.

Layerwise Analysis. The pre-trained StyleGAN we used in most experiments is to generate images of 256 $\times$ 256 (*i.e.*, size 256), whose has 14 layers of the intermediate vector. For a synthesis network trained to generate images of 512 $\times$ 512, the intermediate vector would be of shape (16, 512) (and (18, 512) for 1024 $\times$ 1024), where the number of the layers L is determined by $2\log_2 \mathrm{R}-2$ and R is the image size. In general, layers in the generator at lower resolutions (*e.g.*, 4 $\times$ 4 and 8 $\times$ 8) control high-level styles such as eyeglasses and head pose, layers in the middle (*e.g.*, as 16 $\times$ 16 and 32 $\times$ 32) control hairstyle and facial expression, while the final layers (*e.g.*, 64 $\times$ 64 to 1024 $\times$ 1024) control color schemes and fine-grained details. Based on empirical observations, we list the attributes represented by different layers of a 14-layer StyleGAN in Table 1. The layers from 11-14 represent micro features or fine structures, such as stubble, freckles, or skin pores, which can be regarded as the stochastic variation. High-resolution images contain lots of facial details and cannot be obtained by simply upsampling from the lower-resolutions, making the stochastic variations especially important as they improve the visual perception without affecting the main structures and attributes of the synthesized image.

Table 1. The Empirical Layerwise Analysis of a 14-layer StyleGAN Generator. The 13-th and 14-th layers are omitted since there is basically no visible difference.

n-th	attribute	n-th	attribute
1	eye glasses	7	hair color
2	head pose	8	face color
3	face shape	9	age
4	hair length, nose, lip	10	gender
5	cheekbones	11	micro features
6	chin	12	micro features

4. Experiments

4.1. Experiments Setup

Datasets and Baseline Models. To achieve text-guided image generation and manipulation, the first step is to build a dataset that contains photo-realistic facial images and corresponding descriptions. We introduce the Multi-Modal CelebA-HQ dataset, a large-scale face image dataset that

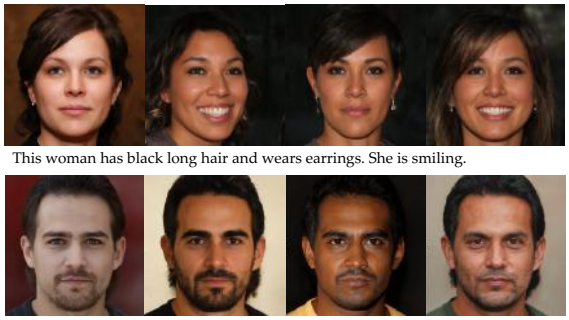


Figure 7. Diverse Text-to-image Generation.

Table 2. Quantitative Comparison of Text-to-Image Generation. We use FID, LPIPS, accuracy (Acc.), and realism (Real.) to compare the state of the art and our method on the proposed Multi-modal CelebA-HQ dataset. ↓ means the lower the better while ↑ means the opposite.

Method	FID ↓	LPIPS ↓	Acc. (%) ↑	Real. (%) ↑
AttnGAN [39]	125.98	0.512	14.2	20.3
ControlGAN [19]	116.32	0.522	18.2	22.5
DFGAN [32]	137.60	0.581	22.8	25.5
DM-GAN [47]	131.05	0.544	19.5	12.8
TediGAN	**106.37**	**0.456**	**25.3**	**31.7**

has 30,000 high-resolution face images, each having a high-quality segmentation mask, sketch, and descriptive text. We evaluate our proposed method on text and image partitions, comparing with state-of-the-art approaches AttnGAN [39], ControlGAN [19], DM-GAN [47], and DFGAN [32] for image generation, and comparing with ManiGAN [20] for image manipulation using natural language descriptions. All methods are retrained with the default settings on the proposed Multi-Modal CelebA-HQ dataset.

Evaluation Metric. For evaluation, there are four important aspects: image quality, image diversity, accuracy, and realism [19, 21]. The quality of generated or manipulated images is evaluated through Fréchet Inception Distance (FID) [10]. The diversity is measured by the Learned Perceptual Image Patch Similarity (LPIPS) [44]. For image generation, the accuracy is evaluated by the similarity between the text and the corresponding generated image. For manipulation, the accuracy is evaluated by whether the modified visual attributes of the synthetic image are aligned with the given description and text-irrelevant contents are preserved. The accuracy and realism are evaluated through a user study, where the users are asked to judge which one is more photo-realistic, and more coherent with the given texts. We test accuracy and realism by randomly sampling 50 images with the same conditions and collect more than 20 surveys from different people with various backgrounds.

Table 3. Quantitative Comparison of Text-Guided Image Manipulation. We use FID, accuracy (Acc.), and realism (Real.) to compare with the state of the art ManiGAN [20].

Method	CelebA		Non-CelebA	
	ManiGAN [20]	Ours	ManiGAN [20]	Ours
FID ↓	117.89	**107.25**	143.39	**135.47**
Acc. (%) ↑	40.9	**59.1**	12.8	**87.2**
Real. (%) ↑	36.2	**63.8**	21.7	**78.3**

4.2. Comparison with State-of-the-Art Methods

4.2.1 Text-to-Image Generation

Quantitative Comparison. In our experiments, we evaluate the FID and LPIPS on a large number of samples generated from randomly selected text descriptions. To evaluate accuracy and realism, we generate images from 50 randomly sampled texts using different methods. In a user study, users are asked to judge which one is the most photo-realistic and most coherent with the given texts. The results are demonstrated in Table 2. Compared with the state-of-the-arts, our method achieves better FID, LPIPS, accuracy, and realism values, which proves that our methods can generate images with the highest quality, diversity, photorealism, and text-relevance.

Qualitative Comparison. Most existing text-to-image generation methods, as shown in Figure 4, can generate photo-realistic and text-relevant results. However, some attributes contained in the text do not appear in the generated image, and the generated image looks like featureless paint and lacks details. This "featureless painterly" look [16] would be significantly obvious and irredeemable when generating higher resolution images using the multi-stage training methods [39, 19, 47]. Furthermore, most existing solutions have limited diversity of the outputs, even if the provided conditions contain different meanings. For example, "*has a beard*" might mean a goatee, short or long beard, and could have different colors. Our method can not only generate results with diversity but also realise the expectation to change where you want by using the control mechanism. To produce diverse results, with the layers related to the text unchanged, the other layers could be replaced by any values sampled from the prior distribution. For example, as shown in the first row of Figure 7, the key visual attributes (*women, black long hair, earrings, and smiling*) are preserved, while the other attributes, like haircuts, makeups, face shapes, and head poses, show a great degree of diversity. The images in the second row illustrate more precise control ability. We keep the layers representing face shape and head pose the same and change the others. Figure 3 shows high-quality and diverse results with resolution at 1024×1024.

4.2.2 Text-Guided Image Manipulation

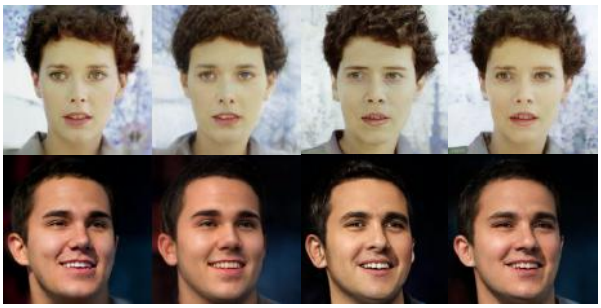
Quantitative Comparison. In our experiments, we evaluate the FID and conduct a user study on randomly selected images from both CelebA and Non-CelebA datasets with randomly chosen descriptions. The results are shown in Table 3. Compared with ManiGAN [20], our method achieves better FID, accuracy, and realism. This indicates that our method can produce high-quality synthetic images, and the modifications are highly aligned with the given descriptions, while preserving other text-irrelevant contents.

Qualitative Comparison. Figure 5 shows the visual comparisons between the recent method ManiGAN [20] and ours. As shown, the second row is to add earrings and change the face shape and hair style of the woman, our method completes this difficult case while ManiGAN fails to produce required attributes. ManiGAN produces less satisfactory modified results: in some cases, the text-relevant regions are not modified and the text-irrelevant ones are changed. Furthermore, since the StyleGAN we used is pre-trained on a very large face dataset [16], the latent space almost covers the full space of facial attributes, which makes our method robust for real images in the wild. The images in last two columns are results of out-of-distribution (Non-CelebA), *i.e.*, images from other face dataset such as [4, 6, 41], which illustrate that our method is prepared to produce pleasing results with images in the wild.

5. Ablation Study and Discussion

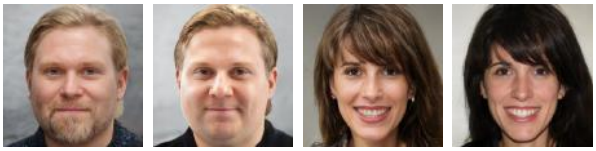
Instance-Level Optimization. The comparison of with or without instance-level optimization is shown in Figure 8. As shown, the inversion results of the image encoder preserve all attributes of the original images, which is sufficient for text-to-image generation since there is no identity to preserve (Figure 8 (c)). Manipulating a given image according to a text, however, should not change the unrelated attributes especially one's identity, which is preserved after the instance-level optimization (Figure 8 (d)). We also compare with a recent inversion-based image synthesis method pSp [29] that incorporates a dedicated recognition loss [7] during training. Despite both preserving the identity, the optional instance-level optimization provides a non-deterministic way to refine the final results accordingly.

Visual-Linguistic Similarity. The text encoder is trained using our visual-linguistic similarity and a very simple pairwise ranking loss [17, 8] to align text and image embedding. Although the learned text embedding can handle near-miss cases, as shown in Figure 9, we found this plain strategy sometimes may lead to insufficient disentanglement of attributes and mismatching of image-text alignment, leaving some room for improvement.



(a) (b) (c) (d)

Figure 8. Inversion Results. (a) original image; (b) inversion result of pSp [29]; (c) inversion result of our image encoder (Section 3.1); (d) inversion results after optimization (Section 3.3).



He has / has no beard. She has brown / black hair.

Figure 9. Illustration of Near-miss Cases.

Potential Issue with StyleGAN. In our experiments, we found that some unrelated attributes are unwantedly changed when we manipulate a given image according to a text description. We thought it might be the problem of visual-linguistic similarity learning in the first place. However, when performing layer-wise style mixing on the inverted codes of two real images, the interference still occurs. This means some facial attributes remain entangled in the $\mathcal{W}$ space, where different attributes should be orthogonal (meaning without affecting other attributes). Another inherent defect of StyleGAN is that some attributes, such as hats, necklaces and earrings, are not well represented in its latent space. This makes our method perform less satisfactorily sometimes when adding or removing jewelry or accessories through natural language descriptions.

6. Conclusion

We have proposed a novel method for image synthesis using textual descriptions, which unifies two different tasks (text-guided image generation and manipulation) into the same framework and achieves high accessibility, diversity, controllability, and accurateness for facial image generation and manipulation. Through the proposed multi-modal GAN inversion and large-scale multi-modal dataset, our method can effectively synthesize images with unprecedented quality. Extensive experimental results demonstrate the superiority of our method, in terms of the effectiveness of image synthesis, the capability of generating high-quality results, and the extendability for multi-modal inputs.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *ICCV*, 2019. 2

[2] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Inverting layers of a large generator. In *ICLR Workshop*, 2019. 2

[3] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a GAN cannot generate. In *ICCV*, pages 4502–4511, 2019. 4

[4] Olga Chelnokova, Bruno Laeng, Marie Eikemo, Jeppe Riegels, Guro Løseth, Hedda Maurud, Frode Willoch, and Siri Leknes. Rewards of beauty: the opioid system mediates social motivation in humans. *Molecular psychiatry*, 19(7):746–747, 2014. 8

[5] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In *CVPR*, pages 10911–10920, 2020. 2

[6] Rémi Courset, Marine Rougier, Richard Palluel-Germain, Annique Smeding, Juliette Manto Jonte, Alan Chauvin, and Dominique Muller. The Caucasian and North African French Faces (CaNAFF): A face database. *International Review of Social Psychology*, 31(1), 2018. 8

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 5, 8

[8] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *ICCV*, pages 5706–5714, 2017. 1, 2, 3, 5, 8

[9] Arnab Ghosh, Richard Zhang, Puneet K. Dokania, Oliver Wang, Alexei A. Efros, Philip H. S. Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *ICCV*, 2019. 1

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017. 7

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 4

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 3

[14] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*, pages 1889–1897, 2014. 3

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018. 2

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 4, 7, 8

[17] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. 3, 5, 8

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 3

[19] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. In *NeurIPS*, 2019. 2, 3, 7

[20] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *CVPR*, 2020. 2, 3, 7, 8

[21] Bowen Li, Xiaojuan Qi, Philip H. S. Torr, and Thomas Lukasiewicz. Lightweight generative adversarial networks for text-guided image manipulation. In *NeurIPS*, 2020. 2, 3, 7

[22] 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 *ECCV*, pages 740–755, 2014. 2

[23] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, pages 2623–2631, 2015. 3

[24] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks. In *ICLR*, 2015. 3

[25] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017. 3

[26] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *NeurIPS*, 2018. 1, 2

[27] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *NeurIPS*, pages 42–51, 2018. 3

[28] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 1, 3

[29] 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. *arXiv preprint arXiv:2008.00951*, 2020. 2, 5, 8

[30] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, 2020. 4

[31] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, pages 1979–1988, 2019. 3

[32] Ming Tao, Hao Tang, Songsong Wu, Fei Sebe, Nicu Wu, and Xiao-Yuan Jing. DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 3, 7

[33] Evangelos Ververas and Stefanos Zafeiriou. Slidergan: Synthesizing expressive face images by sliding 3d blendshape parameters. *IJCV*, pages 1–22, 2020. 5

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2

[35] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, 2020. 5

[36] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 1

[37] Weihao Xia, Yujiu Yang, and Jing-Hao Xue. Cali-sketch: Stroke calibration and completion for high-quality face image generation from poorly-drawn sketches. *arXiv preprint arXiv:1911.00426*, 2019. 1

[38] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv: 2101.05278*, 2021. 2

[39] 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 *CVPR*, 2018. 1, 2, 3, 7

[40] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *arXiv preprint arXiv:1911.09267*, 2019. 4

[41] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. ApdrawingGAN: Generating artistic portrait drawings from face photos with hierarchical gans. In *CVPR*, pages 10743–10752, 2019. 8

[42] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 3

[43] 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. *TPAMI*, 41(8):1947–1962, 2018. 3

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 4, 7

[45] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *ECCV*, 2020. 2, 4

[46] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 4

[47] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pages 5802–5810, 2019. 2, 3, 7