CLIP2GAN: Toward Bridging Text With the Latent Space of GANs
Yixuan Wang, Wengang Zhou, Senior Member, IEEE, Jianmin Bao, Weilun Wang, Graduate Student Member, IEEE, Li Li, Member, IEEE, and Houqiang Li, Fellow, IEEE

Abstract—In this work, we are dedicated to text-guided image generation and propose a novel framework, i.e., CLIP2GAN, by leveraging CLIP model and StyleGAN. The key idea of our CLIP2GAN is to bridge the output feature embedding space of CLIP and the input latent space of StyleGAN, which is realized by introducing a mapping network. In the training stage, we encode an image with CLIP and map the output feature to a latent code, which is further used to reconstruct the image. In this way, the mapping network is optimized in a self-supervised learning way. In the inference stage, since CLIP can embed both image and text into a shared feature embedding space, we replace CLIP image encoder in the training architecture with CLIP text encoder, while keeping the following mapping network as well as StyleGAN model. As a result, we can flexibly input a text description to generate an image. Moreover, by simply adding mapped text features of an attribute to a mapped CLIP image feature, we can effectively edit the attribute to the image. Extensive experiments demonstrate the superior performance of our proposed CLIP2GAN compared to previous methods.

Index Terms—Text-guided image generation, image editing, pre-trained models, generative adversarial nets.

I. INTRODUCTION

Generative models based on GAN [1] have achieved remarkable success in various tasks including image-to-image translation [2], [3], image inpainting [4], [5], [6], image restoration [7], [8], etc. Specifically, in image generation, the quality of images generated by GANs has been improving with the emergence of several advanced GANs [9], [10], [11], [12] that are capable of generating high-resolution and high-fidelity images. Among them, StyleGAN [9] disentangles image attributes in the intermediate latent space, which allows for various image editing and manipulation tasks [13], [14], [15], [16], [17], [18], [19], [20], [21]. Previous work tends to achieve image applications by controlling the latent code features directly, which makes it difficult to involve explicit intentions. Since text can express what people need precisely, it is natural to explore whether text can be utilized to control the latent space directly to achieve text-guided image generation and image editing tasks.

In this paper, we propose a novel framework, i.e., CLIP2GAN, for text-guided image generation and editing. Technically, our task can be decomposed into two subtasks. First, the input text is transferred to an embedding feature space. Second, the text feature is used to generate an image. For each subtask, there are successful solutions in literature, such as CLIP [22] for the first and StyleGAN [9] for the second. However, CLIP and StyleGAN are decoupled since the CLIP feature is not aligned with the latent code of StyleGAN. To bridge CLIP and StyleGAN, we introduce a mapping network, which transfers the feature embedding of CLIP to the latent space of StyleGAN. Specifically, in the training stage, we encode an image with CLIP and map the output feature to a latent code, which is further used to reconstruct the image. In this way, the mapping network is optimized in a self-supervised learning way. In the inference stage, since CLIP can embed both image and text into a shared space, we replace CLIP image encoder in the training architecture with CLIP text encoder. Consequently, we can flexibly input a text description to generate a face image.

In CLIP model, an image is encoded into a 512-dimensional feature, which inevitably loses some detailed visual clue and results in the generated images missing information such as hair details, skin texture, background, etc. To improve the quality of generated images, we introduce an additional discriminator trained adversarially with the mapping network, thus ensuring that the mapping network can generate as realistic and high-resolution images as possible. Besides, we also add noise to the CLIP image feature. On the one hand, it further supplements the details lost by CLIP. On the other hand, inspired by the mode seeking regularization [23], we maximize the ratio of the distance between the output images after adding noise to the distance between their corresponding latent codes, so that the generated images are diverse while maintaining high fidelity.

Unlike previous methods [24], [25], [26], [27], [28], [29] using image-text pairs for training, our framework generates high-quality face images in a zero-shot way, which means we...
Fig. 1. Text-guided image generation and editing results of our proposed CLIP2GAN. The left shows diverse generation results given a text description. The right shows the result of image editing using text. The images generated by our framework are realistic and accurate.

use image data as input for training and text data as input for testing. Thanks to CLIP’s multi-modal embedding space, our text-free training approach is not constrained by large-scale image-text datasets that require precise manual annotation and are not easily available in practice [26], [28], [30], [31], [32], and can generate text-guided realistic and accurate images. Compared with those who use image-text pairs for training for text-guided image generation, our model is implemented in a minimal-cost text-free training approach, but still maintains extremely high fidelity and quality.

Besides image generation, we explore CLIP2GAN on image editing task, where real images can be edited directly using textual attributes to adjust their expressions, hair characteristics, age, etc. Taking advantage of StyleGAN, we locate different feature orientations of different text descriptions in the latent space. By imposing different orientations in the latent code of the image, the semantics of the image is controlled and changed in a fine-grained manner with high quality. Specifically, CLIP2GAN is convenient and flexible for image editing, because instead of optimizing the model for specific text like StyleCLIP [33], we can simply use arithmetic operations to add attributes to images by adding mapped CLIP features of text describing the attributes to mapped CLIP features of images.

To evaluate our proposed method, we perform experiments on the CelebA-HQ [34], AFHQ [35] and LSUN [36] dataset. Both quantitative and qualitative results validate that comparing with previous methods, our text-free training model can generate high-fidelity and diverse results and realize image manipulation, achieving superior performance over most existing models trained using full image-text pairs. Some example results are shown in Fig. 1.

In summary, our contributions are as follows:

- We propose a new framework, i.e., CLIP2GAN, that enables text-guided generation tasks with text-free training, generating diverse and high-quality images given the same input text.
- We apply our framework to image manipulations that allow the editing of real images directly with textual attributes.
- Extensive experiments on public datasets demonstrate the validity and superiority of our framework. The generated images of our method have higher evaluation quality and better visual performance.

II. RELATED WORK

A. High-Quality Image and Face Generation

Due to the great potential of GAN in generating realistic and high-resolution images, it is widely used in image generation and applications [37], [38], [39], [40]. Li et al. [41] explore and leverage semantic information to generate realistic textures in synthesized images. Li et al. [42] propose a staged semi-supervised GAN-based method for sketch-to-image synthesis, which can directly generate realistic images from novice sketches. The survey [43] summarizes the improvements, taxonomy, methods, and challenges in multi-modal data-guided visual content synthesis, which are mostly based on GANs. In particular, high-quality face generation has been an attractive problem in image generation. PGGAN [34] first proposes the idea of resolution progressive generation to generate high-definition face images, which first discovers large-scale structures and then focuses on fine details. StyleGAN [9], [10], [11] introduces a novel style-based generator architecture that generates face images with high fidelity and high resolution. It controls the visual features represented in each layer individually, which can be coarse features influenced by style, e.g., pose, face shape, identity features, etc. or detailed features influenced by noise, e.g., pupil, hair, wrinkles, etc..

B. Joint Vision-Language Models

With the remarkable progress in both computer vision and natural language processing, researchers turn their attentions to joint vision-language (VL) models for many task-specific VL problems, including image captioning [44], [45], [46], visual question and answer (VQA) [47], [48], image text matching [45], [49], etc., which are tailored for specific problems and each model only solves one task. After the introduction of the transformer [50], BERT [51] has achieved unprecedented success in various language tasks. A recent development,
CLIP [22], pre-trained using over 400 million image-text pairs based on contrastive learning, learns a multi-modal co-embedding space and estimates the semantic similarity between texts and images. The robustness of learned joint representation enables CLIP to offer high performance and excellent generalization on various tasks.

C. Text-Guided Image Generation

Text-guided image generation is an interesting topic in image generation, where GAN-based models show better quality. StackGAN [24], [25] improves image resolution in multiple stages. AttGan [26] introduces a cross-modal attention mechanism to explore fine-grained representations. ALR-GAN [52] proposes a multi-stage model which includes an Adaptive Layout Refinement module. CGL-GAN [53] incorporates the local linguistic representation to address the lack of fine-grained information. KD-GAN [54] draws an image according to reference knowledge and introduces a new evaluation system. MA-GAN [55] explores the semantic correlation to guarantee the generation similarity of related sentences. TediGAN [27] trains an encoder to map the text into the latent space of StyleGAN [9]. DF-GAN [29] proposes a one-stage backbone for the direct synthesis of high-resolution images. In addition, some recent works also extend to text-guided video generation. Pan et al. [56] propose TGANs-C to generate videos with semantic and temporal coherence. Li et al. [57] employ VAE and GAN to extract static and dynamic information from text to generate videos.

The above methods are based on end-to-end training of text-image pairs. Some zero-shot methods have been proposed to solve the problem without text-image pairs. LAFITE [58] uses image pseudo-text features to construct conditional style space for training generators and discriminators. FuseDream [59] is an optimization-based approach that uses AugCLIP for data enhancement. Our proposed CLIP2GAN is also a zero-shot text-guided image generation method that does not require text-image pairs. Different from previous zero-shot methods, CLIP2GAN is a feed-forward method based on a lightweight mapping network, which is more practical and achieves better performance by leveraging pre-trained models.

With the rapid development of diffusion models, some diffusion-based methods, e.g., GLIDE [60], DALL-E [31], Stable Diffusion [61], Imagen [62], etc., have been proposed for text-guided image generation. These diffusion-based methods are trained on large-scale datasets and contain strong generalization capabilities. However, compared with GAN-based methods, these methods consume more inference time due to iterative refinement. Also, since these methods are trained for generating open-world images, the generation results obtain inferior performance to GAN-based methods on specific categories, such as faces, and animals. In this paper, we compare our CLIP2GAN with several large-scale diffusion models and demonstrate the superiority of our proposed method.

D. Text-Guided Image Manipulation

Similar to text-guided generation, manipulating a given image using text produces results containing the desired properties. The difference is that the edited result should change the parts related to the text and retain the rest of it. For instance, Dong et al. [63] propose an encoder-decoder structure for text-guided manipulation, and Li et al. [64] generate high-quality images through a multi-stage network. Liu et al. [65] propose ‘Style Intervention’ to improve the visual fidelity of manipulation results. Several recent methods have explored image editing using the pre-trained visual language model, i.e., CLIP. HairCLIP [66] leverages the text and image encoders of CLIP to alter the latent of StyleGAN for hair editing. StyleCLIP [33] and StyleMC [67] use CLIP as a loss function to discover meaningful directions in the latent space of GANs. However, these methods either focus on editing specific semantic attributes or require independent optimization for different texts. To this end, we propose CLIP2GAN, a novel text-to-image editing framework that bridges the output feature embedding space of CLIP and the input latent space of StyleGAN to handle the image editing for arbitrary text.

III. Method

In this paper, we propose a novel framework, CLIP2GAN, for text-guided image generation without text training (see Fig. 2). Our framework is capable of generating accurate and high-quality images under fine-grained text control without training on paired image-text data. Benefiting from the diversity loss we designed, the multi-modal generation of face images that matches a specific text description is also supported. Furthermore, we explore our framework for image manipulation tasks, where real images are edited using text to adjust their attributes, e.g., expressions, hair color, and age,
with high fidelity and reliability. The rest of this section is organized as follows. We first introduce the overall structure of CLIP2GAN. Then the loss functions utilized in our framework are discussed. Finally, we present the image manipulation application implemented using our framework.

A. CLIP2GAN

Fig. 2 gives an overview of our framework, which consists of a pre-trained vision-language model (CLIP), a mapping network, and a pre-trained generation model (StyleGAN). Unlike previous work that uses a large number of image-text pairs for model training, our approach achieves text-free training by establishing a mapping relationship between CLIP multi-modal embedding space and StyleGAN latent space.

On the one hand, to achieve text-guided image generation without text training, we generate pseudo-text features by leveraging the image-text feature alignment of a pre-trained model. We require a universal multi-modal embedding space where the paired text and image features can be well aligned. The recent vision-language model CLIP achieves this by pre-training a large number of image-text pairs through Contrastive Learning, which is exactly what we need. On the other hand, given that StyleGAN has excellent latent space, we can perform a series of manipulations on the generated images by changing its latent code. We take advantage of the other hand, given that StyleGAN has excellent latent space, we can perform a series of manipulations on the generated images by changing its latent code. We take advantage of the other hand, given that StyleGAN has excellent latent space, we can perform a series of manipulations on the generated images by changing its latent code. We take advantage of the other hand, given that StyleGAN has excellent latent space, we can perform a series of manipulations on the generated images by changing its latent code. We take advantage of the other hand, given that StyleGAN has excellent latent space, we can perform a series of manipulations on the generated images by changing its latent code. We take advantage of the other hand, given that StyleGAN has excellent latent space, we can perform a series of manipulations on the generated images by changing its latent code.

To generate images from text, we build a bridge between CLIP and StyleGAN through a mapping network. With this mapping network, it is possible to obtain feature representations of images in the latent space of StyleGAN, which can be equated to the pseudo-features of the corresponding text in that space, and thus generate images using StyleGAN. The source image \(x\) is taken as the input of CLIP and the image encoder of CLIP is used to obtain the image features \(f_{img}\), i.e., pseudo-text features \(\hat{f}_{text}\), in the multi-modal embedding space of CLIP. It is formulated as follows,

\[
\hat{f}_{text} = f_{img} = C_{img}(x),
\]

where \(C_{img}(\cdot)\) denotes the image encoder of the CLIP model [22]. The image features \(f_{img}\) are mapped to latent codes \(z\) of StyleGAN in \(w^+\) space by the mapping network as the input of the pre-trained StyleGAN, and the image \(x'\) is generated by StyleGAN. \(x'\) is expressed as follows,

\[
x' = G(M(C_{img}(x))),
\]

where \(M(\cdot)\) denotes the mapping network and \(G(\cdot)\) denotes the pre-trained StyleGAN model. By learning the consistency of the source image \(x\) and the generated image \(x'\), our generative model is implemented.

The mapping network is capable of mapping the multi-modal embedding space of CLIP into the \(w^+\) latent space of StyleGAN, which makes it possible to invert CLIP features back into the source images using StyleGAN. Our proposed mapping network is a 12-layer fully connected layer with each layer post-connected to the activation layer Leaky ReLU [68]. The mapping network converts a 512-dimensional text feature \(f_{text}\) or image feature \(f_{img}\) from the multi-modal embedding space of CLIP into the \(w^+\) space of StyleGAN for obtaining an 18 × 512-dimensional latent code \(z\), which is formulated as follows,

\[
f^{i+1} = g(w_i \cdot f^i + b_i), \quad f^0 = f_{text} \text{ or } f_{img},
\]

where \(f^{i+1}, f^i \in \mathbb{R}^{H \times W \times C}\) are the input and output features, respectively. \(g(\cdot)\) denotes the Leaky ReLU activation.

However, we discover that when using only a simple mapping network, CLIP suffers from missing details in both the inversion of the CLIP image features and generated images. Although it can ensure that the major features are presented, the image details, especially the hair and skin textures, backgrounds, etc., are lost to varying degrees, which is attributed to the fact that CLIP only extracts major 512-dimensional features of images and ignores the others. To tackle this issue, we introduce a discriminator to determine the truthfulness of the obtained images. It is trained adversarially with the mapping network to complement the image details and improve the generation quality without affecting the feature representations.

Furthermore, when constraints are applied only between the source images and the reconstructed generated images, it is observed that the generated image examples lack diversity. This is because the loss between pixels mainly focuses on the reconstructed identical images rather than the diverse images, although diversity contributes to the performance. To this end, we apply diversity loss on the latent space to generate diverse images as shown in Fig. 8. Inspired by the mode seeking regularization [23], the diversity loss is achieved by maximizing the ratio of the distance between the output images to the distance between the corresponding latent codes, and it encourages the generation of distinctive results when different noise vectors are brought in. Inputting an arbitrary specific text description, we are able to generate multiple images that match the text features but are unique with an additional Gaussian noise of \(\mu = 1, \sigma^2 = 0.36\).

Considering the image-text alignment property of CLIP’s multi-modal embedding space, we input the source image and learn the mapping of CLIP image features, i.e., pseudo-text features, to the latent space by image reconstruction. That means training using image data, as shown in Fig. 2. Meanwhile, during testing, the target text description is input and text-guided image generation is performed utilizing CLIP text features, which is shown in Fig. 3. Through the above process, a solution for text-guided image generation without text training is achieved.

B. Loss Functions

We utilize several objective functions, i.e., reconstruction loss, perceptual loss, adversarial loss, identity loss, and diversity loss to optimize our framework.

1) Reconstruction Loss: For a given image \(x\), the reconstructed image \(x'\) is generated by our framework. We develop a reconstruction loss to guarantee the pixel alignment between
x and x’. The reconstruction loss is formulated as follows,
\[ L_{\text{rec}} = \|x - x'\|_2, \]
where \(\|\cdot\|_2\) denotes the \(l_2\) distance.

2) Perceptual Loss: The reconstruction loss using \(l_2\) distance assumes that the data fit a Gaussian distribution, which leads to producing smoother images. We introduce a perceptual loss, i.e., LPIPS loss [69] to measure the difference between the source image x and the generated image x’ to improve the smoothing problem caused by the reconstruction loss. LPIPS learns to reconstruct the reverse mapping of x from x’ and prioritizes their perceptual similarity, which is formulated as follows,
\[ L_{\text{LPIPS}} = \sum_i \frac{1}{H_i W_i} \sum_k \|f_i[k, w] \odot (y_{i,hw}^d - y_{i,hw}^t)^2\],
where \(y^d\) and \(y^t\) are the feature extracted from the \(l\)-th layer of a pre-trained AlexNet [70].

3) Adversarial Loss: The discriminator is trained adversarially with the mapping network which is considered as a generator. For generator \(G()\) and discriminator \(D()\), we use the WGAN-GP losses [71], [72], which are formulated as follows,
\[ L_G = E_{x}[\log(1 - D(x'))], \]
\[ L_D = -[E_{x}[\log(D(x))] + E_{x'}[\log(1 - D(x'))]] + \lambda E_{x}||\nabla_x D(x)||_2 - 1^2\],
where x and x’ denote the source image and reconstructed image, respectively.

4) Identity Loss: To ensure that the identity features of the faces are unchanged, we present an identity loss. Using an effective arcface model [73], our identity loss calculates the similarity of the identity features of the faces of x and x’, which is formulated as follows,
\[ L_{\text{id}} = 1 - \frac{f_{\text{arc}}(x) \cdot f_{\text{arc}}(x')}{[f_{\text{arc}}(x) \cdot f_{\text{arc}}(x)][f_{\text{arc}}(x') \cdot f_{\text{arc}}(x')]} \],
where \(f_{\text{arc}}()\) denotes an arcface classifier.

5) Diversity Loss: As shown in Fig. 2, in the multi-modal embedding space of CLIP the standard normally distributed noise of \(\mu = 0, \sigma^2 = 1\) is added to the obtained CLIP image feature \(f_{\text{img}}\) to derive the image feature \(f_{\text{img}1}\). The reconstructed images x’, x’’ are generated respectively, and then the diversity loss is formulated as follows,
\[ L_{\text{div}} = \frac{d_f(f_{\text{img}}, f_{\text{img}1})}{d_f(x', x'')} \],
where \(d_f(\cdot, \cdot)\) denotes the distance calculation and we use \(l_1\) distance.

6) Overall Loss: The overall loss for CLIP2GAN is the weighted summation of the above losses, which is formulated as follows,
\[ \min_{M} L_M = L_{\text{rec}} + \lambda_{\text{LPIPS}} L_{\text{LPIPS}} + \lambda_G L_G + \lambda_{\text{id}} L_{\text{id}} + \lambda_{\text{div}} L_{\text{div}} \],
where \(\lambda_{\text{LPIPS}}, \lambda_G, \lambda_{\text{id}}\) and \(\lambda_{\text{div}}\) are the trade-off parameters balancing different losses.

C. Text-Guided Image Editing

With the pre-trained CLIP2GAN model, we further apply the network to text-guided image editing applications. Given a source image x, we are interested in editing certain regions of it by manipulating its latent codes \(z = [z_i]_{i=1}^{18}\) to \(z' = [z'_i]_{i=1}^{18}\) and getting a target image x’ that meets the editing requirements, which is expressed as follows,
\[ z'_i = z_i + \beta n_i, \]
\[ x' = G(z'), \]
where \(n = [n_i]_{i=1}^{18}\) corresponds to the normal direction of a particular semantics of the latent space, and \(\beta\) denotes the degree of editing semantics. That is, if the latent code moves in a certain direction, the semantics contained in the output image should vary accordingly. This requires our framework to locate the semantic direction \(n\) of the text and the latent code \(z\) of the image.

As shown in Fig. 4, the text \(t\), i.e., simple descriptions on age, gender, hair, expression, etc., is fed into the CLIP [22] text encoder to get CLIP text features. Then the vector \(n\) in the StyleGAN latent space is derived by the pre-trained mapping network, which is considered as the normal direction of a particular semantics due to the mapping network bridging the CLIP feature space and the StyleGAN latent space. Meanwhile, by putting the source image x through the CLIP image encoder and the mapping network, the representation \(z\) of \(x\) in the StyleGAN latent space is obtained, which are as follows,
\[ n = M[C_{\text{text}}(t)], \]
\[ z = M[C_{\text{img}}(x)], \]
where $C_{\text{text}}(\cdot)$ and $C_{\text{img}}(\cdot)$ denote the CLIP text and image encoder, respectively.

Finally, the semantic direction $\mathbf{n}$ and the latent feature $\mathbf{z}$ of StyleGAN are weighted together. The weight of $\mathbf{n}$, i.e., $\beta$, ranges from 0 to 1, leading to the modified feature $\mathbf{z}'$ in the latent space of StyleGAN, and thus StyleGAN generates the corresponding images. This allows the text description to control the degree of semantic modification of the image while not changing the identity features of the image itself. Our approach achieves high-quality text-guided image editing by simple arithmetic operations without optimization and additional network structures.

IV. EXPERIMENTS

A. Experiments Setup

1) Datasets: To demonstrate the superiority of our method for text-guided face generation and face editing, we train on the CelebA-HQ [34] dataset and test using text descriptions from the Multi-modal CelebA-HQ (MM-CelebA-HQ) [27] dataset. The CelebA-HQ dataset is a high-quality version of the CelebA [75] dataset, consisting of 30,000 images with a resolution of $1024^2$. We learn the image reconstruction in the CelebA-HQ dataset. The MM-CelebA-HQ dataset creates 10 unique text descriptions for each image in CelebA-HQ. Using the MM-CelebA-HQ dataset, we can test the effectiveness of the text-guided face image generation of our method.

Besides, we have conducted experiments on AFWH [35] dataset and LSUN [36] dataset. The AFWH dataset is an animal faces dataset consisting of 15,000 high-quality images at $512^2$ resolution, which includes three domains of cat, dog, and wildlife, each providing 5000 images. We choose the cat and dog datasets of AFWH for training. The LSUN dataset contains around one million labeled images for each of the 10 scene categories and 20 object categories. We choose the car and church datasets of LSUN for training. For these two datasets, the identity loss of faces is removed. The experiments on AFWH and LSUN datasets also validate the superior performance of our method with the generality of its text-guided image generation function.

2) Implementation Details: The hyper-parameters in the framework are set as follows: The trade-off parameter $\lambda_{\text{rec}}, \lambda_{\text{LPIPS}}, \lambda_{\text{G}}, \lambda_{\text{id}}, \lambda_{\text{div}}$ and $\lambda$ are set to 1, 1, 0.1, 1, 1 and 1, respectively, to ensure the training stability. For the whole framework, we utilize Adam optimizer [76]. The training lasts 100 epochs in total. The learning rate is set to $2 \times 10^{-3}$ and linearly reduces after 50 epochs. The whole framework is implemented by Pytorch and we perform experiments on NVIDIA RTX 3090.

3) Evaluation Metrics: We aim to assess visual quality, image accuracy, and realism for evaluation. The visual quality of generated or manipulated images is evaluated through the widely-used Frchet Inception Distance (FID) [77] metrics. FID measures the distance between two sets of images, computed by the mean value and covariance of the generated image set $(\mu_Y, \Sigma_Y)$ and the ground-truth image set $(\mu_{\hat{Y}}, \Sigma_{\hat{Y}})$, which is formulated as follows,

$$\text{FID}(Y, \hat{Y}) = \|\mu_Y - \mu_{\hat{Y}}\|^2 + \text{tr}(\Sigma_Y + \Sigma_{\hat{Y}} - 2(\Sigma_Y \Sigma_{\hat{Y}})^{1/2}).$$

(15)

To evaluate the perceptual similarity between generated images and real images, we compute the average distance between them by the Learned Perceptual Image Patch Similarity (LPIPS) [69] metrics, which is a weighted perceptual similarity between two images, computing on the features extracted from a pre-trained network.

In addition, accuracy and realism are evaluated through a user study. For image generation, the accuracy is evaluated by the similarity between the text and the generated image. For image manipulation, accuracy is assessed by whether the visual properties of the modified image are aligned with the given description and whether contents unrelated to the text are preserved. Realism is required to be judged as to which is more realistic and consistent with reality. We tested accuracy and realism by collecting surveys on a random sample of 50 images from 20 people.

B. Comparison With State-of-the-Art Methods

We compare our method with several challenging methods in text-guided image generation, i.e., GAN-based methods, AttnGAN [26], ControlGAN [74], DF-GAN [29], DM-GAN [32], TediGAN [27], StyleCLIP [33], LAFITE [58], FuseDream [59], CLIP-GEN [78], and diffusion-based methods, DALL-E [31] and Stable Diffusion [61]. We evaluate FID and LPIPS on a large number of samples generated from randomly selected text descriptions and assess Accuracy and Realism by user studies. To compare with some GAN-based methods that require image-text pairs for training, we use BLIP [79] to produce pseudo text label for CAT of AFWH [35], DOG of AFWH [35] and CHURCH of LSUN [36] dataset. For large-scale pre-trained text-to-image diffusion models, i.e., Stable Diffusion and DALL-E, we directly apply their pre-trained model or API for testing. The quantitative results are shown in Tab. I. From the table, it is observed that our method achieves state-of-the-art performance on four datasets.

Furthermore, we make a qualitative evaluation with several competitive methods, i.e., AttnGAN, ControlGAN, DF-GAN, DM-GAN, TediGAN, StyleCLIP, LAFITE, FuseDream, CLIP-GEN, DALL-E and Stable Diffusion in Fig. 5. It is observed that GAN-based methods based on text-image pairs, e.g., AttnGAN, DM-GAN, and TediGAN, produce unrealistic results; text-free methods based on CLIP, e.g., LAFITE, FuseDream and CLIP-GEN, fail to create images that fully match the complex description, e.g., “curly hair” in CLIP-GEN in the first line and “wide noise” in FuseDream in the second line. Diffusion-based models, i.e., DALL-E and Stable Diffusion, generate images with few details whose scale differs. This is because these models are focused on the specific class, i.e., human faces, and their ability on human faces generalize from the open-set image generation. Different from these methods, the images by CLIP2GAN correspond very closely to the text description and exhibit fine-grained details. Benefiting from the CLIP model [22], we obtain an effective text-driven
TABLE I

QUANTITATIVE COMPARISON WITH EXISTING TEXT-GUIDED IMAGE GENERATION METHODS ON FID AND LPIPS METRICS ON MM-CELEBA-HQ [27] DATASET, THE CAT AND DOG DATASETS OF AFHQ [35], AND THE CHURCH DATASETS OF LSUN [36].

| Method               | MM-CelebA-HQ | CAT of AFHQ | DOG of AFHQ | CHURCH of LSUN |
|----------------------|--------------|-------------|-------------|----------------|
|                      | FID          | LPIPS       | FID          | LPIPS          | FID           | LPIPS         |
| AttnGAN [26]         | 125.98       | 0.512       | 203.46       | 0.587          | 199.28        | 0.579         | 249.34        | 0.595         |
| ControlGAN [74]      | 116.32       | 0.522       | 189.61       | 0.580          | 191.49        | 0.592         | 242.56        | 0.604         |
| DF-GAN [29]          | 137.60       | 0.581       | 178.44       | 0.607          | 192.78        | 0.612         | 246.12        | 0.615         |
| DM-GAN [32]          | 131.05       | 0.544       | 182.36       | 0.574          | 185.24        | 0.584         | 209.11        | 0.591         |
| TediGAN [27]         | 106.37       | 0.456       | 182.62       | 0.532          | 180.61        | 0.531         | 194.92        | 0.549         |
| StyleCLIP [33]       | 101.75       | 0.439       | 123.11       | 0.518          | 126.76        | 0.529         | 147.86        | 0.542         |
| LAFITE [58]          | 72.95        | 0.502       | 78.42        | 0.495          | 75.28         | 0.493         | 86.72         | 0.516         |
| FuseDream [59]       | 84.36        | 0.517       | 75.88        | 0.511          | 77.27         | 0.508         | 84.39         | 0.519         |
| CLIP-GEN [33]        | 78.26        | 0.521       | 71.37        | 0.518          | 73.24         | 0.509         | 84.41         | 0.525         |
| DALL-E [31]          | 98.65        | 0.527       | 84.35        | 0.503          | 87.12         | 0.519         | 114.87        | 0.551         |
| Stable Diffusion [61]| 102.36       | 0.569       | 87.47        | 0.543          | 84.38         | 0.559         | 121.22        | 0.573         |
| Ours                 | 34.25        | 0.408       | 62.33        | 0.491          | 68.17         | 0.487         | 82.64         | 0.516         |

Fig. 5. Qualitative comparison with existing generation methods, i.e., AttnGAN [26], ControlGAN [74], DF-GAN [29], DM-GAN [32], TediGAN [27], StyleCLIP [33], LAFITE [58], FuseDream [59], CLIP-GEN [33], DALL-E [31], and Stable Diffusion [61]. Our generated images using the text on the left show superior performance on fidelity and quality.

capability to generate face images with more text features in the category without being limited to text descriptions in the dataset. We also conduct a user study to evaluate the visual performance of CLIP2GAN. The voting results are reported in Tab. II. It can be observed that our method is clearly preferred over the competitors more than 60% of the time.

When some specific features in the text are changed, our model ensures that the image is modified only in the corresponding features, while other features, including identity features, are guaranteed to remain invariant. This shows that our method is able to decouple different features with excellent robustness. The text description and visual results are presented in Fig. 7. It should be noted that we add constraint words for the background to further demonstrate a certain decoupling ability between the identity features and the background.

The other advantage is that our model can inherently generate diverse results given an arbitrary specific text description. With our approach, the generated images are guaranteed to be consistent with the given text features while other irrelevant features are varied to get multiple unique results. We present the text-guided image generation results in Fig. 8 and Fig. 6.
TABLE II
PAIR OF USER STUDY BETWEEN OUR METHOD AND EXISTING TEXT-GUIDED IMAGE GENERATION METHODS. OUR METHOD OUTPERFORMS PREVIOUS ALGORITHMS ON ACC. AND REAL. METRICS

| Method                  | Acc. (%) | Real. (%) |
|-------------------------|----------|-----------|
| Ours v.s. AttnGAN [26]  | 84.3     | 86.1      |
| Ours v.s. ControlGAN [74] | 81.7     | 82.5      |
| Ours v.s. DfGAN [29]    | 83.9     | 86.5      |
| Ours v.s. DS-GAN [32]   | 83.8     | 93.5      |
| Ours v.s. TedGAN [27]   | 72.2     | 71.9      |
| Ours v.s. StyleCLIP [33] | 68.3     | 94.6      |
| Ours v.s. LAFITE [58]   | 61.4     | 76.6      |
| Ours v.s. FuseDream [59] | 71.7     | 80.2      |
| Ours v.s. CLIP-GEN [76] | 75.6     | 68.3      |
| Ours v.s. DALL-E [31]   | 61.5     | 91.4      |
| Ours v.s. Stable Diffusion [61] | 64.8 | 95.3 |

A middle-aged man with short black hair in a white background. His eyes are dark. dark  → short  → man
black → blond  → woman

middle-aged  → old  → He has beard.
black  → blond

Fig. 7. Generated result after modifying the text. The text on the top left is used as input to generate the image on the bottom left. By modifying the words marked in the text or adding descriptions of expressions as shown in the right half, the image generated from the newly input text only changes the corresponding features.

Fig. 8. Diverse generation results from our framework. It is observed that our method can generate diverse results with high quality.

which demonstrates that our method can generate diverse results with high quality.

C. Ablation Studies

There are several ablation experiments performed to demonstrate the effectiveness of the framework. To evaluate the effectiveness and necessity of the network design, we modified the mapping network with a different number of network layers and pre-trained models of CLIP. From Tab. III, it is observed that the framework performs worse in several metrics if the number of network layers is less. Whereas, with a higher number of layers, it is less significant for performance improvement, while greatly increasing the complexity of the network. In addition, the feature space of CLIP has semantic meanings between images and text, thus using a stronger joint space (ViT/B-16) can improve the generated results.

We also investigate the impact of each component in the objective function. Ensuring that reconstruction loss, perceptual loss, and identity loss are preserved, we ablate by excluding adversarial loss and diversity loss one by one. The results are shown in Tab. IV, and the performance decreases after each removal. This is because removing the adversarial loss loses image realism while removing the diversity loss leads to pattern convergence. In addition to the quantitative analysis, we specifically show the effect of diversity loss on the qualitative generation results. As shown in Fig. 9, we obtain a fixed and average result without the diversity loss. With diversity loss, CLIP2GAN generates diverse images that satisfy the text attributes.

Furthermore, we conduct experiments with different loss weights. In this paper, we mainly use reconstruction loss, perceptual loss, adversarial loss, identity loss, and diversity loss, where reconstruction loss and perceptual loss both control the pixel similarity of images and can have the same proportional weight \( \lambda_1 \). The weights of adversarial loss, identity loss, and diversity loss are \( \lambda_{adv} \), \( \lambda_{id} \), and \( \lambda_{div} \), respectively. Results are shown in Tab. VI and Fig. 10. \( \lambda_1 \) is the joint weight of reconstruction loss and perceptual loss, which determines the quality and fidelity. \( \lambda_{adv} \) encourages the network to improve the generated image realism. \( \lambda_{id} \) benefits for identity
Fig. 9. Some qualitative ablation studies for diversity loss. Input the text description on the left to get the result on the right before and after adding diversity loss. Diverse and high-quality images are generated.

| Model          | Generation | Editing |
|----------------|------------|---------|
|                | FID | LPIPS | Operating Space | FID | LPIPS |
| DCGAN [80]     | 183.37 | 0.562 | CLIP [22]      | n/a | n/a |
| StyleGAN2 [10] | 34.25 | 0.408 | CLIP [22]      | 196.76 | 0.583 |
|                |      |       | StyleGAN       | 60.54 | 0.505 |
|                |      |       | DCGAN          | 35.66 | 0.417 |

Table V

Impact of Different Loss Weights on the Generated Results

| \(\lambda_1\) | \(\lambda_{div}\) | \(\lambda_{l1}\) | \(\lambda_{div}\) |
|---------------|-----------------|-----------------|-----------------|
| 2             | 0.1             | 1               | 1               |
| 0.5           | 0.5             | 1               | 1               |
| 0.1           | 0.02            | 1               | 2               |
| 0.1           | 0.5             | 1               | 1               |
| 0.1           | 0.1             | 1               | 1               |

FID | 41.18 | 55.46 | 151.72 | 47.25 | 63.43 | 38.69 | 43.51 | 36.87 | 34.25

Table VI

consistency. \(\lambda_{div}\) enables the network to generate diverse images of high quality that match the textual features.

Finally, we provide additional ablation studies on using different pre-trained GANs in our framework and different locations for arithmetic operations during editing. To demonstrate that the mapping network in CLIP2GAN can map the feature space of CLIP [22] to the latent space of any GAN, we replace StyleGAN [9] in the framework with DCGAN [80] and perform learning. The text-guided generated images look blurry and of poor quality because of the weak capability of DCGAN, but they are still able to satisfy the requirements of the attributes mentioned in the text, which indicates our model is effective for the latent space of any GAN. For image editing, we choose to perform arithmetic operations on text features and image features at different locations, i.e., the feature space of CLIP, and the latent space of StyleGAN. Better editing results can be obtained by performing arithmetic operations in the StyleGAN space. As shown in Tab. V, we demonstrate the effectiveness of our model to generate and edit with high quality and high fidelity.

D. Image Editing

To evaluate the performance of image editing, we compare CLIP2GAN with several challenging GAN-based methods, i.e., ManiGAN [64], TediGAN [27], StyleCLIP [33], HairCLIP [66] and StyleMC [67] and diffusion-based methods, i.e., InstructPix2Pix [81] and Imagic [82]. From Tab. VII, it is observed that CLIP2GAN achieves state-of-the-art performance on the FID metric on both CelebA-HQ dataset and arbitrary images that are not within the CelebA-HQ dataset. Compared with GAN-based methods, CLIP2GAN achieves a remarkable improvement to the most challenging method, i.e., AFHQ and LSUN datasets. Compared with diffusion-based models, i.e., InstructPix2Pix and Imagic, CLIP2GAN also obtains some improvement on FID metrics. Noticeable, these diffusion-based models are of great parameters and pre-trained on large-scale datasets, which endow them with strong generalization capabilities. Different from them, CLIP2GAN trains a light-weighted mapping network with low time consumption. However, CLIP2GAN still obtains better performance compared with these diffusion-based models by leveraging CLIP and StyleGAN, which further demonstrates the effectiveness of our CLIP2GAN.

Furthermore, we show the qualitative comparison results with other methods on CelebA-HQ. As shown in Fig. 11, for GAN-based methods, TediGAN is difficult to guarantee identity feature invariance and reduces quality; StyleCLIP lacks rationality in its generation results; HairCLIP is limited by its ability to edit only hair; and StyleMC produces some bias in semantic understanding. For diffusion-based methods, i.e., InstructPix2Pix and Imagic, because they are not specially designed for face editing, they do not understand the face features well, such as unreasonable makeup, changing inconspicuous hair, and unrealistic teeth in the figure. Compared to these methods, CLIP2GAN has higher fidelity and better generation quality while keeping the identity attribute. We also conduct a user study to evaluate the visual performance of CLIP2GAN. Volunteers are asked to select the best one from the samples of all methods. The voting results are reported in Tab. VII. It can be observed that our method is selected more than twice as often as our competitors.

In addition, we show the visual results of our text-guided image editing in Fig. 13. With simple word descriptions, we are able to apply different manipulations to the image without changing other irrelevant features. We also present editing results of other datasets, i.e., AFHQ and LSUN datasets. Fig. 12 shows the results of text-guided image editing on different datasets. The edited results we achieve are diverse, high fidelity, and high quality. For the various text features used for...
TABLE VII

| Method          | CelebA-HQ | Not-in-CelebA-HQ |
|-----------------|-----------|------------------|
|                 | FID ↓     | Acc. (%) ↑       | Real. (%) ↑    |
| ManiGAN [64]    | 117.89    | 4.5              | 2.0           |
| TediGAN [27]    | 107.25    | 7.0              | 9.5           |
| StyleCLIP [33]  | 86.82     | 8.0              | 8.5           |
| HairCLIP [66]   | 71.44     | 12.0             | 10.5          |
| StyleMC [67]    | 51.81     | 11.5             | 12.0          |
| InstructPix2Pix [81] | 59.15   | 12.0             | 8.0           |
| Imagic [82]     | 60.26     | 10.5             | 8.5           |
| Ours            | 35.66     | 34.5             | 41.0          |
|                 | FID ↓     | Acc. (%) ↑       | Real. (%) ↑    |
|                 | 143.39    | 3.5              | 2.5           |
|                 | 135.47    | 5.0              | 11.5          |
|                 | 106.24    | 6.0              | 7.5           |
|                 | 94.25     | 9.0              | 13.0          |
|                 | 56.34     | 10.5             | 10.5          |
|                 | 58.77     | 15.5             | 9.5           |
|                 | 59.49     | 16.5             | 8.0           |
|                 | 48.14     | 34.0             | 37.5          |

Fig. 11. Qualitative comparison with existing methods, i.e., TediGAN [27], StyleCLIP [33], HairCLIP [66], StyleMC [67], InstructPix2Pix [81], and Imagic [82]. Inputting images and editing text, our editing images show superior performance on fidelity and quality.

Fig. 12. Qualitative results of text-guided image editing on CAT of AFHQ [35] dataset, DOG of AFHQ [35] dataset, and CHURCH of LSUN [36] dataset. Given an arbitrary image on the left, we are able to edit the image according to the different textual descriptions provided above, without changing other irrelevant attributes, which are shown on the right.

CLIP is trained on a massive dataset of image-text alignment, which possesses a huge text latent space. Therefore, using the pre-trained CLIP, our method can achieve text editing for
achieves superior performance and shows better visual results. Compared to previous methods, our framework provides control on the modification degree.

Fig. 13. Visual results of text-guided image editing. Inputting the image on the left column and the text description above each of the other columns allows editing of the input image.

Fig. 14. Varying degrees of text-guided image editing results. Notably, our method provides control on the modification degree.

numerous different features without being limited to some specific text descriptions.

V. CONCLUSION

In this paper, we investigate vision-language models (CLIP) to generative models (StyleGAN) for the task of text-guided image generation and propose a novel framework named CLIP2GAN. Specifically, the framework bridges the pre-trained CLIP and StyleGAN and implements training in a text-free way, where the trained model generates high-fidelity and high-quality images corresponding to the text description. With the use of CLIP2GAN, we also achieve text-guided image manipulation that allows the editing of real images. Extensive experiments demonstrate the effectiveness of our method. Compared to previous methods, our framework achieves superior performance and shows better visual results.

REFERENCES

[1] I. J. GoodFellow et al., “Generative adversarial networks,” in Proc. NeurIPS, 2014, pp. 1–9.
[2] E. Richardson et al., “Encoding in style: A styleGAN encoder for image-to-image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2287–2296.
[3] J. Huang, J. Liao, and S. Kwong, “Unsupervised image-to-image translation via pre-trained StyleGAN2 network,” IEEE Trans. Multimedia, vol. 24, pp. 1435–1448, 2022.
[4] S. Xu, D. Liu, and Z. Xiong, “E2E: Generative inpainting from edge to image,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 4, pp. 1308–1322, Apr. 2021.
[5] R. Abdal, Y. Qin, and P. Wonka, “Image2StyleGAN++: How to edit the embedded images?” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8293–8302.
[6] Y. Cheng, C. H. Lin, H. Lee, J. Ren, S. Tulyakov, and M. Yang, “InOut: Diverse image outpainting via GAN inversion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 11421–11430.
[7] Y. Hu, Y. Wang, and J. Zhang, “DEAR-GAN: Degradation-aware face restoration with GAN prior,” IEEE Trans. Circuits Syst. Video Technol., early access, Feb. 14, 2023, doi: 10.1109/TCSVT.2023.3244786.
[8] Q. Duan, L. Zhang, and X. Gao, “Simultaneous face completion and frontalization via mask guided two-stage GAN,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 6, pp. 3761–3773, Jun. 2022.
[9] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4396–4405.
[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8107–8116.
[11] T. Karras et al., “Alias-free generative adversarial networks,” in Proc. NeurIPS, vol. 34, 2021, pp. 852–863.
[12] F. Peng, L. Yin, and M. Long, “BDC-GAN: Bidirectional conversion between computer-generated and natural facial images for anti-forensics,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 10, pp. 6657–6670, Oct. 2022.
[13] R. Abdal, Y. Qin, and P. Wonka, “Image2StyleGAN: How to embed images into the StyleGAN latent space?” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4431–4440.
[14] D. Bau et al., “Inverting layers of a large generator,” in Proc. ICLR Workshop, vol. 2, no. 3, 2019, p. 4.
[15] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, “In-domain GAN inversion for real image editing,” in Proc. ECCV. Cham, Switzerland: Springer, 2020, pp. 592–608.
[16] Y. Shen, C. Yang, X. Tang, and B. Zhou, “InterFaceGAN: Interpreting the latent space of GANs for semantic face editing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9240–9249.
[17] Y. Shen, C. Yang, X. Tang, and B. Zhou, “InterFaceGAN: Interpreting the disentangled face representation learned by GANs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 4, pp. 2004–2018, Apr. 2022.
[18] Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of GANs for semantic face editing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9240–9249.
[19] E. Richardson et al., “Encoding in style: A styleGAN encoder for image-to-image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2287–2296.
[20] P. Zhou, L. Xie, B. Ni, L. Liu, and Q. Tian, “HRInversion: High-resolution GAN inversion for cross-domain image synthesis,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 5, pp. 2147–2161, May 2023.
[21] X. Hou, X. Zhang, Y. Li, and L. Shen, “TextFace: Text-to-style mapping based face generation and manipulation,” IEEE Trans. Multimedia, early access, Mar. 17, 2022, doi: 10.1109/TMM.2022.3160360.
[22] A. Radford et al., “Learning transferable visual models from natural language supervision,” in Proc. ICML, 2021, pp. 8748–8763.
Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3730–3738.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

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 Proc. NeurIPS, vol. 30, 2017.

Z. Wang, W. Liu, Q. He, X. Wu, and Z. Yi, “CLIP-GEN: Language-free training of a text-to-image generator with CLIP,” 2022, arXiv:2203.00386.

J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in Proc. ICML, 2022, pp. 12888–12900.

A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2015, arXiv:1511.06434.

T. Brooks, A. Holynski, and A. A. Efros, “InstructPix2Pix: Learning to follow image editing instructions,” in Proc. CVPR, Jun. 2023, pp. 18392–18402.

B. Kawar et al., “Imagic: Text-based real image editing with diffusion models,” in CVPR, 2023, pp. 6007–6017.

Yixuan Wang received the B.E. degree in computer science and technology from the University of Science and Technology of China (USTC), Hefei, China, in 2021, where she is currently pursuing the master’s degree in information and communication engineering with the Department of Electronic Engineering and Information Science. Her research interests include computer vision and deep generative models (e.g., GANs and diffusion models).

Wengang Zhou (Senior Member, IEEE) received the B.E. degree in electronic information engineering from Wuhan University, China, in 2006, and the Ph.D. degree in electronic engineering and information science from the University of Science and Technology of China (USTC), China, in 2011. From September 2011 to September 2013, he was a Post-Doctoral Researcher with the Computer Science Department, The University of Texas at San Antonio. He is currently a Professor with the EEIS Department, USTC. His research interests include multimedia information retrieval, computer vision, and computer game. In these fields, he has published over 100 articles in IEEE/ACM TRANSACTIONS and CCF Tier-A International Conferences. He was a winner of the National Science Funds of China (NSFC) for Excellent Young Scientists. He was a recipient of the Best Paper Award for ICIMCS 2012. He received the Award for the Excellent Ph.D. Supervisor of Chinese Society of Image and Graphics (CSIG) in 2021 and the Award for the Excellent Ph.D. Supervisor of Chinese Academy of Sciences (CAS) in 2022. He won the First Class Wu-Wenjun Award for Progress in Artificial Intelligence Technology in 2021 and the 2021 ICME Outstanding Service Award. He served as the Publication Chair for IEEE ICME 2021. He is currently an Associate Editor and a Lead Guest Editor of IEEE TRANSACTIONS ON MULTIMEDIA.

Jiannin Bao received the Ph.D. degree from the University of Science and Technology of China in 2019. He is a Senior Researcher with Microsoft Research Asia. He has published more than 20 papers in top conferences and journals, such as CVPR, ICCV, ECCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, and NeurIPS. His research interests include large-scale pretraining, general representation learning, and deep generative models. He is a reviewer of many top conferences (including CVPR, ICCV, ECCV, AAAI, NeurIPS, PRCV, ICME, and ACCV) and top journals (including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and MVA). He has reviewed more than 100 papers.

Weiulin Wang (Graduate Student Member, IEEE) received the B.E. degree in electronic information engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2019, where he is currently pursuing the Ph.D. degree in information and communication engineering with the Department of Electronic Engineering and Information Science. His research interests include computer vision and deep generative models (e.g., GANs and denoising diffusion probabilistic models).

Li Li (Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China (USTC), Hefei, Anhui, China, in 2011 and 2016, respectively. He was a Visiting Assistant Professor with the University of Missouri–Kansas City, from 2016 to 2020. He joined the Department of Electronic Engineering and Information Science, USTC, as a Research Fellow in 2020 and became a Professor in 2022. His research interests include computer vision and deep generative models (e.g., GANs and denoising diffusion probabilistic models).

Houqiang Li (Fellow, IEEE) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively. He is a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China. He has authored or coauthored over 200 articles in journals and conferences. His research interests include multimedia search, image/video analysis, video coding, and communication. He was a winner of the National Science Funds (NSFC) for Distinguished Young Scientists, the Distinguished Professor of the Changjiang Scholars Program of China, and the Leading Scientist of the Ten Thousand Talent Program of China. He was a recipient of the National Technological Invention Award of China (second class) in 2019, the National Natural Science Award of China (second class) in 2015, the Best Paper Award for VCIP 2012, the Best Paper Award for ICIMCS 2012, and the Best Paper Award for ACM MUM in 2011. He served as the TPC Co-Chair for VCIP 2010 and the General Co-Chair for ICME 2021. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, from 2010 to 2013.