

# Multimodal Image Synthesis and Editing: A Survey

Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu<sup>§</sup>, Lingjie Liu  
 Adam Kortylewski, Christian Theobalt, Eric Xing, *Fellow, IEEE*

**Abstract**—As information exists in various modalities in real world, effective interaction and fusion among multimodal information plays a key role for the creation and perception of multimodal data in computer vision and deep learning research. With superb power in modelling the interaction among multimodal information, multimodal image synthesis and editing has become a hot research topic in recent years. Instead of providing explicit guidance for network training, multimodal guidance offers intuitive and flexible means for image synthesis and editing. On the other hand, this field is also facing several challenges in alignment of multimodal features, synthesis of high-resolution images, faithful evaluation metrics, etc. In this survey, we comprehensively contextualize the advance of the recent multimodal image synthesis and editing and formulate taxonomies according to data modality and model architectures. We start with an introduction to different guidance modalities in image synthesis and editing. We then describe multimodal image synthesis and editing approaches extensively according to their network architectures. This is followed by a description of benchmark datasets & evaluation metrics and corresponding experimental results. Finally, we provide insights about the current research challenges and possible directions for future research. A project associated with this survey is available at <https://github.com/fnzhan/MISE>.

**Index Terms**—Multimodality, Image Synthesis & Editing, GAN, Auto-regressive Models, Diffusion Models, NeRF.

## 1 INTRODUCTION

HUMANS are naturally capable of imaging a scene according to a piece of visual, text or audio description. However, the underlying processes are not that straightforward to deep neural networks due to the inherent modality gap. This *modality gap* for visual perception can be boiled down to *intra-modal gap* between visual clues and real images and *cross-modal gap* between non-visual clues and real images. Targeting to mimic human imagination and creativity in the real world, the tasks of Multimodal Image Synthesis and Editing (**MISE**) provide profound insights about how deep neural networks correlate multimodal information with image attributes.

Image synthesis and editing aims to create realistic images or edit real images with natural textures. In the last few years, it has witnessed very impressive progress thanks to the advance of deep learning especially Generative Adversarial Networks (GANs) [1]. To achieve more controllable generation, a popular line of research focuses on generating and editing images conditioned on certain guidance. Typically, visual clues, such as segmentation maps and image edge, have been widely adopted for image synthesis and editing [2]–[4]. Beyond these intra-modal guidance of visual clues, cross-modal guidance such as texts, audios, and scene graph provides an alternative but often more intuitive and flexible way of expressing visual concepts. However,

effective retrieval and fusion of heterogeneous information from data of different modalities remains a big challenge in image synthesis and editing.

As a pioneering effort in multimodal image synthesis, [5] shows that recurrent variational auto-encoder could generate novel visual scenes conditioned on image captions. The research of multimodal image synthesis is then greatly advanced with the prosperity of Generative Adversarial Networks (GANs) [1]–[3], [6]–[11]. For example, Isola *et al.* [2] investigate conditional adversarial networks [6] as a generic approach to synthesize images from semantic maps or edge maps. Reed *et al.* [12] further extend conditional GANs [6] to generate natural images based on textual descriptions. Sont *et al.* [13] propose a conditional recurrent generation network to achieve audio-driven talking face generation. In the last few years, the GAN-based methods have achieved notable improvements thanks to the improved multimodal encodings [14], [15], novel architectures [16], [17], and cycle structure [18], [19], etc. With the prosperity of large scale GANs, a bunch of works, such as BigGAN [20] and StyleGAN [21]–[23], have been developed to synthesize images with high quality and diversity from random noise input. As recent studies show that rich semantic information is encoded in the latent space of GANs [24], GAN inversion [25], [26] is introduced to invert a given image back into the latent space of a pretrained GAN, yielding an inverted code that can faithfully reconstruct the given image via the generator. Since GAN inversion allows to control attribute directions in latent spaces, pre-trained GANs become applicable to real image editing, without requiring ad-hoc supervision or cumbersome model training. Specifically, TediGAN [27] maps paired images and texts into the common latent space of StyleGAN to achieve text-driven image generation. With a style-mixing mechanism, it also supports multi-modal

- F. Zhan is with the Max Planck Institute for Informatics, Germany and the S-Lab, Nanyang Technological University, Singapore.
- Y. Yu, R. Wu, J. Zhang and S. Lu are with the Nanyang Technological University, Singapore.
- L. Liu, A. Kortylewski and C. Theobalt are with the Max Planck Institute for Informatics, Germany.
- E. Xing is with the Carnegie Mellon University, USA and the Mohamed bin Zayed University of Artificial Intelligence, UAE.
- <sup>§</sup> denotes corresponding author, E-mail: shijian.lu@ntu.edu.sg.

image editing. Leveraging the power of recent Contrastive Language-Image Pre-training (CLIP) [28], StyleClip [29] achieves text-driven image manipulation without requiring cumbersome manual effort on discovering the desired latent direction or a paired collection of texts and images.

Currently, a CNN architecture is still widely adopted in GANs, which hinders GANs from handling multimodal data in a unified manner. With the prevalence of Transformer model [30] which naturally allows cross-modal input, impressive improvements have been achieved in different modality, such as language models [31], image generative pre-training [32], and audio generation [33]. These recent advances fueled by Transformer suggest a possible route for auto-regressive Transformer [31] in multimodal image synthesis and editing by accommodating the long-range dependency of sequences. Specifically, leveraging discrete VAE [34] to learn discrete compressed and discrete image representation, DALL-E [35] demonstrates that training a large-scale auto-regressive Transformer on numerous image-text pairs can produce a high-fidelity generative model with controllable results through text prompts. By introducing a VQGAN with perceptual loss [36], [37] to learn compressed yet rich image representation, Taming Transformer [38] demonstrates the effectiveness of combining the expressivity of auto-regressive Transformers and the inductive bias of CNNs in high-resolution image synthesis. As auto-regressive models suffer from the well-known exposure bias, ImageBART [39] proposes to tackle it by learning to invert a multinomial diffusion process with the introduction of contextual information. Powered by a 3D Transformer encoder-decoder framework and a 3D Nearby Attention (3DNA) mechanism, the recent work NUWA [40] allows to generate or manipulate visual data (i.e., images and videos) with a unified multimodal pre-trained model.

Recently, diffusion models have emerged as a popular line of likelihood-based generative models which possess several desirable properties like stationary training objective and easy scalability. Specifically, Dhariwal *et al.* [41] show that diffusion models [42], [43] can achieve superior image synthesis quality compared with the current state-of-the-art generative models like GANs. With pre-trained diffusion models, Kim *et al.* [44] introduce a novel DiffusionCLIP model that performs text-driven image manipulation by inverting the diffusion models. Building on the power of diffusion models in high-fidelity image synthesis, the text-to-image generation are advanced significantly by the recent effort of DALL-E 2 [45] and Imagen [46].

Above methods mainly work for 2D images regardless the 3D essence of real world. With the recent advance of neural scene representation, such as Neural Radiance Fields (NeRF) [47], 3D-aware image synthesis and editing have attracted increasing attention from the community. Distinct from synthesis and editing on 2D images, 3D-aware MISE poses a bigger challenge as it has to consider the multi-view consistency during image synthesis or editing. Edit-NeRF [48] takes the first step to edit the shape and color of NeRF given user scribbles. However, due to its limited capacity in shape manipulation, it only allows to add or remove local parts of the object. Aiming to achieve more flexible and complicated manipulation, CLIP-NeRF [49] is introduced to edit NeRF in a more intuitive way by using a text prompt

or a single reference image; As an extension to audio-driven image editing, AD-NeRF [50] employs the feature of input audio to generate a dynamic neural radiance field, from which the corresponding high-fidelity talking-head images can be synthesized via volume rendering.

The contributions of this survey can be summarized in the following aspects:

- This survey covers extensive literature with regard to multimodal image synthesis and editing, and formulates existing methods in a rational and structured framework.
- We provide a foundation of different types of guidance modality underlying image synthesis & editing tasks and elaborate the specifics of encoding approach associated with the guidance modalities.
- By focusing on the cross-modal guidance in image synthesis and editing, we develop a taxonomy of the recent approaches according to the essential architectures and highlight the major strengths and weaknesses of existing methods.
- This survey provides an overview of various datasets and evaluation metrics in multimodal image synthesis and editing, and critically evaluate the performance of contemporary methods.
- We summarize the open challenges in the current research and share our humble opinions on promising areas and directions for future research.

The remainder of this survey is organized as follows. Section 2 presents the foundation of popular guidance modalities in image synthesis and editing. Section 3 provides a comprehensive overview and description of MISE methods with detailed pipelines. Section 4 reviews the popular datasets and evaluation metrics, with quantitative experimental results of typical methods. In Section 5, we discuss the main challenges and future directions for multimodal image synthesis and editing. Some social impacts analysis and concluding remarks are drawn in Section 6 and Section 7, respectively.

## 2 MODALITY FOUNDATIONS

Each source or form of information can be called a modality. For example, people have the sense of touch, hearing, sight, and smell; the medium of information includes voice, video, text, etc.; and data recorded by various sensors (e.g., radar, infrared, and accelerometer). Each above data form can be called a modality. In terms of image synthesis and editing, we group the guidance modality as visual guidance, text guidance, audio guidance, and other modalities. Detailed description of each modality together with related processing methods will be presented in the following subsections.

### 2.1 Visual Guidance

Visual guidance has attracted broad attention in image synthesis and editing thanks to its wide applications. Typically, visual guidance represents certain image properties in pixel space such as segmentation maps [2], [3], keypoints [55]–[57], rendered geometry [58]–[63], edge maps [64], [65], and scene layouts [66]–[70] as illustrated in Fig. 1. Besides, series of works also investigate image synthesis and editing conditioned on depth map [38], mouse trace [71], illumination

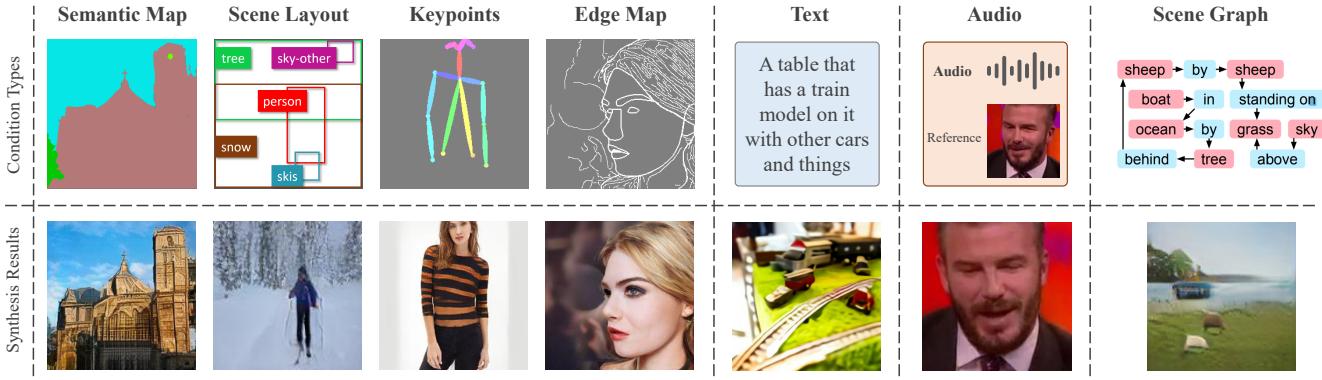


Fig. 1. Typical multimodal guidance in image synthesis and editing: The first row shows intra-modal (i.e., visual) guidance including semantic maps, scene layouts, keypoints and edge maps, and cross-modal guidance including text, audio and scene graph. The second row shows the corresponding image synthesis and editing (the sample images in the first four columns are from [51] and those in last three columns are from [52]–[54]).

map [72]–[75], etc. Realistic images can be naturally induced from visual guidance, as visual clues can be regarded as certain type of images which allow directly encoding with convolution layers for yielding the target generation or editing. Thanks to the accurate and clear guidance in visual information, visual guidance can be paired or unpaired with real images in image synthesis, e.g., paired image translation [2], [3] and unpaired image translation [18], [76].

By editing the visual guidance, such as semantic maps, image synthesis methods can be directly adapted for image manipulation tasks. In addition, visual guided image synthesis and editing can be applied in many low-level vision tasks. For example, we can achieve image colorization by putting grayscale images as visual guidance and the corresponding color images as ground truth. Other tasks like image super-resolution, image de-haze, image de-rain, etc., can be formulated in the similar way.

## 2.2 Text Guidance

Compared to visual guidance, such as edges and object masks, text prompt provides a more flexible way to express visual concepts. The text-to-image synthesis task aims to produce clear, photo-realistic images with high semantic relevance to the corresponding text guidance. This task is very challenging as text descriptions are often ambiguous and can lead to numerous images with correct semantics. In addition, images and texts come with heterogeneous features, which makes it hard to learn accurate and reliable mapping across the two modalities. Thus, learning an accurate embedding of text description plays an important role in text-guided image synthesis and editing.

**Text Encoding.** Learning useful encodings from textual representations is a non-trivial task [77]. There are a number of traditional text representations, such as Word2Vec [78] and Bag-of-Words [79]. With the prevalence of deep neural networks, Reed *et al.* [12] propose to encode texts with a character-level convolutional recurrent neural network (char-CNN-RNN) that is pre-trained to learn correspondences between texts and images. In addition, AttnGAN [14] learns text encoding with a bi-directional LSTM [80]

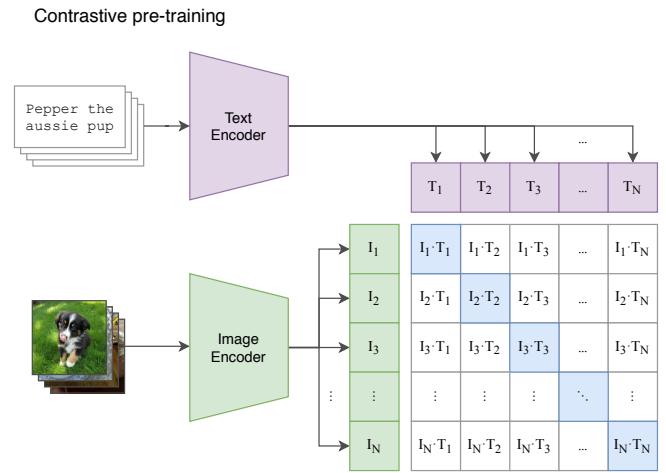


Fig. 2. The framework of the CLIP (The image is from [28]).

by concatenating its hidden states. Instead of obtaining the embedding with a pre-trained network, StackGAN [16] randomly samples latent variables from a Gaussian distribution defined by the text embedding. This encoding technique is widely adopted as it encourages smoothness over the textual guidance manifold. With the development of pre-trained models in natural language processing field, several studies [81], [82] also explore to perform text encoding by leveraging large-scale pre-trained language models such as BERT [83].

Recently, Contrastive Language-Image Pre-training (CLIP) [28] achieves SOTA image representation performance by learning the alignment of images and the corresponding captions from a large amount of image and text pairs. As illustrated in Fig. 2, CLIP jointly optimizes a text encoder and an image encoder to maximize the cosine similarity between positive pairs and minimizing that of negative pairs, yielding informative text embeddings.

## 2.3 Audio Guidance

Hearing helps humans to sense the world. The relation between auditory contents and visual contents has been

explored in previous cross-modal learning and generation research [84], [85], demonstrating that specific objects can be attended while the corresponding audio are pronounced. In addition, Harwath *et al.* [86] explore to learn neural network embeddings from natural images and the corresponding speech waveforms describing the images. With natural image embedding as an interlingua, the experiments in [86] show that the learnt models allow to perform cross-lingual speech-to-speech retrieval. Sounds can not only interact with visual contents but also capture rich semantic information. For example, by transferring the knowledge from other pre-trained scene and object recognition models, SoundNet [87] (a deep model for sound recognition) can learn to identify scenes and objects by using auditory contents only.

**Audio Encoding.** An audio sequence can be generated from given videos where deep convolution network is employed to extract features from video screenshots followed by LSTM [88] to generate audio waveform of the corresponding input video [89]. An input audio segment can also be represented by a sequence of features which can be spectrograms, fBanks, and mel-frequency cepstral coefficients (MFCCs), and the hidden layer outputs of the pre-trained SoundNet model [87]. For talking face generation [13], Action Units (AUs) [90] is also widely adopted to convert the driving audio into coherent visual signals for talking face generation.

## 2.4 Other Modality Guidance

Several other types of guidance have also been investigated to guide image synthesis and editing.

**Scene Graph.** Scene Graphs represent scenes as directed graphs, where nodes are objects and edges give relationships between objects. Image generation conditioned on scene graphs allows to reason explicit object relationships and synthesize faithful images with complex scene relationships. The guided scene graph can be encoded through a graph convolution network [54] which predicts object bounding boxes to yield a scene layout. To derive each individual subject-predicate-object relation, Vo *et al.* [91] propose to further predict relation units between objects, which is converted to a visual layout via convolutional LSTM [92].

**Coordinate.** Some work aims to synthesize or edit images conditioned on certain specific parameters. For example, Liu *et al.* [93] explore to generate one-hot images conditioned on point coordinates in  $(x, y)$  Cartesian space; with spatial coordinates of image patches as the condition, Lin *et al.* [94] propose to achieve high-fidelity image synthesis by generating images by parts.

## 2.5 Unified Representation

Multimodal data exists in different forms, unifying the data representation regardless of data modalities is of great significance for MISE and has attracted increasing attention from the research communities. Aiming for a generic representation across modalities, discrete representation learning [34], [95] plays a significant role in unified MISE.

**VQ-VAE.** Inspired by vector quantisation (VQ), Oord *et al.* propose VQ-VAE [95] to quantize image, audio and video

into discrete tokens with a learnt vector codebook. Specifically, VQ-VAE consists of an encoder, a feature quantizer, and a decoder. The image is fed into the encoder to learn a continuous representation. Then the continuous feature is quantized via the feature quantizer which assigns the feature to the nearest codebook entry. Then the decoder aims to reconstruct the original image from the quantized feature, yielding a reconstruction result. To make the quantization process differentiable, the gradient is approximated with the straight-through estimator [96] by copying the gradient from decoder to encoder [95].

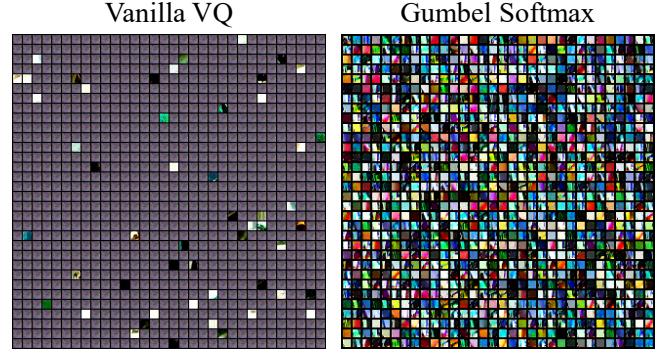


Fig. 3. The visualization of the first 1024 codes in vanilla vector quantization (VQ) and Gumbel Softmax quantization (both with downsampling factor  $f=8$ ).

**Gumbel Vector Quantization** However, the vanilla VQ-VAE with argmin operation (to get the nearest codebook entry) suffers from severe codebook collapse, *e.g.*, only few codebook entries are effectively utilized for quantization. As shown in Fig. 3 (images are from<sup>1</sup>), most codes in the original VQ-VAE are invalid or not utilized for feature encoding. Recently, vq-wav2vec [97] introduces Gumbel Softmax [98] to replace argmin for quantization. The Gumbel-Softmax allows to sample discrete representation in a differentiable way through straight-through gradient estimator [96], which boosts the codebook utilization significantly.

## 3 METHODS

We broadly categorize the methods for multimodal image synthesis and editing (MISE) into five categories: the GAN-based methods (Sec. 3.1), the Auto-regressive methods (Sec. 3.2), the Diffusion-based methods (Sec. 3.3), the Nerf-based methods (Sec. 3.4), and other methods (Sec. 3.5). We first discuss the GAN-based methods, which generally rely on generative adversarial networks and its inversion. We then discuss the prevailing Auto-regressive and Diffusion-based frameworks comprehensively. After that, we introduce neural radiance fields for the challenging task of 3D-aware MISE. Later, we present several other methods for visual synthesis and editing under the context of multimodal guidance. Finally, we compare and discuss the strengths and weaknesses of different generation methods.

1. <https://github.com/CompVis/taming-transformers/issues/67>

### 3.1 GAN-based Methods

GAN-based methods have been widely adopted for various MISE tasks by either developing conditional GANs with certain multimodal inputs or inverting unconditional GANs to yield target latent codes. Notably, these conditional GANs are usually designed for specific modalities in the corresponding tasks. We therefore group GAN-based methods into three categories including methods with **Intra-modal Conditions**, methods with **Cross-modal Conditions**, and **GAN Inversion** methods.

#### 3.1.1 Intra-modal Conditions

Intra-modal guidance provides certain visual clues for image synthesis and editing. Compared with cross-modal guidance, these visual clues can be deemed as stronger guidance which make it possible to achieve synthesis or editing without paired training data. Thus, we further group the methods with intra-modal condition into methods with paired data and methods with unpaired data.

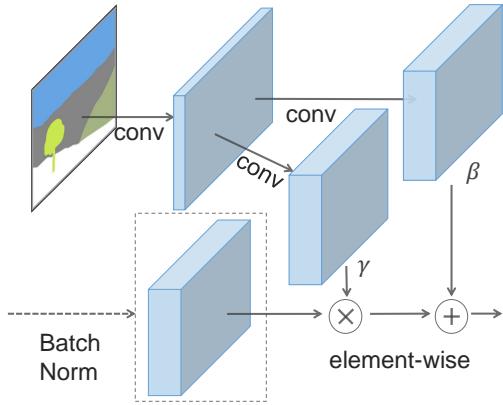


Fig. 4. Illustration of the spatially-adaptive de-normalization [3]. The image is from [3].

**Paired Data.** Paired visual guidance means the provided guidance is accompanied with corresponding ground truth images to provide certain direct supervision. Besides adversarial loss, image synthesis with paired visual guidance is usually trained with certain supervised loss between the generated image and the ground truth. Isola *et al.* [2] first investigate conditional GAN as a general framework named Pix2Pix for various image translation tasks (e.g., edge-to-image, day-to-night, and semantic-to-image). To mitigate the constraint in high-resolution image synthesis in Pix2Pix [2], Wang *et al.* [99] propose Pix2PixHD that allows to synthesize images of  $2048 \times 1024$ . However, Pix2Pix [2] as well as its variant [99] cannot encode complex scene structural relationships between the guidance and real images when there exist very different views or severe deformations. Therefore, Tang *et al.* [100] propose an attention selection module to align the cross-view guidance with the target view.

On the other hand, previous methods directly encode the visual guidance with deep networks for further generation, which is suboptimal as part of the guidance information tends to be lost in normalization layers. SPADE [3] is designed to inject the guided feature effectively through a spatially-adaptive de-normalization as shown in Fig. 4.

SEAN [101] introduces a semantic region-adaptive normalization layer to achieve region-wise style injection. Claiming that traditional image translation networks [2], [3], [99] suffer from high computational cost while handling high-resolution images, Shaham *et al.* [102] propose ASAPNet which is a lightweight yet efficient network for the translation of high-resolution images. Recently, Zhang *et al.* [103] and Zhan *et al.* [51] introduce exemplar-based image translation frameworks which build dense correspondences between exemplars and condition inputs to provide accurate guidance. However, building the dense correspondences incurs quadratic memory cost. Thus, Zhou *et al.* [104] propose to leverage PatchMatch [105] with GRU assistance to build correspondence at high resolution efficiently. In addition, Zhan *et al.* [106] introduce a bi-level alignment scheme to reduce memory cost while building dense correspondences.

**Unpaired Data** Unpaired image synthesis utilizes unpaired training images to convert images from one domain to another. The generation of realistic images mainly relies on adversarial learning [1] with certain constraint losses. Specially, Zhu *et al.* [18] design a cycle-consistency loss to preserve the image content by ensuring the input image can be recovered from the translation result. However, cycle-consistency loss is too restrictive for image translation as it assumes a bi-junctional relationship between the two domains. Several studies [107]–[109] thus aim to explore one-way translation and bypass the bijection constraint of cycle-consistency. With the emergence of contrastive learning, Park *et al.* [76] proposes to maximize the mutual information of positive pairs via noise contrastive estimation [110] for the preservation of contents in unpaired image translation. Andonian *et al.* [111] introduce contrastive learning to measure the inter-image similarity in unpaired image translation. However, heterogeneous domains often have mappings where an individual image in one domain may not share any characteristics with its representation in another domain after mapping. Therefore, TravelGAN [108] proposes to preserve the intra-domain vector transformations in a latent space learned by a siamese network, which allows to learn mappings across heterogeneous domains.

#### 3.1.2 Cross-modal Conditions

In terms of cross-modal guidance, GANs are usually developed according the specific tasks. Thus, we formulate this section with typical GAN-based multimodal tasks, i.e., text-to-image synthesis and audio-driven image editing.

**Text-to-Image Synthesis.** Text-to-image synthesis [77], [113] aims to generate an image that faithfully reflects the semantic of a textual description. Reed *et al.* [12] are the first who extend conditional GANs [6] to achieve text-to-image synthesis. Empowered by the advance of GANs for image synthesis, the task has made significant progress with the employment of stacked architecture, attention mechanism, cycle consistency.

**Stacked Architectures:** Targeting to synthesize high-resolution images, stacked architectures are widely adopted in GAN-based methods. Specially, StackGAN [16] generates a coarse image of  $64 \times 64$  at the first stage, followed by a second generator to further output an image of  $256 \times 256$  at the second stage. StackGAN++ [17] further improves

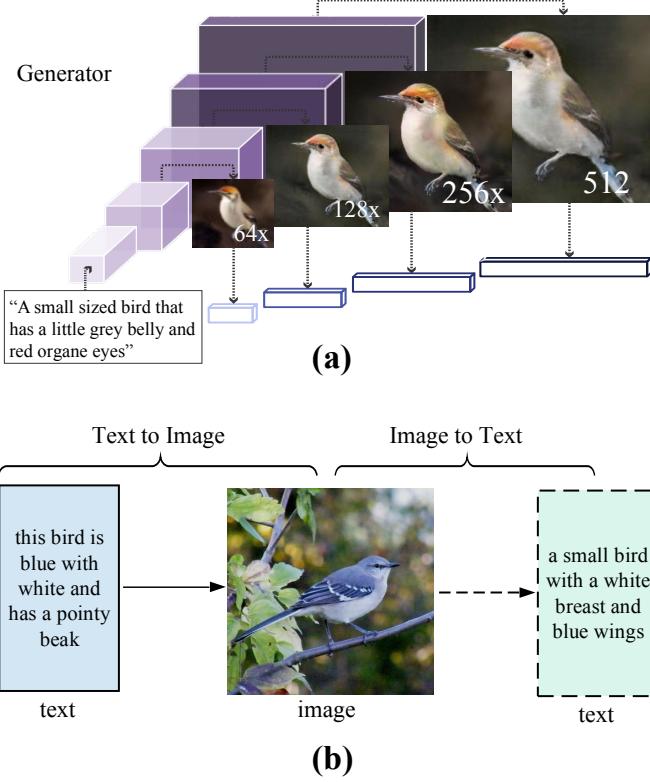


Fig. 5. (a) illustrates the hierarchically-nested adversarial network in HDGAN [112]; (b) illustrates the cycle structure in MirrorGAN [19]. The image is adapted from [112] and [19].

StackGAN [16] by jointly training three generators and discriminators. Instead of using multiple generators, HDGAN [112] propose to employ hierarchically-nested discriminators at multi-scale layers to generate high-resolution images as shown in Fig. 5 (a).

**Cycle Consistency:** To ensure cycle consistency for the text prompt or encoded text features, some works explore to pass the generated images through an image captioning [19], [114] or image encoding network [115] as shown in Fig. 5 (b). Specifically, PPGN [114] employs an image captioning model to iteratively retrieve a latent code which maximizes a feature activation of the corresponding image according to a feedback network. Inspired by CycleGAN [18], cycle-consistent re-description architectures [19], [116] allows to learn a consistent feature embedding between images and the corresponding text description. Specially, MirrorGAN [19] aims to re-describe the generated images via a semantic text regeneration and alignment module.

**Attention Mechanism:** By allowing the model to focus on specific part of an input, attention mechanisms have proven to be beneficial to language and vision models [30], [117]. In terms of text guided image synthesis, AttnGAN [14] incorporates attention mechanisms in a multi-stage manner to synthesize fine-grained details based on both relevant words and global sentence. Huang *et al.* [118] introduce an attention mechanism between text words and object regions obtained from bounding boxes. SEGAN [119] introduces an attention regularization term [120] that only preserves the weights for keywords with zero weight for other words. As the spatial attention in [14] mainly focuses on color information, ControlGAN [121] proposes to correlate the words with the corresponding semantic region through a word-level spatial attention. In order to incorporating additional visual details and avoid conflicting, RiFeGAN [122] employs an attention-based caption matching model to select candidate captions from prior knowledge.

Large scale GANs [20], [21] have achieved remarkable progress in unconditional image synthesis with high-resolution and high-fidelity. With a pre-trained GAN model, a series of studies explore to invert a given image back into the latent space of the GAN, which is termed as GAN inversion [25]. Specifically, a pre-trained GAN learns a mapping



Fig. 6. Illustration of the audio-driven talking face generation task. With one frame as the identity reference, the task aims to synthesize talking faces that are aligned with the given audio source. The image is adapted from [53].

**Audio-driven Image Editing.** Thanks to the value in practical applications, current audio-driven image editing methods focus on the talking face generation [11], [13], [53], [123]–[127]. As illustrated in Fig. 6, the task of audio-driven talking face generation aims to synthesize talking faces that speak the given audio clips [123], which has wide applications in digital face animation, film production, visual dubbing, etc. One fundamental challenge in audio-driven talking face generation is how to accurately convert audio contents into visual information. Leveraging generative adversarial models [1], researchers develop different techniques to address this challenge. For instance, Chung *et al.* [11] learn the joint embedding of raw audio and video data and project it to image plane with a decoder to generate talking faces. Following [11], Zhou *et al.* [125] propose DAVS that learns a disentangled audio-visual representation which helps improve the quality of the synthesized talking faces. Song *et al.* [13] introduce a conditional RNN network for adversarial generation of talking face. Chen *et al.* [123] design a hierarchical structure that maps the audio clip into facial landmarks which are further leveraged to generate talking faces. Zhou *et al.* [128] introduce MakeItTalk that predicts speaker-aware facial landmarks from the speech contents for better preserving the characteristic of the speaker. However, the head pose is almost fixed in the talking faces generated by these approaches. To improve the perceptual realism, recent approaches [53], [124], [129], [130] take head pose into consideration while generating talking face. For example, Yi *et al.* [129] propose to map audio contents to 3DMM parameters [131] for guiding the pose-controllable generation talking faces; Zhou *et al.* [53] present PC-AVS that achieves pose-controllable talking face generation by learning disentangled feature spaces of poses, identities, and speech contents.

### 3.1.3 GAN Inversion

Large scale GANs [20], [21] have achieved remarkable progress in unconditional image synthesis with high-resolution and high-fidelity. With a pre-trained GAN model, a series of studies explore to invert a given image back into the latent space of the GAN, which is termed as GAN inversion [25]. Specifically, a pre-trained GAN learns a mapping

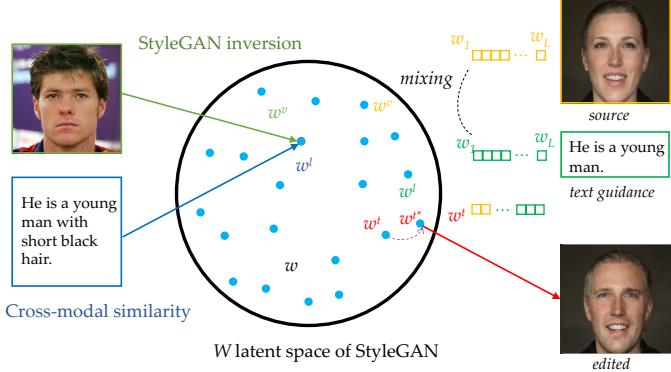


Fig. 7. GAN inversion method with cross-modal matching in latent space: Both image and guidance embeddings are projected into the StyleGAN [21] latent space  $\mathcal{W}$ . The cross-modal similarity learning aims to pull the visual embedding  $w^v$  and guidance embedding  $w^l$  to be closer. For cross-modal image editing (e.g., text guidance), the cross-modal embedding  $w^v$  and  $w^l$  be first obtained through the corresponding encoders. Then image editing can be performed through style mixing to get the edited latent code  $w^t$  which is further updated through instance-level optimization. The edited latent code is fed into the StyleGAN generator to yield the edited image. The illustration is from [27].

from latent codes to real images, while the GAN inversion maps images back to latent code, which is achieved by feeding the latent code into the pre-trained GAN to reconstruct the image through optimization. Typically, the reconstruction metrics are based on  $\ell_1$ ,  $\ell_2$ , perceptual [36] loss or LPIPS [37]. Some constraints on face identity [132] or latent codes [26] could also be included during optimization. With the obtained latent, we can faithfully reconstruct the original image and conduct image manipulation in the latent space.

In terms of multimodal image synthesis and editing, the key of GAN-inversion based methods lies in how to edit or generate latent code according to the corresponding guidance. Notably, the inversion-based methods tend to be less sensitive to the modalities as the inversion is conducted on the unstructured 1-dimension latent space.

**Cross-modal Matching in Latent Space.** Multimodal image synthesis and editing can be achieved by matching the embeddings of images and cross-modal inputs (e.g., semantic maps, texts) in a common embedding space [27], [133] as shown in Fig. 7. Specifically, a cross-modal encoder is trained to learn the embeddings with a visual-linguistic similarity loss and a pairwise ranking loss [134], [135]. To preserve the identity after editing, an instance-level optimization module can be employed in the objective which allows to modify the target attributes according to text descriptions. As optimization are performed in the StyleGAN latent space, this framework inherently allows image generation from given multimodal conditions. Conditional image manipulation can also be performed via style mixing in the shared latent space.

**Image Code Optimization in Latent Space.** Instead of mapping the text into the latent space, a popular line of research aims to optimize the latent code of the original image directly, guided by certain losses that measure cross-modal consistency. Particularly, Jiang *et al.* [136] propose to optimize image latent codes through a pre-trained fine-grained attribute predictor which pushes the output latent code to

change in a direction consistent with the text description. However, this attribute predictor is specially designed for face editing with fine-grained attribute annotations, which makes it hard to generalize to other scenes.

Rather than employing specific attribute predictor, several concurrent projects use Contrastive Language-Image Pre-training (CLIP) [28] to guide the inversion process for text-to-image synthesis. Aiming for text-guided image inpainting, Bau *et al.* [137] define a semantic consistency loss based on CLIP that optimizes latent codes inside the inpainting region to achieve semantic consistency with the given text. StyleClip [29] and StyleMC [138] use pre-trained CLIP as the loss supervision to match the manipulated results with the text condition as illustrated in Fig. 8. The framework of StyleCLIP is further extend and improved by the introduction of CLIP-based contrastive loss [139] for counterfactual image manipulation, AugCLIP score [140] to robustify the standard CLIP score, over-parameterization strategy [140] to navigate the optimization in latent space.

**Domain Generalization.** However, StyleCLIP requires to train a separate mapper for each specific text description which is not flexible in real applications. HairCLIP [141] presents a hair editing framework that supports different texts by introducing a shared condition embedding strategy which unifies the text and image conditions into the same domain. Instead of generalization on text description, StyleGAN-NADA [142] presents a text-guided image editing method that allows to shift a generative model to new domains, without having to collect even a single image from the new domains. The domain shift is achieved by adapting the generator's weights to be aligned with the driving text, along certain textually-prescribed paths in CLIP's embedding space.

### 3.2 Auto-regressive Methods

#### 3.2.1 Auto-regression Preliminary

Leveraging their powerful attention mechanisms, Transformer [30], [32], [35], [143] models have emerged as a paradigm in sequence-dependent modeling. Inspired by the success of GPT model [31] in natural language modeling, image GPT (iGPT) [32] employs Transformer for auto-regressive image generation, by treating the flattened image sequences as discrete tokens. The plausibility of generated images demonstrates that Transformer models are able to model the spatial relationships between pixels and high-level attributes (texture, semantic, and scale).

As Transformer models inherently support multimodal inputs, a series of studies have been proposed to explore multimodal image synthesis with Transformer-based auto-regressive models [38]–[40], [144]. Overall, the pipeline for Transformer-based image synthesis consists of a *vector quantization* process to achieve discrete representation and data dimensionality compression, and an *auto-regressive modeling* process which establishes the dependency between discrete tokens in a raster-scan order as illustrated in Fig. 9.

#### 3.2.2 Image Vector Quantization

Directly treating all image pixels as a sequence for auto-regressive modeling with Transformer is expensive in terms of memory consumption as the self-attention mechanism

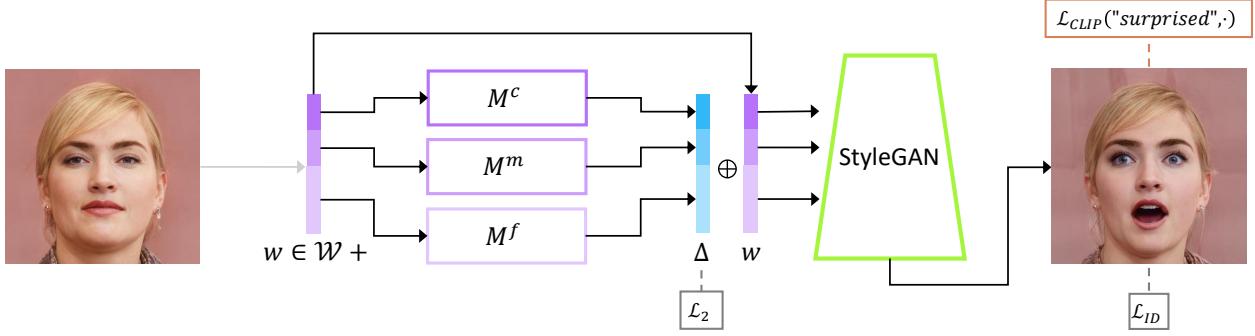


Fig. 8. The architecture of text-guided mapper in StyleCLIP [29]. The source image (left) is inverted into a latent code  $w$ . Three separate mapping functions are trained to generate residuals (in blue) that are added to  $w$  to yield the target code, from which a pre-trained StyleGAN (in green) generates an image (right), assessed by the CLIP and identity losses. The image is from [29].

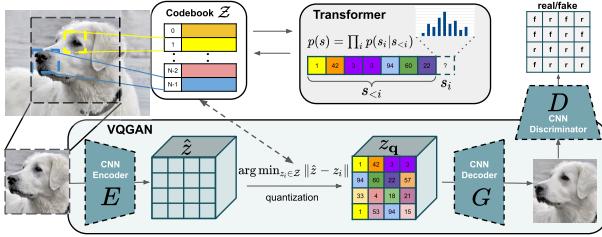


Fig. 9. Taming Transformer [38] first learn discrete and compressed representation which can reconstruct the original image faithfully, followed by an autoregressive Transformer to model the dependency of discrete sequence. The image is from [38].

in Transformer incurs quadratic memory cost. Thus, compressed and discrete representation of image is essential and significant for Transformer-based image synthesis and editing. Chen *et al.* [32] adopt a color palette to reduce the dimensionality to 512 (generated by k-means clustering of RGB pixel values of ImageNet [145] dataset, with  $k=512$ ) while faithfully preserving the main structure of original images. However, k-means clustering only reduces the size of codebook dimensionality but the sequence length is still unchanged. Thus, the Transformer model still cannot be scaled to higher resolutions, due to the quadratically increasing cost in sequence length. To this end, Vector Quantised VAE [95] is widely adopted to learn discrete and compressed representation of images. Targeting for learning superior discrete and compressed image representation, a series of efforts have been devoted to improve VQ-VAE in terms of loss design, network architecture and regularization.

**Loss Design.** To achieve good perceptual quality for the reconstructed image, Esser *et al.* [38] propose VQGAN which incorporates an adversarial loss with a patch-based discriminator and a perceptual loss [36], [146], [147] for image reconstruction as shown in Fig. 9. Instead of using VGG features pre-trained on ImageNet, Dong *et al.* [148] leverage self-supervised network [83], [149] for the learning of deep visual features to enforce perceptual similarity during the dVAE training. With the extra adversarial loss and perceptual loss, the image quality is clearly improved compared with the original pixel loss in image reconstruction. Besides, to emphasize reconstruction quality in face region, Gafni *et al.* [150] employ a feature-matching loss over the activations of a pre-trained face-embedding network [151]

during vector quantization. The face-matching loss can also be generalized to common objects to increase awareness and perceptual knowledge of objects reconstruction, just replacing the face-embedding network with a pre-trained VGG [152] network.

**Network Architecture.** In above approaches, a convolution neural network is learned to quantize and generate images. Instead, Yu *et al.* [153] propose ViT-VQGAN which replaces the CNN encoder and decoder with Vision Transformer (ViT) [154]. Given sufficient data (for which unlabeled image data is plentiful), ViT-VQGAN is shown to be less constrained by the inductive priors imposed by convolutions and is able to yield better computational efficiency with higher reconstruction quality. ViT-VQGAN also presents a factorized code architecture which introduces a linear projection from the encoder output to a low dimensional latent variable space for code index lookup and boosts the codebook usage substantially. Besides, NUWALIP [155] explores a multi-perspective encoding which enhances visual information by including both low-level pixels and high-level tokens; DiVAE [156] employs a diffusion-based decoder to learn discrete image representation with superior image reconstruction performance.

**Regularization.** Shin *et al.* [157] validate that the vanilla VQ-VAE doesn't satisfy translation equivariance during quantization, resulting in degraded performance for text-to-image generation. A simple but effective TE-VQGAN [157] is thus proposed to achieve translation equivariance by regularizing orthogonality in the codebook embeddings. To achieve joint quantization in multiple domains for conditional image generation, Zhan *et al.* [158] design an Integrated Quantization VAE (IQ-VAE) with a variational regularizer to regularize feature quantization in cross-domain spaces.

### 3.2.3 Auto-regressive Modeling

Autoregressive model has been widely explored for building sequence dependency. Previous auto-regressive models such as PixelCNN [159] struggle in modeling long term relationships within an image due to the limited receptive field. With the prevailing of Transformer [30], Parmar *et al.* [160] develop a Transformer-based generation model with enhanced receptive field which allows to sequentially predict each pixel conditioned on previous prediction results.

Auto-regressive (AR) modeling is representative paradigm to accommodate sequence dependencies, complying with the chain rule of probability. The probability of each token in the sequence is conditioned on all previously predictions, yielding a joint distribution of sequences as the product of conditional distributions:  $p(x) = \prod_{t=1}^n p(x_t|x_1, x_2, \dots, x_{t-1}) = \prod_{t=1}^n p(x_t|x_{<t})$ . During inference, each token is predicted auto-regressively in a raster-scan order. A top- $k$  sampling strategy is adopted to randomly sample from the  $k$  most likely next tokens, which naturally enables diverse sampling results. The predicted tokens are then concatenated with the previous sequence as conditions for the prediction of next token. This process repeats iteratively until all the tokens are sampled.

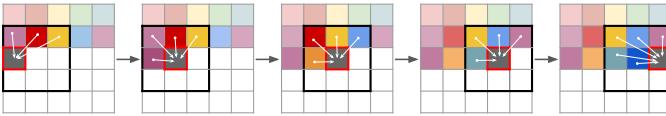


Fig. 10. The Sliding window strategy for image sampling in auto-regressive models. The image is from [38].

**Sliding Window Sampling.** To speed up auto-regressive image generation, Esser *et al.* [38] employ a sliding-window strategy to conduct sampling from the trained Transformer model as illustrated in Fig. 10. Instead of estimating current results leveraging all previous predictions, sliding window strategy only utilizes the predictions within a local window which reduces the inference time significantly. As long as that the spatial conditioning information is available or the dataset statistics are approximately spatially invariant, the local context in the sliding window is sufficient for the faithful modeling of images sequences.

**Bidirectional Context.** On the other hand, previous methods incorporate image context in a raster-scan order by attending only to previous generation results. This strategy is unidirectional and suffers from sequential bias as it disregards much context information until autoregression is nearly complete. It also ignores much contextual information in different scales as it only processes the image on a single scale. Grounded in above observations, ImageBART [39] presents a coarse-to-fine approach in a unified framework that addresses the unidirectional bias of autoregressive modelling and the corresponding exposure bias. With compressed contextual information of images, a diffusion process is applied to successively eliminate information, yielding a hierarchy of representations which is further compressed via a multinomial diffusion process [161], [162]. By modeling the Markovian transition autoregressively with attending to the preceding hierarchical state, crucial global context can be leveraged for each individual autoregressive step. As an alternative, bidirectional Transformer are also widely explored to incorporate bidirectional context, accompanied with a Masked Visual Token Modeling (MVTM) [163] or Masked Language Modeling (MLM) [143], [164] mechanisms.

**Better Self-Attention.** To handle language, image, and video in different tasks in a unified manner, NUWA [40] presents a 3D Transformer framework with a unified 3D Nearby Self-Attention (3DNA) which not only reduces the complexity of full attention but also shows superior per-

formance. With a focus on semantic image editing at high resolution, ASSET proposes to sparsify the Transformer’s attention matrix at high resolutions guided by dense attention at lower resolutions, leading to reduced computational cost.

**Model Architecture.** To explore the limits of auto-regressive text-to-image synthesis, Parti [165] scales the parameter size of Transformer up to 20B, and observe consistent quality improvements in terms of image quality and text-image alignment. Instead of unidirectionally modeling from text to image, Huang *et al.* [166], [167] is the first to present a bi-directional image-and-text framework with Transformer, which generates both multiple diverse captions and images.

### 3.3 Diffusion-based Methods

**Diffusion Model Preliminary.** Recently, diffusion models such as denoising diffusion probabilistic models (DDPM) [42], [162] have achieved great successes in generative image modelling [42], [168]–[170]. Denoising diffusion probabilistic models are a type of latent variable models that consist of a forward diffusion process and a reverse diffusion process. The forward process is a Markov chain where noise is gradually added to the data when sequentially sampling the latent variables  $\mathbf{x}_t$  for  $t = 1, \dots, T$ . Each step in the forward process is a Gaussian transition  $q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I})$ , where  $\{\beta_t\}_{t=0}^T$  are fixed or learned variance schedule. The reverse process  $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$  is parametrized by another Gaussian transition  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)\mathbf{I})$ .  $\mu_\theta(\mathbf{x}_t, t)$  can be decomposed into a linear combination of  $\mathbf{x}_t$  and a noise approximation model  $\epsilon_\theta(\mathbf{x}_t, t)$  that can be learned through optimization. After training  $\epsilon_\theta(\mathbf{x}, t)$ , the sampling process of DDPM can be achieved by following a reverse diffusion process.

Song *et al.* [168] propose an alternative non-Markovian noising process that has the same forward marginals as DDPM but allows using different samplers by changing the variance of the noise  $\sigma_t$ . Especially, by setting this noise to 0, which is a DDIM sampling process [168], the sampling process becomes deterministic, enabling full inversion of the latent variables into the original images with significantly fewer steps [41], [168]. Notably, the latest work [41] has demonstrated even higher quality of image synthesis performance compared to variational autoencoders (VAEs) [171], flows [172], [173], auto-regressive models [174], [175] and generative adversarial networks (GANs) [1], [21]. To achieve image generation and editing conditioned on multimodal guidance, a conditional variant of diffusion model and inversion of pre-trained diffusion model are both extensively studied in the literature.

#### 3.3.1 Conditional Diffusion Models

To launch the MISE tasks, a conditional diffusion model can be naturally derived by directly concatenate the condition information with the noisy image as input of the denoising network. Recently, the performance of conditional diffusion models is significantly pushed forward by a series of designs.

**Classifier-free Guidance.** A downside of classifier guidance lies its requirement of an additional classifier model

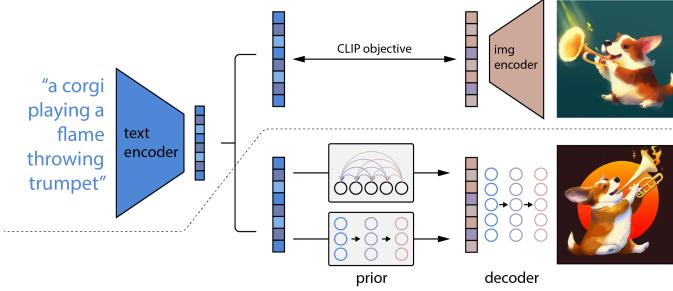


Fig. 11. The overall architecture of DALL-E 2. The CLIP training process above the dotted line aims to a joint representation between text and images. The text-to-image generation process below the dotted line feeds a CLIP text embedding to a prior network to produce an image embedding which is used to yield the final images via a diffusion decoder. The image is from [45].

which leads to a complicated training pipeline. Recently, Ho *et al.* [176] achieve compelling results without a separately trained classifier through the use of classifier-free guidance, a form of guidance that interpolates between predictions from a diffusion model with and without labels. Within this line of research, GLIDE [177] compares CLIP guidance and classifier-free guidance in diffusion models for the text-guided image synthesis, and concludes that classifier-free guidance yields better performance and a diffusion model of 3.5 billion parameters outperforms DALL-E in terms of human evaluations. Besides, Tang *et al.* [178] explore classifier-free guidance sampling for discrete denoising diffusion model with the introduction of an effective implementation of classifier-free guidance.

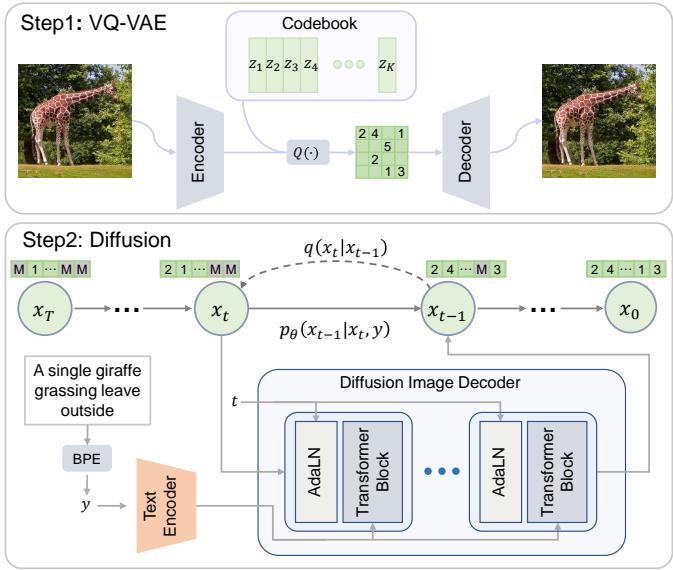


Fig. 12. Overall framework of discrete Diffusion process. With a VQ-VAE for discrete representation, discrete Diffusion process models the discrete latent space by reversing a forward diffusion process that gradually corrupts the input via a fixed Markov chain. The image is from [179].

**Condition Incorporation.** Instead of conditioning on the conditional embedding directly, DALL-E 2 [45] propose to generate a CLIP image embedding from a text caption at first, followed by a diffusion decoder that generates an

image conditioned on the image embedding, as illustrated in Fig. 11. The prior network explicitly generates image representations which improves image diversity with minimal loss in photorealism and caption similarity. To fully leverage the conditional information for semantic image synthesis, Wang *et al.* [180] propose to feed the semantic maps to the decoder via a spatially-adaptive normalization, which improves both the quality and semantic coherence of generated images. With a different route, Zhu *et al.* [181] propose to ensure the correspondence between the condition and generated output by maximizing their mutual information using contrastive learning.

**Model Architecture.** Building on the power of large Transformer language models in understanding text, Imagen [46] achieves SOTA text-to-image generation performance among diffusion-based methods and discovers that generic large language models (*e.g.* T5 [182]), pre-trained on text-only corpora, are surprisingly effective at encoding text for image synthesis. Aiming to understand the composition concepts in scenes, Liu *et al.* [183] propose a compositional architecture for diffusion-based image synthesis which generates a image by composing a set of diffusion models. Observing much of the success of diffusion models is due to the dramatic increase of training cost, Blattmann *et al.* [184] propose to complement the diffusion model with a retrieval-based approach which incurs low computational cost.

**Discrete Diffusion.** To enable diffusion models training on limited computational resources while retaining their quality and flexibility, several works explore to conduct diffusion process in the latent space of a VQ-VAE [95] as shown in Fig. 12. By training diffusion models on the latent space of a VQGAN [38] variant with the quantization layer absorbed by the decoder, Latent Diffusion Models (LDMs) [185] is the first work that achieves a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. Similarly, Gu *et al.* [179] present a vector quantized diffusion (VQ-Diffusion) model for text-to-image generation by learning a parametric model using a conditional variant of the Denoising Diffusion Probabilistic Model (DDPM). Tang *et al.* [178] further improve VQ-Diffusion by introducing a high-quality inference strategy to alleviate the joint distribution issue. Following VQ-Diffusion, Text2Human [186] is introduced to achieve high-quality text-driven human generation by employing a diffusion-based Transformer to model a hierarchical discrete latent space.

### 3.3.2 Pre-trained Diffusion Models

For each type of conditional input, above conditional diffusion models are required to be re-trained from scratch, leading to high computational cost. Therefore, a popular line of research explores to invert pre-trained diffusion models to achieve conditional generation like GAN inversion. As an early exploration, Dhariwal *et al.* [41] augment diffusion models with classifier guidance which allows conditional generation from classifier's labels. Following [41], Kim *et al.* [44] propose a DiffusionCLIP for text-driven image manipulation with pre-trained diffusion models which uses CLIP loss to steer the editing towards the given text prompt as shown in Fig. 13. The input image is first converted to the latent using the pretrained diffusion model. Then, the

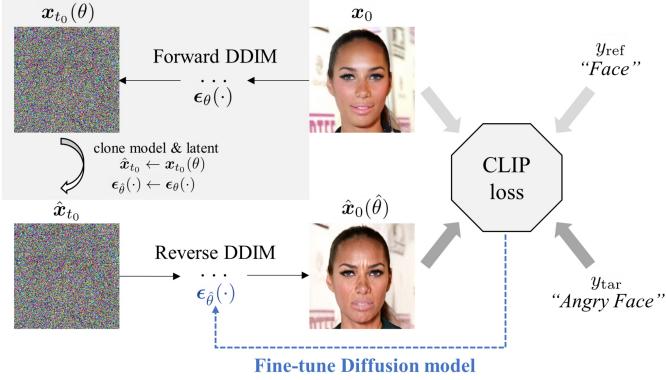


Fig. 13. Overview of DiffusionCLIP [44]. The input image is first converted to the latent via diffusion models. Then, guided by directional CLIP loss, the diffusion model is fine-tuned, and the updated sample is generated during reverse diffusion. The image is from [44].

diffusion model at the reverse path is fine-tuned to generate samples driven by the target text and the CLIP loss.

Since CLIP is trained on clean images, a way of estimating a clean image from the noisy latent is expected during the denoising diffusion process. Specifically, Liu *et al.* [187] introduce a self-supervised finetuning of CLIP without text annotations to force an alignment between features extracted from clean and noised images. Avrahami *et al.* [188] present a way of estimating a clean image from the noisy latent during the denoising diffusion process. Then, a CLIP-based loss can be defined as the cosine distance between the CLIP embedding of the text prompt and the embedding of the estimated clean image. Besides, to achieve the local guidance for image editing, Avrahami *et al.* [188] only consider the gradients of CLIP under a input mask. A similar estimation approach is presented in **CLIP-guided diffusion**, where a linear combination the noisy latent and the estimated embedding of clean image is used to provide global guidance for the diffusion.

### 3.4 NeRF-based Methods

**NeRF Preliminary.** Neural radiance fields (NeRF) [47] achieves impressive performance for novel views synthesis by using neural networks to define an implicit scene representation. Specially, a fully-connected neural network is adopted in NeRF, by taking a spatial location ( $x, y, z$ ) with the corresponding viewing direction ( $\theta, \phi$ ) as input, and the volume density with the corresponding emitted radiance as output. To render 2D images from the implicit 3D representation, differentiable volume rendering is performed with a numerical integrator [47] to approximate the intractable volumetric projection integral. Then, NeRF can be optimized driven by a photometric loss between the rendered image and corresponding ground truth image.

Powered by NeRF for 3D scene representation, 3D-aware MISE can be achieved by optimizing per-scene NeRF with multiview supervision, training generative NeRF on monocular images, or inverting a pre-trained generative NeRF.

#### 3.4.1 Per-scene Optimization NeRF

Consistent with the original NeRF, a number of works focus on the implicit representation of a single scene. To achieve

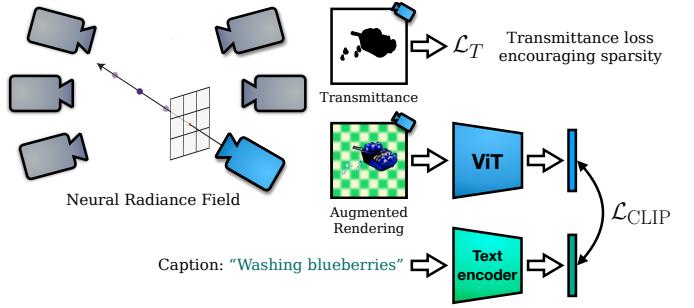


Fig. 14. The framework of Dream Field [189]. Given a caption, Dream Field yields a continuous volumetric representation of an object's geometry and appearance learned with guidance from a pre-trained CLIP model. The image is from [189].

text-driven 3D-aware image synthesis, Jain *et al.* [189] introduce Dream Fields which leverages pre-trained image-text models to optimizes a Neural Radiance Field as shown in Fig. 14. Specifically, the NeRF is optimized to render multi-view images that score highly with a target text description according to a pre-trained CLIP model. To improve fidelity and visual quality, Dream Field introduces simple geometric priors, including sparsity-inducing transmittance regularization, scene bounds, and new MLP architectures. As an extension to audio-driven image editing, AD-NeRF [50] achieves high-fidelity talking-head synthesis by training neural radiance fields on a video sequence with the audio track of one target person. Different from previous methods which bridge audio inputs and video outputs based on the intermediate representations, AD-NeRF directly feeds the audio features into an implicit function to yield a dynamic neural radiance field, which is further exploited to synthesize high-fidelity talking-face videos accompanied with the audio via volume rendering.

Some recent works such as NeuS [190] explore to combine the advantages of SDF and NeRF to achieve high-quality surface reconstruction. With NeuS as the base representations of avatars, Hong *et al.* propose AvatarCLIP [191] to achieve zero-shot text-driven framework for 3D avatar generation and animation. A pre-trained CLIP is exploited to supervise the neural human generation including 3D geometry, texture and animation.

#### 3.4.2 Generative NeRF

Distinct from per-scene optimization NeRFs which require posed multi-view images of a scene for training, generative NeRF can be trained from a collection of unposed 2D images alone. GRAF [193] is the first to introduce an adversarial framework for the generative training of radiance fields by employing a multi-scale patch-based discriminator. Lot of efforts have recently been devoted to improve the generative NeRF, *e.g.*, GIRAFFE [194] for introducing volume rendering at the feature level and separating the object instances in a controllable way; Pi-GAN [195] for the FiLM-based conditioning scheme [196] with a SIREN architecture [197]; StyleNeRF [198] for the integration of style-based generator to achieve high-resolution image synthesis; EG3D [199] for an efficient tri-plane representation.

Based on these prior explorations of unconditional generative NeRF, building generative NeRF conditioned on cer-

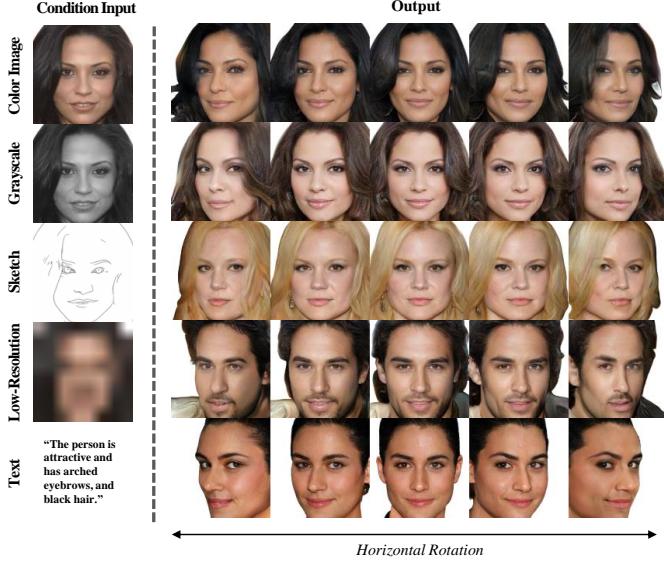


Fig. 15. CG-NeRF is capable of producing 3D-aware output images that are faithful to the corresponding condition inputs. The image is from [192].

tain guidance has attracted increasing interest. Recently, Jo *et al.* [192] propose a conditional generative neural radiance fields (CG-NeRF), which can generate multi-view images reflecting extra input conditions (*e.g.*, images or texts) as shown in Fig. 15. Specifically, the pre-trained CLIP model is employed to extract the conditional image and text features as the input of the NeRF. Although the generative NeRFs can maintain view consistency, their generated images are not locally editable. To overcome these limitations, FENeRF [200] presents a 3D-aware generator that can produce view-consistent and locally-editable portrait images. FENeRF employs two decoupled latent codes to generate corresponding facial semantics and texture in a spatial-aligned 3D volume with shared geometry. Benefiting from such underlying 3D representation, FENeRF can jointly render the boundary-aligned image and semantic mask and use the semantic mask to edit the 3D volume via GAN inversion.

### 3.4.3 Generative NeRF Inversion

In light of recent advances in generative NeRFs for 3D-aware image synthesis, some works explore the inversion of pre-trained generative NeRFs.

Accompanied with the new task that aims to construct a NeRF conditioned on a semantic mask, Chen *et al.* [201] adopt encoder-based inversion to map the semantic map into the latent space of a pretrained pi-GAN [195]. To further improve the accuracy of the inverse mapping, Sem2NeRF integrates a new region-aware learning strategy into the design of both the encoder and the decoder and augments the input semantic masks with extracted contours and distance field representations [202]. Targeting for interactive 3D-aware image editing with local control of the shape and texture, IDE-3D [203] propose to train 3D-semantics-aware generative model that produces view-consistent face images and semantic masks concurrently. Then, two inverse encoders are employed to yield latent codes from the semantic

map and images, followed by a canonical encoder to enables efficient manipulation of semantic masks in canonical view.

Different from above encoder-based inversion, CLIP-NeRF [49] employs optimization-based inversion to achieve the manipulation of neural radiance fields according to a short text prompt. Driven by a CLIP-based matching loss as described in StyleCLIP [29], CLIP-NeRF bridges the generative latent space and the CLIP embedding space by employing two code mappers to optimize the latent codes towards the targeted manipulation.

On the other hand, the inversion of generative NeRF is still challenging thanks to the including of camera pose. To achieve stable text-guided image editing, StyleNeRF [198] explores to combine encoder-based inversion and optimization-based inversion, where the encoder predicts a camera pose and a coarse style code which is further refined through inverse optimization.

### 3.5 Other Methods

With the development of generative models and neural rendering, other up-to-date methods are also explored for multimodal image synthesis and editing.

**Style Transfer.** Without requiring training or inversion of generative models, CLVA [204] manipulates the style of a content image through text prompts by comparing the contrastive pairs of content image and style instruction to achieve the mutual relativity. However, CLVA is constrained as it requires style images accompanied with the text prompts during training. Instead, CLIPstyler [205] leverages pre-trained CLIP model to achieves text guided style transfer by training a lightweight network which transforms a content image to follow the text condition. Aiming for style transfer of 3D scenes, Mu *et al.* [206] propose to learn geometry-aware content features from a point cloud representation of the scene, followed by point-to-pixel adaptive attention normalization (AdaAttN) to transfer the style of a given image. As an extension to video, Loeschcke *et al.* [207] harness the power of CLIP to stylize the object in a video according to two target texts.

**3D Object Generation & Editing.** With style transfer network to predict color and displacement, Text2Mesh [208] aims to modifies an input mesh to conform to the target text. The neural style network is optimized by rendering multiple 2D images, which are optimized to be consistent with the target text description via a CLIP-based loss. Similarly, CLIPMesh [209] presents a technique for zero-shot generation of a 3D model using only a target text prompt. Relying on a pre-trained CLIP model, CLIPMesh directly performs optimization on mesh parameters to generate shape and texture that enables to render images matching the input text prompt. Instead of using CLIP, Liu *et al.* [210] adopt a BERT-based text encoder and achieve text-guided 3D shape generation with high-fidelity.

### 3.6 Comparison and Discussion

All generation methods possess their own strength and weakness. GAN-based methods can achieve high-fidelity image synthesis in terms of FID and Inception Score and also have fast inference speed, while GANs are notorious for unstable training and are prone to mode collapse. Moreover,

it has been shown that GANs focus more on fidelity rather than capturing the diversity of the training data distribution compared with likelihood-based models like diffusion models and auto-regressive models [41]. Besides, GANs usually adopt a CNN architecture (although Transformer structure is explored in some works [211]–[214]), which make them struggling to handle multimodal data in a unified manner and generalize to new MISE tasks. With wide adoption of Transformer backbone, auto-regressive and diffusion models can handle different MISE tasks in a unified manner. However, they both suffer from slow inference speed, due to the auto-regressive prediction of token and large numbers of diffusion steps. Overall, Transformer-based auto-regressive models and diffusion models are more favored in SOTA methods compared with GANs, especially for text-to-image synthesis.

Auto-regressive models and diffusion models are likely-based generative models which is equipped with stationary training objective and easy scalability. The comparison of generative modeling capability between auto-regressive models and diffusion is still inconclusive. DALL-E 2 [45] shows that diffusion models are slightly better than auto-regressive models in modeling the diffusion prior. However, the recent work Parti [165] which adopts an auto-regressive structure presents superior performance over the SOTA work of diffusion-based methods (i.e., Imagen). On the other hand, the exploration of two different families of generative models may open exciting opportunities to combine the merits of the two powerful models.

Different from above generation methods which mainly work on 2D images and have few requirements for the training datasets, NeRF-based methods handle the 3D scene geometry and thus have relatively high requirements for training data, e.g., per-scene optimization NeRFs require multiview images with camera pose annotation, generative NeRFs require the scene geometry of the dataset to be simple. Thus, the application of NeRF in multimodal image synthesis and editing is still quite constrained. Nevertheless, the 3D-aware modeling of real world with NeRF opens a new door for MISE tasks in the future research.

Besides, state-of-the-art methods prone to combine different generative models to yield superior performance. For example, Taming Transformer [38] incorporates VQ-GAN and Auto-regressive models to achieve high-resolution image synthesis, StyleNeRF [198] combines NeRF with GAN to enable adversarial image synthesis with both high-fidelity and 3D-awareness; ImageBart [39] combines the autoregressive formulation with a multinomial diffusion process to incorporate a coarse-to-fine hierarchy of context information; X-LXMERT [215] integrates GAN into the framework of cross-modality representation to achieve text-guided image generation.

## 4 EXPERIMENTAL EVALUATION

### 4.1 Datasets and Evaluation

Datasets are at the core of image synthesis and editing tasks. To give an overall picture of the datasets in MISE, we tabulate the detailed annotation types in popular datasets in Table 1. Notably, ADE20K [216], COCO-Stuff [218], and

Cityscapes [219] are common benchmark datasets for semantic image synthesis; Oxford-120 Flowers [226], CUB-200 Birds [227], and COCO [217] are widely adopted in text-to-image synthesis; VoxCeleb2 [231] and Lip Reading in the Wild (LRW) [232] are usually used for the benchmark of taking face generation. Please refer to the supplementary material for more details of the widely adopted datasets in different modalities.

Precise evaluation metrics are also of great importance in driving progress of research. On the other hand, the evaluation of MISE tasks is challenging as multiple attributes account for a fine generation result and the notion of image evaluation is often subjective. To achieve faithful evaluation, comprehensive metrics are adopted to evaluate MISE tasks from multiple aspects including image quality (e.g., IS [233] and FID [234]), image diversity (e.g., LPIPS [37]), guidance alignment (R-precision [14], VS similarity [112], Captioning Metrics [235], Semantic Object Accuracy [236], Audio-lip Synchronization [237]), and user studies. Besides, some evaluation metrics are specially designed for certain generation tasks, e.g., mIoU and mAP for semantic image synthesis, cumulative probability blur detection (CPBD) [238] for talking face generation. Please refer to the supplementary material for more details of the corresponding evaluation metrics.

### 4.2 Experimental Results

We quantitatively compare the image synthesis performance of different models in terms of visual guidance, text guidance, and audio guidance.

#### 4.2.1 Visual Guidance

For visual guidance, we mainly conduct comparison on semantic image synthesis as there are numbers of works for benchmarking. As shown in Table 2, the experimental comparison is conducted on four challenging datasets: ADE20K [216], ADE20K-outdoors [216], COCO-stuff [218] and Cityscapes [219], following the setting of [3]. The evaluation is performed with Fréchet Inception Distance (FID) [255] and mean Intersection-over-Union (mIoU). Specially, the mIoU aims to assess the alignment between the generated image and the ground truth segmentation via a pre-trained semantic segmentation network. Pre-trained UperNet101 [256], multi-scale DRN-D-105 [257], and DeepLabV2 [258] are adopted for Cityscapes, ADE20K & ADE20K-outdoors, and COCO-Stuff, respectively.

As shown in Table 2, diffusion-based method (i.e., SDM [180]) achieves superior generation quality as evaluated by FID and comparable semantic consistency as evaluated by mIoU, compared with GAN-based methods and auto-regressive methods. Although the comparison may not be fair as the model sizes are varied, diffusion-based method still demonstrates its powerful modeling capability for image synthesis. With a large model size, the auto-regressive method (i.e., Taming [38]) doesn't show a clear advantage over other methods. We conjecture that the Taming Transformer [38] is a versatile framework for various conditional generation tasks without specific design for semantic image synthesis, while other methods in Table 2 mainly focus on the task of semantic image synthesis. Notably, auto-regressive method and diffusion method inherently support

TABLE 1

Annotation types in popular datasets for MISE. Note, only currently available annotations are labeled with checkmarks, although some off-the-shelf models (e.g., segmentation models, edge detectors, image caption models) can be employed to annotate the corresponding datasets. Part of the information is retrieved from [113].

Datasets	Semantic Map	Keypoint	Sketch	Bounding Box	Depth	Attribute	Text	Audio	Scene Graph
ADE20K [216]	✓	✗	✗	✗	✗	✗	✗	✗	✗
COCO [217]	✓	✓	✗	✓	✗	✗	✓	✗	✓
COCO-Stuff [218]	✓	✗	✗	✗	✗	✗	✗	✗	✗
Cityscapes [219]	✓	✗	✗	✗	✗	✗	✗	✗	✗
CelebA [220]	✗	✓	✗	✗	✗	✓	✗	✗	✗
CelebA-HQ [221]	✗	✓	✗	✗	✗	✓	✗	✗	✗
CelebAMask-HQ [4]	✓	✓	✗	✗	✗	✓	✗	✗	✗
CelebA-Dialog [136]	✗	✗	✗	✗	✗	✓	✓	✗	✗
MM-CelebA-HQ [27]	✓	✓	✓	✗	✗	✓	✓	✗	✗
DeepFashion [222]	✗	✓	✗	✗	✗	✓	✗	✗	✗
DeepFashion-MM [186]	✓	✓	✗	✗	✗	✓	✓	✗	✗
Chictopia10K [223]	✓	✗	✗	✗	✗	✗	✗	✗	✗
NYU Depth [224]	✓	✗	✗	✗	✓	✗	✗	✗	✗
Stanford’s Cars [225]	✗	✗	✗	✓	✓	✓	✗	✗	✗
Oxford-102 [226]	✗	✗	✗	✗	✗	✓	✗	✗	✗
CUB-200 [227]	✓	✗	✗	✓	✗	✓	✗	✗	✗
Visual Genome [228]	✗	✗	✗	✓	✗	✓	✓	✗	✓
VoxCeleb [229]	✗	✗	✗	✗	✗	✗	✗	✓	✗
LRS [230]	✗	✗	✗	✗	✗	✗	✓	✓	✗

TABLE 2

Visual guided (semantic map) image synthesis performance of on different benchmark datasets. The parameter number (#param) is evaluated on ADE20K. The rows in grey and cyan denote the results of Transformer-based and Diffusion-based methods, respectively. Others are the results of GAN-based methods. Part of the results is retrieved from [239].

Methods	# param	VGG	ADE20K		ADE-outdoor		Cityscapes		COCO-stuff	
			FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CRN [240]	84M	✓	73.3	22.4	99.0	16.5	104.7	52.4	70.4	23.7
SIMS [241]	56M	✓	-	-	67.7	13.1	49.7	47.2	-	-
Pix2pixHD [99]	183M	✓	81.8	20.3	97.8	17.4	95.0	58.3	111.5	14.6
LGGAN [242]	-	✓	31.6	41.6	-	-	57.7	68.4	-	-
CC-FPSE [243]	131M	✓	31.7	43.7	-	-	54.3	65.5	19.2	41.6
SPADE [3]	102M	✓	33.9	38.5	63.3	30.8	71.8	62.3	22.6	37.4
GroupDNet [244]	68M	✓	39.1	26.09	-	-	41.1	59.2	-	-
INADE [245]	91M	✓	29.6	34.96	-	-	38.0	61.0	-	-
OASIS [239]	94M	✗	28.3	48.8	48.6	40.4	47.7	69.3	17.0	44.1
Taming [38]	465M	✓	35.5	-	-	-	-	-	-	-
SDM [180]	-	✓	27.5	39.2	-	-	42.1	77.5	-	-

diverse conditional generation results, while GAN-based methods usually require additional modules (e.g., Variational Autoencoders (VAEs) [171]) or designs to achieve diverse generation.

#### 4.2.2 Text Guidance

We benchmark text-to-image generation methods on COCO dataset as tabulated in 3 (The results are extracted from relevant papers). As shown in Table 3, GAN-based, auto-regressive, and diffusion-based methods can all achieve SOTA performance in terms of FID, e.g., 8.12 in GAN-based method LAFITE [253], 7.23 in auto-regressive method Parti [165], 7.27 in diffusion-based method Imagen [46]. However, auto-regressive and diffusion-based methods are still preferred in recent SOTA works, thanks to their stationary training objective and easy scalability [41].

#### 4.2.3 Audio Guidance

In terms of audio guided image synthesis and editing, we conduct quantitative comparison in the task of audio-driven talking face generation as there are numbers of benchmark methods for comparison. Notably, current development of

talking face generation methods mainly rely on GANs, and auto-regressive or diffusion-based method for talking face generation still remains for exploration. The quantitative results of talking face generation on LFW [232] and VoxCeleb2 [231] datasets are shown in Table 4. For some evaluation metrics, the listed methods may outperform the result of ground truth. It is reasonable and just proves that the result of the method is nearly comparable to the ground truth.

## 5 OPEN CHALLENGES & FUTURE EXPLORATION

Though multimodal image synthesis and editing have made notable progress and achieved superior performance in recent years, there exist several challenges for future exploration. In this section, we overview the typical challenges, share our humble opinions on possible solutions, and highlight the future research directions.

### 5.1 Towards Large-Scale Multi-Modality Datasets

As current datasets mainly provide annotations in a single modality (e.g., visual guidance), most existing methods focus on image synthesis and editing conditioned on

TABLE 3

Text-to-Image generation performance on the COCO dataset. <sup>†</sup> denotes the results obtained by using the corresponding open-source code. The rows in grey and cyan denote the results of Transformer-based and Diffusion-based methods, respectively. Others are the results of GAN-based methods. Part of the results is retrieved from [77].

Model	IS $\uparrow$	FID $\downarrow$	R-Prec. $\uparrow$
Real Images [236]	34.88	6.09	68.58
StackGAN [16]	8.450	74.05	-
StackGAN++ [17]	8.300	81.59	-
AttnGAN [14]	25.89	35.20	85.47
MirrorGAN [19]	26.47	-	74.52
AttnGAN+OP [236]	24.76	33.35	82.44
OP-GAN [236]	27.88	24.70	89.01
SEGAN [119]	27.86	32.28	-
ControlGAN [121]	24.06	-	82.43
DM-GAN [246]	30.49	32.64	88.56
DM-GAN [246] <sup>†</sup>	32.43	24.24	92.23
Obj-GAN [247]	27.37	25.64	91.05
Obj-GAN [247] <sup>†</sup>	27.32	24.70	91.91
TVBi-GAN [248]	31.01	31.97	-
Wang et al. [249]	29.03	16.28	82.70
Rombach et al. [250]	34.70	30.63	-
CPGAN [251]	<b>52.73</b>	-	<b>93.59</b>
Pavllo et al. [82]	-	19.65	-
XMC-GAN [252]	30.45	9.330	-
LAFITE [253]	32.34	8.120	-
CogView [144]	18.20	27.10	-
CogView2 [254]	22.40	24.10	-
DALL-E [35]	17.90	27.50	-
NUWA [40]	27.20	12.90	-
DiVAE [156]	-	11.53	-
Make-A-Scene [150]	-	11.84	-
Parti [165]	-	<b>7.230</b>	-
VQ-Diffusion [179]	-	13.86	-
LDM [185]	30.29	12.63	-
GLIDE [177]	-	12.24	-
DALL-E 2 [45]	-	10.39	-
Imagen [46]	-	7.270	-

guidance from a single modality (*e.g.*, text-to-image synthesis, semantic image synthesis). However, humans possess the capability of creating visual contents with guidance of multiple modalities concurrently. Targeting to mimic the human intelligence, multimodal inputs are expected to be effectively fused and leveraged in image generation concurrently. Recently, Maka-A-Scene [150] explores to include semantic segmentation tokens in auto-regressive modeling, which helps to achieve better quality in image synthesis; Xia *et al.* [27] introduce a MM-CelebA-HQ dataset which is annotated with semantic segmentation, edge maps, and text descriptions. With MM-CelebA-HQ [27], COCO [217], and COCO-Stuff [218] as the training set, PoE-GAN [264] achieves image generation conditioned on multi-modal including segmentation, sketch, image, and text. However, the size of MM-CelebA-HQ [27], COCO [217], and COCO-Stuff [218] is still far from narrowing the gap with real-world distributions. Thus, to incorporate comprehensive modalities into image generation, a natural scene dataset which is equipped with annotations from a wide spectrum of modalities (*e.g.*, semantic segmentation, text description,

scene graph) is expected to be created and made publicly available in the future.

## 5.2 Towards Faithful Evaluation Metrics

The evaluation of multimodal image synthesis and editing is still an open problem. Leveraging pre-trained models to conduct evaluations (*e.g.*, FID) is constraint to the pre-trained datasets, which tends to pose discrepancy with the target datasets. User study recruits human subjects to assess the synthesized images directly, which is however often subjective. Designing accurate yet faithful evaluation metrics is thus very meaningful and critical to the development of multimodal image synthesis and editing.

Recently, to evaluate the cross-modal alignment in text-to-image synthesis, pre-trained CLIP has been used to measure the similarity between the texts and the corresponding generated images. Notably, current advance in multimodal learning [265] enables a more accurate cross-modal coherence, which may contribute to the future development of faithful evaluation metrics.

## 5.3 Towards Efficient Network Architecture

With inherent support for multimodal input and powerful generative modeling, Transformer-based auto-regressive models or diffusion models have been a new paradigm for unified multimodal image synthesis and editing. However, both auto-regressive models and diffusion models suffer from slow inference speed, which is more severe in high-resolution image synthesis. Some recent works [266], [267] explore to accelerate auto-regressive models and diffusion models, while the experiments are constrained to toy datasets with low resolution like CIFAR-10 [268]. How to accelerate the inference speed of auto-regressive models and diffusion models for practical applications remains a grand challenge for future exploration.

On the other hand, GAN-based methods have a better inference speed while the widely adopted CNN architecture hinders it from unified handling of multimodal input. Recently, some works [269], [270] have explored to adopt Transformer-based architecture in GANs, which may provide some insights into developing GANs for unified handling of MISE tasks.

## 5.4 Towards 3D Awareness

With the emergence of neural scene representation models especially NeRF, 3D-aware image synthesis and editing has the potential to be the next breaking point for MISE as it models the 3D geometry of real world. With the incorporation of adversarial loss, generative NeRF is notably appealing for MISE as it is associated with a latent space. Current generative NeRF models (*e.g.*, StyleNeRF, EG3D) have enabled to model scenes with simple geometry (*e.g.*, faces, cars) from a collection of unposed 2D images, just like the training of unconditional GANs (*e.g.*, StyleGAN). Powered by these efforts, some 3D-aware MISE tasks have been explored, *e.g.*, text-to-NeRF [192] and semantic-to-NeRF [201]. However, current generative NeRFs still struggle on datasets with complex geometry variation, *e.g.*, DeepFashion [222] and ImageNet [145].

TABLE 4

The audio guided image editing (talking-head) performance on LRW [232] and VoxCeleb2 [231] under four metrics.  $\dagger$  denotes that the model is evaluated by directly using the authors' generated samples under their setting. Part of the results is retrieved from [53].

Method	LRW [232]				VoxCeleb2 [231]			
	SSIM $\uparrow$	CPBD $\uparrow$	LMD $\downarrow$	Sync <sub>conf</sub> $\uparrow$	SSIM $\uparrow$	CPBD $\uparrow$	LMD $\downarrow$	Sync <sub>conf</sub> $\uparrow$
ATVG [123]	0.810	0.102	5.25	4.1	0.826	0.061	6.49	4.3
Wav2Lip [259]	<b>0.862</b>	0.152	5.73	<b>6.9</b>	0.846	0.078	12.26	4.5
MakeitTalk [128]	0.796	0.161	7.13	3.1	0.817	0.068	31.44	2.8
Rhythmic Head $\dagger$ [124]	-	-	-	-	0.779	0.802	14.76	3.8
PC-AVS [53] (Fix Pose)	0.815	0.180	6.14	6.3	0.820	<b>0.084</b>	7.68	5.8
PC-AVS [53]	<b>0.861</b>	<b>0.185</b>	3.93	6.4	<b>0.886</b>	<b>0.083</b>	6.88	<b>5.9</b>
GC-AVT [260]	-	-	-	-	0.710	-	<b>3.03</b>	5.3
EAMM [261]	0.740	-	2.08	5.5	-	-	-	-
SyncTalkFace [262]	0.893	-	<b>1.25</b>	-	-	-	-	-
AVCT [263]	-	-	-	-	-	0.564	0.525	6.98
Ground Truth	1.000	0.173	0.00	6.5	1.000	0.090	0.00	5.9

Only relying on adversarial loss to learn the complex scene geometry from unposed 2D images is indeed intractable and challenging. A possible solution is to provide more prior knowledge of the scene, *e.g.*, obtaining prior scene geometry with off-the-shelf reconstruction models, providing skeleton prior for generative human modeling. Notably, the power of prior knowledge has been explored in some recent works of 3D-aware style transfer [206] and NeRF generalization [271]. Another possible approach is to provide more supervision, *e.g.*, creating a large dataset with multiview annotations or geometry information. Once the 3D-aware generative modelling succeeds to work on complex natural scenes, some interesting multimodal applications will become possible, *e.g.*, 3D version of DALL-E.

## 6 SOCIAL IMPACTS

The rapid advancements in multi-modal image synthesis and editing offer unprecedented generation realism and editing possibilities, which have influenced and will continue to influence our society in both positive and potentially negative ways. In this section, we will discuss these potential social impacts.

### 6.1 Applications

The multi-modal image synthesis and editing techniques can be applied in artistic creation and content generation, which could widely benefit designers, photographers, and content creators [272]. Moreover, they can be democratized in everyday applications as image generation or editing tools for popular entertainment. In addition, the various conditions as intermediate representations for synthesis & editing greatly ease the use of the methods and improve the flexibility of user interaction. In general, the techniques greatly lower the barrier for the public and unleash their creativity on content generation and editing.

### 6.2 Misuse

On the other hand, the increasing editing capability and generation realism also offers opportunities to generate or manipulate images for malicious purposes. The misuse of synthesis & editing techniques may spread fake or nefarious information and lead to negative social impacts. To prevent potential misuses, one potential way is to develop detection

techniques for automatically identifying GAN-generated or fake images, which has been actively researched by the community [273]. Meanwhile, sufficient guardrails, labelling, and access control should be carefully considered when deploying the multi-modal image synthesis and editing techniques to minimize the risk of misuses.

## 6.3 Environment

As deep-learning-based methods, the current multi-model generative methods inevitably require GPUs and considerable energy consumption for training and inference, which may negatively influence the environment and global climate before the large-scale use of renewable energy. One direction to soften the need for computational resources lies in the active exploration of model generalization. For example, a pretrained model generalized in various datasets could greatly accelerate the training process or provide semantical knowledge for downstream tasks.

## 7 CONCLUSION

This review has covered main approaches for multimodal image synthesis and editing. We provide an overview of different guidance modalities including visual guidance, text guidance, audio guidance, and other modal guidance (*e.g.*, scene graph). We then provided a detailed introduction of the main image synthesis & editing paradigms: GAN-based methods, Auto-regressive methods, Diffusion-based methods, and NeRF-based methods. The corresponding strengths and weaknesses were comprehensively discussed to inspire new paradigm that takes advantage of the strengths of existing frameworks. We also conduct a comprehensive survey of datasets and evaluation metrics for MISE conditioned on different guidance modalities. Then, we tabularize and compare the performance of existing approaches in different MISE tasks. Last but not least, we provided our perspective on the current challenges and future directions related to integrating all modalities, comprehensive datasets, evaluation metrics, model architecture, and 3D awareness.

## ACKNOWLEDGMENTS

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

## REFERENCES

- [1] I. Goodfellow et al. Generative adversarial nets. *NeurIPS*, 27, 2014.
- [2] P. Isola et al. Image-to-image translation with conditional adversarial networks. In *CVPR*, pp. 1125–1134, 2017.
- [3] T. Park et al. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pp. 2337–2346, 2019.
- [4] C.-H. Lee et al. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pp. 5549–5558, 2020.
- [5] E. Mansimov et al. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.
- [6] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [7] M. Arjovsky et al. Wasserstein generative adversarial networks. In *ICML*, pp. 214–223, 2017.
- [8] C.-H. Lin et al. St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*, pp. 9455–9464, 2018.
- [9] F. Zhan et al. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *ICCV*, pp. 9105–9115, 2019.
- [10] A. Shrivastava et al. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, pp. 2107–2116, 2017.
- [11] J. S. Chung et al. You said that? *arXiv preprint arXiv:1705.02966*, 2017.
- [12] S. Reed et al. Generative adversarial text to image synthesis. In *ICML*, pp. 1060–1069. PMLR, 2016.
- [13] Y. Song et al. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018.
- [14] T. Xu et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pp. 1316–1324, 2018.
- [15] A. Dash et al. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*, 2017.
- [16] H. Zhang et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.
- [17] H. Zhang et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 41(8):1947–1962, 2018.
- [18] J.-Y. Zhu et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pp. 2223–2232, 2017.
- [19] T. Qiao et al. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, pp. 1505–1514, 2019.
- [20] A. Brock et al. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [21] T. Karras et al. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4401–4410, 2019.
- [22] T. Karras et al. Analyzing and improving the image quality of stylegan. In *CVPR*, pp. 8110–8119, 2020.
- [23] T. Karras et al. Alias-free generative adversarial networks. In *NeurIPS*, 2021.
- [24] L. Goetschalckx et al. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, pp. 5744–5753, 2019.
- [25] W. Xia et al. Gan inversion: A survey. *TPAMI*, 2022.
- [26] J. Zhu et al. In-domain gan inversion for real image editing. In *ECCV*, pp. 592–608. Springer, 2020.
- [27] W. Xia et al. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, pp. 2256–2265, 2021.
- [28] A. Radford et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.
- [29] O. Patashnik et al. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pp. 2085–2094, 2021.
- [30] A. Vaswani et al. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.
- [31] A. Radford et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [32] M. Chen et al. Generative pretraining from pixels. In *ICML*, pp. 1691–1703. PMLR, 2020.
- [33] P. Dhariwal et al. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.
- [34] J. T. Rolfe. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016.
- [35] A. Ramesh et al. DALL-E: Creating images from text. Technical report, OpenAI, 2021.
- [36] J. Johnson et al. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pp. 694–711. Springer, 2016.
- [37] R. Zhang et al. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.
- [38] P. Esser et al. Taming transformers for high-resolution image synthesis. *arXiv:2012.09841*, 2020.
- [39] P. Esser et al. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. In *NeurIPS*, 2021.
- [40] C. Wu et al. N\” uwa: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021.
- [41] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.
- [42] J. Ho et al. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.
- [43] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.
- [44] G. Kim and J. C. Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*, 2021.
- [45] A. Ramesh et al. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [46] C. Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.
- [47] B. Mildenhall et al. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pp. 405–421. Springer, 2020.
- [48] S. Liu et al. Editing conditional radiance fields. In *ICCV*, pp. 5773–5783, 2021.
- [49] C. Wang et al. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. *arXiv preprint arXiv:2112.05139*, 2021.
- [50] Y. Guo et al. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. *arXiv preprint arXiv:2103.11078*, 2021.
- [51] F. Zhan et al. Unbalanced feature transport for exemplar-based image translation. In *CVPR*, 2021.
- [52] A. Ramesh et al. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.
- [53] H. Zhou et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, pp. 4176–4186, 2021.
- [54] J. Johnson et al. Image generation from scene graphs. In *CVPR*, pp. 1219–1228, 2018.
- [55] L. Ma et al. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017.
- [56] Y. Men et al. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020.
- [57] C. Zhang et al. Deep monocular 3d human pose estimation via cascaded dimension-lifting. *arXiv preprint arXiv:2104.03520*, 2021.
- [58] J. Thies et al. Deferred neural rendering: image synthesis using neural textures. *TOG*, 38, 2019.
- [59] H. Kim et al. Deep video portraits. *TOG*, 37, 2018.
- [60] L. Liu et al. Neural rendering and reenactment of human actor videos. *TOG*, 2019.
- [61] L. Liu et al. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 05 2020.
- [62] M. Meshry et al. Neural rerendering in the wild. In *CVPR*, 2019.
- [63] K.-A. Aliev et al. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*, 2019.
- [64] J.-Y. Zhu et al. Toward multimodal image-to-image translation. In *NeurIPS*, pp. 465–476, 2017.
- [65] H.-Y. Lee et al. Diverse image-to-image translation via disentangled representations. In *ECCV*, pp. 35–51, 2018.
- [66] W. Sun and T. Wu. Image synthesis from reconfigurable layout and style. In *ICCV*, pp. 10531–10540, 2019.
- [67] B. Zhao et al. Image generation from layout. In *CVPR*, pp. 8584–8593, 2019.
- [68] Y. Li et al. Bachgan: High-resolution image synthesis from salient object layout. In *CVPR*, 2020.
- [69] Z. Li et al. Image synthesis from layout with locality-aware mask adaption. In *ICCV*, pp. 13819–13828, 2021.
- [70] S. Frolov et al. Attrlostgan: Attribute controlled image synthesis from reconfigurable layout and style. *arXiv preprint arXiv:2103.13722*, 2021.
- [71] J. Y. Koh et al. Text-to-image generation grounded by fine-grained user attention. In *WACV*, pp. 237–246, 2021.
- [72] T. Sun et al. Single image portrait relighting. *TOG*, 38(4):79–1, 2019.

- [73] P. P. Srinivasan et al. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, pp. 7495–7504, 2021.
- [74] F. Zhan et al. Emlight: Lighting estimation via spherical distribution approximation. In *AAAI*, pp. 3287–3295, 2021.
- [75] F. Zhan et al. Gmlight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244*, 2021.
- [76] T. Park et al. Contrastive learning for unpaired image-to-image translation. In *ECCV*, pp. 319–345. Springer, 2020.
- [77] S. Frolov et al. Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209, 2021.
- [78] T. Mikolov et al. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pp. 3111–3119, 2013.
- [79] Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.
- [80] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [81] T. Wang et al. Faces à la carte: Text-to-face generation via attribute disentanglement. In *WACV*, pp. 3380–3388, 2021.
- [82] D. Pavllo et al. Controlling style and semantics in weakly-supervised image generation. In *ECCV*, pp. 482–499. Springer, 2020.
- [83] J. Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [84] D. Harwath and J. R. Glass. Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*, 2017.
- [85] J. Li et al. Direct speech-to-image translation. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):517–529, 2020.
- [86] D. Harwath et al. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4969–4973. IEEE, 2018.
- [87] Y. Aytar et al. Soundnet: Learning sound representations from unlabeled video. *NeurIPS*, 29:892–900, 2016.
- [88] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [89] A. Owens et al. Visually indicated sounds. In *CVPR*, pp. 2405–2413, 2016.
- [90] P. Ekman et al. Facial action coding system (facs) a human face. *Salt Lake City*, 2002.
- [91] D. M. Vo and A. Sugimoto. Visual-relation conscious image generation from structured-text. In *ECCV*, pp. 290–306. Springer, 2020.
- [92] X. Shi et al. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *NeurIPS*, 28, 2015.
- [93] R. Liu et al. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018.
- [94] C. H. Lin et al. Coco-gan: Generation by parts via conditional coordinating. In *ICCV*, pp. 4512–4521, 2019.
- [95] A. v. d. Oord et al. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.
- [96] Y. Bengio et al. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
- [97] A. Baevski et al. Vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.
- [98] E. Jang et al. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- [99] T.-C. Wang et al. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pp. 8798–8807, 2018.
- [100] H. Tang et al. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, pp. 2417–2426, 2019.
- [101] P. Zhu et al. Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR*, June 2020.
- [102] T. R. Shaham et al. Spatially-adaptive pixelwise networks for fast image translation. *arXiv preprint arXiv:2012.02992*, 2020.
- [103] P. Zhang et al. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, pp. 5143–5153, 2020.
- [104] X. Zhou et al. Full-resolution correspondence learning for image translation. *arXiv preprint arXiv:2012.02047*, 2020.
- [105] C. Barnes et al. Patchmatch: A randomized correspondence algorithm for structural image editing. *TOG*, 28(3):24, 2009.
- [106] F. Zhan et al. Bi-level feature alignment for versatile image translation and manipulation. *arXiv preprint arXiv:2107.03021*, 2021.
- [107] S. Benaim and L. Wolf. One-sided unsupervised domain mapping. *NeurIPS*, 30, 2017.
- [108] M. Amadio and S. Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *CVPR*, pp. 8983–8992, 2019.
- [109] H. Fu et al. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *CVPR*, pp. 2427–2436, 2019.
- [110] A. v. d. Oord et al. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [111] A. Andonian et al. Contrastive feature loss for image prediction. In *ICCV*, pp. 1934–1943, 2021.
- [112] Z. Zhang et al. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, pp. 6199–6208, 2018.
- [113] Y. Xue et al. Deep image synthesis from intuitive user input: A review and perspectives. *Computational Visual Media*, 8(1):3–31, 2022.
- [114] A. Nguyen et al. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, pp. 4467–4477, 2017.
- [115] Q. Lao et al. Dual adversarial inference for text-to-image synthesis. In *ICCV*, pp. 7567–7576, 2019.
- [116] Z. Chen and Y. Luo. Cycle-consistent diverse image synthesis from natural language. In *ICMEW*, pp. 459–464. IEEE, 2019.
- [117] D. Bahdanau et al. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [118] W. Huang et al. Realistic image generation using region-phrase attention. In *Asian Conference on Machine Learning*, pp. 284–299. PMLR, 2019.
- [119] H. Tan et al. Semantics-enhanced adversarial nets for text-to-image synthesis. In *ICCV*, pp. 10501–10510, 2019.
- [120] Z. Lin et al. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.
- [121] B. Li et al. Controllable text-to-image generation. *arXiv preprint arXiv:1909.07083*, 2019.
- [122] J. Cheng et al. Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In *CVPR*, pp. 10911–10920, 2020.
- [123] L. Chen et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pp. 7832–7841, 2019.
- [124] L. Chen et al. Talking-head generation with rhythmic head motion. In *ECCV*, pp. 35–51. Springer, 2020.
- [125] H. Zhou et al. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, pp. 9299–9306, 2019.
- [126] S. Suwajanakorn et al. Synthesizing obama: learning lip sync from audio. *TOG*, 36(4):1–13, 2017.
- [127] S. Wang et al. One-shot talking face generation from single-speaker audio-visual correlation learning. *arXiv preprint arXiv:2112.02749*, 2021.
- [128] Y. Zhou et al. Makelttalk: speaker-aware talking-head animation. *TOG*, 39(6):1–15, 2020.
- [129] R. Yi et al. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020.
- [130] S. Wang et al. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.
- [131] V. Blanz et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, pp. 187–194, 1999.
- [132] E. Richardson et al. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020.
- [133] H. Wang et al. Cycle-consistent inverse gan for text-to-image synthesis. In *MM*, pp. 630–638, 2021.
- [134] H. Dong et al. Semantic image synthesis via adversarial learning. In *ICCV*, pp. 5706–5714, 2017.
- [135] R. Kiros et al. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [136] Y. Jiang et al. Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV*, pp. 13799–13808, 2021.
- [137] D. Bau et al. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021.
- [138] U. Kocasari et al. Stylemc: Multi-channel based fast text-guided image generation and manipulation. *arXiv preprint arXiv:2112.08493*, 2021.

- [139] Y. Yu et al. Towards counterfactual image manipulation via clip. *arXiv preprint arXiv:2207.02812*, 2022.
- [140] X. Liu et al. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021.
- [141] T. Wei et al. Hairclip: Design your hair by text and reference image. *arXiv preprint arXiv:2112.05142*, 2021.
- [142] R. Gal et al. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.
- [143] Y. Yu et al. Diverse image inpainting with bidirectional and autoregressive transformers. In *MM*, pp. 69–78, 2021.
- [144] M. Ding et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021.
- [145] J. Deng et al. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [146] A. Lamb et al. Discriminative regularization for generative models. *arXiv preprint arXiv:1602.03220*, 2016.
- [147] A. B. L. Larsen et al. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, pp. 1558–1566. PMLR, 2016.
- [148] X. Dong et al. Poco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.
- [149] H. Bao et al. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [150] O. Gafni et al. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.
- [151] Q. Cao et al. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pp. 67–74. IEEE, 2018.
- [152] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [153] J. Yu et al. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.
- [154] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
- [155] M. Ni et al. N\wua-lip: Language guided image inpainting with defect-free vqgan. *arXiv preprint arXiv:2202.05009*, 2022.
- [156] J. Shi et al. Divae: Photorealistic images synthesis with denoising diffusion decoder. *arXiv preprint arXiv:2206.00386*, 2022.
- [157] W. Shin et al. Translation-equivariant image quantizer for bi-directional image-text generation. *arXiv preprint arXiv:2112.00384*, 2021.
- [158] F. Zhan et al. Auto-regressive image synthesis with integrated quantization. In *ECCV*, 2022.
- [159] A. Van den Oord et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016.
- [160] N. Parmar et al. Image transformer. In *ICML*, 2018.
- [161] E. Hoogeboom et al. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. *arXiv preprint arXiv:2102.05379*, 2021.
- [162] J. Sohl-Dickstein et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pp. 2256–2265. PMLR, 2015.
- [163] H. Chang et al. Maskgit: Masked generative image transformer. In *CVPR*, pp. 11315–11325, 2022.
- [164] Z. Zhang et al. M6-ufc: Unifying multi-modal controls for conditional image synthesis. *arXiv preprint arXiv:2105.14211*, 2021.
- [165] J. Yu et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.
- [166] Y. Huang et al. A picture is worth a thousand words: A unified system for diverse captions and rich images generation. In *MM*, pp. 2792–2794, 2021.
- [167] Y. Huang et al. Unifying multimodal transformer for bi-directional image and text generation. In *MM*, pp. 1138–1147, 2021.
- [168] J. Song et al. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [169] Y. Song et al. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [170] A. Jolicoeur-Martineau et al. Adversarial score matching and improved sampling for image generation. *arXiv preprint arXiv:2009.05475*, 2020.
- [171] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [172] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *ICML*, pp. 1530–1538. PMLR, 2015.
- [173] L. Dinh et al. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.
- [174] J. Menick and N. Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. *arXiv preprint arXiv:1812.01608*, 2018.
- [175] A. Van Oord et al. Pixel recurrent neural networks. In *ICML*, pp. 1747–1756. PMLR, 2016.
- [176] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- [177] A. Nichol et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [178] Z. Tang et al. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022.
- [179] S. Gu et al. Vector quantized diffusion model for text-to-image synthesis. *arXiv preprint arXiv:2111.14822*, 2021.
- [180] W. Wang et al. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022.
- [181] Y. Zhu et al. Discrete contrastive diffusion for cross-modal and conditional generation. *arXiv preprint arXiv:2206.07771*, 2022.
- [182] C. Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.
- [183] N. Liu et al. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022.
- [184] A. Blattmann et al. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022.
- [185] R. Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.
- [186] Y. Jiang et al. Text2human: Text-driven controllable human image generation. *TOG*, 41(4):1–11, 2022.
- [187] X. Liu et al. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021.
- [188] O. Avrahami et al. Blended diffusion for text-driven editing of natural images. *arXiv preprint arXiv:2111.14818*, 2021.
- [189] A. Jain et al. Zero-shot text-guided object generation with dream fields. In *CVPR*, pp. 867–876, 2022.
- [190] P. Wang et al. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.
- [191] F. Hong et al. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *TOG*, 41(4):1–19, 2022.
- [192] K. Jo et al. Cg-nerf: Conditional generative neural radiance fields. *arXiv preprint arXiv:2112.03517*, 2021.
- [193] K. Schwarz et al. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 33:20154–20166, 2020.
- [194] M. Niemeyer and A. Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pp. 11453–11464, 2021.
- [195] E. R. Chan et al. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pp. 5799–5809, 2021.
- [196] E. Perez et al. Film: Visual reasoning with a general conditioning layer. In *AAAI*, volume 32, 2018.
- [197] V. Sitzmann et al. Implicit neural representations with periodic activation functions. *NeurIPS*, 33:7462–7473, 2020.
- [198] J. Gu et al. Stylenet: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*, 2022.
- [199] E. R. Chan et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pp. 16123–16133, 2022.
- [200] J. Sun et al. Fenerf: Face editing in neural radiance fields. In *CVPR*, pp. 7672–7682, 2022.
- [201] Y. Chen et al. Sem2nerf: Converting single-view semantic masks to neural radiance fields. *arXiv preprint arXiv:2203.10821*, 2022.
- [202] W. Chen and J. Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *CVPR*, pp. 9416–9425, 2018.
- [203] J. Sun et al. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022.
- [204] T.-J. Fu et al. Language-driven image style transfer. *arXiv preprint arXiv:2106.00178*, 2021.
- [205] G. Kwon and J. C. Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021.
- [206] F. Mu et al. 3d photo stylization: Learning to generate stylized novel views from a single image. In *CVPR*, pp. 16273–16282, 2022.
- [207] S. Loeschke et al. Text-driven stylization of video objects. *arXiv preprint arXiv:2206.12396*, 2022.

- [208] O. Michel et al. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, pp. 13492–13502, 2022.
- [209] N. Khalid et al. Text to mesh without 3d supervision using limit subdivision. *arXiv preprint arXiv:2203.13333*, 2022.
- [210] Z. Liu et al. Towards implicit text-guided 3d shape generation. In *CVPR*, pp. 17896–17906, 2022.
- [211] Y. Jiang et al. Transgan: Two pure transformers can make one strong gan, and that can scale up. *NeurIPS*, 34:14745–14758, 2021.
- [212] J. Park and Y. Kim. Styleformer: Transformer based generative adversarial networks with style vector. In *CVPR*, pp. 8983–8992, 2022.
- [213] D. A. Hudson and C. L. Zitnick. Generative adversarial transformers. *ICML*, 2021.
- [214] D. Arad Hudson and L. Zitnick. Compositional transformers for scene generation. *NeurIPS*, 34, 2021.
- [215] J. Cho et al. X-lmert: Paint, caption and answer questions with multi-modal transformers. *arXiv preprint arXiv:2009.11278*, 2020.
- [216] B. Zhou et al. Scene parsing through ade20k dataset. In *CVPR*, pp. 633–641, 2017.
- [217] T.-Y. Lin et al. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.
- [218] H. Caesar et al. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pp. 1209–1218, 2018.
- [219] M. Cordts et al. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.
- [220] Z. Liu et al. Deep learning face attributes in the wild. In *ICCV*, pp. 3730–3738, 2015.
- [221] T. Karras et al. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [222] Z. Liu et al. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pp. 1096–1104, 2016.
- [223] X. Liang et al. Deep human parsing with active template regression. *TPAMI*, 37(12):2402–2414, 2015.
- [224] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *ICCV workshops*, pp. 601–608. IEEE, 2011.
- [225] J. Krause et al. 3d object representations for fine-grained categorization. In *ICCV workshops*, pp. 554–561, 2013.
- [226] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.
- [227] P. Welinder et al. Caltech-ucsd birds 200. *California Institute of Technology*, 2010.
- [228] R. Krishna et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- [229] A. Nagrani et al. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.
- [230] J. Son Chung et al. Lip reading sentences in the wild. In *CVPR*, pp. 6447–6456, 2017.
- [231] J. S. Chung et al. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.
- [232] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pp. 87–103. Springer, 2016.
- [233] T. Salimans et al. Improved techniques for training gans. *NeurIPS*, 29:2234–2242, 2016.
- [234] M. Heusel et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pp. 6626–6637, 2017.
- [235] S. Hong et al. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, pp. 7986–7994, 2018.
- [236] T. Hinz et al. Semantic object accuracy for generative text-to-image synthesis. *arXiv preprint arXiv:1910.13321*, 2019.
- [237] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.
- [238] N. D. Narvekar and L. J. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *IEEE Transactions on Image Processing*, 20(9):2678–2683, 2011.
- [239] V. Sushko et al. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020.
- [240] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, pp. 1511–1520, 2017.
- [241] X. Qi et al. Semi-parametric image synthesis. In *CVPR*, pp. 8808–8816, 2018.
- [242] H. Tang et al. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, pp. 7870–7879, 2020.
- [243] X. Liu et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv preprint arXiv:1910.06809*, 2019.
- [244] Z. Zhu et al. Semantically multi-modal image synthesis. In *CVPR*, pp. 5467–5476, 2020.
- [245] Z. Tan et al. Diverse semantic image synthesis via probability distribution modeling. In *CVPR*, pp. 7962–7971, 2021.
- [246] M. Zhu et al. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pp. 5802–5810, 2019.
- [247] W. Li et al. Object-driven text-to-image synthesis via adversarial training. In *CVPR*, pp. 12174–12182, 2019.
- [248] Z. Wang et al. Text to image synthesis with bidirectional generative adversarial network. In *ICME*, pp. 1–6. IEEE, 2020.
- [249] M. Wang et al. End-to-end text-to-image synthesis with spatial constraints. *TIST*, 11(4):1–19, 2020.
- [250] R. Rombach et al. Network-to-network translation with conditional invertible neural networks. *arXiv preprint arXiv:2005.13580*, 2020.
- [251] J. Liang et al. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In *ECCV*, pp. 491–508. Springer, 2020.
- [252] H. Zhang et al. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, pp. 833–842, 2021.
- [253] Y. Zhou et al. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021.
- [254] M. Ding et al. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.
- [255] M. Heusel et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.
- [256] T. Xiao et al. Unified perceptual parsing for scene understanding. In *ECCV*, pp. 418–434, 2018.
- [257] F. Yu et al. Dilated residual networks. In *CVPR*, pp. 472–480, 2017.
- [258] L.-C. Chen et al. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.
- [259] K. Prajwal et al. A lip sync expert is all you need for speech to lip generation in the wild. In *MM*, pp. 484–492, 2020.
- [260] B. Liang et al. Expressive talking head generation with granular audio-visual control. In *CVPR*, pp. 3387–3396, 2022.
- [261] X. Ji et al. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. *arXiv preprint arXiv:2205.15278*, 2022.
- [262] S. J. Park et al. SyncTalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *AAAI 22. Association for the Advancement of Artificial Intelligence*, 2022.
- [263] S. Wang et al. One-shot talking face generation from single-speaker audio-visual correlation learning. In *AAAI*, 2022.
- [264] X. Huang et al. Multimodal conditional image synthesis with product-of-experts gans. *arXiv preprint arXiv:2112.05130*, 2021.
- [265] J. Summairia et al. Recent advances and trends in multimodal deep learning: A review. *arXiv preprint arXiv:2105.11087*, 2021.
- [266] V. Jayaram and J. Thickstun. Parallel and flexible sampling from autoregressive models via langevin dynamics. In *ICML*, pp. 4807–4818. PMLR, 2021.
- [267] T. Dockhorn et al. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.
- [268] A. Krizhevsky et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009.
- [269] X. Wang et al. SceneFormer: Indoor scene generation with transformers. *arXiv preprint arXiv:2012.09793*, 2020.
- [270] B. Zhang et al. Styleswin: Transformer-based gan for high-resolution image generation. In *CVPR*, pp. 11304–11314, 2022.
- [271] Q. Xu et al. Point-nerf: Point-based neural radiance fields. In *CVPR*, pp. 5438–5448, 2022.
- [272] J. Bailey. The tools of generative art, from flash to neural networks. *Art in America*, 8, 2020.
- [273] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.