Neurocomputing 497 (2022) 110–128

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A survey on multimodal-guided visual content synthesis

Ziqi Zhang [a], Zeyu Li [a], Kun Wei [a], Siduo Pan [a], Cheng Deng [a,*]

[a] School of Electronic Engineering, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 30 September 2021
Revised 14 April 2022
Accepted 24 April 2022
Available online 2 May 2022

Keywords:
Deep learning
Multimodal
Visual content synthesis
GAN

ABSTRACT

With the increasing interest in various creative scenes such as social media, film production, and intelligence courses, people expect to be able to compile rich visual content according to their subjective ideas and actual needs. In this context, visual content synthesis technique based on multimodal data has attracted much attention in recent years. Compared to traditional generative methods, multimodal data offer more flexible and concrete clues that provide an interactive and controllable way to generate the desired visual content. In this survey, we comprehensively summarize the improvements in multimodal-guided visual content synthesis. We first formulate the taxonomy of visual content synthesis and divide it into four different subfields depending on the input modality, including visual-guided visual content synthesis, text-guided visual content synthesis, audio-guided visual content synthesis, and visual content synthesis guided by other modalities. In each subfield, we describe the paradigm of different modality-guided visual content synthesis, and also discuss the signature methods mainly based on Generative Adversarial Networks (GANs). Next, we present commonly used benchmark datasets and metrics for evaluating models, as well as detailed comparisons between different methods. Finally, we provide insight into current research challenges and possible future research directions.

1. Introduction

As one of the fundamental research areas of Deep Learning [1], visual content synthesis aims to generate or beautify images or videos that matches the target distribution based on certain inputs (text, image, video, audio, *etc.*). It is widely used in our daily life, because the essence of many creative scenes is a practical extension of visual synthesis, such as art painting, film production, and entertainment advertising. In such scenes, the creative process can be regarded as a process that constantly fitting people's subjective imagination. The development of deep visual content synthesis can help people better transform their subjective imagination into visual creations, and provide more comfort and inspiration. Therefore, the research of multimodal-guided visual content synthesis has become a topic in the field of artificial intelligence.

In recent decades, artificial intelligence technology has developed rapidly thanks to the powerful representation learning capabilities [11–14,220,227]. Various models [15–20] based on deep neural networks have achieved remarkable success in the field of computer vision. Among them, deep generative networks [21–23] have broken new ground driven by the development of hardware and availability of larger datasets. The central idea of generative modeling stems around training a generative network to fit the training data distribution $x \sim p_\theta(x)$. Early generative methods were implemented by the Restricted Boltzmann Machine (RBM) based method, which attracted considerable attention due to its strong interpretability and progress in the generation effect. However, these models suffer from a significant timing problem caused by the sampling of the data. To circumvent this issue, Variational AutoEncoders (VAEs) [22] have been proposed to fit data distribution through variational inference [24]. Subsequently, autoregressive models, such as PixelRNN [25], PixelCNN [25], and other methods have also appeared. In the last few years, with the advent of Generative Adversarial Network (GAN) [21], GAN-based generative methods have become dominant. The vanila GAN uses Jensen-Shannon divergence as loss function to measure the generative distribution and the real data distribution. To improve training instability of GAN model, a number of GAN variants with better metrics have emerged, such as LSGAN [26], WGAN [26], and WGAN-GP [27]. In addition, a number of large-scale GANs have been proposed to synthesize photorealistic and high-resolution visual content from noise inputs, such as Progressive GAN [28], BigGAN [29], StyleGAN [30,31].

The development of above models has laid the foundation for processing more complex data, making it possible to generate visual content guided by multimodal data. Like other outstanding

* Corresponding author.
E-mail address: chdeng.xd@gmail.com (C. Deng).

https://doi.org/10.1016/j.neucom.2022.04.126
0925-2312/

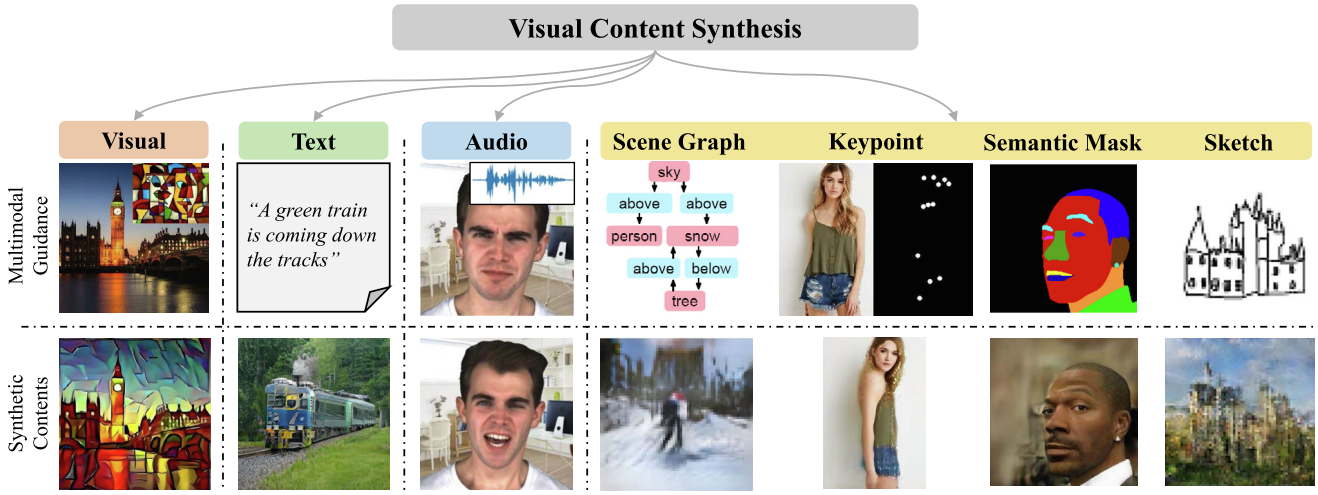

Fig. 1. A taxonomy of visual content synthesis techniques. Different forms of inputs (shown in first row) and models are adopted for visual content synthesis tasks. We categorize visual content synthesis into four categories according to the different input forms (visual, text, audio, and other modalities). The second row illustrates the corresponding synthetic contents which are from [4–10]., respectively.

multimodal learning methods [32–39,221], progress has been made in extending generative methods to multimodal settings [3,222,223]. Mansimov et al. propose the first multimodal synthesis method [40], which leverages recurrent variational autoencoders to generate visual scenes guided by captions. With the rise of powerful GANs, the research of multimodal visual content synthesis is then greatly advanced by its variants [41,42]. On the one hand, this field has achieved remarkable improvements thanks to the enhancement of multimodal data processing capabilities [43–45]. Reed et al. [46] extend conditional GANs [47] to generate natural images based on text descriptions. Chen et al. [48] introduce conditional GANs to achieve cross-modal audio-visual generation of musical performances. Johnson et al. develope a novel GAN-based model [7] that takes as input a scene graph to generate a realistic image. On the other hand, novel architectures have been developed to adapt to high-fidelity multimodal visual content synthesis, such as stack architecture [49,50], cycle architecture [51,52], and attention mechanism [43,53]. In addition to exploration of the latent space [54] of GANs, many methods use a pretrained GAN to process images through GAN inversion [55]. With the development of large-scale natural language processing models such as Transformer [56], methods like DALL-E [57] can perform high-fidelity and controllable cross-modal generation. Obviously, the study of visual content synthesis has important implications for artificial intelligence.

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

- This survey covers contemporary literature with respect to multimodal-guided visual content synthesis, and provides a comprehensive overview of the recent efforts in terms of the modalities, datasets, evaluation metrics, and future research directions.
- In order to make the survey more organized, we formulate a taxonomy of multimodal-guided synthesis methods according to its guidance forms. In each subsection, we summarize the related models under the aspects of structure, idea, application, and limitation. Then, the commonly used datasets and evaluation metrics are provided, followed by the performance of existing methods.
- Finally, this survey summarizes the typical challenges with an outlook towards promising areas and directions for future research in this field.

The reminder of this survey is organized as follows: Section 2 discusses the taxonomy of visual content synthesis. In order to make a clear roadmap, we categorize the tasks according to different input forms. In Section 3–6, we summarize the typical task of visual content synthesis, and discuss the ideas and structures of each model. Section 7 summarizes benchmark datasets and common evaluation metrics. Finally, challenges and future directions are given in Section 8.

2. The Taxonomy of Visual Content Synthesis

Automatically synthesizing lively visual content conditioned on different modal information, is of great value in many real-world applications, such as social media, film production and entertainment advertising. It is interesting that people could ask the machines to synthesize the desired images or videos, depending on the information they get from seeing, reading, hearing or other ways. The data obtained from different senses has a different form, which is called multimodal data. Based on the powerful generation capabilities of neural networks, the synthesis of diverse visual content driven by multimodal guidance has evolved rapidly. In the following, we propose a taxonomy to summarize multimodal-guided visual content synthesis methods, and divide them into four categories, including *visual-guided*, *text-guided*, *audio-guided* and *other-modal-guided*, depending on the form of input data. Our proposed taxonomy of visual content synthesis is illustrated in Fig. 1.

2.1. Visual-Gudied Visual Content Synthesis

Visual information makes up most of the information people receive in their daily lives. Thanks to the development of unconditional synthesis methods [21,28,31], conditional visual content synthesis methods could train a powerful generative network that uses additional visual information as a condition. Compared to other modal data, visual modality could provide clearer cues of texture and structure [42,58–61]. We discuss the visual-gudied visual content synthesis in Section 3.

2.2. Text-Guided Visual Content Synthesis

Text descriptions are more flexible modal data for conditional visual content synthesis. The common task of text-guided visual

content synthesis is general text-to-image synthesis [46,63–66], which uses the text description as semantic guidance. Moreover, text-guided visual content manipulation [67,68] is another common task for visual content synthesis, which combines both visual and text modalities as conditional input. Since the text descriptions usually contain objects and their corresponding relationships, it is crucial to extract explicit guidance information from it. In addition, due to the heterogeneous features, different researchers have proposed several approaches for cross-modal fusion. We discuss text-guided visual content synthesis in Section 4.

2.3. Audio-Guided Visual Content Synthesis

Sound is a special modal for conditional visual content synthesis that helps people recognize the real world. Audio-to-visual cross-modal generation and manipulation have attracted considerable attention recently [69–71]. It involves separating semantic information from audio signals and creating a cross-modal generation network. Although it is easy for human to perceive the natural correlation between sounds and appearance, this task is still challenging for machines due to heterogeneity of modalities. We discuss audio-guided visual content synthesis in Section 5.

2.4. Other-Modal-Guided Visual Content Synthesis

Beyond regular modalities, there also exists other forms of data that provides more refined objective relationships and more direct way for interaction. Such modalities also can be treated as conditional guidance, including semantic segmentation [72,73], scene graph [7], facial mask [74], keypoints [8], line art drawing [75]. We give a brief review of such methods in Section 6.

3. Visual-Guided Visual Content Synthesis

For visual content synthesis, the common guidance modalities are natural images and videos. We group the visual-guided visual content synthesis as image-to-image (Section 3.1), image-to-video (Section 3.2), and video-to-video (Section 3.3) synthesis.

3.1. Image-to-Image Translation

Many problems in image processing, computer graphics, and computer vision can be posed as "translating" an input image into a corresponding output image [41]. Image-to-image translation aims to learn the mapping from a reference image to a target image, which can be regarded as domain transfer problem [2,76,77]. The key idea of this task goes back to Hertzmann et al. [78]. In order to better transfer cross-domain information, image-to-image translation needs to disentangle the domain-dependent features and domain-independent features. Due to that collecting a pair of images which belongs to different domains may be unreachable, it is challenging to learn the mapping tranformation betweent multiple domains. In the following we discuss the way of supervision, then we list three common tasks.

3.1.1. Paired & Unpaired Supervision

Ideally, image-to-image translation requires paired data containing the original images and its corresponding ground truth images (x, y) belonging to different domains $(\mathscr{X}, \mathscr{Y})$ respectively, which is called **paired supervision**. During training, the models learn a mapping network to translate x to y. Isola et al. propose Pix2Pix [41], which is the first supervised image-to-image translation approach using conditional GAN. Pix2Pix framework consists of a U-Net-based [80] generator and a markovian discriminator. The objective function of Pix2Pix uses conditional GAN (cGAN) loss

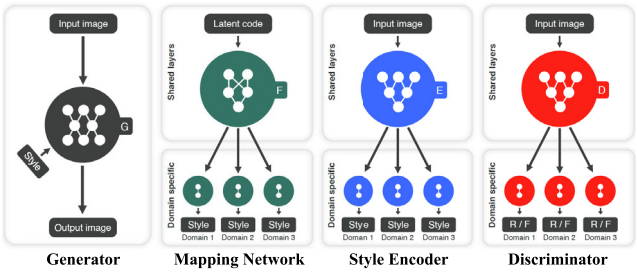

Fig. 2. The overview of StarGAN2 model. Through different networks, the model can achieve multiple domain transaltion. This figure comes from [62].

with L1 norm, which leads to ideal image tranlation. However, the Pix2Pix method [41] is still limited to being of low resolution. Besides, directly using L1 loss as constrain leads to blurry images [41,81]. Wang et al. propose Pix2pix-HD [82], which increases the resolution of the tranlation images to 2048 × 1024. Pix2pixHD utilizes a coarse-to-fine generator with three multiscale discriminators. The robust adversarial objective function helps to synthesize high-resolution images.

Usually paired data for training image-to-image translation is unreachable, so dominant method in this field gradually adopts **unpaired supervision**. CycleGAN [42] and DualGAN [83] are two methods in the field of unsupervised image translation research. These two methods simultaneously use the cycle consistency loss, which can learn the mapping of one domain to another without the aligned paired image data. Liu et al. [84] propose an UNsupervised Image-to-image Translation framework (UNIT), which uses VAE to project different domain images into the shared latent space. Aiming at the problem that the existing methods cannot generate multiple styles of images from a given source domain image, Hu et al. [85] propose Multimodal UNsupervised Image-to-image Translation (MUNIT). This method trains two autoencoders to encode the content and the style respectively, which can disentangle the image representation into domain-invariant and domain-specific parts. By combining the different domain codes, this method can realize high-quality and diverse image translation. At the same time, Lee et al. propose DRIT [86], which uses the content discriminator and cross-cycle consistency loss to achieve image representation disentanglement. Subsequently, on the basis of DRIT, DRIT++ [87] adds regularization items to alleviate the problem of pattern collapse in DRIT. To learn multi-domain image translation, Choi et al. propose StarGAN [88] which allows simultaneous training of multiple datasets within different domains of a single network by leveraging three loss functions. After that, StarGAN2 [62] is proposed, which is able to use a single generator and discriminator with multiple branches to map between multiple domains and produce a diverse set of images, as illustrated in Fig. 2.

3.1.2. Common Tasks

There are many meaningful tasks in the field of image-to-image translation. We give a review of three common tasks in the following, including image super-resolution, image dehazing, and image style transfer.

Image Super-Resolution (SR) aims to recover a high-resolution image from a single low-resolution image. SRCNN [89] is the earliest deep image SR method, which directly use a convolutional neural network to learn the mapping relationship between low-resolution interpolated images and high-resolution images. After that, image super-resolution developed rapidly [90–92]. With the development of GANs, SRGAN [93] first applies it to solve single image SR task. RCAN [94] introduces an attention mechanism into the network model for single image SR task to improve the expressive ability of the network. Since the training data distribution in

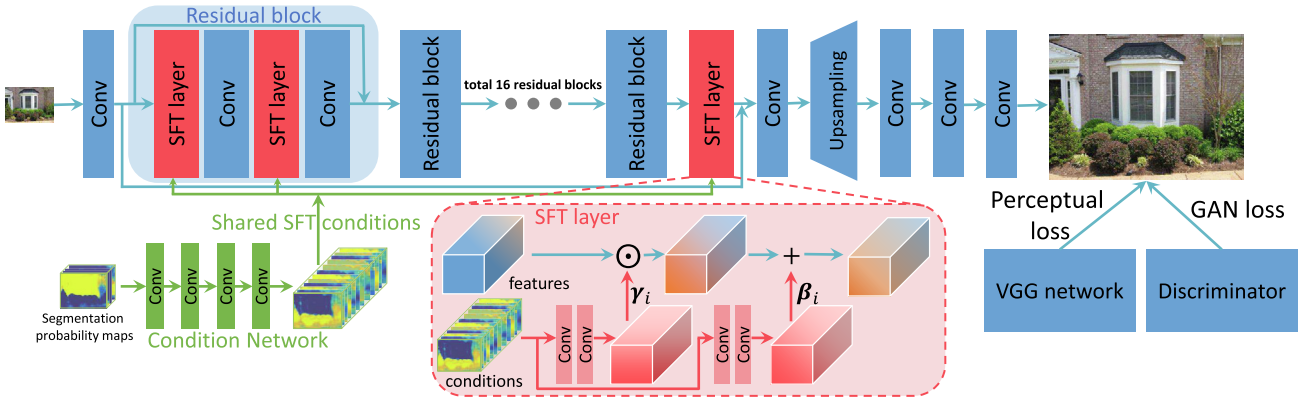

Fig. 3. The overview of SFTGAN model. The method proposed a novel Spatial Feature Transform (SFT) layer based on affine transforation to perform image super resolution. This figure comes from [79].

super-resolution tasks is generally uniform, the batch normalization layer will not improve the model effect but will complicate the mapping which the network needs to learn, resulting in a poor model effect. ESRGAN [95] removes the batch normalization layer in the residual module to improve the effect of the model. Wang et al. [79] propose a method to recover natural and realistic texture named SFTGAN, which leverages a novel Spatial Feature Transform (SFT) layer based on affine transforation, as shown in Fig. 3. Besides, AdderSR [96] leverages additive neural networks to deal with the single image SR task. Chan et al. [97] propose an image super-resolution method based on GAN inversion, which is named GLEAN. Reference-based SR (Ref-SR) is a technical route of image SR which is different from single-image super-resolution. Aiming at the problem that the existing Ref-SR method has less consideration of the scene structure and cannot complete the high-quality super-resolution in the case of a large resolution gap, Zhou et al. [98] propose Cross-MPI. It proposes plane-aware attention mechanism to make full use of the hidden scene structure for efficient attention-based correspondence search. In order to solve the transformation gap (representing scale and rotation transformation) and resolution gap (the mapping between high-resolution and low-resolution), Jiang et al. [99] propose the C2-Matching method, which uses contrastive learning and teacher-student knowledge distillation to enhance transformation mapping relationship.

Image Dehazing has received a great deal of research focus in image restoration field. Various end-to-end CNN-based methods have been proposed [100–104]. Zhang et al. [105] propose an end-to-end image dehazing method (DCPDN). It adopts the encoder-decoder structure of densely connected edge reservation based on the multi-level pyramid pooling module. Zhang et al. [106] propose a multi-scale image dehazing method based on a deep perceptual pyramid network, which uses CNN to learn the nonlinear relationships between the blurred image and the corresponding clear image. Due to inaccurate parameter estimation, the performance of dehazing will further reduce, resulting in color distortion. To solve this problem, Dong et al. [107] propose an end-to-end image dehazing method dubbed FD-GAN which is based on GANs with fusion-discriminator. By using frequency information as the additional priors, this model can generate more realistic dehazing images with less color distortion and fewer artifacts. Wu et al. propose AECR-Net [108] with autoencoder-like architecture and contrastive regularization to further enhance dehazing ability.

Image Style Transfer has become an active research topic in computer vision fields that aims to map a content image into the style of a different reference image [109,110,4,111]. Gatys et al. [112] propose the method which is the first to study how to use the convolution neural network to reproduce the well-known painting style on natural images. The method successfully produces a stylized image with a given artistic style and opens up a new field called image style transformation, which is the initial work of rendering different styles for content images using convolution neural networks. The current image style transfer methods can be categorized as being based on either image optimization or model optimization [113]. The pioneering work of the method based on iterative optimization of images is proposed [114], which changes the style of the input image through iterative optimization. The key idea of image optimization is to match feature statistics of intermediate layers in a CNN [115–118]. However, iterative optimization per image is comparatively slow. Model optimization-based approach trains feed-forward networks offline on datasets, which have the advantage that it can realize real-time image style transfer [81,119]. Johnson et al. [81] construct perceptual loss by using specific scale features extracted from the pre-trained network. Then, the perceptual loss is used as the objective function to train the feed-forward network for style transfer tasks. However, these methods were restricted to a fixed set of styles [120,119]. To solve this problem and realize multi-style transfer, Dumoulin et al. [121] propose Conditional Instance Normalization (CIN) based on Instance Normalization (IN). Subsequently, Huang et al. [122] propose Adaptive Instance Normalization (AdaIN) on CIN which makes that arbitrary image style transfer models can be realized.

3.2. Image-to-Video Synthesis

Generating video from image refers to changing static image into dynamic visual frequency, which can be used in time-lapse photography to make video animation from images [123–127]. In video generation, an important problem is how to obtain timing information. Most methods rely on the color appearance change information provided by the reference video, but it is very difficult to find the reference video with similar semantic information to the input image. Nam et al. propose Time-lapse [128] that can generate a continuous video with timing information from a single outdoor image to achieve the effect of time-lapse photography by learning the correlation between lighting changes and time of outdoor scene. Zhao et al. propose a method to generate a painting video according to the created painting [126].

3.3. Video-to-Video Synthesis

Different from image, video modality includes temporal information which is more complicated for machines to process. The original synthesis researches are aimed at the task of video

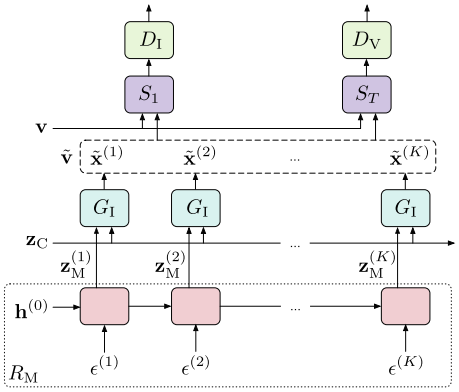

Fig. 4. The overview of MoCoGAN model. The generator produces a frame using the content and the motion vectors which are mapped from random variables via recurrent neural network. This figure comes from [58].

generation. Among them, the Video Generative Adversarial Network (VGAN) [130], the Time Generative Adversarial Network (TGAN) [131], and the Motion and Content decomposed Generative Adversarial Network (MoCoGAN) [58] are proposed for unconditional video synthesis. The overview of MoCoGAN model is shown in Fig. 4. Thanks to the development of such models, video-to-video synthesis [132,133] has attracted more attention. Video-to-video synthesis aims to convert the input semantic video into realistic video, but it is difficult to ensure the consistency of frames. Wang et al. propose a video-to-video synthesis model (Vid2Vid) [132], which takes the optical flow information as constraints to generate coherent and high-quality videos. Aiming at the problems of lack of data and limited generalization ability of model, Wang et al. [133] propose a few-shot video-to-video synthesis framework, which uses a small number of samples to synthesize videos of objects or scenes that have not been seen before.

3.3.1. Common Tasks

The common tasks of video-to-video synthesis mainly includes video SR, style transfer, and video prediction [134–136]. Earlier, Shechtman et al. [137] use the multiple spatio-temporal information at the same time to perform **video super-resolution**. However, because the super-resolution operation is performed in the high-resolution space, the calculation complexity is relatively large. Therefore, Shi et al. [91] reduce the video super-resolution process by extracting feature maps in the low-resolution space. In addition, Wang et al. [132] introduce spatio-temporal adversarial loss to realize that videos can be generated from inputs of different formats to achieve high-resolution video-to-video synthesis.

In the direction of **video style transfer**, Chan et al. propose a model [134] that transfers the dance pose from source video to target video. The model first uses a detector to create a pose estimation model for the input video, and then designs a system to learn the image mapping from the normalized pose to the target person. Nonetheless, it is still difficult to establish an accurate model to describe the complex nonlinear motions of human body. Yang et al. propose Trans-MoMo [136] model, which uses 2D keypoint information to train the network end-to-end, so as to better generate human action videos.

Video prediction is another task which predicts future frames based on existing frames. Video prediction trains a video prediction model to predict future frames based on existing frames. Convolutional Dynamic Neural Convection Network (CDNA) [138] and Video Ladder Network (VLN) [139] use the Long Short-Term Memory [225] network and stacked autoencoder respectively to model the motion through the transformation of pixels, which realize the prediction of future frames. Chiappa et al. [140] design an action condition generative model to accurately predict future frames based on the actions in the known frames. The above research only considers motion information, so the DRNET [141] learns and combines the potential representation of content and motion in the video. However, these methods do not explicitly model the inherent pixel motion trajectory, which can lead to blurred predictions. Therefore, Liang et al. [142] develop a dual motion generative adversarial network architecture, which uses a dual adversarial learning mechanism to make the synthesized pixel value in the future frame consistent with the pixel motion trajectory, thereby generating a clear prediction.

4. Text-Guided Visual Content Synthesis

Text is the most common guidance for cross-modal visual content synthesis in the real world. The goal of text guidance synthesis is to generate high-realistic visual content matching with the semantics of the given textual descriptions. Due to the flexibility and multiformity of text modal, it is essential to extract the correct guidance information influenced by ambiguous semantics. What's more, there still exists heterogeneous differences between modalities, which makes it hard for deep models to generate precise visual content. Although significant progress has been made, it is still challenging to maintain the semantic consistency between the text description and the generated visual content. Thus, the keys can be measured from two aspects, namely, text encoding and consistent generation. In the following sections, we first give a brief of common text encoding methods in Section 4.1. Then, we discuss text-guided visual content synthesis tasks in Section 4.2.

4.1. A Brief of Text Encoding Approaches

As mentioned above, a good semantic representation encoded from text descriptions is essential to guide visual synthesis. In the begining, traditional methods, such as Bag-of-Words [143] and Word2Vec [144], have been used to encode texts. With the development of natural language processing, more efficient methods are leveraged in text guidance synthesis. The milestone work, GAN-CLS [46] uses a hybrid character-level convolutional-recurrent neural network to encode texts. StackGAN [49] leverages Conditioning Augmentation (CA), which samples latent codes from a Gaussian distribution based on text embedding, to yield more training image-text pairs. AttnGAN [43] learns text encoding with a bidirectional LSTM [145] by concatenating its hidden states. MirrorGAN [51] uses recurrent neural networks to extract word feature and sentence feature. Other methods [146] leverage the large-scale pre-trained models, such as Transformer [56] and BERT [147], as powerful text encoder. As one of the cross-modal pre-trained models, CLIP [129] has attracted more and more attention. It jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples, as depicted in Fig. 5, achieving SOTA performance on both image and text representations. Recently, some methods [148,45,149,150] leverage the pre-trained text encoder integrated in CLIP to yield accurate text embeddings.

4.2. Text-Guided Visual Content Synthesis Methods

In the following, we discuss two tasks in the field of text-guided visual content synthesis. Each tasks we list different efficient models with various basic networks.

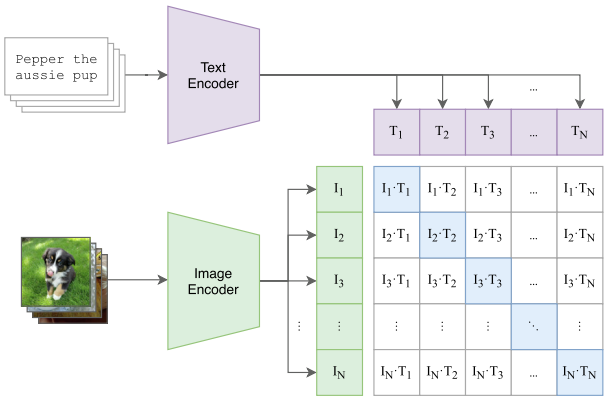

Fig. 5. The contrastive pre-training overview of CLIP model. The pre-trained text encoder of CLIP model can be leveraged to extract textual guidance. This figure comes from [129].

4.2.1. Text-to-Visual Generation

Given a desire text description *T*, this task aims to generate visual content *V* matching in the semantic space. Through the text encoding network, the textual semantic information is extracted to guide the deep generator network. In the following, we categorize and discuss the different methods according to their characteristics.

Pioneer. Reed et al. [46] propose GAN-CLS, which is the pioneer of text-to-image generation method. It concatenates text encoding with randomly sampled gaussian noise to synthesize final images, and leverages matching-aware discriminator to score three types of input. Mittal et al. proposed Sync-DRAW [226] that is the first method for text-to-video generation. It creates the sequence of time frames through VAE combined with attention mechanism, and effectively learns the spatio-temporal property from the video.

Stacked Architectures. Aimming to synthesize high-resolution images, stached architectures are adopted in text-to-image synthesis. Zhang et al. [49] propose StackGAN, which uses an iterative method to generate high-resolution images. The first generator ouputs a low-resolution image, which is further improved resolution by the second generator. Most of the subsequent text-to-image synthesis methods follow the two-stage model. StachGAN++ [50] leverages three generators and discriminators jointly to perform high-resolution synthesis. Instead of using multiple generators, HDGAN [152] utilizes hierarchically nested discriminators at multi-scale layers.

Attention Mechanisms. To fucus on the synthesis of detailed local region, attention mechanism was employed in text-to-image synthesis. Xu et al. [43] propose an Attention Generation Adversarial Network (AttnGAN). It learns the relationship between related words and images through the attention mechanism so that each area of the synthesized image has finer-grained details. Tan et al. [153] propose the Semantic Enhance Generative Adversarial Network (SEGAN) which constructs adaptive attention weights to distinguish between keywords and unimportant words to improve the stability and accuracy of the network. ControlGAN [53] proposes a word-level spatial attention which allows to correlate the words with the corresponding semantic region.

Cycle Consistency. Cycle consistency could enforce a strong connection between domains by constraining the models (*e.g.*, encoder and decoder) to be consistent with one another [42,83,154]. Qiao et al. [51] propose MirrorGAN, which uses a text-to-image generation framework with global–local attention and semantic preserving mechanism to deal with text-to-image generation problems. Lao et al. propose [155] to disentangle the content and style by augmenting current text-to-image synthesis frameworks with a dual adversarial inference mechanism. Liu et al. propose CMDL [228], which through the dual learning mechanism to learn the bidirectional mapping between sentences and videos simultaneously.

Contrastive Mechanisms. Zhang et al. [5] use contrastive learning to build a cross-modal generative adversarial network (XMC-GAN) to deal with text-to-image generation problems. XMC-GAN uses multiple contrastive loss to maximize the mutual information between image and text, captures the correspondences of inter-modality and intra-modality of text and images. The attentional self-modulation generator and contrast discriminator are used to force the learning of text-image correspondence to achieve the continuity, rationality, semantic relevance of text-to-image synthesis.

Language-Free. Inspired by the recent progress in large-scale cross-modality pre-training model CLIP [129], various methods attempt to leverage the latent space of CLIP to optimize the matching score between textual prompt and visual generation content. Along with this approach, a new generation paradigm, language-free text-to-image generation, is rising. Wang et al. propose CLIP-GEN [45], which uses the cross-modal semantic consistency in CLIP space to realize language-free generation. The framework is trained to leverage the extracted visual embedding from CLIP to synthesize images with language-free. During inference, the framework can generate images driven by the given description.

4.2.2. Text-Guided Visual Content Manipulation

Visual content manipulation is a challenging task in computer vision. Given a desire textual prompt *T* and visual content *V*, the manipulator learns to edit the visual content to meet the requirements. Take image manipulation as an example, it mainly manipulates images through color and geometric interaction and completes tasks such as image deformation and mixing. However, it is challenging for this task to manipulate relevant visual content, while keep the unrelevant parts unchanged. In the following, we discuss this task in two ways.

GAN-Based manipulation methods are dominant which leverages the cross-modal alignment. Dong et al. propose a Semantic Image Synthesis GAN (SISGAN) [156] for synthesize realistic images directly with language description. SISGAN concatenates text and image representations to synthesize final manipulated images while discriminator performing the distinguishing task conditioned on text semantic features. However, such methods could not disentangle the relevant and irrelevant regions, resulting in undesirable modification of text-irrelevant parts. To overcome this problem, Nam et al. propose the Text-Adaptive Generative Adversarial Network (TAGAN) [157], which forces the generator to disentangle differenet regions of image. The key idea is to split a single sentence-level discriminator into a number of word-level discriminators so that each word-level discriminator is attached to a specific type of visual attribute. Li et al. propose ManiGAN [67], as illustrated in Fig. 6. It is composed of text-image affine combination module (ACM) and correction module (DCM). The former enforces text and image features to collaborate to select text-relevant regions, and correlate those regions with corresponding semantic words for generating new visual attributes semantically aligned with the given text description. The latter rectifies mismatched attributes and complete missing contents.

GAN inversion manipulation methods have been proposed recently to bridge real and fake image domains, which can be used to perform text-guided image manipulation. We briefly introduce the preliminary of GAN inversion.

The unconditional GAN learns to map a latent vector $\mathbf{z}$ to image x. By contrast, GAN inversion is to map real image x back to latent representation $\mathbf{z}^*$ through a well-trained generator. Formally,

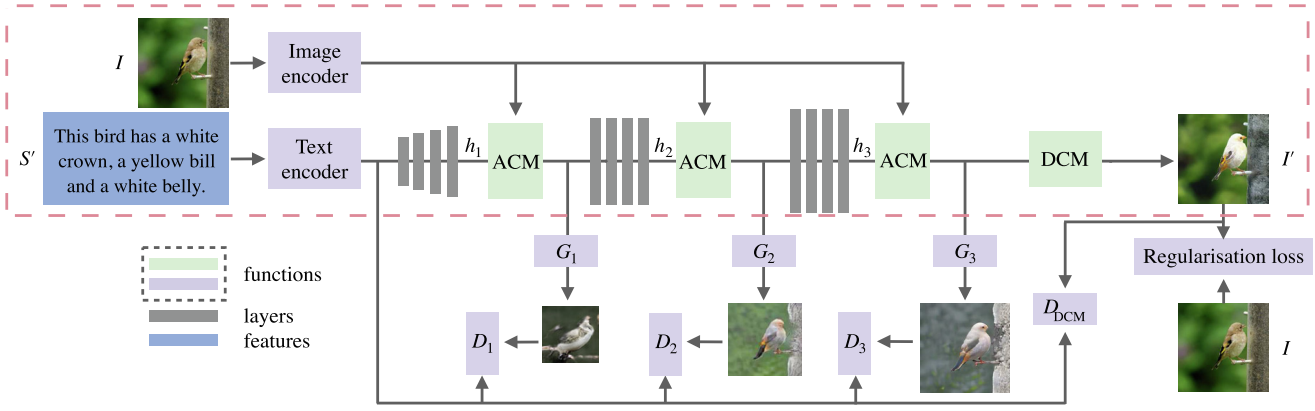

Fig. 6. The architecture of ManiGAN model. The network could effectively select image regions corresponding to the given text, and correlate those regions with text information for accurate manipulation. This figure comes from [67].

denoting the signal to be inverted as x, the well-trained generator as G, latent vector as $\mathbf{z}$, GAN inversion can be formulated as below:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \ell(G(\mathbf{z}), x), \tag{1}$$

where $\ell()$ is a distance metric in the image or feature space. Typically, $\ell()$ can be based on ℓ_1, ℓ_2, perceptual [81] or LPIPS metrics [158]. After obtaining the latent vector $\mathbf{z}^*$, we can edit the corresponding image throuth latent manipulation, the schematic diagram is shown in Fig. 7.

Through GAN inversion, text-guided manipulation can be achieved in a latent way. Xia et al. [160] propose TediGAN for multimodal image generation and manipulation with textual descriptions. It consists of three parts: a model inversion module based on StyleGAN [30,31], visual-linguistic similarity learning, and instance-level optimization. The inversion module maps the real image to the latent space of the pre-trained StyleGAN. Visual language similarity learns the text-image matching relationship by mapping images and text to a common space. The instance-level optimization is for identity preservation in manipulation. By using a control mechanism based on style mixing, TediGAN can support image synthesis with multi-modal input. Inspired by the recent progress in cross-modality language-vision pre-training of CLIP model [129], various optimization-based methods attempt to search in image space based on a query text by optimizing the text-image matching score of a pre-trained CLIP model. StyleCLIP [148] leverages the latent codes inverted by pre-trained StyleGAN model. By optimizing the semantic distance between text and image in CLIP space, StyleCLIP can achieve image manipulation through latent mapping, as shown in Fig. 8.

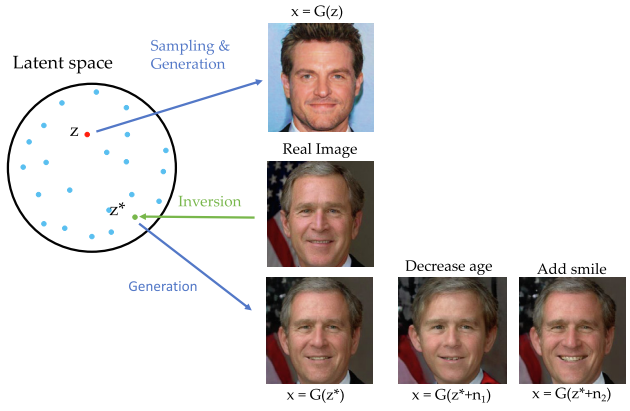

Fig. 7. Illustration of GAN inversion, which maps a real image x back into the latent space and obtains the latent vector $\mathbf{z}^*$. By varying the interpretable directions of $\mathbf{z}^*$ (e.g. $\mathbf{z}^* + \mathbf{n}$), the image manipulation can be achieved in a latent manner. This figure comes from [151].

5. Audio-Guided Visual Content Synthesis

Humans can imagine the scenes corresponding to sounds and vice versa. Researchers have tried to endow machines with this kind of imagination for many years. Due to the difference between audio and visual modalities, the potential correlation between them is nonetheless difficult for machines to discover. Thanks to the development of deep generative networks, many methods [70,161,69] succeed in synthesizing visual content guided by audio modal. In this section, we first provide a brief of audio encoding in Section 5.1. Then, we give a review of audio-to-visual generation tasks in Section 5.2.

5.1. A Brief of Audio Encoding Approaches

To perform perfect generation, it is essential for machines to obtain powerful audio embeddings from audio signals. Chen et al. [48] explore a set of representations including the Short-Time Fourier Transform (STFT), Constant-Q Transform (CQT), Mel-Frequency Cepstral Coefficients (MFCC), Mel-Spectrum (MS) and Log-amplitude of Mel-Spectrum (LMS). Besides, Wang et al. propose [162] a audio-to-image generation network given pairs of sound segments and images. To perform better cross-modality synthesis, the method evaluates four sound feature representation approaches, including Spectrogram, MFCC, Fbank, and pre-trained SoundNet [163]. After encoding, all the features in the sequence are averaged into a single vector which is taken as the condition of generator.

5.2. Audio-Guided Visual Content Synthesis

In the following, we give a brief review of several audio-driven visual synthesis tasks, incluing audio-to-visual generation and audio-guided visual manipulations.

5.2.1. Audio-to-Visual Generation.

Speech-to-image Generation. Chen et al. [48] first introduce the problem of cross-modal audio-visual generation. The method defines a Sound-to-Image (S2I) network and an Image-to-Sound (I2S) network. Each of them contains an encoder, a generator,

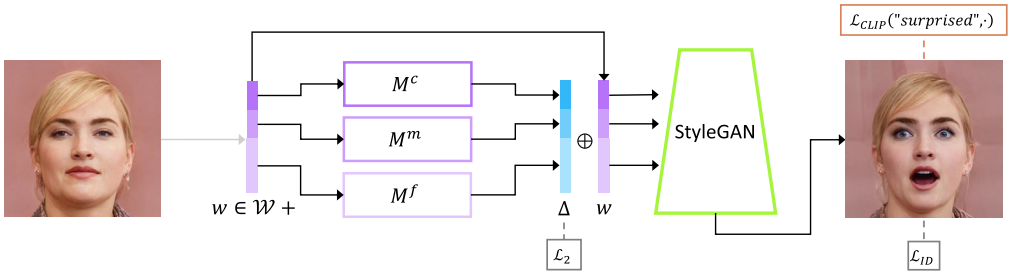

Fig. 8. The training strategy of StyleCLIP model. The image is inverted into latent codes first. Then mapping network manipulates the latent code to meet the requirements guided by CLIP model. This figure comes from [148].

and a discriminator. Most recently, some studies try to generate images conditioned on the speech description. Li et al. [159] propose a speech-to-image model that trains the speech encoder via teacher-student learning, so that the knowledge in a pre-trained image encoder is transferred to speech encoder. Then, the speech feature is leveraged to synthesize images through conditional GAN, as illustrated in Fig. 9.

Body Motion Generation. Body motion generation aims to animate avatars motions using audio signals as input. Alemi et al. propose a real-time GrooveNet [164]. It is trained on a set of recordings of dance movements performed with dance music. The model is based on conditional restricted Boltzmann machines and recurrent neural neworks to generate dance movements from music. Shlizerman et al. propose a method [165] that predicts body skeleton and uses the skeleton to animate an avatar given as input a music of violin. First it builds a LSTM network that learns the correlation between audio features and body skeleton landmarks. Then, it animates an avatar using predicted landmarks.

Talking Face Generation. Talking face generation aims to synthesize people's faces from speech or music, which abstracted great interest in cross-modality generation. Kumar et al. propose ObamaNet [166] that uses LSTM network with time-delay to predict the representation of the mouth shape given the audio features as input and further generates photo-realistic lip-sync videos. Eskimez et al. propose a method [167] that trains a neural network to process the waveform with 1D convolutional layers, and predict the active shape model (ASM) parameters of 3D face landmarks with a following fully connected (FC) layer. Chen et al. [161] design a hierarchical structure that leverages facial landmarks as intermediate representation and further generates talking faces based on the landmarks. Wang et al. [168] release the MEAD dataset and proposes a method that generates emotional talking faces by manipulating the upper and lower part of the face respectively. Song et al. [169] propose a method that factorize each target video frame into orthogonal parameter spaces, *i.e.*, expression, geometry, and pose, via monocular 3D face reconstruction to construct a photo-realistic video. Ji et al. [6] propose cross-reconstructed emotion disentanglement to decompose content and emotion of the audio, and achieves emotional video portraits generation.

5.2.2. Audio-Guided Visual Manipulation.

Sound provides polyphonic information of the scene and contains multiple sound events [170]. Audio-guided visual manipulation aims to use polyphonic information as an imagery source for visual content editing. Some methods mainly focus on music-guided cross-modal generation with no sound semantics. Lee et al. propose a music-to-visual style transfer method [171]. The transfer system contains two major networks, including the Music Visualization Net (MVNet) and the Style Transfer Net (STNet). The former translates an input audio to an image which resembles the style of that image paired with the audio. Then, the style image generated by the MVNet and the target image are fed into the STNet to synthesize the modified image which resembles the style of the style image. Jeong et al. propose TräumerAI [172] that generates a visually appealing video that responds to the input music. The author manually labeled the music and image paires in a subjective manner. The music covers various genres including classi-

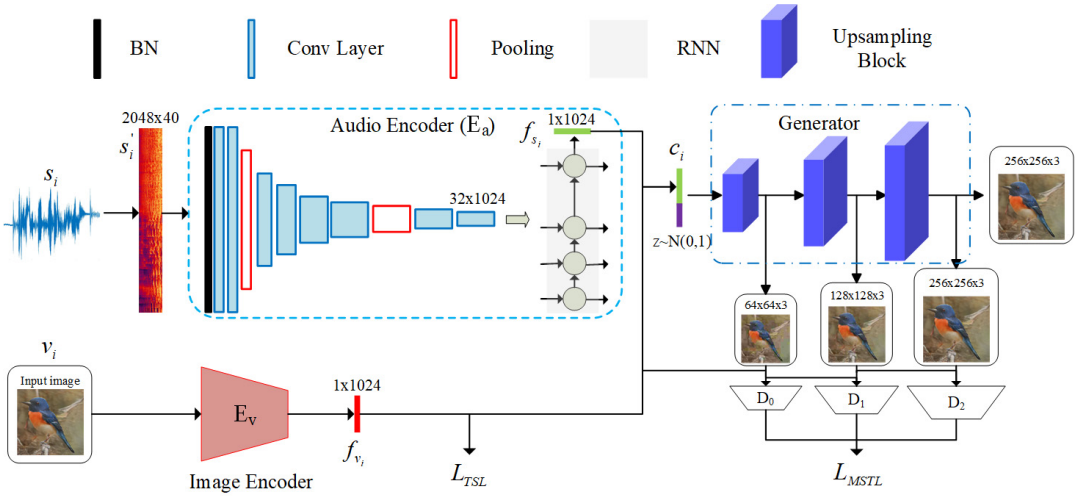

Fig. 9. The speech-to-image generative network translates the audio signals to photo-realistic images, which is trained via teacher-student learning. This figure comes from [159].

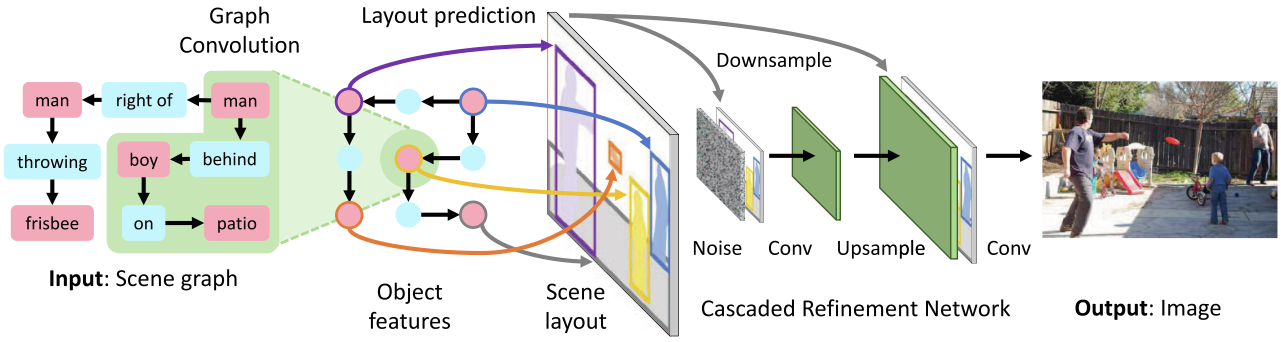

Fig. 10. The input scene graph is used to predict a scene layout. After that the layout is converted to an image using a cascaded refinement network (CRN). This figure comes from [7].

cal, jazz, pop, and so on. Based on the collected data, TräumerAI trained a simple transfer function that converts an audio embedding to a style embedding which is generated as a sequence of images by StyleGAN. The method sampled the 30 audio embeddings per second so that each frame of video is generated from the corresponding audio embedding. Besides, Lee et al. [173] leverage GAN inversion to maninpulate images guided by sound.

6. Other-Modality-Guided Visual Content Synthesis

Inspired by the rapidly growing of deep generative networks, many approaches try to synthesize realistic visual content with the help of its expressive generating ability. Thus, other modalities, such as semantic segment, scene graph, facial mask, sketch and so on, are leveraged as special input to synthesize more diverse visual content.

Some methods try to leverage the **semantic segmentation** or **layout** as more meticulous conditions. Karacan et al. propose ALCGAN [72], which takes semantic layout and scene attributes integrated as conditioning variables. The model is based on a cGAN architecture which learns the layout and the content of the scene using ground truth semantic layouts and transient attributes. Qi et al. propose a semi-parametric method to photographic image synthesis from semantic layouts, dubbed as SIMS [73]. It combines the complementary strengs of parametric and nonparametric techniques. Park et al. propose SPADE [174] with spatially-adaptive normalization to convert semantic segmentation mask to a photorealistic image. The proposed normalization utilizes input semantic layout while performing the affine transformation in the normalization layers. Zhu et al. propose Semantic Region-Adaptive Normalization (SEAN) [9], which extended SPADE [174]. The method proposed semantic region-adaptive normalization for GANs conditioned on segmentation masks that describe the semantic regions in the desired output image. The SEAN normalization can extract style from a given reference image, and processes the style information to bring it in the form of spatially-varying normalization parameters. To perform fast and efficient high-resolution synthesis, Shaham et al. propose ASAP-Net [175].

Besides, some methods utilize **scene graph** as condition to synthesize visual content. Johnson et al. develop a model [7] which takes as input a scene graph to generate a realistic image, as shown in Fig. 10. The scene graph is processed with a graph convolution network to compute embedding vectors for all objects. These vectors are used to predict bounding boxes and segmentation masks for objects, which are combined to form a scene layout. Finally, the layout is converted to an image using a cascaded refinement network [176].

What's more, many methods leverages other modalities as input, such as **sketch**, **facial mask**. Jo et al. [177] propose a GAN-based face editing method, which named SC-FEGAN. By using the end-to-end trainable convolutional network and free-form user input with colors and shapes as a guide, the image is generated by the guidance of masks, sketches, and colors. Dong et al. [178] propose FE-GAN, which first uses the parsing network with multi-scale attention normalization to generate human parsing from sketches and color. Subsequently, the generated human parsing is used as the input of the image inpainting network, and a fashion image with detailed texture is generated under the semantic guidance from the human parsing. Gu et al. [179] propose a portrait editing method based on mask-guided cGAN, which is guided by the facial mask and can generate various high-quality images. By learning the feature embedding of each face component to control the synthesis and editing of face images, it is helpful to improve the performance of image translation and partial editing of face images. Li et al. propose a novel framework [10] that explores and leverages semantic information to generate realistic textures in sketch-to-image synthesis. Xu et al. [180] propose FaceShapeGene, which learns the disentangled shape representation of the face image to achieve editing. FaceShapeGene realizes the task of face editing by encoding the shape information of each semantic part of the face into one-dimensional latent vectors while preserving the identity of the input face image at the same time.

7. The Experimental Evaluation

7.1. Datasets

It is obvious that high-quality and sufficient data is essential for visual content synthesis. In this section, we list benchmark datasets for visual content synthesis under multimodal settings.

7.1.1. Visual-Guided Datasets

MNIST. The MNIST dataset [181] is a widely used dataset in deep learning methods. It contains 70,000 labeled handwritten digital images of which 6,000 are used as the training set and 1000 are used for testing.

SVHN. The Street View House Numbers (SVHN) digit database [182] is a color house number dataset, which is similar to the MNIST dataset. The dataset contains a total of 99,289 images. Among them, 73,257 images are used as the training set and 26,032 are used for testing.

CIFAR. The CIFAR-10 dataset [183] is composed of color natural scene graphs with a pixel size of 32×32. The 60000 images can be divided into 10 categories. CIFAR-100 is composed of 100 types of images, each category contains 600 images. Among them, 500 images in each category are used for training, and 100 images are used as test set.

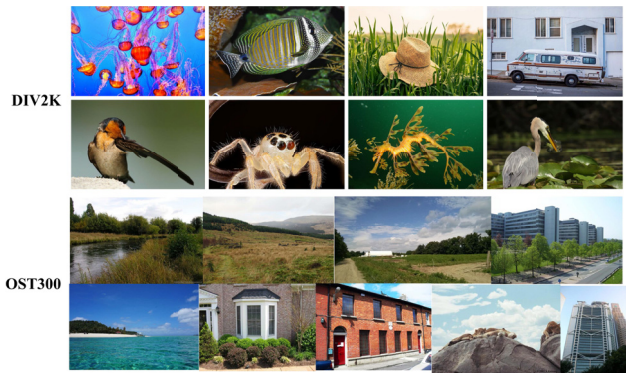

Fig. 11. Examples images of mentioned datasets for image super-resolution, include DVI2K[184] dataset and OST300[79] dataset.

BSD. The BSD300 [188] have collected 12,000 hand-labeled segmentations of 1,000 corel dataset images from 30 human subjects. Half of the segmentations were obtained from presenting the subject with a color image; the other half from presenting a grayscale image. The public benchmark based on this data consists of all of the grayscale and color segmentations for 300 images. The images are divided into a training set (BSD200) of 200 images, and a test set (BSD100) of 100 images. The BSDS500 is an extended version of the BSDS300 that includes 200 fresh test images. The BSD100, together with Set5 [189], Set14 [190], Urban100 [191], and Manga109 [192], is often used as a test set for single-image super-resolution task.

Transient Attributes. The transient attributes dataset [193] contains 8,571 images taken from 101 webcams. This database uses a taxonomy of 40 attributes labels related to weather, lighting, time of the day, season and more subjective impressions.

DIV2K. The DIV2K [184] dataset is divided into training set, validation set, and testing set. The training set consists low resolution images with 2, 3 and 4 downscaling factors obtaining from 800 corresponding high resolution images. Validation set and the testing set each contain 100 images for testing. Examples are shown in Fig. 11.

OST300. This outdoor scene dataset [79] is divided into OutdoorSceneTrain and OutdoorSceneTest for training and testing respectively. For OutdoorSceneTrain, each image is cropped so that only one category exists, resulting in 1 k to 2 k images for each category. Background images are randomly sampled from ImageNet. The total number of training images is 10,324. The OutdoorSceneTest partition consists of 300 images and they are not pre-processedIn particular. Examples are shown in Fig. 11.

LSUN. The Large-scale Scene Understanding (LSUN) [194] challenge aims to provide a benchmark for large-scale scene classification and understanding. The LSUN classification dataset contains 10 scene categories, such as dining room, bedroom, chicken, outdoor church, and so on. For training data, each category contains a huge number of images, ranging from around 120,000 to 3,000,000. The validation data includes 300 images, and the test data has 1000 images for each category.

FFHQ. Flickr-Faces-HQ (FFHQ) [30] consists of 70,000 high-quality facial images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity, background and other attributes.

Vimeo90K. The Vimeo90K [195] is the most widely used data set in the field of video super-resolution. It is used for video super-resolution, video denoise, video artifact removal and video interpolation. But the resolution of this data set is relatively small.

REDS. The realistic and dynamic scenes (REDS) [196] was proposed in the NTIRE19 Challenge. The dataset is composed of 300 video sequences with resolution of 720×1280, and each video has 100 frames, where the training set, the validation set and the testing set have 240, 30, and 30 videos.

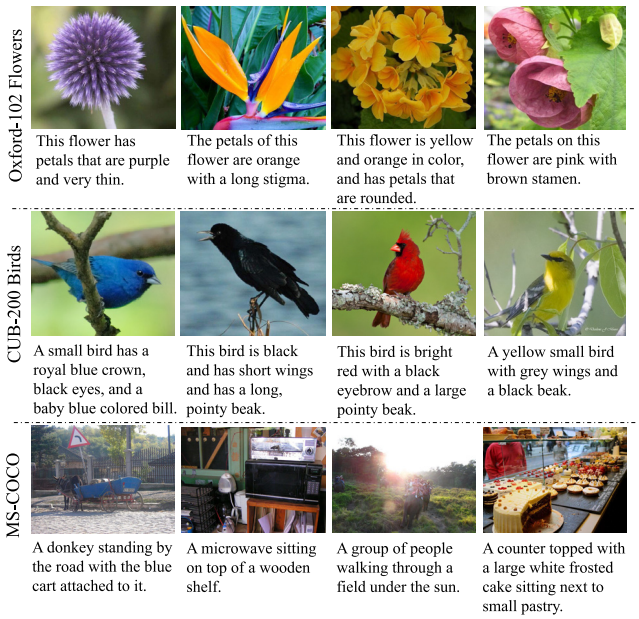

Fig. 12. Examples images and corresponding captions of mentioned datasets for text-to-image generation, including CUB-200 [185], Oxford-102 [186], and MS-COCO [187].

CelebA-HQ. CelebA-HQ [28] dataset, which consists of 30,000 high quality facial images picked from the original CelebA [197] dataset. The size of each high quality image is 1024×1024. In the original dataset, each image has 40 attributes annotations inherited from the original CelebA.

7.1.2. Text-Guided Datasets

CUB. CUB [198] is a widely used text-to-image generation dataset. CUB-200–2011 [185] is an extended version of CUB-200 which has a total of 11,788 bird images in 200 categories. Each image only contains a single object associated with 10 captions, as shown in Fig. 12.

Oxford-102. The Oxford-102 dataset [186] is a flower dataset proposed by Oxford University in 2008, which is mainly used for image classification. The dataset contains 8,189 flower images in 102 categories. Each image contains a single object associated with 10 captions, as shown in Fig. 12.

MS-COCO. Microsoft Common Objects in Context dataset (MS-COCO) [187] is built by Microsoft in 2014, which contains 91 object categories. It contains captions that can be used for text-to-image generation, as shown in Fig. 12. The 2014 split (COCO-14) is used for evaluation in most methods. LN-COCO [199], which contains localized narratives for images in the 2017 split of MS-COCO (COCO-17), is a more challenging than MS-COCO for text-to-image synthesis.

Visual Genome. Visula genome [200] dataset contains 108 K images densely annotated with scene graphs containing objects, attributes and relationships, as well as 1.7 million visual question answers. It contains 5.4 million region descriptions, which can be used for text-guided image generation.

ImageNet. ImageNet database contains more than 14 million images, and a little more than 21 thousand classes. To evaluate conditonal generation tasks, Wang et al. [45] construct the input descriptions.

7.1.3. Audio-Guided Datasets

GRID. There are 34 native English speakers in this dataset [201], with 16 female and 18 male speakers, who are ranging from 18 to 49 years old. Each speaker has 1000 recordings that are 3 s in duration. The recordings contain sentences that are identical for each speaker. The videos in GRID dataset have a frame rate of 25 FPS and a resolution of 720×576 pixels. Since each recording is 3 s in duration, each video has a total of 75 frames. The video files contain the corresponding audio that has a sampling rate of 44.1 kHz.

VoxCeleb2. This dataset [202] contains over 1 million utterances for over 6,000 celebrities, extracted from videos uploaded to YouTube. Videos included in the dataset are shot in a large number of challenging visual and auditory environments.

MEAD. MEAD [168] is a large-scale, highquality emotional audio-visual dataset that contains 60 actors and actresses talking with eight different emotions at three different intensity levels. This large-scale emotional dataset can be applied to many fields, such as conditional generation, cross-modal understanding, and expression recognition.

7.1.4. Other Datasets for Visual Content Synthesis

Facial Mask. CelebAMask-HQ dataset [74] is a large-scale face image dataset that has 30,000 high-resolution face images selected from the CelebA [197] dataset with their corresponding segmentation mask of facial attributes. The masks of CelebAMask-HQ are manually-annotated with the size of 512×512 and 19 classes including all facial components and accessories.

Scene Graph. Visual Genome [200] version 1.4 (VG) contains annotated scene graph, which can be used for visual content synthesis.

Semantic Segmentation. ADE20K [203] dataset is annotated with a 150-class semantic segmentation. Besides, COCO-Stuff [204] and Cityscapes [205] also serve as the benchmark datasets for semantic image synthesis.

Keypoints. DeepFashion [206] dataset, Faces dataset [207], and Market-1501 [208] dataset can be used for human keypoint guided image generation.

7.2. Evaluation Metrics

In the field of visual content synthesis, a variety of evaluation metrics have been adopted for various tasks. The following is the commonly used evaluation metrics for visual content synthesis.

RMSE. The Root-Mean-Square Error (RMSE) is a frequently used metric that measures the differences between samples predicted by the model. The RMSE represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. RMSE is defined as follows:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)}, \tag{2}$$

where $\hat{\theta}$ is estimator respect to parameter θ.

SSIM. The Structural Similarity Index Measure (SSIM) [209] is a method for predicting the perceived quality of images and videos. SSIM is used for measuring the similarity between two images, consisting of luminance, contrast and structure. A more advanced form of SSIM, called Multiscale SSIM (MS-SSIM) [210] is conducted over multiple scales through a process of multiple stages of subsampling, reminiscent of multiscale processing in the early vision system. Suppose x and y are two nonnegative image signals, SSIM is defined as follows:

$$\begin{aligned} &\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}, \\ &l(x,y) = \tfrac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \\ &c(x,y) = \tfrac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \\ &s(x,y) = \tfrac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \end{aligned} \tag{3}$$

where $l(x,y), c(x,y)$, and $s(x,y)$ denote comparisons of luminance, contrast and structure, respectively. In order to simplify the expression, the weights are generally setting as $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$, which results in a simplified form of the SSIM:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where μ_x and μ_y are the averages of x and y, σ_x^2 and σ_y^2 are the variances of x and y.

PSNR. Peak Signal-to-Noise Ratio (PSNR) denotes the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. PSNR is commonly used to quantify reconstruction quality for images and video subject to lossy compression. PSNR is defined via the mean squared error (MSE) as follows:

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}_i^2}{\mathrm{MSE}}\right), \tag{5}$$

where MAX_i is the maximum possible pixel value of the image.

LPIPS. Learning Perceptual Image Patch Similarity (LPIPS) [211] measures the differences between two images. The lower the value of LPIPS is, the more similar the two images are.

Inception Score. The Inception Score (IS) [212] is a popular metric for image generation tasks. The IS uses the Inception v3 [229] network pre-trained on ImageNet, and calculates the statistics of the network output when applied to the generated images, as follows:

$$\mathrm{IS} = \exp\left(\mathbb{E}_{\mathbf{x}\in p_g} D_{KL}(p(y|\mathbf{x}) \| p(y))\right), \tag{6}$$

where $\mathbf{x} \in p_g$ denotes an image $\mathbf{x}$ sampled from p_g, while D_{KL} represents the KL-divergence, $p(\mathbf{x}|y)$ is the conditional class distribution, and $p(y)$ indicates the marginal class distribution. An approximation to the expected KL-divergence can be calculated as follows:

$$\mathrm{IS} \approx \exp\frac{1}{N}\sum_{i=1}^{N} D_{KL}\left(p(y|\mathbf{x}^i) \| \hat{p}(y)\right). \tag{7}$$

Fréchet Inception Distance. The Fréchet Inception Distance (FID) [213] evaluates the quality of an image generator by measuring the difference between two distributions. The distributions of these two types of images are regarded as multivariate Gaussian distributions with different parameters. The lower the FID value of a model, the better the performance. The FID between real images and generated images is calculated as follows:

$$\mathrm{FID} = \left\|\mu_1 - \mu_2\right\|^2 + Tr(C_1 + C_2 - 2(C_1C_2)^{1/2}), \tag{8}$$

where μ_1, μ_2 denote the mean vectors of the features of the real and generated images, respectively. Correspondingly, C_1 and C_2 are the covariance matrices, while $\|\cdot\|$ represents the norm operator on vectors and $Tr(\cdot)$ indicates the trace operator on a matrix. Compared with the Inception score, FID is a more reasonable evaluation metric: specifically, this is because FID compares the generated

image directly with the real images, and not with images of ImageNet in the Inception score.

R-precision. R-precision [43] is an evaluation metric for the text-to-image synthesis task. First the cosine similarities between the global image features and the global text features are computed. Then, rank candidate text descriptions for each image in descending similarity. If there are R documents for a query with r relevant descriptions, for the top R ranked retrieval results the R-precision is r/R.

Landmark Distance & **Landmark Velocity Difference.** Landmark Distance (LD) [161] and Landmark Velocity Difference [217] are utilized to evaluate facial motions. LD represents the average Euclidean distance between generated and recorded landmarks and LVD represents the average velocity differences of landmark motions between two sequences.

7.3. Experimental Results

We list experimental results of multimodal-guided visual content synthesis collected from the related literature.

7.3.1. Visual-Guided

For visual-guided visual content synthesis, we conduct comparison on image-to-image translation and image SR. For image-to-image translation, the experimental comparison is conducted on transient attributes [193] and CelebA [197] datasets, which is shown in Table 1. Evaluation is performed with F1 Score and Fréchet Inception Distance (FID) [213]. Meanwhile, qualitative experiment conducted on CelebA-HQ dataset for image-to-image translation [41,42,85,62] is shown in Fig. 13. Methods of comparison include MUNIT [85], DRIT [86], MSGAN [218], and StarGAN2 [62]. By contrast, StarGAN2 model outperforms other methods in both experimental results.

For image super-resolution, the experimental comparison is conducted on Set5 [189] dataset, as illustrated in the top half of Table 2. The rest of the Table 2 is experimental results on other datasets: GLEAN is evaluated on Face [28] dataset, Cross-MPI is evaluated on RealEstate10K [219] dataset, C2-Matching is evaluated on Manga109 [192].

7.3.2. Text-Guided

For text-to-image generation task, we conduct comparison on several datasets, including MS-COCO [187], Oxford-102 [186], CUB [198]. The quantitative results are shown in Table 3. The evaluation metrics include IS [212], Human Rank (HR), FID [213], and R-precision [43]. The selected qualitative results are shown in Fig. 14. Compared to other models [214–216], the images generated by XMC-GAN [5] are much higher fidelity.

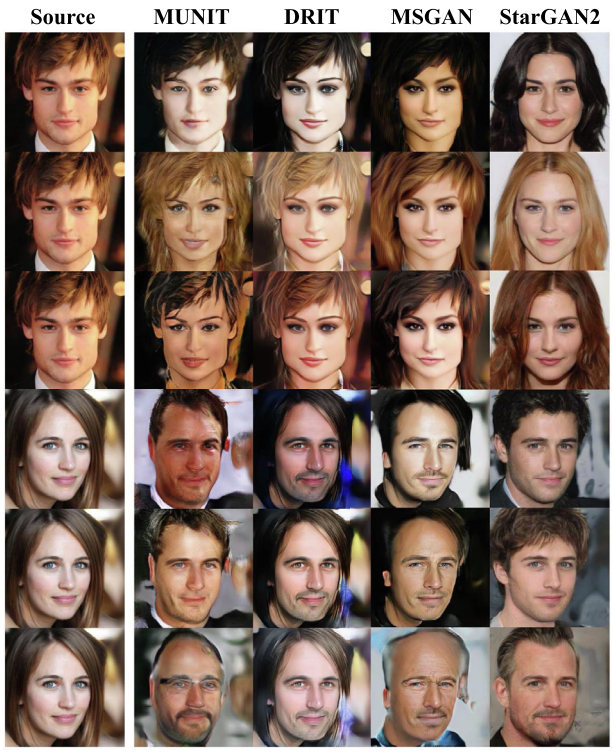

Fig. 13. Qualitative comparison of latent-guided image synthesis results on the CelebA-HQ [28]. Each method translates the source images (left-most column) to target domains using randomly sampled latent codes. The top three rows correspond to the results of converting male to female and vice versa in the bo. ttom three rows.

Table 1
Results of image-to-image translation methods on two datasets. The evaluation metrics include F1 Score and Fréchet Inception Distance (FID).

Dataset	Application	Models	F1 Score ↑	FID ↓
Transient Attributes [193]	Night to Day	Pix2Pix	0.025831	202.146
		CycleGAN	0.107669	199.840
		MUNIT	-	268.827
		StarGAN2	0.212981	197.913
	Day to Night	Pix2Pix	0.012000	228.808
		CycleGAN	0.010880	177.672
		MUNIT	0.003818	240.590
		StarGAN2	0.089745	180.301
CelebA [197]	Male to Female	CycleGAN	0.721799	47.529
		MUNIT	0.939948	40.630
		StarGAN2	0.787322	36.249
	Female to Male	CycleGAN	0.717887	48.420
		MUNIT	0.814733	19.486
		StarGAN2	0.871022	30.120

Table 2
Above the horizontal line is the summary of performance of image super-resolution models on Set5 [189] dataset. The rest is the performance on other datasets. The evaluation metrics include peak signal-to-noise ratio(PSNR), structural similarity (SSIM).

Model	PSNR ↑	SSIM ↑	Max Scale
SRCNN[89]	30.49	0.8628	×4
VDSR[90]	31.35	0.8838	×4
ESPCN[91]	30.90	-	×4
EDSR[92]	32.60	0.8982	×4
SRGAN[93]	29.40	0.8472	×4
RCAN[94]	27.47	0.7913	×4
SFTGAN[79]	29.82	0.8400	×4
AdderSR[96]	32.13	0.8864	×4
GLEAN $_{\text{Face}}$[97]	26.84	-	×16
Cross-MPI$_{\text{RealEstate10K}}$[98]	32.878	0.937	×8
C2-Matching$_{\text{Manga109}}$[99]	30.47	0.911	×8

7.3.3. Audio-Guided

We conduct the comparison in the task of audio-guided talking face generation on MEAD dataset [168]. The metrics of Landmark Distance (LD) and Landmark Velocity Difference (LV-D) [161,217] are utilized to evaluate facial motions. The quantitative results are shown in Table 4. By contrast, model [6] outperforms others [161,168,169] in audio-visual synchronization (M-LD, M-LVD), facial expressions (F-LD, F-LVD), and video quality (SSIM, PSNR, FID).

Table 3
A summary of performance of text-to-image synthesis models mentioned above with regard to evaluation metrics. The evaluation metrics include Inception Score (IS), Human Rank (HR), Fréchet Inception Distance (FID), and R-precision (RP) [43].

Model	Dataset	Metrics	Performance
GAN-CLS[49]	CUB	IS	2.88 ± .04
		HR	2.81 ± .03
	Oxford	IS	2.66 ± .03
		HR	1.87 ± .03
	COCO-14	IS	7.88 ± .07
		HR	1.89 ± .04
StackGAN[49]	CUB	IS	3.70 ± .04
		HR	1.37 ± .02
	Oxford	IS	3.20 ± .01
		HR	1.13 ± .03
	COCO-14	IS	8.45 ± .03
		HR	1.11 ± .03
AttnGAN[43]	CUB	IS	4.36 ± .03
		RP	67.82 ± 4.43
	COCO-14	IS	25.89 ± .47
		RP	85.47 ± 3.69
StackGAN++[50]	CUB	IS	4.04 ± .05
		FID	15.30
		HR	1.19 ± .02
	Oxford	IS	3.29 ± .01
		FID	48.68
		HR	1.30 ± .03
	COCO-14	IS	8.30 ± .1
		FID	81.59
		HR	1.55 ± .05
MirrorGAN[51]	CUB	IS	4.56 ± .05
		RP	60.42
	COCO-14	IS	26.47 ± .41
		RP	80.21
SEGAN[153]	CUB	IS	4.67 ± .04
		FID	18.167
	COCO-14	IS	27.86 ± .31
		FID	32.276
OPGAN[214]	COCO-14	IS	27.88 ± .12
		FID	24.70
SDGAN[215]	CUB	IS	4.67 ± .09
	COCO-14	IS	35.69 ± .50
CPGAN[216]	COCO-14	IS	52.73 ± .61
XMC-GAN[5]	COCO-14	IS	30.45
		FID	9.33
		RP	71.00
	LN-COCO	IS	28.37
		FID	14.12
		RP	66.92
	LN-OpenImages	IS	24.90
		FID	26.91
		RP	57.55
CLIP-GEN[45]	COCO-14	IS	21.4
		FID	20.7
	ImageNet	IS	45.16
		FID	16.74

8. Challenges & Future Directions

Multimodal-guided visual content synthesis has achieved impressive success in the field of deep learning, which comes with higher requirements and challenges. In this section, we overview the typical challenges in this field. Then, we highlight the future directions through a comprehensive view.

8.1. Challenges

It is gratifying to see that visual content synthesis methods have made great strides at present. However, since visual synthesis is still in the development stage, there exist several challenges for their practical applicability. To this end, we delve deep into this field, and discuss the challenges for future development including the following aspects.

Fig. 14. Qualitative experiment on MS-COCO dataset [187] for text-to-i.mage generation.

Interpretability & **Controllability.** In order to enable the specific visual content generated by the deep networks to meet people's requirements, we need to analyze the interpretability of the model and improve its controllability. However, due to the high complexity, most generative models lack the interpretability of the generation process. Although some methods try to increase the interpretability of the model to some extent, they cannot be used effectively to guide the synthesis process because they do not perform a quantitative analysis of each dimension of the hidden space.

Training Stability. The vast majority of current visual synthesis methods are based on GANs, which achieve outstanding success in this field. For further development, the open challenges of GANs are still worthy of attention, such as "mode collapse" and unstable training problems. Therefore, how to stabilize the training process and improve the performance of GANs are still challenging in the future research,.

Novel Generative Models. In order to process higher dimensions and more complex multimodal input, novel efficient generative models are needed. With the emergence of Transformer model [56] that supports for multimodal data processing, new generative paradigm has been established for visual content synthesis. However, due to the quadratical complexity, it is hard for Transformer model to perform real-time inference. Therefore, the design of novel generative models remains a grand challenge in this field.

Evaluation Metrics. In addition, the evaluation metrics are another important aspect that needs further improvement. Currently, there is no consensus on which evaluation metric best assesses the strengths and limitations of models and can be used for fair comparisons. Leveraging pre-trained models (*e.g.*, FID) to conduct evaluations does not adapt well to the new datasets. Besides, user study is the most direct evaluations for visual content synthesis, which is however too subjective.

8.2. Future Directions

The development of multimodal-guided visual content synthesis technology is a concentrated manifestation of the development of computer vision and multimodal information processing, and

Table 4
Quantitative comparisons with the state-of-the-art methods on MEAD dataset [168], including the results of landmark accuracies and video qualities. M- represents mouth and F- stands for face region.

Model	M-LD ↓	M-LVD ↓	F-LD ↓	F-LVD ↓	SSIM ↑	PSNR ↑	FID ↓
Chen et al. [161]	3.27	2.09	3.82	1.71	0.60	28.55	67.60
Wang et al. [168]	2.52	2.28	3.16	2.01	0.68	28.61	22.52
Song et al. [169]	2.54	1.99	3.49	1.76	0.64	29.11	36.33
Ji et al. [6]	2.45	1.78	3.01	1.56	0.71	29.53	7.99

aims to generate visual content as realistically as possible. Continuous advances in hardware and software have expanded access to information sources and utilization of information. In this case, more multimodal data can be obtained for visual content synthesis. Therefore, novel algorithms are expected to be able to handle guidance from multiple modalities concurrently. Moreover, the interpretability and controllability that exist in visual content synthesis are still worthy research directions. Although some methods, such as GAN inversion methods, have been explored in this regard, but are still in their infancy. Further exploration the interpretability and controllability of models is needed in the future to generate higher quality visual content. Currently the vast majority of visual synthesis methods are based on GANs which suffer from unstable training problems. For multimodal-guided visual content synthesis, it is valueable to find the way to improve the stability and efficiency of model training. In addition, the use of different evaluation metrics can lead to conflicting conclusions about the quality of image synthesis. The evaluation metrics of multimodal visual content synthesis are another directions of future development. By making improvements in the above directions, the quality of visual content synthesis can be further improved in the future, which will facilitate further applications of this research in other fields such as education, business, and human–computer interaction.

9. Conclusion

Understanding the world better from a human perspective is a perennial topic in the deep learning community. In this survey, we provide an overview of visual content synthesis from four perspectives depending on input form, including *visual-guided*, *text-guided*, *audio-guided*, and *other-modal-guided*. Furthermore, we detail the advantages and motivation of these methods for various tasks, such as image-to-image generation, text-to-image generation, and audio-to-image generation. After introducing the methods, we give a comprehensive overview of common datasets and evaluation metrics for visual content synthesis considering different control modalities. Then, we compare the performance of existing methods on several tasks. Last but not least, we give a comprehensive overview of the challenges and future directions in visual content synthesis.

CRediT authorship contribution statement

Ziqi Zhang: Formal analysis, Writing - review & editing. **Zeyu Li:** Methodology, Software, Writing - original draft. **Kun Wei:** Writing - review & editing. **Siduo Pan:** Investigation. **Cheng Deng:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Our work was supported in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171–343, Grant 62071361; in part by Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03; and in part by the Fundamental Research Funds for the Central Universities ZDRC2102.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[2] S. Saxena, M.N. Teli, Comparison and analysis of image-to-image generative adversarial networks: A survey, arXiv preprint arXiv:2112.12625.
[3] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, Multimodal image synthesis and editing: A survey, arXiv preprint arXiv:2112.13592..
[4] P. Wang, Y. Li, N. Vasconcelos, Rethinking and improving the robustness of image style transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 124–133.
[5] H. Zhang, J.Y. Koh, J. Baldridge, H. Lee, Y. Yang, Cross-modal contrastive learning for text-to-image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 833–842.
[6] X. Ji, H. Zhou, K. Wang, W. Wu, C.C. Loy, X. Cao, F. Xu, Audio-driven emotional video portraits, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14080–14089.
[7] J. Johnson, A. Gupta, L. Fei-Fei, Image generation from scene graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1219–1228.
[8] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, Advances in Neural Information Processing Systems 30.
[9] P. Zhu, R. Abdal, Y. Qin, P. Wonka, Sean: Image synthesis with semantic region-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5104–5113.
[10] Z. Li, C. Deng, K. Wei, W. Liu, D. Tao, Learning semantic priors for texture-realistic sketch-to-image synthesis, Neurocomputing 464 (2021) 130–140.
[11] C. Deng, E. Yang, T. Liu, D. Tao, Two-stream deep hashing with class-specific centers for supervised image search, IEEE Transactions on Neural Networks and Learning Systems 31 (6) (2019) 2189–2201.
[12] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, D. Tao, Unsupervised semantic-preserving adversarial hashing for image search, IEEE Transactions on Image Processing 28 (8) (2019) 4032–4044.
[13] E. Yang, T. Liu, C. Deng, W. Liu, D. Tao, Distillhash: Unsupervised deep hashing by distilling data pairs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2946–2955.
[14] J. Dong, X. Wang, L. Zhang, C. Xu, G. Yang, X. Li, Feature re-learning with data augmentation for video relevance prediction, IEEE Transactions on Knowledge and Data Engineering 33 (5) (2021) 1946–1959.
[15] X. Yang, X. Liu, M. Jian, X. Gao, M. Wang, Weakly-supervised video object grounding by exploring spatio-temporal contexts, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1939–1947.
[16] J. Xiao, X. Shang, X. Yang, S. Tang, T.-S. Chua, Visual relation grounding in videos, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 447–464.
[17] Y. Li, X. Yang, X. Shang, T.-S. Chua, Interventional video relation detection, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4091–4099.
[18] Y. Tan, Y. Hao, X. He, Y. Wei, X. Yang, Selective dependency aggregation for action classification, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 592–601.
[19] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, T.-S. Chua, Annotating objects and relations in user-generated videos, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 279–287.
[20] Z. Li, C. Deng, E. Yang, D. Tao, Staged sketch-to-image synthesis via semi-supervised generative adversarial networks, IEEE Transactions on Multimedia 23 (2020) 2694–2705.
[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in Neural Information Processing Systems (2014) 2672–2680.

[22] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114.
[23] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: International Conference on Machine Learning, PMLR, 2015, pp. 2256–2265.
[24] D.J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: International Conference on Machine Learning, PMLR, 2014, pp. 1278–1286.
[25] A. Van Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: International Conference on Machine Learning, PMLR, 2016, pp. 1747–1756.
[26] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.
[27] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein gans, arXiv preprint arXiv:1704.00028.
[28] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196..
[29] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, in: International Conference on Learning Representations, 2018.
[30] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.
[31] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8110–8119.
[32] E. Yang, C. Deng, W. Liu, X. Liu, D. Tao, X. Gao, Pairwise relationship guided deep hashing for cross-modal retrieval, in: Poceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, 2017.
[33] C. Deng, X. Xu, H. Wang, M. Yang, D. Tao, Progressive cross-modal semantic network for zero-shot sketch-based image retrieval, IEEE Transactions on Image Processing 29 (2020) 8892–8902.
[34] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, M. Wang, Dual encoding for video retrieval by text, IEEE Transactions on Pattern Analysis and Machine Intelligence.
[35] J. Dong, X. Li, C.G. Snoek, Predicting visual features from text for image and video caption retrieval, IEEE Transactions on Multimedia 20 (12) (2018) 3377–3388.
[36] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, T.-S. Chua, Tree-augmented cross-modal encoding for complex-query video retrieval, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1339–1348.
[37] X. Yang, F. Feng, W. Ji, M. Wang, T.-S. Chua, Deconfounded video moment retrieval with causal intervention, in: SIGIR, 2021, pp. 1–10.
[38] J. Dong, Z. Ma, X. Mao, X. Yang, Y. He, R. Hong, S. Ji, Fine-grained fashion similarity prediction by attribute-specific embedding learning, IEEE Transactions on Image Processing 30 (2021) 8410–8425.
[39] X. Liu, X. Yang, M. Wang, R. Hong, Deep neighborhood component analysis for visual similarity modeling, ACM Transactions on Intelligent Systems and Technology 11 (3) (2020) 1–15.
[40] E. Mansimov, E. Parisotto, J.L. Ba, R. Salakhutdinov, Generating images from captions with attention, arXiv preprint arXiv:1511.02793.
[41] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[42] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[43] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine-grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1316–1324.
[44] A. Dash, J.C.B. Gamboa, S. Ahmed, M. Liwicki, M.Z. Afzal, Tac-gan-text conditioned auxiliary classifier generative adversarial network, arXiv preprint arXiv:1703.06412.
[45] Z. Wang, W. Liu, Q. He, X. Wu, Z. Yi, Clip-gen: Language-free training of a text-to-image generator with clip, arXiv preprint arXiv:2203.00386.
[46] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: International Conference on Machine Learning PMLR, 2016, pp. 1060–1069.
[47] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.
[48] L. Chen, S. Srivastava, Z. Duan, C. Xu, Deep cross-modal audio-visual generation, in: Proceedings of the on Thematic Workshops of ACM Multimedia 2017, 2017, pp. 349–357.
[49] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907–5915.
[50] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (8) (2019) 1947–1962.
[51] T. Qiao, J. Zhang, D. Xu, D. Tao, Mirrorgan: Learning text-to-image generation by redescription, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1505–1514.
[52] Z. Chen, Y. Luo, Cycle-consistent diverse image synthesis from natural language, in: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2019, pp. 459–464.
[53] B. Li, X. Qi, T. Lukasiewicz, P. Torr, Controllable text-to-image generation, Advances in Neural Information Processing Systems 32.
[54] L. Goetschalckx, A. Andonian, A. Oliva, P. Isola, Ganalyze: Toward visual definitions of cognitive image properties, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5744–5753.
[55] J. Zhu, Y. Shen, D. Zhao, B. Zhou, In-domain gan inversion for real image editing, in: Proceedings of the European Conference on Computer Vision Springer, 2020, pp. 592–608.
[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30.
[57] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning, PMLR, 2021, pp. 8821–8831.
[58] S. Tulyakov, M.-Y. Liu, X. Yang, J. Kautz, Mocogan: Decomposing motion and content for video generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1526–1535.
[59] X. Nie, J. Jia, H. Ding, E.K. Wong, Gigan: Gate in gan, could gate mechanism filter the features in image-to-image translation?, Neurocomputing 462 (2021) 376–388
[60] J. Lin, Y. Xia, S. Liu, S. Zhao, Z. Chen, Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation, Neurocomputing 461 (2021) 327–335.
[61] K. Lyu, S. Pan, Y. Li, Z. Zhang, Jsenet: A deep convolutional neural network for joint image super-resolution and enhancement, Neurocomputing.
[62] Y. Choi, Y. Uh, J. Yoo, J.-W. Ha, Stargan v2: Diverse image synthesis for multiple domains, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8188–8197.
[63] Y. Yang, L. Wang, D. Xie, C. Deng, D. Tao, Multi-sentence auxiliary adversarial networks for fine-grained text-to-image synthesis, IEEE Transactions on Image Processing 30 (2021) 2798–2809.
[64] S. Pande, S. Chouhan, R. Sonavane, R. Walambe, G. Ghinea, K. Kotecha, Development and deployment of a generative model-based framework for text to photorealistic image generation, Neurocomputing 463 (2021) 1–16.
[65] Z. Zhang, L. Schomaker, Divergan: An efficient and effective single-stage framework for diverse text-to-image generation, Neurocomputing 473 (2022) 182–198.
[66] Z. Qi, J. Sun, J. Qian, J. Xu, S. Zhan, Pccm-gan: Photographic text-to-image generation with pyramid contrastive consistency model, Neurocomputing 449 (2021) 330–341.
[67] B. Li, X. Qi, T. Lukasiewicz, P.H. Torr, Manigan: Text-guided image manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7880–7889.
[68] Y. Shen, J. Gu, X. Tang, B. Zhou, Interpreting the latent space of gans for semantic face editing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9243–9252.
[69] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, O. Scharenborg, Generating images from spoken descriptions, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 850–865.
[70] N. Yalta, S. Watanabe, K. Nakadai, T. Ogata, Weakly-supervised deep recurrent neural networks for basic dance step generation, in: 2019 International Joint Conference on Neural Networks, IEEE, 2019, pp. 1–8.
[71] P. Zhao, Y. Chen, L. Zhao, G. Wu, X. Zhou, Generating images from audio under semantic consistency, Neurocomputing.
[72] L. Karacan, Z. Akata, A. Erdem, E. Erdem, Learning to generate images of outdoor scenes from attributes and semantic layouts, arXiv preprint arXiv:1612.00215.
[73] X. Qi, Q. Chen, J. Jia, V. Koltun, Semi-parametric image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8808–8816.
[74] C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, IEEE Conference on Computer Vision and Pattern Recognition (2020) 5549–5558.
[75] S. Wu, W. Liu, Q. Wang, S. Zhang, Z. Hong, S. Xu, Reffacenet: Reference-based face image generation from line art drawings, Neurocomputing.
[76] X. Yang, C. Deng, T. Liu, D. Tao, Heterogeneous graph attention network for unsupervised multiple-target domain adaptation, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (4) (2022) 1992–2003.
[77] Y. Zhu, C. Deng, H. Cao, H. Wang, Object and background disentanglement for unsupervised cross-domain person re-identification, Neurocomputing 403 (2020) 88–97.
[78] A. Hertzmann, C.E. Jacobs, N. Oliver, B. Curless, D.H. Salesin, Image analogies, in: Proceedings of the 28th annual Conference on Computer Graphics and Interactive Techniques, 2001, pp. 327–340.
[79] C.D. Xintao Wang, Yu. Ke, C.C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 606–615.

[80] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.
[81] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[82] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8798–8807.
[83] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: Unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2849–2857.
[84] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, Advances in Neural Information Processing Systems (2017) 700–708.
[85] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 172–189.
[86] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, M.-H. Yang, Diverse image-to-image translation via disentangled representations, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 35–51.
[87] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, M.-H. Yang, Drit ++: Diverse image-to-image translation via disentangled representations, International Journal of Computer Vision 128 (10) (2020) 2402–2417.
[88] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[89] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2) (2016) 295–307.
[90] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[91] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.
[92] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.
[93] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[94] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 286–301.
[95] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision Workshops, 2018.
[96] D. Song, Y. Wang, H. Chen, C. Xu, C. Xu, D. Tao, Addersr: Towards energy efficient image super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15648–15657.
[97] K.C. Chan, X. Wang, X. Xu, J. Gu, C.C. Loy, Glean: Generative latent bank for large-factor image super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14245–14254.
[98] Y. Zhou, G. Wu, Y. Fu, K. Li, Y. Liu, Cross-mpi: Cross-scale stereo for image super-resolution using multiplane images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14842–14851.
[99] Y. Jiang, K.C. Chan, X. Wang, C.C. Loy, Z. Liu, Robust reference-based super-resolution via c2-matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2103–2112.
[100] Y. Qu, Y. Chen, J. Huang, Y. Xie, Enhanced pix2pix dehazing network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8160–8168.
[101] X. Liu, Y. Ma, Z. Shi, J. Chen, Griddehazenet: Attention-based multi-scale network for image dehazing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7314–7323.
[102] X. Qin, Z. Wang, Y. Bai, X. Xie, H. Jia, Ffa-net: Feature fusion attention network for single image dehazing, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11908–11915.
[103] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, M.-H. Yang, Multi-scale boosted dehazing network with dense feature fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2157–2167.
[104] M. Hong, Y. Xie, C. Li, Y. Qu, Distilling image dehazing with heterogeneous task imitation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3462–3471.
[105] H. Zhang, V.M. Patel, Densely connected pyramid dehazing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3194–3203.
[106] H. Zhang, V. Sindagi, V.M. Patel, Multi-scale single image dehazing using perceptual pyramid deep network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 902–911.
[107] Y. Dong, Y. Liu, H. Zhang, S. Chen, Y. Qiao, Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10729–10736.
[108] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, L. Ma, Contrastive learning for compact single image dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10551–10560.
[109] L. Zhang, Y. Ji, X. Lin, C. Liu, Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier gan, in: 2017 4th IAPR Asian Conference on Pattern Recognition, IEEE, 2017, pp. 506–511.
[110] M. Lu, H. Zhao, A. Yao, Y. Chen, F. Xu, L. Zhang, A closed-form solution to universal style transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5952–5961.
[111] N. Kalischek, J.D. Wegner, K. Schindler, In the light of feature distributions: moment matching for neural style transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9382–9391.
[112] L.A. Gatys, A.S. Ecker, M. Bethge, A neural algorithm of artistic style, arXiv preprint arXiv:1508.06576.
[113] F. Shen, S. Yan, G. Zeng, Neural style transfer via meta networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8061–8069.
[114] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.
[115] E. Risser, P. Wilmot, C. Barnes, Stable and controllable neural texture synthesis and style transfer using histogram losses, arXiv preprint arXiv:1701.08893.
[116] R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 768–783.
[117] C. Li, M. Wand, Combining markov random fields and convolutional neural networks for image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2479–2486.
[118] L.A. Gatys, A.S. Ecker, M. Bethge, A. Hertzmann, E. Shechtman, Controlling perceptual factors in neural style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3985–3993.
[119] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022..
[120] D. Ulyanov, V. Lebedev, A. Vedaldi, V.S. Lempitsky, Texture networks: Feed-forward synthesis of textures and stylized images., in: International Conference on Machine Learning, Vol. 1, 2016, p. 4..
[121] V. Dumoulin, J. Shlens, M. Kudlur, A learned representation for artistic style, arXiv preprint arXiv:1610.07629..
[122] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.
[123] G. Shen, W. Huang, C. Gan, M. Tan, J. Huang, W. Zhu, B. Gong, Facial image-to-video translation by a hidden affine transformation, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2505–2513.
[124] S.W. Kim, Y. Zhou, J. Philion, A. Torralba, S. Fidler, Learning to simulate dynamic environments with gamegan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1231–1240.
[125] N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, S. Berretti, Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets, IEEE Transactions on Pattern Analysis and Machine Intelligence.
[126] A. Zhao, G. Balakrishnan, K.M. Lewis, F. Durand, J.V. Guttag, A.V. Dalca, Painting many pasts: Synthesizing time lapse videos of paintings, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8435–8445.
[127] M. Maximov, I. Elezi, L. Leal-Taixé, Ciagan: Conditional identity anonymization generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5447–5456.
[128] S. Nam, C. Ma, M. Chai, W. Brendel, N. Xu, S.J. Kim, End-to-end time-lapse video synthesis from a single outdoor image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1409–1418.
[129] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[130] C. Vondrick, H. Pirsiavash, A. Torralba, Generating videos with scene dynamics, Advances in Neural Information Processing Systems 29 (2016) 613–621.

[131] M. Saito, E. Matsumoto, S. Saito, Temporal generative adversarial nets with singular value clipping, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2830–2839.
[132] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro, Video-to-video synthesis, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 1152–1164.
[133] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, B. Catanzaro, Few-shot video-to-video synthesis, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 5013–5024.
[134] C. Chan, S. Ginosar, T. Zhou, A.A. Efros, Everybody dance now, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5933–5942.
[135] R. Xu, X. Li, B. Zhou, C.C. Loy, Deep flow-guided video inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3723–3732.
[136] Z. Yang, W. Zhu, W. Wu, C. Qian, Q. Zhou, B. Zhou, C.C. Loy, Transmomo: Invariance-driven unsupervised video motion retargeting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5306–5315.
[137] E. Shechtman, Y. Caspi, M. Irani, Space-time super-resolution, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (4) (2005) 531–545.
[138] C. Finn, I. Goodfellow, S. Levine, Unsupervised learning for physical interaction through video prediction, Advances in Neural Information Processing Systems 29 (2016) 64–72.
[139] F. Cricri, X. Ni, M. Honkala, E. Aksu, M. Gabbouj, Video ladder networks, arXiv preprint arXiv:1612.01756.
[140] S. Chiappa, S. Racaniere, D. Wierstra, S. Mohamed, Recurrent environment simulators, arXiv preprint arXiv:1704.02254.
[141] E. Denton, V. Birodkar, Unsupervised learning of disentangled representations from video, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4417–4426.
[142] X. Liang, L. Lee, W. Dai, E.P. Xing, Dual motion gan for future-flow embedded video prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1744–1752.
[143] Z.S. Harris, Distributional structure, Word 10 (2–3) (1954) 146–162.
[144] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems 26.
[145] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (11) (1997) 2673–2681.
[146] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, et al., Cogview: Mastering text-to-image generation via transformers, Advances in Neural Information Processing Systems 34.
[147] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
[148] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, D. Lischinski, Styleclip: Text-driven manipulation of stylegan imagery, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2085–2094.
[149] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, B. Guo, Vector quantized diffusion model for text-to-image synthesis, arXiv preprint arXiv:2111.14822.
[150] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, T. Sun, Lafite: Towards language-free training for text-to-image generation, arXiv preprint arXiv:2111.13792.
[151] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, M.-H. Yang, Gan inversion: A survey, arXiv preprint arXiv:2101.05278..
[152] Z. Zhang, Y. Xie, L. Yang, Photographic text-to-image synthesis with a hierarchically-nested adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6199–6208.
[153] H. Tan, X. Liu, X. Li, Y. Zhang, B. Yin, Semantics-enhanced adversarial nets for text-to-image synthesis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10501–10510.
[154] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1857–1865.
[155] Q. Lao, M. Havaei, A. Pesaranghader, F. Dutil, L.D. Jorio, T. Fevens, Dual adversarial inference for text-to-image synthesis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7567–7576.
[156] H. Dong, S. Yu, C. Wu, Y. Guo, Semantic image synthesis via adversarial learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5706–5714.
[157] S. Nam, Y. Kim, S.J. Kim, Text-adaptive generative adversarial networks: manipulating images with natural language, Advances in Neural Information Processing Systems 31.
[158] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[159] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, W. Gao, Direct speech-to-image translation, IEEE Journal of Selected Topics in Signal Processing 14 (3) (2020) 517–529.
[160] W. Xia, Y. Yang, J.-H. Xue, B. Wu, Tedigan: Text-guided diverse face image generation and manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2256–2265.
[161] L. Chen, R.K. Maddox, Z. Duan, C. Xu, Hierarchical cross-modal talking face generation with dynamic pixel-wise loss, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7832–7841.
[162] C.-H. Wan, S.-P. Chuang, H.-Y. Lee, Towards audio to scene image synthesis using generative adversarial network, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, 2019, pp. 496–500.
[163] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, Advances in Neural Information Processing Systems 29.
[164] O. Alemi, J. Françoise, P. Pasquier, Groovenet: Real-time music-driven dance movement generation using artificial neural networks, Networks 8 (17) (2017) 26.
[165] E. Shlizerman, L. Dery, H. Schoen, I. Kemelmacher-Shlizerman, Audio to body dynamics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7574–7583.
[166] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, Y. Bengio, Obamanet: Photo-realistic lip-sync from text, arXiv preprint arXiv:1801.01442.
[167] S.E. Eskimez, R.K. Maddox, C. Xu, Z. Duan, Noise-resilient training method for face landmark generation from speech, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2019) 27–38.
[168] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, C.C. Loy, Mead: A large-scale audio-visual dataset for emotional talking-face generation, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 700–717.
[169] L. Song, W. Wu, C. Qian, R. He, C.C. Loy, Everybody's talkin': Let me talk as you want, IEEE Transactions on Information Forensics and Security.
[170] A. Mesaros, T. Heittola, T. Virtanen, M.D. Plumbley, Sound event detection: A tutorial, IEEE Signal Processing Magazine 38 (5) (2021) 67–83.
[171] C.-C. Lee, W.-Y. Lin, Y.-T. Shih, P.-Y. Kuo, L. Su, Crossing you in style: Cross-modal style transfer from music to visual arts, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3219–3227.
[172] D. Jeong, S. Doh, T. Kwon, Träumerai: Dreaming music with stylegan, arXiv preprint arXiv:2102.04680.
[173] S.H. Lee, W. Roh, W. Byeon, S.H. Yoon, C.Y. Kim, J. Kim, S. Kim, Sound-guided semantic image manipulation, arXiv preprint arXiv:2112.00007.
[174] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.
[175] T.R. Shaham, M. Gharbi, R. Zhang, E. Shechtman, T. Michaeli, Spatially-adaptive pixelwise networks for fast image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14882–14891.
[176] Q. Chen, V. Koltun, Photographic image synthesis with cascaded refinement networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1511–1520.
[177] Y. Jo, J. Park, Sc-fegan: Face editing generative adversarial network with user's sketch and color, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1745–1753.
[178] H. Dong, X. Liang, Y. Zhang, X. Zhang, X. Shen, Z. Xie, B. Wu, J. Yin, Fashion editing with adversarial parsing learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8120–8128.
[179] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, L. Yuan, Mask-guided portrait editing with conditional gans, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3436–3445.
[180] S.-Z. Xu, H.-Z. Huang, F.-L. Zhang, S.-H. Zhang, Faceshapegene: a disentangled shape representation for flexible face image editing, Graphics and Visual, Computing (2021) 200023.
[181] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[182] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning.
[183] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images.
[184] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim, et al., Ntire 2017 challenge on single image super-resolution: Methods and results, in: The IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017.
[185] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset..
[186] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE, 2008, pp. 722–729.
[187] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[188] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, Proc. 8th Int'l Conf. Computer Vision, vol. 2, 2001, pp. 416–423.
[189] M. Bevilacqua, A. Roumy, C. Guillemot, M.L. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding.
[190] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International Conference on Curves and Surfaces, Springer, 2010, pp. 711–730.

[191] J.-B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197–5206.
[192] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, K. Aizawa, Sketch-based manga retrieval using manga109 dataset, Multimedia Tools and Applications 76 (20) (2017) 21811–21838.
[193] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, J. Hays, Transient attributes for high-level understanding and editing of outdoor scenes, ACM Transactions on Graphics 33 (4) (2014) 1–11.
[194] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, J. Xiao, Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365.
[195] T. Xue, B. Chen, J. Wu, D. Wei, W.T. Freeman, Video enhancement with task-oriented flow, International Journal of Computer Vision 127 (8) (2019) 1106–1125.
[196] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, K.M. Lee, Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[197] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision, 2015, pp. 3730–3738.
[198] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-ucsd birds 200.
[199] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, V. Ferrari, Connecting vision and language with localized narratives, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 647–664.
[200] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision 123 (1) (2017) 32–73.
[201] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, The Journal of the Acoustical Society of America 120 (5) (2006) 2421–2424.
[202] J.S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, arXiv preprint arXiv:1806.05622.
[203] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, International Journal of Computer Vision 127 (3) (2019) 302–321.
[204] H. Caesar, J. Uijlings, V. Ferrari, Coco-stuff: Thing and stuff classes in context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1209–1218.
[205] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[206] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.
[207] O. Langner, R. Dotsch, G. Bijlstra, D.H. Wigboldus, S.T. Hawk, A. Van Knippenberg, Presentation and validation of the radboud faces database, Cognition and emotion 24 (8) (2010) 1377–1388.
[208] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[209] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.
[210] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2, IEEE, 2003, pp. 1398–1402.
[211] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[212] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Advances in Neural Information Processing Systems (2016) 2234–2242.
[213] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30.
[214] T. Hinz, S. Heinrich, S. Wermter, Semantic object accuracy for generative text-to-image synthesis, arXiv preprint arXiv:1910.13321.
[215] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, J. Shao, Semantics disentangling for text-to-image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2327–2336.
[216] J. Liang, W. Pei, F. Lu, Cpgan: full-spectrum content-parsing generative adversarial networks for text-to-image synthesis, arXiv preprint arXiv:1912.08562.
[217] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, D. Li, Makelttalk: speaker-aware talking-head animation, ACM Transactions on Graphics (TOG) 39 (6) (2020) 1–15.
[218] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, M.-H. Yang, Mode seeking generative adversarial networks for diverse image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1429–1437.
[219] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, N. Snavely, Stereo magnification: Learning view synthesis using multiplane images, arXiv preprint arXiv:1805.09817.
[220] C. Deng, Y. Xue, X. Liu, C. Li, D. Tao, Active transfer learning network: A unified deep joint spectral-spatial feature learning model for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 57 (3) (2018) 1741–1754.
[221] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, IEEE Transactions on Image Processing 27 (8) (2018) 3893–3903.
[222] H. Tan, X. Liu, M. Liu, B. Yin, X. Li, KT-GAN: Knowledge-Transfer Generative Adversarial Network for Text-to-Image Synthesis, IEEE Transactions on Image Processing 30 (2021) 1275–1290.
[223] H. Tan, X. Liu, B. Yin, X. Li, Cross-Modal Semantic Matching Generative Adversarial Networks for Text-to-Image Synthesis, IEEE Transactions on Multimedia 24 (2022) 832–845.
[225] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
[226] G. Mittal, T. Marwah, V.N. Balasubramanian, Sync-draw: Automatic video generation using deep recurrent attentive architectures, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1096–1104.
[227] E. Yang, C. Deng, T. Liu, W. Liu, D. Tao, Semantic structure-based unsupervised deep hashing, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 1064–1070.
[228] Y. Liu, X. Wang, Y. Yuan, W. Zhu, Cross-modal dual learning for sentence-to-video generation, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1239–1247.
[229] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

Ziqi Zhang received the B.E. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Electronic Engineering. His research interests lie primarily in computer vision and machine learning.

Zeyu Li received the B.E. degree in Optoelectronic Information Science and Engineering from Xi'an University of Post & Telecommunications, China, in 2013. He is currently pursuing his Ph.D. degree at School of Electronic Engineering, Xidian University. His research interests lie primarily in computer vision and machine learning.

Kun Wei received the B.E. degree in Electronic and Information Engineering from Xidian University, China, in 2017. He is currently pursuing his Ph.D. degree at School of Electronic Engineering, Xidian University. His research interests lie primarily in computer vision and machine learning.

Siduo Pan received the B.E. degree from the School of Electronic Engineering, Xidian University, Xi'an, China, in 2021. She is currently pursuing the degree of Master at the School of Electronic Engineering, Xidian University. Her research interests include computer vision and image synthesis.

Cheng Deng (SM'20) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Full Professor with the School of Electronic Engineering at Xidian University. His research interests include computer vision, pattern recognition, and information hiding. He is the author and coauthor of more than 100 scientific articles at top venues, including IEEE TNNLS, TIP, TCYB, TMM, TSMC, ICCV, CVPR, ICML, NIPS, IJCAI, and AAAI.