Scripted Video Generation With a Bottom-Up Generative Adversarial Network

Qi Chen, Qi Wu, Jian Chen, *Member, IEEE*, Qingyao Wu, *Member, IEEE*, Anton van den Hengel, *Member, IEEE*, and Mingkui Tan, *Member, IEEE*

***Abstract*— Generating videos given a text description (such as a script) is non-trivial due to the intrinsic complexity of image frames and the structure of videos. Although Generative Adversarial Networks (GANs) have been successfully applied to generate images conditioned on a natural language description, it is still very challenging to generate realistic videos in which the frames are required to follow both spatial and temporal coherence. In this paper, we propose a novel Bottom-up GAN (BoGAN) method for generating videos given a text description. To ensure the coherence of the generated frames and also make the whole video match the language descriptions semantically, we design a bottom-up optimisation mechanism to train BoGAN. Specifically, we devise a region-level loss via attention mechanism to preserve the local semantic alignment and draw details in different sub-regions of video conditioned on words which are most relevant to them. Moreover, to guarantee the matching between text and frame, we introduce a frame-level discriminator, which can also maintain the fidelity of each frame and the coherence across frames. Last, to ensure the global semantic alignment between whole video and given text, we apply a video-level discriminator. We evaluate the effectiveness of the proposed BoGAN on two synthetic datasets (i.e., SBMG and TBMG) and two real-world datasets (i.e., MSVD and KTH).**

***Index Terms*— Generative adversarial networks, video generation, semantic alignment, temporal coherence.**

I. Introduction

VISION is one of the most important ways in which humans experience, interact with, understand, and learn about the world around them. Intelligent systems that can generate videos for human users has tremendous potential application, such as video editing, video games, and computer-aided design. Unfortunately, many modern and creative works are now generated or edited using digital graphic design tools. The complexity of these tools may lead to inaccessibility issues, particularly for people with insufficient technical knowledge or resources. Thus, a system that has the ability to follow speech- or text-based instructions and then perform a corresponding video editing task could improve accessibility substantially. These benefits can easily extend to other domains of video generation such as gaming, animation, creating visual teaching material, *etc*. In this paper, we take a step in this exciting research direction by the text to video generation task.

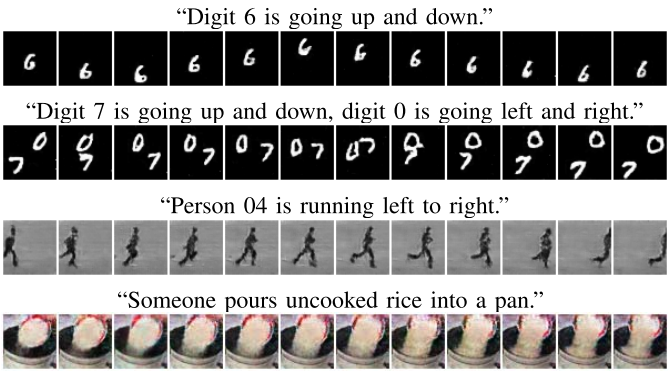


Fig. 1. Generated videos conditioned on given scripts by our BoGAN model trained on SBMG, TBMG, KTH and MSVD datasets, respectively.

Specifically, we focus on video generation from text, which aims to generate a video semantically aligned with some given descriptive scripts. Compared to image generation [1]–[4], the video generation task is much more difficult since the video is often a complex sequence of many frames which should follow strong spatial and temporal dependencies. More critically, generating video conditioned on given text is even more complicated due to the requirement of semantic alignment between video and text at both frame and video levels. Thus, although there are already a lot of existing models for text-to-image generation, simply using the image generator to synthesise videos may incur poor performance (see the experimental results in Sec. IV and the results in [5]).

In video generation from text, there are two main challenges: 1) semantic alignment between given text and video content; 2) realistic video generation with temporal coherence across frames. Recently, some existing works have tried to address these challenges individually. For example,

Manuscript received April 2, 2019; revised January 9, 2020 and April 2, 2020; accepted May 31, 2020. Date of current version July 13, 2020. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2018B010107001, in part by the National Natural Science Foundation of China (NSFC) through the Key Project under Grant 61836003, in part by the Guangdong Project under Grant 2017ZT07X183, in part by the Fundamental Research Funds for the Central Universities under Grant D2191240, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515130001, in part by the Guangdong Special Branch Plans Young Talent with Scientific and Technological Innovation under Grant 2016TQ03X445, and in part by the Guangzhou Science and Technology Planning Project under Grant 201904010197. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Junsong Yuan. *(Qi Chen, Qi Wu, and Jian Chen contributed equally to this work.) (Corresponding author: Mingkui Tan.)*

Qi Chen is with the School of Software Engineering, South China University of Technology, Guangzhou 510640, China, and also with the Pazhou Laboratory, Guangzhou 510335, China (e-mail: sechenqi@mail.scut.edu.cn).

Qi Wu and Anton van den Hengel are with the School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia (e-mail: qi.wu01@adelaide.edu.au; anton.vandenhengel@adelaide.edu.au).

Jian Chen, Qingyao Wu, and Mingkui Tan are with the School of Software Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: ellachen@scut.edu.cn; qyw@scut.edu.cn; mingkuitan@scut.edu.cn).

Digital Object Identifier 10.1109/TIP.2020.3003227

Li *et al*. [5] propose a two-stage VAE-based generator to yield a 'gist' of video from the input text first, where the gist is an image that gives the background colour and object layout. The content and motion of the video is then generated by conditioning on the gist. However, since they neglect the relation between consecutive frames, the motions of generated video are incoherent. Conversely, Pan *et al*. [6] consider the temporal coherence across frames and generate video with the given text with a carefully designed discriminator. Nonetheless, due to the sweeping alignment between text and video via the classical conditional loss [1], the generated videos ignore some subtle semantics of words, which is vital to synthesise the details.

To address the above issues, we propose a novel Bottom-Up Generative Adversarial Network (BoGAN), which ensures the coherence between consecutive frames and preserve the semantic matching between video and the corresponding language description at different levels. More specifically, we design a sophisticated bottom-up mechanism with multiple losses in terms of three levels: 1) region-level, 2) frame-level and 3) video-level. To keep the local semantic alignment, we devise a *region-level* loss via attention mechanism to draw different sub-regions of video conditioned on the words which are relevant to them. Moreover, to guarantee the match between given script and each frame, we design a *frame-level* discriminator, which maintains the fidelity of each frame and the coherence across frames. Finally, to ensure the global semantic alignment between the entire video and corresponding description, we apply a *video-level* discriminator with 3D convolutional filers. On the other hand, for video generator, relying on the classical encoder-decoder design, we build the encoder with a Long-Short Term Memory (LSTM) network to convert the input script into an embedding while devise the decoder with 3D deconvolutional layers, which generates a sequence of frames from given vector.

We test our BoGAN on two synthetic datasets (SBMG [6] and TBMG [6]) and two real-world datasets (MSVD [7] and KTH [8]). We employ quantitative evaluations by using the commonly used generative adversarial metrics, including Fréchet Inception Distance (FID) [9] and its variant in video (FID2vid) [10]. BoGAN outperforms the baseline model and state-of-the-art models by a large margin. Several variants of our model are tested to validate the contribution of each component. To measure whether the generated video semantically matches the script, several human studies are performed.

We highlight our principal contributions as follows:

- We propose a novel Bottom-Up Generative Adversarial Network (BoGAN) to produce videos according to natural language descriptions, in which the generated frames well follow the semantic alignment with given scripts.
- In synthesising videos, to ensure the fidelity of frames and multi-scale semantic alignment with given text, we design a sophisticated bottom-up mechanism to optimise our video generator. To be specific, the mechanism consists of multiple losses at three different levels (from local to global), which focus on temporal coherence and semantic match on various granularity.
- We achieve the state-of-the-art performance on both synthetic and real-world datasets in terms of qualitative and quantitative metrics. Moreover, we conduct an ablation study to verify the effectiveness of each component and report some discussions about hyper-parameters.

II. Related Work

A. Video Generation

A spatial-temporal 3D deconvolution-based GANs is proposed for unconditioned video generation in [11]. To learn the semantic representation of unlabeled videos, Saito *et al*. [12] design two different generators (a temporal generator and an image generator), which sequentially transform a single latent variable into a video. Tulyakov *et al*. [13] propose a framework to generate video by decomposing motion and content in an unsupervised manner. On the other hand, many applications focus on generating video conditioned on a stable image (frame), such as [14]–[18]. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are widely used and achieve the great performance in these applications [19]–[23]. The common thread of these works is to use CNN for encoding each frame and then to apply an sequence-to-sequence model for frame prediction. Instead of unconditioned or image conditioned videos generation, we focus on generating videos conditioned on text, which is more challenging due to the requirement of semantic alignment between videos and natural languages.

B. Video Generation Conditioned on Text

Video generation conditioned on text aims to synthesise a video which is semantically aligned with the given descriptive sentence, such as a caption or script. Mittal *et al*. [24] devise a method to generate video from text by combining a Variational Autoencoder with a Recurrent Attention Mechanism, which captures the temporally dependent sequence of frames. Marwah *et al*. [25] propose an improved model to incorporate the long-term and short-term dependencies between frames and generate video in a incremental manner. Most recently, Li *et al*. [5] propose a two-stage VAE-based generator to generate a 'gist' of the video from input text, where the gist is an image that gives the background colour and object layout. The content and motion of the video are then generated by conditioning on the gist. Meanwhile, due to the success of Generative Adversarial Networks (GANs) [26], Pan *et al*. [6] consider the temporal coherence across frames and the semantic matching between text the whole video with a carefully designed discriminator. Besides the frame coherence and video-text semantic matching considered in the above methods, we also develop a bottom-up mechanism to ensure the spatial-temporal coherence and semantic matching between text and video at multiple levels, including region, frame and video.

C. Generative Adversarial Network

Generative Adversarial Network (GAN) has been the subject of significant attention in the last few years in light of

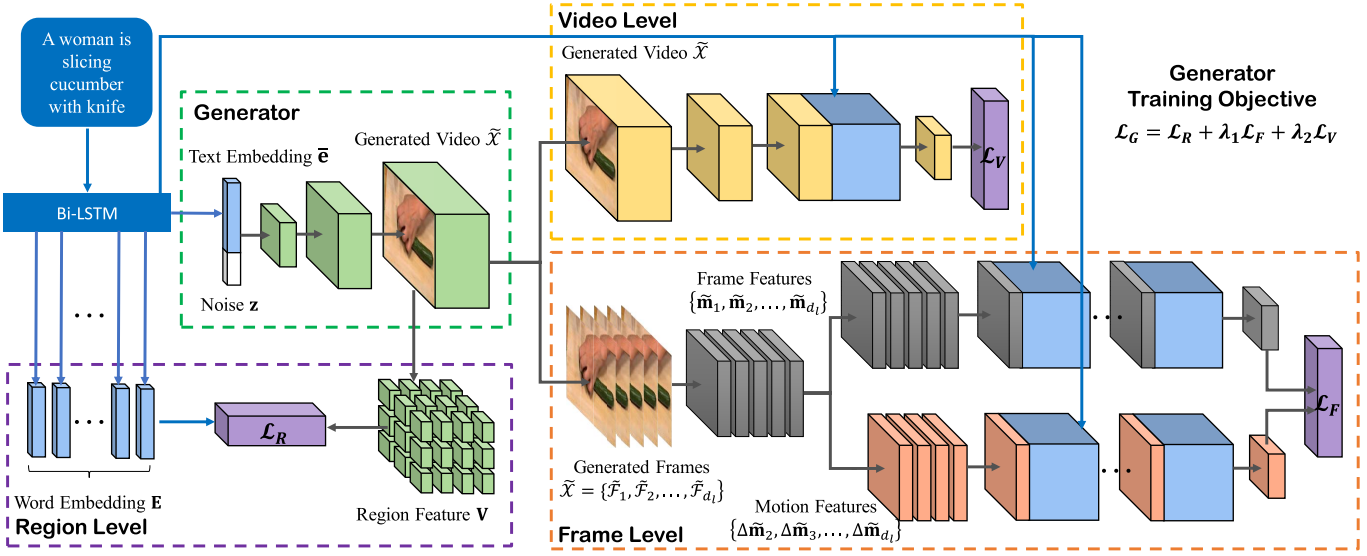


Fig. 2. The architecture of our proposed BoGAN. The region-level model seeks to draw different sub-regions of the video according to the most relevant words; the frame-level model ensures the fidelity in each frame and the coherence across frames; the last video-level model makes the generated video more natural and more consistent with the script.

advances in deep learning. Original GAN [26] is proposed to generate natural images from input noise. The application domain has broadened significantly, however, particularly into image generation [27]–[34]. Besides, GAN has also been applied to range of other interesting applications including image inpainting [35]–[38], image/video super-resolution [39]–[42], image/video deblurring [43], [44] and facial attribute manipulation [45]–[47], *etc*. Specifically, GAN learns a generative model by playing a two-player minimax game to match the underlying data distribution. A GAN is made up of a generator and a discriminator. The generator struggles to produce samples whose distribution is indistinguishable from that of the training samples. On the other hand, the discriminator acts as a judge to distinguish the generated samples and real samples. Our model extends the original GAN structure, having one generator, one semantic matching module and two discriminators.

III. Proposed Method

The main challenges for video generation from script lie on how to capture both spatial and temporal coherence, as well as the semantic dependency between text and video. To solve these problems, we propose a novel architecture named Bottom-Up Generative Adversarial Network (BoGAN). From Fig. 2, our proposed BoGAN consists of four parts: a video generator, a region-level semantic alignment module, a frame-level coherence-aware discriminator and a video-level semantic-aware discriminator. To be specific, the motivation of each module is illustrated as follows:

Video Generator. We first build a video generator, which aims to synthesise the video from input script via a typical encoder-decoder design. Given a text, we use a LSTM-based encoder to produce hidden states and then generate a sequence of frames by a 3D convolutional decoder.

Region-level Semantic Alignment Module. To enable the generator to exploit the local semantic alignment between video and words, we devise a region-level semantic alignment module, which draws different sub-regions of the video conditioned on the words that are most relevant to those sub-regions.

Frame-level Coherence-aware Discriminator. In order to keep the match between text and each frame, we propose a frame-level coherence-aware discriminator, which can also enhances the realism in each generated frame and the temporal coherence between two consecutive frames.

Video-level Semantic-aware Discriminator. Finally, to produce a natural video with strong semantic alignment with the natural language description, we design a video-level semantic-aware discriminator to exploit the global information over the entire video.

A. Video Generator

As shown in Fig. 2, to synthesise video from given text, we build a video generator, which follows the widely used encoder-decoder design. More specifically, the encoder is a bi-directional Long Short-Term Memory (LSTM) [48] that extracts semantic vector from the input text $\mathcal{Q}$. In the bi-directional LSTM, each word corresponds to two hidden states, one for the forward and one for the backward direction. Thus, we concatenate its two hidden states to represent the semantic meaning of a word. The feature matrix of all words is indicated by $\mathbf{E} \in \mathbb{R}^{D\times T}$. Its i^{th} column $\mathbf{e}_i$ represents the embedding of i^{th} word. D is the dimension of the extracted word embeddings and T denotes the number of words in a script. The global sentence representation $\bar{\mathbf{e}} \in \mathbb{R}^D$ is the concatenation of last hidden states of the bi-directional LSTM.

To synthesise video that embodies the same semantic content as the input text, we incorporate the semantic text embedding $\bar{\mathbf{e}}$ with random noise $\mathbf{z} \in \mathbb{R}^{d_z} \sim \mathcal{N}(0, 1)$. Then,

we use a fully-connected layer to learn a unified embedding:

$$\mathbf{p} = \mathbf{W}_e\,[\mathbf{z}; \bar{\mathbf{e}}] \in \mathbb{R}^{d_p}, \tag{1}$$

where $\mathbf{W}_e \in \mathbb{R}^{d_p \times (d_z + D)}$ is the embedding weight, d_p is the embedding size, and $[\cdot;\cdot]$ is the concatenation operation. Conditioned on input vector $\mathbf{p}$, we generate the corresponding video by

$$\widetilde{\mathcal{X}} = \mathcal{G}(\mathbf{p}) \in \mathbb{R}^{d_c \times d_l \times d_h \times d_w}, \tag{2}$$

where $\widetilde{\mathcal{X}} = \{\widetilde{\mathcal{F}}_1, \widetilde{\mathcal{F}}_2, \ldots, \widetilde{\mathcal{F}}_{d_l}\}$ denotes the synthetic video, and $\widetilde{\mathcal{F}}_i \in \mathbb{R}^{d_c \times d_h \times d_w}$ is the i^{th} synthetic frame. Here d_c, d_l, d_h, d_w denote the channels number, sequence length, and frame height and width, respectively. To preserve the spatial-temporal information, we devise the generative model $\mathcal{G}$ using the 3D convolution filters [49] as deconvolutions [50], which is able to simultaneously capture the spatial structural information in each frame and the temporal information across frames.

To generate video from script, the proposed model must: 1) preserve the semantic alignment between given text and video content; and 2) ensure the realism of each frame while maintain the coherence across frames.

To satisfy the above requirements, we propose a bottom-up mechanism consisting of three losses $\mathcal{L}_R$, $\mathcal{L}_F$ and $\mathcal{L}_V$, which reflect the region-level, frame-level and video-level suitability, respectively. The $\mathcal{L}_R$ is a novel region-level loss, which focuses on exploiting the match between words and sub-regions of video. The adversarial losses $\mathcal{L}_F$ and $\mathcal{L}_V$ are used to ensure the semantic alignment and realism of the generated videos at frame-level and video-level, respectively. Overall, the final objective function of the generator $\mathcal{G}$ is

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_R + \lambda_1 \mathcal{L}_F + \lambda_2 \mathcal{L}_V, \tag{3}$$

where λ_1 and λ_2 are trade-off parameters. In experiments, we set the parameters λ_1 and λ_2 to 1 by default. We elaborate on the modules that lead to these losses in the following sections.

B. Region-Level Semantic Alignment Module

The region-level semantic alignment module enables the generator $\mathcal{G}$ to exploit the local information in the video and draw different sub-regions of the video conditioned on words that are most relevant to those sub-regions.

Specifically, the region-level module takes two features as inputs, one from text and the other from video. The text feature is the word feature matrix $\mathbf{e}$ extracted from a bi-directional LSTM (see Sec. III-A), and the video feature is extracted from a 3D convolution neural network (3D CNN). We extract the local feature matrix $\mathbf{U} \in \mathbb{R}^{\hat{D} \times N}$ (reshaped from $\hat{D} \times w \times h$) from one of the intermediate layers of this 3D CNN network. Each column of $\mathbf{U}$ is a feature, which represents a sub-region of the video. $\hat{D}$ is the dimension of the local features, and N is the number of sub-regions in a video. In order to measure the relevance of the words and sub-regions, we convert the video features into the common semantic space of the word embeddings $\mathbf{e}$ using a perceptual layer $\mathbf{W}_p$:

$$\mathbf{V} = \mathbf{W}_p^{\top}\mathbf{U}, \tag{4}$$

where $\mathbf{W}_p \in \mathbb{R}^{\hat{D} \times D}$, $\mathbf{V} \in \mathbb{R}^{D \times N}$ and its i^{th} column $\mathbf{v}_i$ is the converted feature in the common space for the i^{th} sub-region of the video, and D denotes the dimension of both text embeddings and video features.

To measure the alignment of the word-region pairs, we design a region-level loss $\mathcal{L}_R$ to optimise the generator $\mathcal{G}$. The loss $\mathcal{L}_R$ is able to calculate the similarity of all possible pairs between words and sub-regions of video. Firstly, to represent the relevance of words and sub-regions, we calculate the similarity matrix for all possible pairs of words in the text and sub-regions in the video by

$$\mathbf{S} = \mathbf{E}^{\top}\mathbf{V}, \tag{5}$$

where $\mathbf{S} \in \mathbb{R}^{T \times N}$ and $s_{i,j}$ is the similarity between the i^{th} word of the text and the j^{th} sub-region of the video. Second, we build an attentional model to compute a feature vector for each sub-region for each word. The vector $\mathbf{c}_i$ is a representation of the video's sub-regions, which is relative to the i^{th} word of the whole text. We compute $\mathbf{c}_i$ as the weighted sum over all features of sub-region. The function is designed as

$$\mathbf{c}_i = \sum_{j=1}^{N} \alpha_j \mathbf{v}_j, \text{ where } \alpha_j = \frac{\exp(s_{i,j}/\mu_1)}{\sum_{k=1}^{N} \exp(s_{i,k}/\mu_1)}, \tag{6}$$

where μ_1 is used to adjust the smooth of attention in different sub-regions. Then, we establish the relevance between the i^{th} word and the corresponding sub-region of video using the cosine similarity between $\mathbf{c}_i$ and $\mathbf{e}_i$. The function can be defined as

$$\Phi(\mathbf{c}_i, \mathbf{e}_i) = \frac{\mathbf{c}_i^{\top}\mathbf{e}_i}{\|\mathbf{c}_i\|\|\mathbf{e}_i\|}. \tag{7}$$

Motivated by [51] and [52], we design a function $R(\mathcal{V}, \mathcal{Q})$ to measure the magnitude of coherence between the entire video $\mathcal{V}$ and the whole text $\mathcal{Q}$, which is

$$R(\mathcal{V}, \mathcal{Q}) = \log\left(\sum_{i=1}^{T} \exp(\gamma\, \Phi(\mathbf{c}_i, \mathbf{e}_i))\right)^{\frac{1}{\gamma}}, \tag{8}$$

where γ is a factor that determines how much to magnify the significance of the most relevant word-region pair. When $\gamma \to \infty$, $R(\mathcal{V}, \mathcal{Q})$ approximates to $\max_{i=1}^{T} \Phi(\mathbf{c}_i, \mathbf{e}_i)$.

Finally, based on the posterior probability of text $\mathcal{Q}$ matching with video $\mathcal{V}$, we design a loss function as

$$\mathcal{L}_{r_1} = -\frac{1}{M}\sum_{i=1}^{M} \log P(\mathcal{V}_i|\mathcal{Q}_i), \tag{9}$$

where M denotes the size of mini-batch and $P(\mathcal{V}_i|\mathcal{Q}_i)$ can be defined as

$$P(\mathcal{V}_i|\mathcal{Q}_i) = \frac{\exp(R(\mathcal{V}_i, \mathcal{Q}_i)/\mu_2)}{\sum_{j=1}^{M} \exp(R(\mathcal{V}_i, \mathcal{Q}_j)/\mu_2)}, \tag{10}$$

where μ_2 is a hyper-parameter to determine the smooth magnitude. Likewise, we formulate another loss which is symmetrical to Eq. (9), *i.e.*,

$$\mathcal{L}_{r_2} = -\frac{1}{M}\sum_{i=1}^{M} \log P(\mathcal{Q}_i|\mathcal{V}_i), \tag{11}$$

where $P(\mathcal{Q}_i|\mathcal{V}_i) = \frac{\exp(R(\mathcal{Q}_i,\mathcal{V}_i)/\mu_2)}{\sum_{j=1}^{M}\exp(R(\mathcal{Q}_i,\mathcal{V}_j)/\mu_2)}$. Thus, the objective function of our region-level loss is

$$\mathcal{L}_R = \mathcal{L}_{r_1} + \mathcal{L}_{r_2}, \tag{12}$$

where $\mathcal{L}_{r_1}$ and $\mathcal{L}_{r_2}$ play the equivalent role. Notably, unlike the image-text matching in existing researches, such as [53]–[55], our region-level module aims to capture the related regions for each word. To this end, we measure the similarity between the sub-region of video and each word of sentence. In this way, the region-level module ensures the video generator to preserve the fine-grained semantic alignment between video and given descriptions.

C. Frame-Level Coherence-Aware Discriminator

To ensure the alignment between frames and given text and further enhance the realism of generated video, we introduce a discriminative module $\mathcal{D}_F$ at frame-level that enables the generator $\mathcal{G}$ to exploit both the quality of each frame and the coherence of each motion.

As shown in Fig. 2, we first use a shared 2D convolution model to extract the frame-level features $\mathbf{m}$ from each frame of the video. Unlike image generation, which focuses on the realism of single image/frame, we aim to capture both spacial and temporal information among frames. Thus, we propose two different convolution models to handle the frames and motions. In this way, the model $\mathcal{D}_F$ can be separated into two sub-modules: 1) a frame discriminator $\mathcal{D}_f$ to distinguish realistic frames from generated ones and 2) a motion discriminator $\mathcal{D}_m$ used to distinguish whether the motions across frames are generated or not.

The first $\mathcal{D}_f$ discriminates whether each frame of the input video is both real and semantically matched with the text script. Hence, the frame feature $\mathbf{m}$ is augmented by the global text embedding feature $\bar{\mathbf{e}}$ and sent to the discriminator $\mathcal{D}_f$. To align the generated frame with the conditioning information, the frame discriminator $\mathcal{D}_f$ must learn to evaluate whether samples from the generator $\mathcal{G}$ meet this conditioning constraint. However, in the original GAN, the discriminator only contains two kinds of inputs: real samples with matching text and synthetic samples with arbitrary text. This lacks the ability to semantically match the text and generated video. Thus, to ensure the alignment between frames and given text, inspired by [1], we design three kinds of training pairs including $\{\mathcal{X}, \bar{\mathbf{e}}\}$, $\{\mathcal{X}', \bar{\mathbf{e}}\}$ and $\{\widetilde{\mathcal{X}}, \bar{\mathbf{e}}\}$, where $\mathcal{X} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_{d_l}\}$ denotes the matched real video corresponding to the text vector $\bar{\mathbf{e}}$ while $\mathcal{X}' = \{\mathcal{F}'_1, \mathcal{F}'_2, \ldots, \mathcal{F}'_{d_l}\}$ is the mismatched real one. Likewise, $\widetilde{\mathcal{X}} = \{\widetilde{\mathcal{F}}_1, \widetilde{\mathcal{F}}_2, \ldots, \widetilde{\mathcal{F}}_{d_l}\}$ represent our synthetic video which is conditioned on $\bar{\mathbf{e}}$. And $\mathcal{F}_i$, $\mathcal{F}_i{}'$ and $\widetilde{\mathcal{F}}_i$ denote the i^{th} frame in video $\mathcal{X}$, $\mathcal{X}'$ and $\widetilde{\mathcal{X}}$, respectively. To optimise the discriminator $\mathcal{D}_f$, we define the following loss

$$\mathcal{L}_{\mathcal{D}_f} = -\frac{1}{3d_l}\Bigg[\sum_{i=1}^{d_l}\log\mathcal{D}_f\left(\mathcal{F}_i, \bar{\mathbf{e}}\right) + \sum_{i=1}^{d_l}\log\left(1-\mathcal{D}_f\left(\mathcal{F}_i{}', \bar{\mathbf{e}}\right)\right) + \sum_{i=1}^{d_l}\log(1-\mathcal{D}_f(\widetilde{\mathcal{F}}_i, \bar{\mathbf{e}}))\Bigg], \tag{13}$$

where d_l is the number of frames of the input video.

For the motion discriminator $\mathcal{D}_m$, to ensure the temporal coherence across frames, we calculate the difference between two consecutive frames according to the distance between their frame-level features since the high-level representations contain more semantic information. We define the distance as

$$\Delta\mathbf{m}_i = Dist(\mathbf{m}_i, \mathbf{m}_{i-1}) = \|\mathbf{m}_i - \mathbf{m}_{i-1}\|_1, \tag{14}$$

where $\mathbf{m}_i$ and $\mathbf{m}_{i-1}$ denote the extracted features from frame $\mathcal{F}_i$ and $\mathcal{F}_{i-1}$, respectively. $\|\cdot\|_1$ is the L1-norm. $\Delta\mathbf{m}_i$ is the difference between consecutive frames, which represents the magnitude of the motion between frames $\mathcal{F}_i$ and $\mathcal{F}_{i-1}$. To incorporate this information, we optimise this discriminator by minimising the loss

$$\mathcal{L}_{\mathcal{D}_m} = -\frac{1}{3(d_l-1)}\Bigg[\sum_{i=2}^{d_l}\log\mathcal{D}_m\left(\Delta\mathbf{m}_i, \bar{\mathbf{e}}\right) + \sum_{i=2}^{d_l}\log\left(1-\mathcal{D}_m\left(\Delta\mathbf{m}'_i, \bar{\mathbf{e}}\right)\right) + \sum_{i=2}^{d_l}\log\left(1-\mathcal{D}_m\left(\Delta\widetilde{\mathbf{m}}_i, \bar{\mathbf{e}}\right)\right)\Bigg], \tag{15}$$

where $\Delta\mathbf{m}_i$, $\Delta\mathbf{m}'_i$ and $\Delta\widetilde{\mathbf{m}}_i$ denote the motion features between the i^{th} and $(i\text{-}1)^{th}$ frames in video $\mathcal{X}$, $\mathcal{X}'$ and $\widetilde{\mathcal{X}}$, respectively.

Thus, we optimise the frame-level discriminative model $\mathcal{D}_F$ end-to-end by minimising the objective function, which can be defined as

$$\mathcal{L}_{\mathcal{D}_F} = \mathcal{L}_{\mathcal{D}_f} + \mathcal{L}_{\mathcal{D}_m}. \tag{16}$$

The corresponding adversarial loss for optimising $\mathcal{G}$ at frame-level is defined as

$$\mathcal{L}_F = -\frac{1}{3d_l}\sum_{i=1}^{d_l}\log(\mathcal{D}_f(\widetilde{\mathcal{F}}_i, \bar{\mathbf{e}})) - \frac{1}{3(d_l-1)}\sum_{i=2}^{d_l}\log\left(\mathcal{D}_m\left(\Delta\widetilde{\mathbf{m}}_i, \bar{\mathbf{e}}\right)\right). \tag{17}$$

D. Video-Level Semantic-Aware Discriminator

To generate a natural video containing strong semantic alignment with the given text, we propose a video-level discriminator $\mathcal{D}_V$ to distinguish between real and generated videos, which enhances the ability of the generator $\mathcal{G}$ by capturing the global information over the entire video.

Specifically, we implement the model $\mathcal{D}_V$ using a 3D convolutional neural network (3D CNN) to translate the input video into a global feature. Then, we augment the extracted feature with corresponding text embedding $\bar{\mathbf{e}}$ to match the conditioning text script, thus can verify whether the video is semantically matched with the given script.

To align the generated video with the conditioning information, the video discriminator $\mathcal{D}_V$ must learn to evaluate

Algorithm 1 Training Algorithm for BoGAN

Step 1:
Train region-level module by Eq. (12).
Step 2:
for $k = 1$ **to** K **do**
  Obtain random noise $\mathbf{z} \sim \mathcal{N}(0, 1)$.
  Get input $\mathbf{p}$ via combining $\mathbf{z}$ with text embedding $\bar{\mathbf{e}}$.
  Produce synthetic video $\widetilde{\mathcal{X}} = \mathcal{G}(\mathbf{p})$.
  Update discriminator $\mathcal{D}_F$ by minimising Eq. (16).
  Update discriminator $\mathcal{D}_V$ by minimising Eq. (18).
  Update generator $\mathcal{G}$ by minimising Eqs. (12), (17) and (19).
end for

whether samples from generator $\mathcal{G}$ meet this conditioning constraint. Hence, the loss for training $\mathcal{D}_V$ is defined as:

$$\mathcal{L}_{\mathcal{D}_V} = -\frac{1}{3}\big[\log \mathcal{D}_V\left(\mathcal{X}, \bar{\mathbf{e}}\right) + \log\left(1 - \mathcal{D}_V\left(\mathcal{X}', \bar{\mathbf{e}}\right)\right) + \log(1 - \mathcal{D}_V(\widetilde{\mathcal{X}}, \bar{\mathbf{e}}))\big], \quad (18)$$

where $\mathcal{X}$, $\mathcal{X}'$ and $\widetilde{\mathcal{X}}$ are real video, mismatched real video and our synthetic video, respectively. We define the corresponding adversarial loss for training the generator $\mathcal{G}$ as

$$\mathcal{L}_V = -\frac{1}{3}\log(\mathcal{D}_V(\widetilde{\mathcal{X}}, \bar{\mathbf{e}})). \quad (19)$$

E. Training Process

In this section, we introduce the training mechanism for the proposed BoGAN. The training performance of video-level and frame-level modules strongly depends on the quality of text and word embeddings extracted from region-level module. Thus, in order to effectively optimise the whole model, we first train the region-level module and try to obtain the satisfying text and word embedding. Specifically, as shown in Algorithm 1, to effectively train the whole model, we divide the training of the proposed model into two steps. In the first step, to capture the relevance of the words and sub-regions of video, we optimise our region-level multi-modal similarity module using Eq. (12). In the second step, we fix the parameters of the region-level model and train the rest architecture in an alternative manner. For the discriminators $\mathcal{D}_F$ and $\mathcal{D}_V$, we update their parameters by minimising Eq. (16) and Eq. (18), respectively. Meanwhile, for the generator $\mathcal{G}$, its parameters are adjusted by Eqs. (12), (17) and (19).

IV. Experiments

We evaluate and compare our proposed BoGAN model with several state-of-the-art approaches, on two synthetic text-to-video datasets (SBMG [6] and TBMG [6]) and two real-world datasets (MSVD [7] and KTH [24]), with both quantitative and qualitative evaluation metrics. A detailed ablation study is performed to test the contribution of each model component and a human study is performed to examine the reality, relevance and coherence of the generated videos. We finally evaluate the generalisation ability of our proposed model.

A. Datasets

Single-Digit Bouncing MNIST GIFs (SBMG) [6] is a synthetic dataset that has single handwritten digit bouncing inside a 64×64 frame. It is composed of $12,000$ GIFs and every GIF is 16 frames long, which contains a single 28×28 digit moving left-right or up-down. Each GIF is accompanied with single sentence describing the digit and its moving direction.

Two-Digit Bouncing MNIST GIFs (TBMG) [6] is an extended synthetic dataset of SBMG which contains two handwritten digits bouncing.

KTH Human Action Dataset [24] consists of 1200 videos with 25 persons performing 3 actions (walking, running and jogging). Each video has 16 frames with size 48×48 and the script describes the action and direction of a person, such as "person 8 is walking left-to-right" or "person 17 is running right-to-left".

MSVD Dataset [7] contains $1,970$ video snippets collected from YouTube. There are roughly 40 available English descriptions for each video. Since the combined visual and text quality and consistency is mixed, following [6], we manually filter out the videos about cooking and generate a subset of 421 cooking videos.

B. Experimental Settings

1) Implementation Details: For a fair comparison, following [6], we only focus on generating each video with size 48×48 and 16 frames, *i.e.*$d_l = 16$, $d_h = d_w = 48$. For sentence encoding, the dimension of the input, hidden layers, output in bi-LSTM are all set to 256, *i.e.*, $D = d_p = 256$. The dimension of random noise variable $\mathbf{z}$, *i.e.*, d_z is 100. All the weights are initialised from a normal distribution with zero-mean and standard deviation of 0.02. For the hyperparameters in Sec. III-B, we set $\gamma = 5.0$, $\mu_1 = 0.2$ and $\mu_2 = 0.1$. The slope of the leak in LeakyReLU is set to 0.2. In the training, we use Adam [56] with $\beta_1 = 0.9$ to update the model parameters. We set the mini-batch and learning rate to 64 and 0.0002, respectively.

2) Evaluation Metrics: For quantitative evaluation, FID [9] is used to evaluate the quality of each frames, while the FID2vid [10] is to measure both quality and temporal consistency of the whole videos. In general, the smaller these two values are, the better performance the method will be. To further evaluate the video generation model quantitatively, we adopt Generative Adversarial Metric (GAM) [57], which is able to directly compare two generative models by having them engage in a "battle" against each other.

C. Compared Methods

To evaluate the performance of our proposed method, several state-of-the-art models are adopted for comparison, including SyncDRAW [24], VGAN [11], GAN-CLS [1], Cap2vid [25], TGANs-C [6], T2V [5] and MoCoGAN [13]. Since the original VGAN and MoCoGAN attempt to generate videos in an unconditional manner, for a fair comparison, here we additionally incorporate the matching-aware loss into the discriminator of basic VGAN and MoCoGAN,

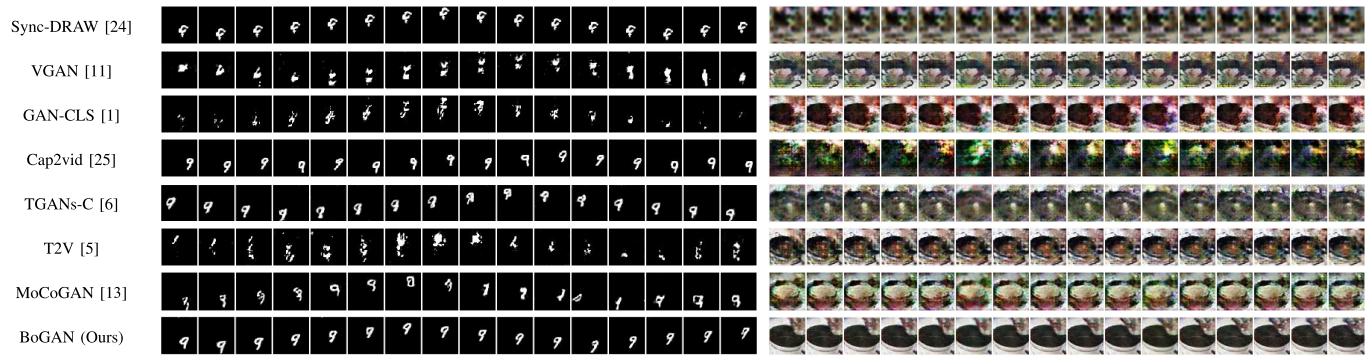


Fig. 3. The experimental results of different methods on the SBMG dataset for the caption "Digit 9 is going up and down", and on the MSVD dataset for the caption "Someone is pouring water into a bowl".

TABLE I

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART ALGORITHMS (SMALLER IS BETTER)

Methods	SBMG		TBMG		KTH		MSVD	
	FID	FID2vid	FID	FID2vid	FID	FID2vid	FID	FID2vid
Sync-DRAW [24]	69.75	4.54	101.87	5.26	58.85	5.72	268.76	16.45
VGAN [11]	170.31	6.59	168.64	5.97	91.86	5.01	105.57	12.53
GAN-CLS [1]	252.74	4.59	247.28	6.19	127.6	7.50	72.74	10.53
Cap2vid [25]	**40.38**	3.13	53.06	5.22	78.09	6.53	186.69	12.51
TGANs-C [6]	63.05	4.84	57.59	5.36	60.05	4.09	149.12	13.68
T2V [5]	130.24	4.81	153.61	6.91	77.17	5.90	176.89	13.40
MoCoGAN [13]	89.14	4.66	95.01	5.54	48.53	4.43	80.29	11.89
BoGAN (Ours)	47.57	**3.12**	**48.31**	**4.22**	**29.84**	**3.63**	**57.01**	**9.38**

and enable them to generate videos conditioning on captions. Likewise, the original GAN-CLS model focuses on image generation from text. For video generation, we directly extend the architecture by replacing 2D convolutions with 3D convolutions.

D. Quantitative Evaluation

1) Comparison via FID and FID2vid: To evaluate the performance of our method, we compute the quantitative evaluation metrics FID and FID2vid on both synthetic and real-world datasets. Shown in Tab. I, our proposed method achieves state-of-the-art performance on the TBMG, KTH, and MSVD datasets, and obtains highly comparable results on the SBMG dataset compared to the best baseline method. That means BoGAN is able to produce natural videos and photo-realistic frames on both synthetic and real-world datasets. Especially on the real video datasets, our model outperforms the previous state-of-the-art methods in a large margin. The results also demonstrate that our method is able to consider not only spatial and temporal information, but also the global and local information very well.

2) Comparison via GAM: To further evaluate the video generation model quantitatively, we adopt GAM [57] to "battle" against with GAN-CLS and TGANs-C in the MSVD dataset. From Tab. II, our BoGAN is able to beat the best performing models.

TABLE II

MODEL EVALUATION WITH THE GAM METRIC ON THE MSVD DATASET. WHEN $r_{test} \simeq 1, r_{sample} < 1$ MEANS THE FORMER BEATS THE LATTER

Battler	r_{test}	r_{sample}	**Winner**
BoGAN vs GAN-CLS [1]	1.07	0.68	BoGAN
BoGAN vs TGANs-C [6]	0.99	0.82	BoGAN

E. Qualitative Evaluation

Fig. 3 depicts example results selected from two datasets, a synthetic one and a real one. Results produced by other state-of-the-art methods are also displayed for comparison. The results demonstrate that our BoGAN is able to generate the whole video semantically matching with the language description while keeping reality and coherence in frame-level. For the synthetic data, although Cap2vid obtains the best FID score on the SBMG dataset, it cannot capture the coherence between each frames so that the visual results of this method seem chaotic across the sequence of generated videos. The results of TGANs-C and Sync-DRAW seem well-organised in temporal, but the generated digit in each frame is distorted. For the photo-realistic examples, VGAN, TGANs-C and MoCoGAN obtain some plausible but blurry results conditioning with the input description. Although the generated video via GAN-CLS seems realistic, it contains many noises in each frame and losses the movement across frames. Instead, our model is able to generate natural and photo-realistic frames with fine-grained details.

F. Ablation Study

1) Quantitative Results: To investigate the effect of each part in our proposed method, we conduct an ablation study to compare the performance by removing some components,

TABLE III

THE EFFECT OF DIFFERENT COMPONENTS OF THE PROPOSED METHOD, EVALUATED ON SBMG AND KTH

Methods	SBMG		KTH	
	FID	FID2vid	FID	FID2vid
Region-Level $\mathcal{L}_R$	332.63	20.68	270.60	24.37
Frame-Level $\mathcal{L}_F$	83.83	5.05	60.25	5.26
Video-Level $\mathcal{L}_V$	118.55	5.54	55.75	5.58
Region $\mathcal{L}_R$ + Frame $\mathcal{L}_F$	80.01	4.61	53.06	4.80
Region $\mathcal{L}_R$ + Video $\mathcal{L}_V$	106.15	4.80	40.20	4.40
Frame $\mathcal{L}_F$ + Video $\mathcal{L}_V$	63.05	4.84	35.64	4.18
$\mathcal{L}_R + \mathcal{L}_F + \mathcal{L}_V$ (our final)	**47.57**	**3.12**	**29.84**	**3.63**

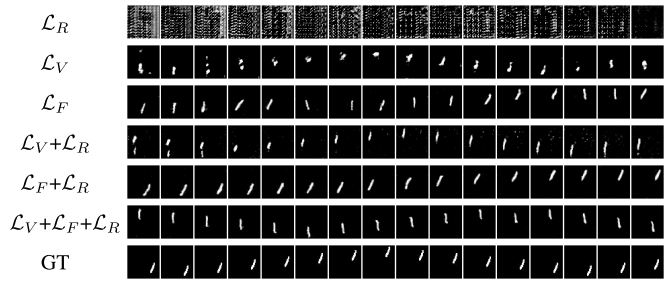


Fig. 4. Visual results for each component on SBMG, conditioned on "Digit 1 is going up and down".

Fig. 5. The results for caption "Digit 3 and 1 is going up and down." generated by our BoGAN training on TBMG.

on the SBMG and KTH datasets. The quantitative results are shown in Tab. III. The model with region-level loss ($\mathcal{L}_R$) only performs worst among all the variant models. This is acceptable because neither frame-level coherence nor video-level semantic matching is considered in the method. Moreover, this method does not adopt an adversarial learning procedure, which is important for producing realistic results. Compare to $\mathcal{L}_R$, $\mathcal{L}_F$ and $\mathcal{L}_V$ perform better, which demonstrates the effects of frame coherence and video semantics. However, compared to $\mathcal{L}_F$ and $\mathcal{L}_V$ individually, we find that incorporating the region-level loss $\mathcal{L}_R$ can obtain better performance, which proves that local region-level information is significant in video generation. Finally, we use all the three losses into our objective model, which has the best results.

2) Qualitative Results: Moreover, we exhibit the qualitative results corresponding to each component in Fig. 4. To be specific, only using $\mathcal{L}_V$, the results can well exploit the semantics (*i.e.*, motion) from the given text, but lack the

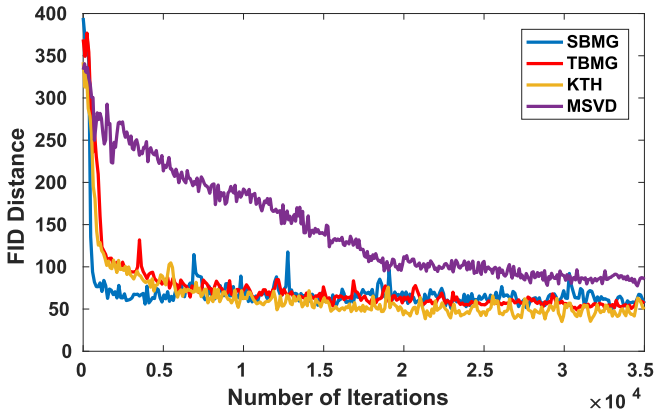


Fig. 6. Convergence performance of BoGAN training on the different datasets.

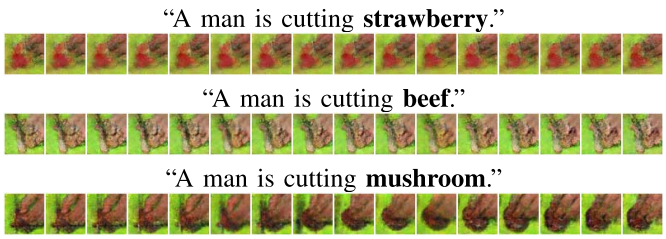
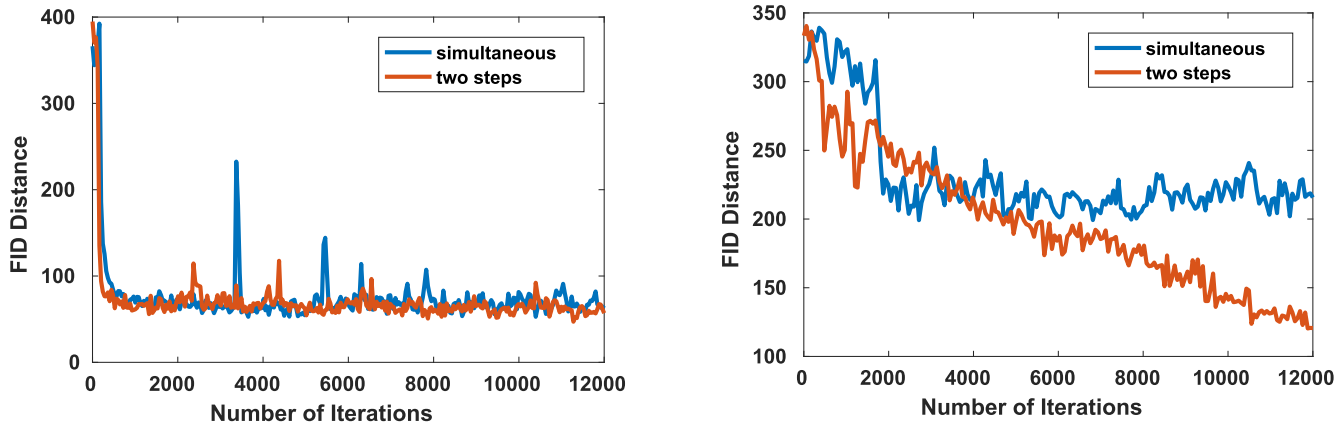


Fig. 7. Example results of proposed BoGAN trained on MSVD dataset while changing some significant words in the script.

detailed information in each frame. Conversely, when using $\mathcal{L}_F$, the generated video obtains more promising visual performance but a poor movement. Notably, if we only use $\mathcal{L}_R$, the synthesised images are meaningless since the details can not be guaranteed when the global semantics and realism are uncertain. Moreover, when combining temporal and spatial information (*i.e.*, $\mathcal{L}_V+\mathcal{L}_R$), the generated video matches better with the given script. Finally, by using all the losses, we obtain the best visual results with promising temporal and spatial information.

G. Human Study

Since FID and FID2vid only focus on measuring the realism of the generated videos while ignoring the semantic alignment between generated videos and descriptions, following [6], we further conduct a human evaluation to compare our method against other state-of-the-art methods. We randomly select 100 generated videos with corresponding descriptions from the MSVD dataset and asked 30 human subjects (university students) to score them. Evaluators measure the generated videos with respect to three criteria: (1) Realism: the realism of generated videos; (2) Relevance: the relevance between generated videos and corresponding description; (3) Coherence: the temporal coherence across consecutive frames. Each criterion contains 10 rankings from 1 to 10 (bad to good). For an objective annotation, each generated video must be scoredwe by three evaluators at least. Then, we average the ranking on each criterion of all the synthetic videos generated by each method and obtain three metrics. Tab. IV shows the results of the human study on the MSVD dataset. Our BoGAN



(a) Results on SBMG (b) Results on MSVD

Fig. 8. The FID distance of different training process on the synthetic (*i.e.*, SBMG) and real-world (*i.e.*, MSVD) datasets.

TABLE IV

THE AVERAGE RANKING (HIGHER IS BETTER) ON EACH CRITERION OF ALL THE GENERATED VIDEOS BY EACH APPROACH ON MSVD

Methods	Realism	Relevance	Coherence
Sync-DRAW [24]	3.47	3.48	4.17
VGAN [11]	4.53	4.54	4.98
GAN-CLS [1]	5.69	5.80	6.06
Cap2vid [25]	4.40	4.59	4.42
TGANs-C [6]	4.87	4.97	5.35
T2V [5]	4.21	4.33	4.78
BoGAN (Ours)	**7.52**	**7.76**	**8.52**

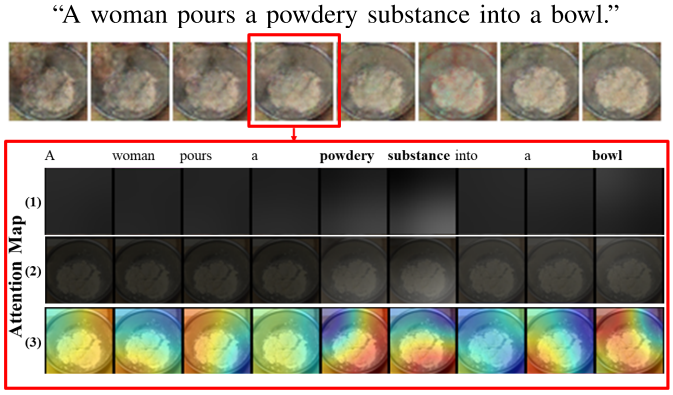


Fig. 9. Relation between each word and its sub-region. We take one of the generated videos on MSVD as example and rank the words by similarity between words and sub-regions of one frame. We highlight the top three words in bold type.

consistently achieves the best performance across all the three criteria.

Large-Scale Human Study: To compare the visual results of our method and the baselines, we conducted a large-scale user study on the Internet. To this end, we randomly select 10 text-video pairs for each method and hence obtain 70 pairs in total. The participants evaluate each sample in terms of three criteria, *i.e.*, realism, relevance, and coherence. Each video snippet is scored by at least 10 participants. In practice, we obtain 3710 results in total, which means

Fig. 10. Failure case conditioned on "Someone scrubbing a zucchini with brush under the running water from a faucet a sink".

TABLE V

THE RESULTS OF LARGE-SCALE HUMAN STUDY. WE REPORT THE AVERAGE RANKING (HIGHER IS BETTER) ON EACH CRITERION OF ALL THE GENERATED VIDEOS BY EACH APPROACH ON MSVD. BESIDES, WE ALSO PROVIDE THE AVERAGE RANKING OF ALL THE CRITERIA

Methods	Realism	Relevance	Coherence	Average
Sync-DRAW	4.24	4.21	4.68	4.38
VGAN	4.96	4.95	5.25	5.05
GAN-CLS	5.86	5.86	6.08	5.93
Cap2vid	4.86	5.00	5.20	5.02
TGANs-C	5.21	5.27	5.56	5.35
T2V	4.74	4.88	5.11	4.91
BoGAN (Ours)	**6.37**	**6.39**	**6.31**	**6.35**

that each video snippet is scored by $\sim$ 53 participants on average. As shown in Table V, our BoGAN achieves the best performance compared with the baseline methods in all three criteria. It demonstrates that our proposed method has the ability to generate more photo-realistic videos from the given descriptions.

H. Convergence and Generalisation Analysis

Since our BoGAN involves a complex adversarial learning procedure, we investigate the convergence performance. Following [6], we show the evolution of the generator network $\mathcal{G}$ at the training stage to illustrate the convergence. Fig. 5 shows the visual quality of our generated video improves as the iteration increases. In addition, we explore the convergence of our proposed BoGAN in FID distance on several benchmark datasets. Fig. 6 shows that although our BoGAN involves a complex adversarial learning procedure, it is still able to converge on both the synthetic (*i.e.*, SBMG and TBMG)

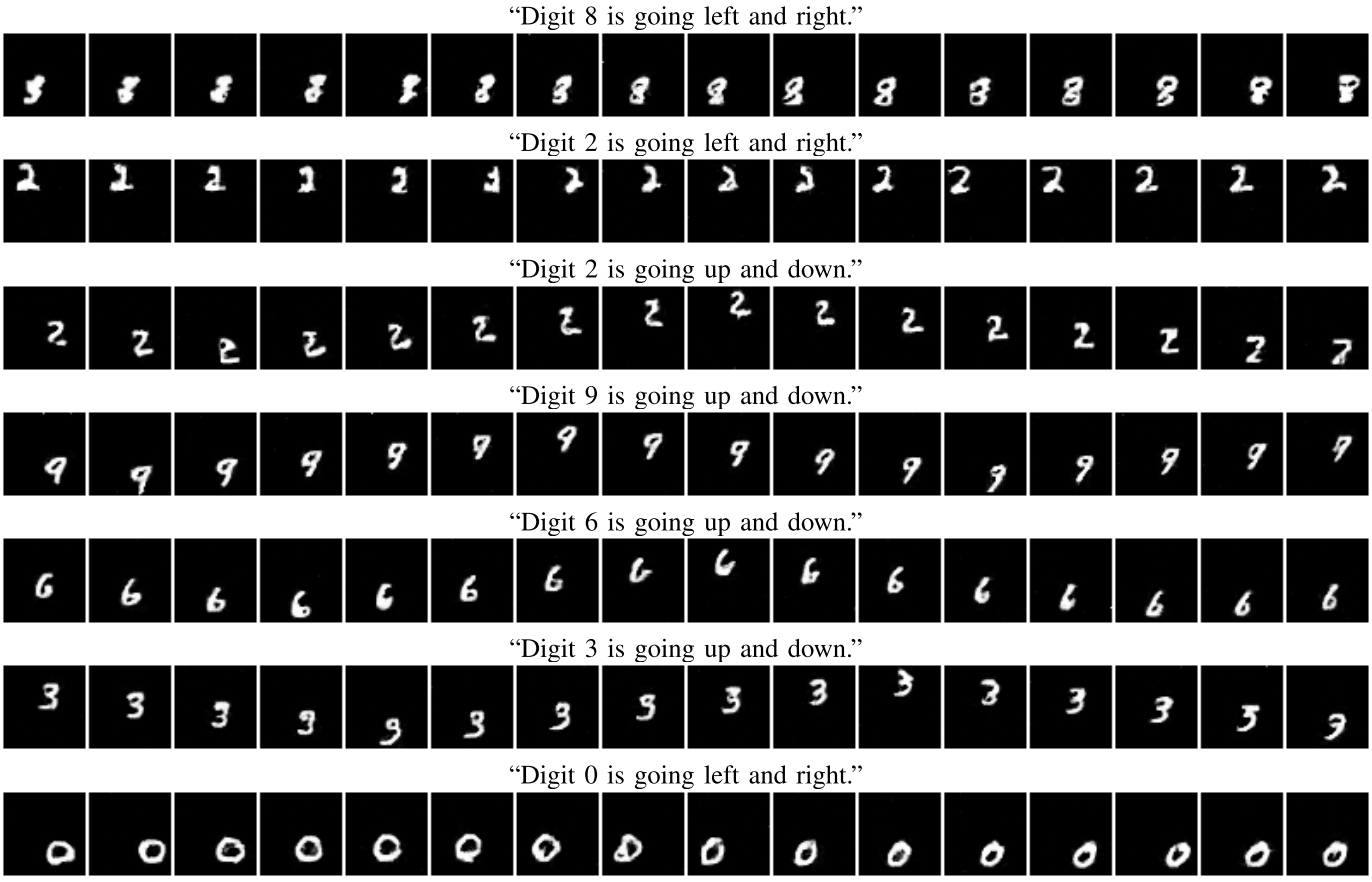


Fig. 11. The generated samples of BoGAN on the SBMG dataset.

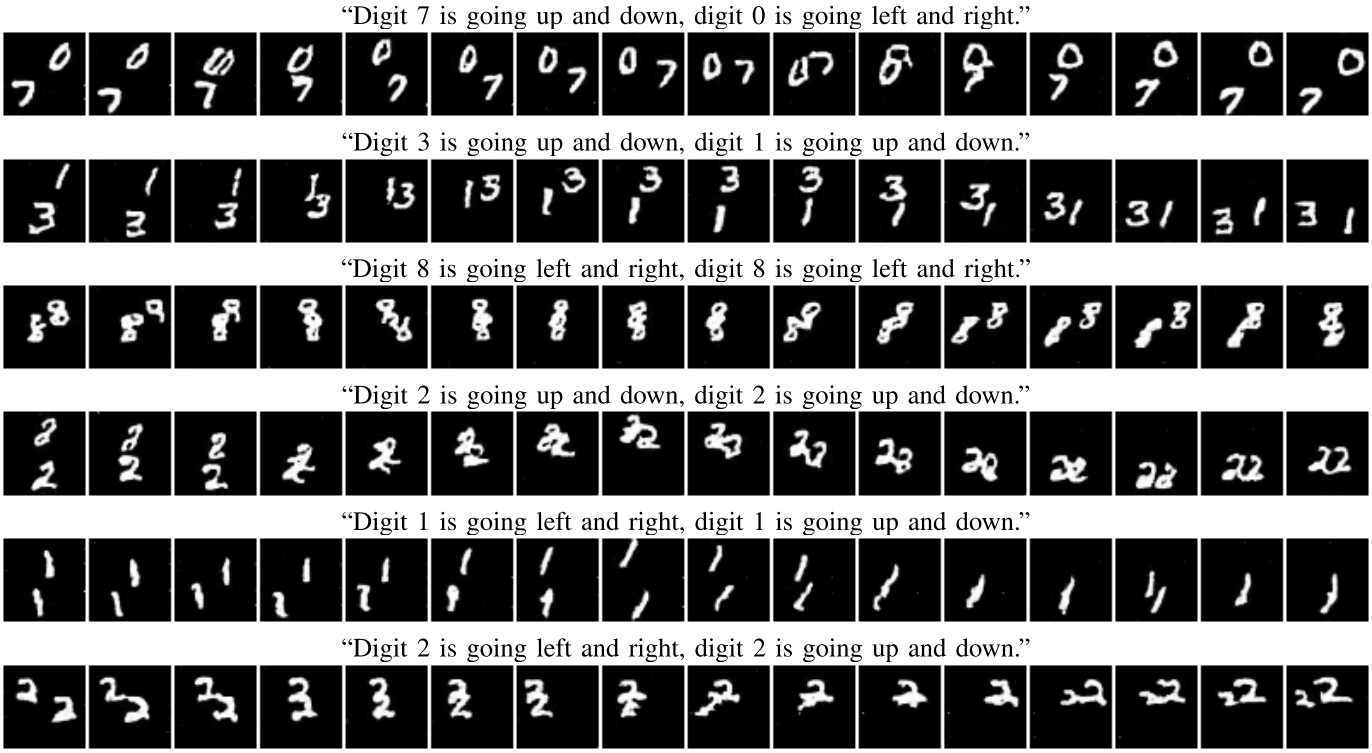


Fig. 12. The generated samples of BoGAN on the TBMG dataset.

and real-world datasets (*i.e.*, KTH and MSVD). Moreover, the performance of FID converges more quickly on simple datasets (*e.g.*, SBMG, TBMG or KTH) than on complex datasets (*e.g.*, MSVD).

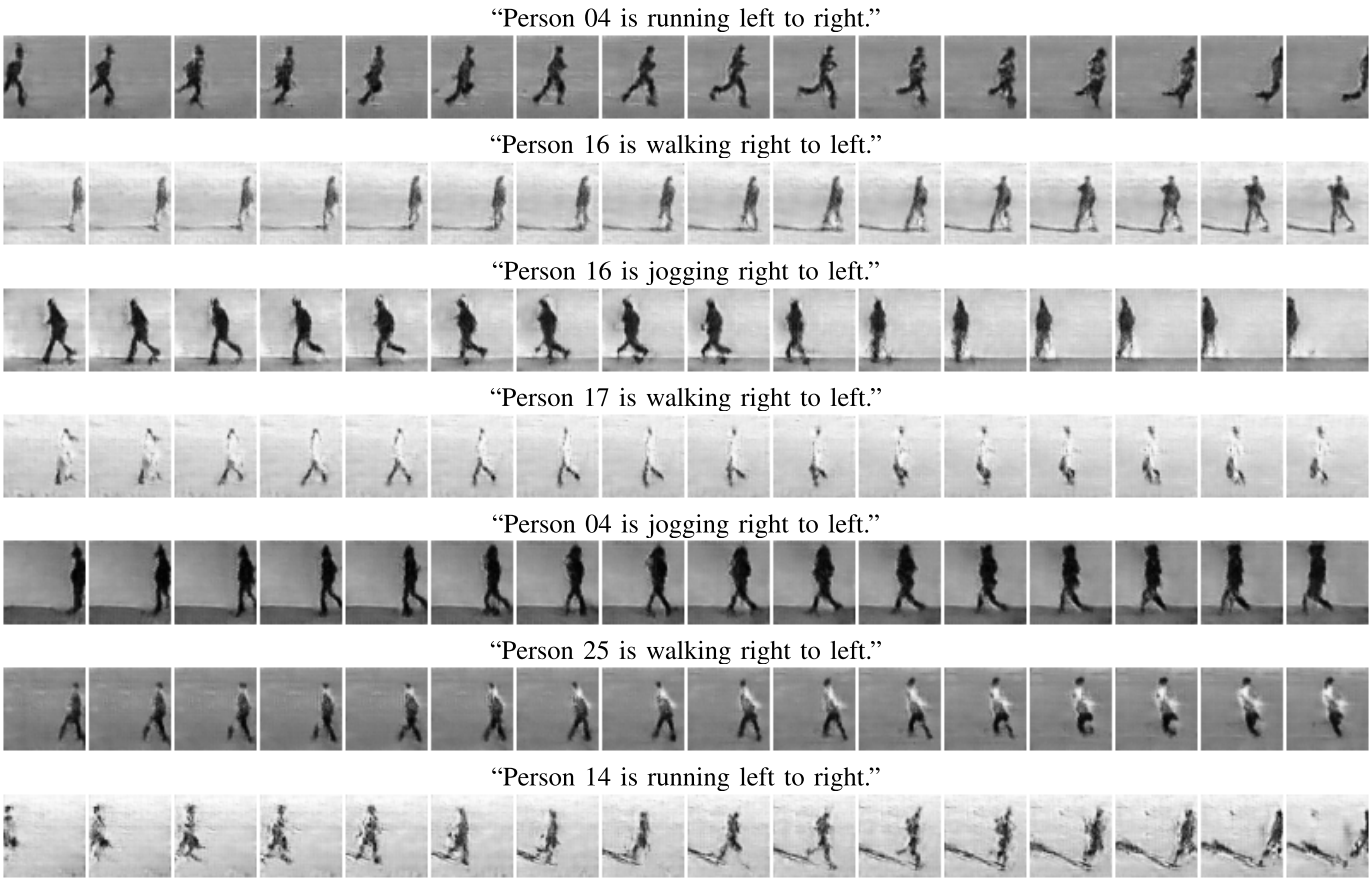


Fig. 13. The generated samples of BoGAN on the KTH dataset.

TABLE VI
RESULTS OF BOGAN WITH DIFFERENT RESOLUTIONS ON MSVD

Video Resolution	48×48	64×64	128×128
FID	57.01	74.55	118.95
FID2vid	9.38	9.85	12.08

TABLE VII
DISCUSSION OF HYPER-PARAMETERS μ_1, μ_2 AND γ ON MSVD

$\mu_2 = 0.1, \gamma = 5.0$	μ_1	0.002	0.02	0.2	2	20
	FID	165.63	169.16	**57.01**	168.96	149.38
	FID2vid	14.59	13.72	**9.38**	14.15	13.73
$\mu_1 = 0.2, \gamma = 5.0$	μ_2	0.001	0.01	0.1	1	10
	FID	185.09	180.46	**57.01**	187.04	163.09
	FID2vid	14.91	13.79	**9.38**	14.45	16.98
$\mu_1 = 0.2, \mu_2 = 0.1$	γ	0.05	0.5	5	50	500
	FID	256.13	217.42	**57.01**	99.67	/
	FID2vid	16.14	14.96	**9.38**	11.80	/

TABLE VIII
DISCUSSION OF HYPER-PARAMETERS λ_1 AND λ_2 ON MSVD

$\lambda_2 = 1$	λ_1	0.001	0.01	0.1	1	10	100
	FID	102.45	184.12	147.39	**57.01**	88.87	173.31
	FID2vid	12.33	14.87	14.46	**9.38**	11.39	16.45
$\lambda_1 = 1$	λ_2	0.001	0.01	0.1	1	10	100
	FID	67.22	**51.54**	58.68	57.01	181.82	194.85
	FID2vid	10.31	**9.22**	9.34	9.38	15.27	17.16

To verify the generalisation ability of our proposed method, we test how sensitive the generated videos are to the unseen input sentences by replacing some significant words in the input descriptions. The generated videos are shown in Fig. 7. When we change some words in the input descriptions, the relevant sub-regions of the generated videos are changed accordingly while other parts are constant. This indicates that our proposed model is able to catch subtle semantic difference of the text description when generating videos.

I. More Discussions

1) Results on Higher-Resolution Videos: Moreover, we extend BoGAN to generate higher-resolution videos (64×64 and 128×128) on MSVD. From Table VI, our model is able to obtain competitive results.

2) Analysis for Hyper-Parameter Settings: From Tab. VII, when changing hyper-parameter values, the performances become worse since the attention modules are impacted by unsuitable values of hyper-parameters μ_1, μ_2 and γ (too small or too large). From Tab. VIII, compared with $\mathcal{L}_F$ (λ_1), the results are more sensitive to $\mathcal{L}_V$ (λ_2).

3) Effectiveness of Two-Steps Training Process: To evaluate the effectiveness of the two-steps training process, we conduct an experiment to compare the performances of generative models when using different training methods, *i.e.,* two-steps

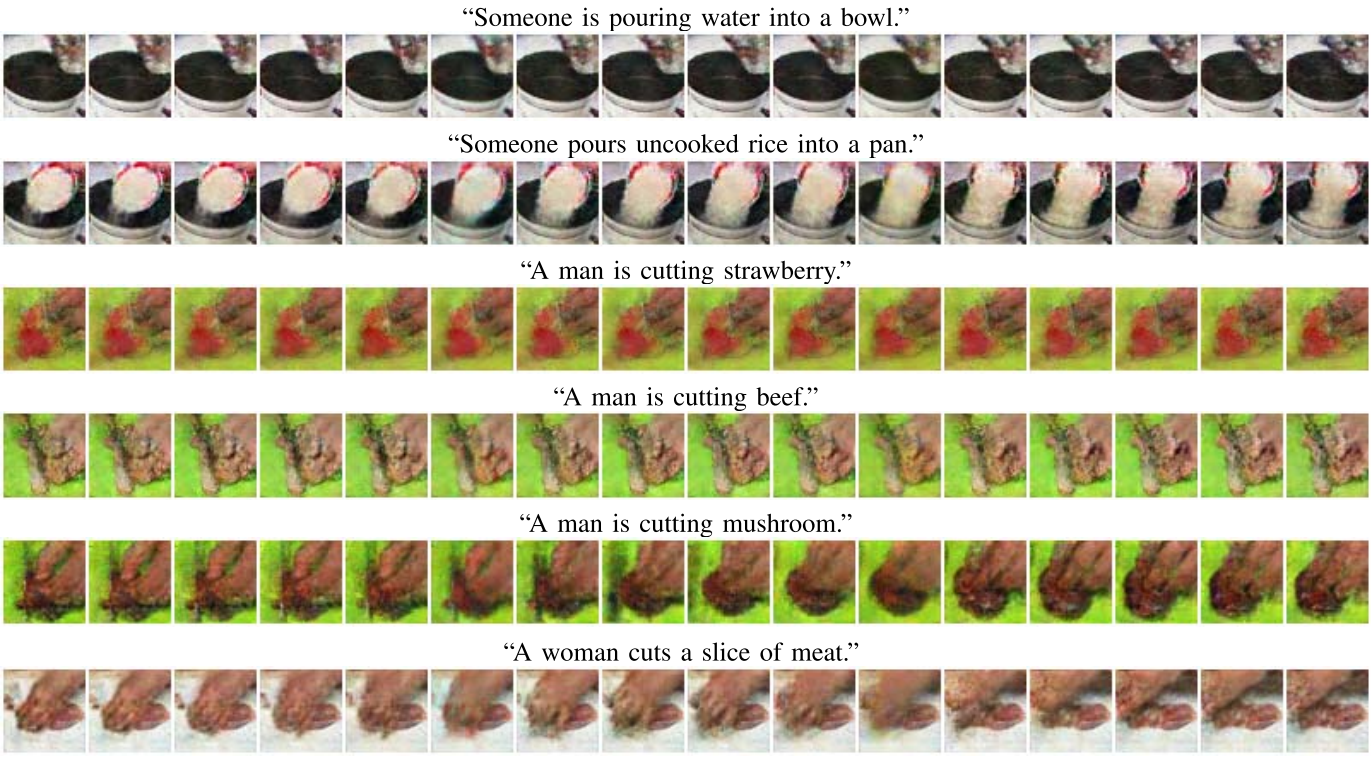


Fig. 14. The generated samples of BoGAN on the MSVD dataset.

and one-step (simultaneous). As shown in Figure 8, the performance improves more quickly when using the two-steps training process on both synthetic (*i.e.*, SBMG) and real-world (*i.e.,* MSVD) datasets. Especially on the more complex dataset MSVD (Figure 8 (b)), our two-steps method obtains a more impressing performance compared with the simultaneous approach, which demonstrates the effectiveness of our proposed training method.

4) Visual Results for Region-Level Module: To further demonstrate the effectiveness of region-level module, we provide some visual results in Fig. 9 to visualise the relation between each word and its relevant sub-region. Specifically, given a description "A woman pours a powdery substance into a bowl", we use BoGAN to generate a video containing a serial of frames. Then, we randomly choose a frame and visualise the corresponding attention maps (see Fig. 9 (1)). For better visualisation, we cover the selected frame by the produced attention maps separately (see Fig. 9 (2)). Besides, we also visualise the feature maps with the pseudocolor to make it more clear (see Fig. 9 (3)). Based on the visual results, we rank the words by the similarity between words and sub-regions in the selected frame. We highlight the top three words in bold type. The results show that our region-level module has the ability to enforce the generator to focus on the most significant word and the corresponding sub-region.

5) Failure Cases: From Fig. 10, given a complex and long sentence with multiple entities, our model may not work well, since the semantic information is hard to be captured.

6) More Qualitative Results: We present more generated intact samples for qualitative evaluation. More results of our BoGAN on SBMG, TBMG, KTH and MSVD are shown in Figs. 11, 12, 13 and 14, respectively.

V. Conclusion

Video generation from text is challenging due to the intrinsic complexity. In this paper, we have proposed a novel Bottom-Up Generative Adversarial Network (BoGAN) to ensure the realism of the generated video and achieve the required multi-scale semantic alignment. Specifically, to ensure the coherence between generated frames and the semantic match between a video and a language description, we have devised a bottom-up optimisation mechanism that includes three levels, from local to global. The proposed method outperforms its competitors on the benchmarks, which demonstrates the power of the architecture we have described. In terms of human study, our proposed method also performs better than the competitors, which is far more indicative of the value of our approach.

References

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: http://arxiv.org/abs/1605.05396

[2] T. Xu *et al.*, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[3] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5907–5915.

[4] H. Zhang *et al.*, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," 2017, *arXiv:1710.10916*. [Online]. Available: http://arxiv.org/abs/1710.10916

[5] Y. Li, M. R. Min, D. Shen, D. Carlson, and L. Carin, "Video generation from text," in *Proc. Conf. AAAI*, 2018, pp. 7065–7072.

[6] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei, "To create what you tell: Generating videos from captions," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1789–1798.

[7] S. Guadarrama *et al.*, "YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2712–2719.

[8] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, 2004, pp. 32–36.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[10] T.-C. Wang *et al.*, "Video-to-video synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–14.

[11] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 613–621.

[12] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2830–2839.

[13] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.

[14] Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng, "Forecasting human dynamics from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 548–556.

[15] B. Chen, W. Wang, and J. Wang, "Video imagination from a single image with transformation generation," in *Proc. Thematic Workshops ACM Multimedia-Thematic Workshops*. New York, NY, USA: ACM, 2017, pp. 358–366.

[16] V. Vukotić, S.-L. Pintea, C. Raymond, G. Gravier, and J. C. Van Gemert, "One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network," in *Proc. Int. Conf. Image Anal. Process.* Springer, 2017, pp. 140–151.

[17] J. Walker, C. Doersch, A. Gupta, and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders," in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 835–851.

[18] T. Xue, J. Wu, K. Bouman, and B. Freeman, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 91–99.

[19] X. Jia, B. D. Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 667–675.

[20] N. Kalchbrenner *et al.*, "Video pixel networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–16.

[21] J. van Amersfoort, A. Kannan, M. Ranzato, A. Szlam, D. Tran, and S. Chintala, "Transformation-based models of video sequences," 2017, *arXiv:1701.08435*. [Online]. Available: http://arxiv.org/abs/1701.08435

[22] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–22.

[23] W. Xiong, W. Luo, L. Ma, W. Liu, and J. Luo, "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2364–2373.

[24] G. Mittal, T. Marwah, and V. N. Balasubramanian, "Sync-DRAW: Automatic video generation using deep recurrent attentive architectures," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1096–1104.

[25] T. Marwah, G. Mittal, and V. N. Balasubramanian, "Attentive semantic video generation using captions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1426–1434.

[26] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[27] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[28] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–35.

[29] J. Cao, Y. Guo, Q. Wu, C. Shen, and M. Tan, "Adversarial learning with local coordinate coding," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–14.

[30] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[31] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," 2019, *arXiv:1903.11250*. [Online]. Available: https://arxiv.org/abs/1903.11250

[32] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.

[33] J. Li, J. Jia, and D. Xu, "Unsupervised representation learning of image-based plant disease with deep convolutional generative adversarial networks," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9159–9163.

[34] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "Improved ArtGAN for conditional synthesis of natural image and artwork," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 394–409, Jan. 2019.

[35] B. Dolhansky and C. C. Ferrer, "Eye in-painting with exemplar generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7902–7911.

[36] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6721–6729.

[37] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[38] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-autoencoders for face de-occlusion in the wild," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 778–790, Feb. 2018.

[39] Y. Guo *et al.*, "Dual reconstruction nets for image super-resolution with gradient sensitive loss," 2018, *arXiv:1809.07099*. [Online]. Available: http://arxiv.org/abs/1809.07099

[40] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[41] A. Lucas, S. Lopez-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3312–3327, Jul. 2019.

[42] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further improving enhanced super-resolution generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3637–3641.

[43] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

[44] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 291–301, Jan. 2019.

[45] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[46] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4030–4038.

[47] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5541–5550.

[48] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[49] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[50] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.

[51] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, Sep. 2008.

[52] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.

[53] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 201–216.
[54] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
[55] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 290–298.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[57] D. Jiwoong Im, C. Dongjoo Kim, H. Jiang, and R. Memisevic, "Generating images with recurrent adversarial networks," 2016, *arXiv:1602.05110*. [Online]. Available: http://arxiv.org/abs/1602.05110

Qi Chen received the bachelor's degree in software engineering from the School of Software Engineering, South China University of Technology, Guangzhou, China, in 2017, where he is currently pursuing the master's degree. His research interests include deep learning and computer vision.

Qi Wu received the M.Sc. and Ph.D. degrees in computer science from the University of Bath, U.K., in 2011 and 2015, respectively. He is currently a Lecturer (Assistant Professor) with The University of Adelaide, where he is also an Associate Investigator with the Australia Centre for Robotic Vision (ACRV). He is also the ARC Discovery Early Career Researcher Award (DECRA) Fellow from 2019 to 2021. His educational background is primarily in computer science and mathematics. He works on the vision and language problems, including image captioning, visual question answering, and visual dialog. His work has been published in prestigious journals and conferences, such as TPAMI, CVPR, ICCV, AAAI, and ECCV.

Jian Chen (Member, IEEE) received the B.S. and Ph.D. degrees in computer science from Sun Yat-sen University, Guangzhou, in 2000 and 2005, respectively. She joined the School of Software Engineering (SSE), South China University of Technology, as a Faculty Member, in 2005, where she is currently a Professor. Her research interests include developing effective and efficient data analysis techniques for complex data and the related applications. She is also interested in various techniques of data mining, Web search, information retrieval, and recommendation techniques as well as their applications.

Qingyao Wu (Member, IEEE) received the Ph.D. degree in computer science from the Harbin Institute of Technology in 2013. He was a Postdoctoral Research Fellow with the School of Computer Engineering, Nanyang Technological University, Singapore, from 2014 to 2015. He is currently an Associate Professor with the School of Software Engineering, South China University of Technology, China. His current research interests include machine learning, data mining, big data research, and bioinformatics.

Anton van den Hengel (Member, IEEE) received the Bachelor of Mathematical Science degree, the Bachelor of Laws degree, the master's degree in computer science, and the Ph.D. degree in computer vision from The University of Adelaide, in 1991, 1993, 1994, and 2000, respectively. He is currently a Professor with The University of Adelaide, where he is also the Founding Director of The Australian Centre for Visual Technologies (ACVT).

Mingkui Tan (Member, IEEE) received the bachelor's degree in environmental science and engineering and the master's degree in control science and engineering from Hunan University, Changsha, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2014. From 2014 to 2016, he has worked as a Senior Research Associate on computer vision with the School of Computer Science, The University of Adelaide, Australia. He is currently a Professor with the School of Software Engineering, South China University of Technology. His research interests include machine learning, sparse analysis, deep learning, and large-scale optimization.