Cross-Modal Dual Learning for Sentence-to-Video Generation

Yue Liu*
liuyuethuer@gmail.com
Tsinghua University

Xin Wang†
xin_wang@tsinghua.edu.cn
Tsinghua University

Yitian Yuan
yyt18@mails.tsinghua.edu.cn
Tsinghua University

Wenwu Zhu†
wwzhu@tsinghua.edu.cn
Tsinghua University

ABSTRACT

Automatic content generation has become an attractive while challenging topic in the past decade. Generating videos from sentences particularly poses great challenges to multimedia community due to its multi-modal characteristics in essence, e.g., difficulties in semantic alignment, and the temporal dependencies in video contents. Existing works resort to Variational Auto-Encoders (VAEs) or Generative Adversarial Networks (GANs) for generating videos given sentences, which may suffer from either blurry generated video frames or unstable training processes as well as difficulties in converging to optimal solutions. In this paper, we propose a cross-modal dual learning (CMDL) algorithm to tackle the challenges in sentence-to-video generation and address the weaknesses in existing works. The proposed CMDL model adopts a dual learning mechanism to simultaneously learn the bidirectional mappings between sentences and videos such that it is able to generate realistic videos which maintain semantic consistencies with their corresponding textual descriptions. By further capturing both global and local video structures, CMDL employs a multi-scale sentence-to-visual encoder to produce sequentially consistent and visually plausible videos. Extensive experiments on various datasets validate the advantages of our proposed CMDL model against several state-of-the-art benchmarks both qualitatively and quantitatively.

KEYWORDS

Video Generation, Dual Learning, Multi-modal Understanding

ACM Reference Format:
Yue Liu, Xin Wang, Yitian Yuan, and Wenwu Zhu. 2019. Cross-Modal Dual Learning for Sentence-to-Video Generation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350986

1 INTRODUCTION

Automatically generating visual contents has been widely studied as a fundamental problem over the past years. These visual contents range from images to videos which may contain rich and complex information, causing great difficulties in quantifying the various contents and then generating high-quality results. There have been some works utilizing Generative Adversarial Networks (GANs) [7] to generate images [23] or videos [33] from Gaussian noise without any specific assumptions, which may run the risk of producing unpredictable and uncontrollable visual contents. Given that many videos now come with textual descriptions such as sentences or captions, generating videos from given sentences is becoming a timely and promising problem in real-world scenarios. In this paper, we aim to generate a realistic video given a descriptive sentence such that the generated video and the sentence describing it have the same semantic meaning.

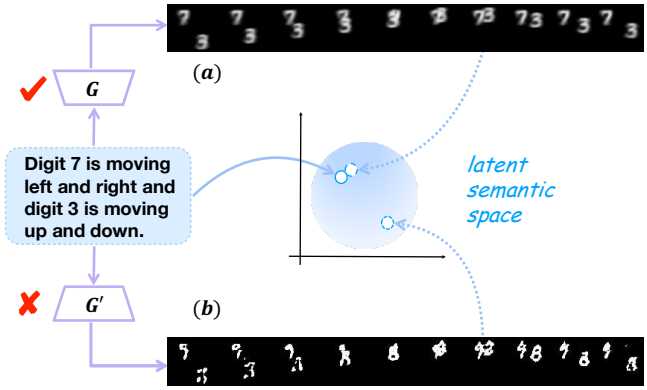

Figure 1: Illustration of the dual architecture in our CMDL model. The generated videos are re-embedded to a latent semantic space, where the paired videos and sentences should have similar embeddings. (a) A semantically consistent example generated by our CMDL model. (b) A semantically inconsistent example generated by one of the baselines GAN-CLS [24].

The task of sentence-to-video generation remains a significant problem to multimedia community. Although there are many works on sentence-to-image generation [36, 39, 40] and video captioning [6, 11, 32], little research has been done on generating videos from sentences. One challenging issue in sentence-to-video generation is maintaining the semantic consistency between the given sentence and the generated video. Different from generating videos from Gaussian noises, sentence-to-video generation seeks for a model which is capable of producing videos semantically aligned with the given textual descriptions. Another challenge is to keep the temporal coherence across video frames. This intrinsic and generic property of video requires that the generated video frames are both visually and semantically coherent, and meanwhile have smooth changes in motion over time.

*Beijing National Research Center for Information Science and Technology (BNRist)
†Corresponding authors.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350986

To tackle the challenges in sentence-to-video generation and address the weaknesses of existing literature, we propose a cross-modal dual learning (CMDL) algorithm which jointly considers the semantic consistencies between descriptive sentences and generated videos as well as the spatial-temporal coherence across video frames. The proposed CMDL model consists of two main components, i.e., sentence-to-video generating component and video-to-sentence re-embedding component. The sentence-to-video generator is a cascaded architecture containing a text-to-visual feature encoder and a conditional video generator, where the former maps sentence embeddings to visual features and the latter is utilized to generate videos based on the corresponding text and visual features. The text-to-visual feature encoder is designed for simultaneously generating visual representations both globally and locally such that the multi-scale visual representations in videos can be explored to keep the temporal consistencies across frames. To make sure that the generated videos and the given descriptive sentences are semantically consistent, CMDL employs a dual learning architecture through the video-to-sentence re-embedding component, where we align the re-embedded textual descriptions with ground truth sentence embeddings in a common latent semantic space. Figure 1 shows a semantically consistent example by the proposed dual procedure in CMDL compared with a semantically inconsistent example generated by GAN-CLS proposed by Reed et al. [24] without video-to-sentence re-embedding module. As such, our proposed CMDL model simultaneously learns the bidirectional mappings between sentences and videos in pair, producing realistic videos and ensuring their semantic meanings are consistent with those of the corresponding textual descriptions.

To the best of our knowledge, we are the first to utilize dual learning mechanism in the problem of video generation, which has advantages in addressing the weaknesses in existing works based on VAEs or GANs. The contributions of this work can be summarized as follows:

- We propose a novel end-to-end cross-modal dual learning (CMDL) approach capable of generating videos which are semantically consistent with their descriptive sentences and are sequentially plausible across frames.
- We introduce a dual mapping structure to learn the bidirectional semantic relations between descriptive sentences and generated videos such that they are semantically consistent.
- We employ a multi-scale text-to-visual feature encoder to obtain both global and local representations in videos so that the generated videos can be temporally consistent and sequentially plausible.
- We conduct extensive experiments on various datasets and compare CMDL with several benchmarks to validate the effectiveness of our proposed model.

2 RELATED WORK

Generative models have been extensively explored by both academia and industry recently, and we group existing image/video generation models into two categories: Variational Auto-encoders (VAEs) [14] based approaches and Generative Adversarial Networks (GANs) [7] based methods. VAEs model characterizes the video features as probability distributions, where the encoder is an inference network mapping the input video to a posteriors distribution, and the decoder is a generative network taking latent variables z as input and projecting z to low-level visual pixels. GANs can be regarded as a minimax game between the generator and the discriminator. Taking video generation from noise as an example, the generator aims to produce fake videos from noises to cheat the discriminator by imitating the real data distribution while the discriminator learns to distinguish between real videos and fake videos generated by the generator. VGAN [33] is one of the early works based on GANs to generate videos from Gaussian noises, which is followed by other subsequent work [19] exploring the video generation problem.

Video generation from sentences. Several existing works investigate the problem of controllable video generation [9, 27], especially generating videos from given sentences [15, 16, 18, 20]. Particularly, two sequence to sequence models Sync-Draw [18] and Attn-VAE [16] adopt VAEs for video generation. Sync-Draw utilizes Recurrent VAE [8] to generate video frame by frame. Attn-VAE proposes to combine Recurrent VAE with attention mechanism to modeling long-term and short-term contexts in video generation. Li et al. [15] and Pan et al. [20] employ GANs to generate videos from sentences, where Li's model fuses the textual information in the generator and discriminator, and Pan's work adds two auxiliary discriminators to determine the frame-level and motion-level plausibility as well as relevance of the generated video with its given caption. However, VAE-based methods tend to produce unrealistic, blurry samples and the training of GANs is unstable and is sometimes hard to converge to an optimal solution, resulting in the mode collapse problem [17]. Instead of using a hard-to-train discriminator, our proposed CMDL model adopts a dual-mapping structure to ensure relevance of the generated video with its descriptive sentence and a text-to-visual encoder to obtain multi-scale visual features, thus is capable of generating more semantically consistent and plausible videos in the experiments.

Dual learning in image generation. There have also been some works utilizing the idea of dual learning to generate images from texts. Being first proposed by Xia et al. [10] for machine translation, dual learning mechanism is designed to generate informative feedback signals through the dual task to train the translation model with less labeled data. CycleGAN [41] and DualGAN [37] adopt the idea of enhancing cycle consistency between two domains in image-to-image translation, where the dual learning is conducted only in the single vision modal. MirrorGAN [22] targets at aligning semantic meanings between images and sentences in image generation task. Our work differs from theirs in the following aspects: (i) We focus on sentence-to-video generation problem, which is more challenging because of the semantic and spatial-temporal diversities between sentences and videos; (ii) Rather than calculating the dual learning loss in corpus level whose performance may be deteriorative due to synonymy of sentences, we introduce a dual mapping loss in the latent semantic space to guarantee the semantic consistencies between videos and sentences, providing more options for the semantic mappings between them; (iii) We train the proposed CMDL model in an end-to-end fashion, jointly optimizing sentence-to-video generating component and video-to-sentence re-embedding component.

In general, our work targets at generating semantically consistent videos from sentences, which is different from existing works

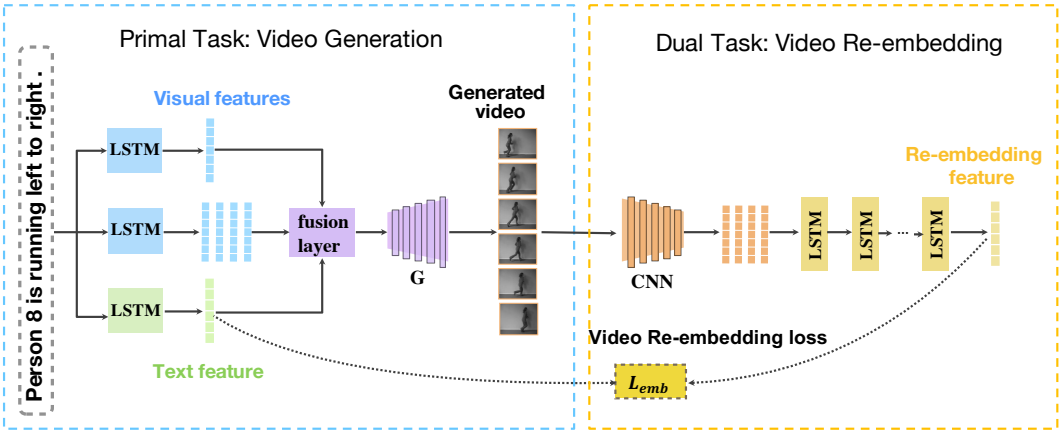

Figure 2: The framework of our proposed CMDL model for sentence-to-video generation. There are two main components in the model. The left part is sentence-to-video generation component and the right one is video-to-sentence re-embedding component. A video re-embedding loss in the latent semantic space is calculated to constrain the semantic consistencies between generated videos and given sentences.

on video generation from noise or sentence-to-image generation. We utilize dual learning structure with a dual mapping loss and a text-to-visual encoder to address the weaknesses in existing works. Furthermore, our proposed model is advantageous in optimizing the whole structure jointly in an end-to-end way.

3 THE CROSS-MODAL DUAL LEARNING MODEL

As is illustrated in Figure 2, CMDL consists of two components: one is the sentence-to-video generating component which aims to generate videos conditioning on given sentences, the other is the video re-embedding module which maps the created videos to a latent space keeping the semantic meaning of the ground truth sentence. These two components correspond to two tasks, i.e., sentence-to-video generation (primary task) and video-to-sentence re-embedding (dual task). We first define the primary task and dual task formally in Sec 3.1, followed by introducing sentence-to-video generation component in Sec 3.2 and video-to-sentence re-embedding component in Sec 3.3. Finally, we describe the training strategy in detail in Sec 3.4.

3.1 Definitions of Primary Task and Dual Task

Primary Task: Our primary task aims to learn a sentence-to-video generator G_v, which can map a descriptive sentence $\mathcal{S}$ to a video $\mathcal{V}_{syn}$ depicting the semantics of sentence. Formally, the mapping is defined as:

$$\mathcal{V}_{syn} = G_v(\mathcal{S}). \tag{1}$$

Suppose the input sentence is $\mathcal{S} = \{w_1, w_2, ..., w_{N_s-1}, w_{N_s}\}$ which includes N_s words and each $w_i \in \mathbb{R}^{d_c}$ is a d_c-dimensional one-hot vector collected from the vocabulary corpus. The target output $\mathcal{V}_{syn} = \{f_1, f_2, ..., f_{T-1}, f_T\}_{syn}$ consists of T sequential frames with spatial-temporal coherence and is semantically consistent with the given sentence $\mathcal{S}$. Here, T is the total number of frames, and $f_i \in \mathbb{R}^{d_h \times d_w}$ represents the i-th synthetic frame, where d_h and d_w denote the height and width of each frame, respectively.

Dual Task: The dual task in our proposed CMDL model is an inverse procedure of the primal task, aiming to map the generated video $\mathcal{V}_{syn} = \{f_1, f_2, .., f_T\}_{syn}$ to a latent semantic embedding space where the descriptive sentence $\mathcal{S}$ can also be mapped to. Thus the video-sentence semantic correlation can be measured in this common space. Specifically, this video-to-sentence mapping is denoted as C_r and implemented in the video re-embedding module of CMDL model. The re-embedding module generates corresponding sentence embeddings given both real and synthetic/generated videos, which are formulated as follows:

$$\begin{aligned} \mathbf{s}_{real} &= C_r(\mathcal{V}_{real}), \\ \mathbf{s}_{syn} &= C_r(\mathcal{V}_{syn}), \end{aligned} \tag{2}$$

where $\mathcal{V}_{real}$ and $\mathcal{V}_{syn}$ denote the real and synthetic/generated video respectively. Accordingly, $\mathbf{s}_{real}$ and $\mathbf{s}_{syn}$ denote the embedded sentence vectors from the real video and synthetic video respectively. The re-embedding module of CMDL generates the embedded sentence vectors instead of the original words in sentences because each video can be described by various sentences with the same semantic meaning due to the ambiguity of natural language.

In brief, the sentence-to-video generation component aims to generate videos given their corresponding descriptive sentences while the video re-embedding module further enhances the semantic consistencies between the generated videos and their descriptive sentences. These two cross-modal components in the proposed CMDL model cooperate jointly to generate more realistic and reliable videos.

3.2 Video Generation Module

Taking a descriptive sentence as input, we first extract the textual feature and then customize two architectures for text-to-visual feature encoder to generate visual features on different scales. The *global visual features* capture the consistent information in videos

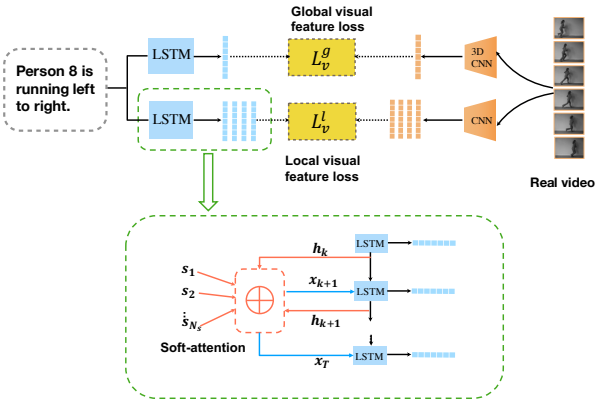

Figure 3: The text-to-visual feature encoder in our CMDL model. A soft-attention mechanism is applied to capture various information in sentences.

and keep it same over time, whereas the *local visual features* represent the varying contextual contents and model the motion parts in videos. The generator **G** is then able to generate videos based on the multi-scale textual and visual features.

Sentence Features. Let $\mathcal{S}$ be the input sentence, which consists of N_s words $\mathcal{S} = \{w_1, w_2, ..., w_{N_s}\}$. Each word w_j is first encoded into a 300-dimentional feature s_j using GloVe [21] algorithm to obtain the word-level features $\{s_1, s_2, ..., s_{N_s}\}$. We follow InferSent [4] by leveraging a pre-trained bidirectional LSTM (bi-LSTM) to contextually embed the input word-level features into a sentence representation **s**. Particularly, for the sequence of N_s words, the bi-LSTM processes the input sequence from two directions. One is the natural order from s_1 to s_{N_s} that computes the sequence flow of forward hidden states $\overrightarrow{h}_t$, the other reads the sentence in a reverse order from s_{N_s} to s_1 and gets the sequence of backward hidden states $\overleftarrow{h}_t$. For $t \in [1, ..., N_s]$, h_t is the concatenation of $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$. The sentence representation **s** is then calculated by selecting the maximum value over each dimension of the hidden states (max pooling) h_t [3].

$$\begin{aligned} \overrightarrow{h}_t &= \overrightarrow{LSTM}(s_1, s_2, ..., s_{N_s}), \\ \overleftarrow{h}_t &= \overleftarrow{LSTM}(s_1, s_2, ..., s_{N_s}), \\ \mathbf{s} &= maxpooling(\overrightarrow{h}_t || \overleftarrow{h}_t), \end{aligned} \tag{3}$$

where || denotes the concatenation of vectors.

Global and Local Visual Features. In general, sentence descriptions are less informative than videos. Sentence embeddings are always low-dimensional while videos normally contain more complex visual information in high dimension, resulting in the fact that directly generating video frames seems to be intractable. To address this issue, CMDL first obtains visual features from the given sentence and then generates videos from visual features. Since the input sentence consists of a sequence of words, an LSTM [29] encoder is used for text-to-visual feature encoder. We apply soft-attention mechanism proposed in [36] to create a word-context input for LSTM at each step. As illustrated in Figure 3, the sentence after word embedding is fed to the LSTM. At t-th step, the word-level embeddings and the hidden units h_{t-1} are first projected into a common space by perception layers U and V. Then the LSTM input is computed by attending to the words embedding features after projection based on the hidden state embedding feature. Since the hidden state h_{t-1} contains the information of previous generated visual features, it can guide the selection of word-context input in the next step. Formal definitions are shown as follows:

$$\begin{aligned} h_t &= LSTM(x_t, h_{t-1}), \\ e &= Us, h_t^e = Vh_t, \\ x_{t+1} &= F^{attn}(e, h_t^e), \end{aligned} \tag{4}$$

where s is the word-level embedding of input sentence, $e = \{e_i\}_{i=1}^{N_s}$ and h_t^e denote the projected word vectors and hidden state vector, respectively. U and V are the transformation matrices to be learned. F^{attn} is the soft-attention model for computing the attention output x_{t+1}. The attentive weights and attention output yielded by the attention function are computed as follows :

$$x_{t+1} = \sum_{i=1}^{N_s} \beta_i e_i, \quad \beta_i = \frac{exp(e_i^T h_t^e)}{\sum_{j=1}^{N_s} exp(e_j^T h_t^e)}.$$

The aforementinoned LSTM outputs a sequence of hidden units $h_1, h_2, ..., h_T$. Inspired by [29], we apply the mean pooling strategy on the hidden states to yield the global visual feature:

$$\mathbf{f}_g = \frac{1}{T} \sum_{t=1}^{T} h_t. \tag{5}$$

Here T is the sequence length of the hidden states. The mean pooling process yields the global visual feature with the same size as h_t.

The global visual feature of a video represents characteristics of the whole sequential frames, omitting the local structure in each frame. Therefore, we design a second LSTM model to encode word embeddings into visual features frame by frame, considering the local structures in a video sequence. The procedure of encoding word embeddings with soft-attention mechanism is the same as Eq (4), where an LSTM model is adopted to construct local-level features except that there is no mean pooling strategy cascaded with the LSTM. In order to distinguish from the previous $\{h_t\}_{t=1}^{T}$ in global visual feature encoder, we define $\{h_t^l\}_{t=1}^{T}$ as the outputs of the local visual LSTM encoder. The hidden units $\{h_1^l, h_2^l, ..., h_T^l\}$ are directly considered as the frame representations of a video, where T is the length of video frames:

$$\mathbf{f}_l = [h_1^l, h_2^l, ..., h_T^l], \tag{6}$$

where $\mathbf{f}_l \in \mathbb{R}^{d_v \times T}$ denotes the local visual features for T frames and d_v is the local visual feature dimension.

Global-Local Collaborative Video Generator. The aforementioned sentence encoder transforms a given sentence into textual features **s**, and the text-to-visual feature encoder creates global visual features $\mathbf{f}_g$ as well as local visual features $\mathbf{f}_l$. Simply concatenating these features leads to a high dimension output and therefore high computational complexity in the video generator. Besides, naive vector concatenations empirically result in an overly reliant usage of either textual or visual information. Therefore, we

utilize a fusion layer to automatically select useful information for the video generator. The fusion layer is designed as a perception layer, which can be mathematically expressed as:

$$\mathbf{f}_m = \mathcal{F}(\mathbf{f}_g || \mathbf{f}_l^i || \mathbf{s}). \tag{7}$$

Here $||$ denotes the column-wise concatenation, $\mathbf{f}_l^i$ denotes the i-th frame feature and $\mathcal{F}$ is a perception layer with weights $W \in \mathbb{R}^{d_{con} \times d_m}$, where d_{con} represents the dimension of the concatenated vectors. The fused feature obtained through the fusion layer is denoted as $\mathbf{f}_m \in \mathbb{R}^{d_m}$.

For the video generator, recent work [33] adopts 3D deconvolution network to generate a fixed-length frame sequence simultaneously and attain the spatial-temporal invariance across frames. Inspired by this work, we also use the generic deconvolution network [38] with 3D filters [31] for video generation. Given the fusion feature $\mathbf{f}_m$, CMDL utilizes a modified generator to produce a sequence of frames $\mathcal{V}_{syn} = \{f_1, f_2, ..., f_T\}_{syn}$ where T is the frame length and f_i is the i-th frame of generated video.

3.3 Video Re-embedding Module

The main challenge for video generation from sentence lies in the alignment between the generated video and its textual description. Previous works either neglect the paired relationship between video and sentence [15], or add an auxiliary discriminator that forces the generated video to be matched with the given sentence [20], which is empirically unstable and fails to converge rapidly in training. To address this problem, we adopt the dual learning mechanism as an auxiliary feedback to the video generator. The video re-embedding module encodes videos into the common latent semantic space where the corresponding descriptive sentences are also embedded to, and then computes the distances between the re-embedded descriptions and the ground-truth sentences. Actually, due to the ambiguity of natural language, a generated video can be described in several ways, and we therefore re-embed it into the latent semantic space instead of a real word sequence to encourage more generalization in CMDL.

More concretely, a pre-trained CNN is first utilized to capture the high-level semantic information about the video. We adopt VGG19 [28] for high-level semantic feature extraction. In this way, the video sequence is encoded as a fixed-length frame by frame feature map $\{v_1, v_2, ..., v_T\}$. Then the features extracted from the previous generated video are fed into the following LSTM to produce the re-embedded sentence embedding vector in the common latent semantic space. The video-to-sentence re-embedding module can be mathematically written as follows:

$$\begin{aligned} h_t^c &= LSTM(v_t, h_{t-1}^c), \\ \mathbf{s}_{syn} &= h_T^c, \end{aligned} \tag{8}$$

where h_t^c is the hidden state of LSTM model and the re-embedding result is denoted as $\mathbf{s}_{syn}$.

To better learn the mappings from videos to the common semantic space, we train the re-embedding module not only using the generated videos but also the real videos. Thus both $\{\mathcal{V}_{real}, \mathbf{s}\}$ and $\{\mathcal{V}_{syn}, \mathbf{s}\}$ are used for training, where $\mathbf{s}$ is the ground truth sentence representation in the common latent space, $\mathcal{V}_{real}$ and $\mathcal{V}_{syn}$ denote the true video and generated video respectively. The overall video-to-sentence re-embedding loss is defined as follows:

$$\mathcal{L}_{emb} = \frac{1}{2}[\psi(\mathbf{s}, \mathbf{s}_{syn}) + \psi(\mathbf{s}, \mathbf{s}_{real})], \tag{9}$$

where $\mathbf{s}$ is the ground truth sentence semantic representation, $\mathbf{s}_{syn}$ and $\mathbf{s}_{real}$ denote the re-embedded representations of the real video and generated video respectively. The Euclidean distance is chosen as the metric ψ to measure the differences between the re-embedding textual descriptions and the real sentences in the common semantic space.

3.4 Training Strategy

Full Objective. Formally, the training data consists of $(\mathcal{V}_{real}, \mathcal{S})$ pairs, where $\mathcal{V}_{real}$ indicates a video and $\mathcal{S}$ indicates the description for the video. The full objective consists of four parts: the traditional pixel-level reconstruction loss $\mathcal{L}_{rec}$, the video re-embedding loss $\mathcal{L}_{emb}$ and two text-to-visual feature encoding losses $\mathcal{L}_v^g$ and $\mathcal{L}_v^l$.

A traditional reconstruction loss, i.e., $L1$ distance, is used to force the generator to generate realistic videos at pixel level. We use $L1$ distance rather than $L2$ distance because $L1$ loss encourages less blurring. Formally, the reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \mathbf{E}\left[\, ||\mathcal{V}_{real} - \mathcal{V}_{syn}||_1 \,\right], \tag{10}$$

where $\mathcal{V}_{real}$ is the ground truth video and $\mathcal{V}_{syn}$ is the generated video. Besides, $\mathbf{E}$ denotes the average loss on each frame.

For the text-to-visual feature encoding losses, the encoded global and local visual features are compared with the ground truth features extracted from pre-trained 3D CNN network (C3D) [12, 31] and VGG19 network [28], respectively. The vision feature encoding losses are defined as:

$$\begin{aligned} \mathcal{L}_v^g &= \psi(\mathbf{f}_g, C3D(\mathcal{V}_{real})), \\ \mathcal{L}_v^l &= \psi(\mathbf{f}_l, VGG19(\mathcal{V}_{real})), \end{aligned} \tag{11}$$

where $C3D(\cdot)$ indicates C3D neural network [13], yielding the ground-truth global feature of the input video. $VGG19(\cdot)$ indicates VGG19 neural network [28] which extracts the ground-truth frame-level features. $\mathbf{f}_g$ and $\mathbf{f}_l$ represent the global and local visual features obtained from the text-to-visual feature encoder. The Euclidean distance is chosen as the metric $\psi(\cdot)$ to measure the encoding losses.

The full objective function of our CMDL model is defined as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_e \mathcal{L}_{emb} + \lambda_g \mathcal{L}_v^g + \lambda_l \mathcal{L}_v^l, \tag{12}$$

where λ_e, λ_g and λ_l are weighting factors that control the relative importance of each term.

We pre-train the video-to-sentence re-embedding module as an initialization such that the training process can be more stable and converge faster. The pre-training process is terminated by the early stopping strategy. The sentence-to-video generation module and the video-to-sentence re-embedding module are jointly trained according to the full objective in Eq (12). Algorithm 1 summarizes the core training procedure of our proposed CMDL model.

4 EXPERIMENTS

To assess the capability of our proposed model, we conduct extensive experiments on four datasets collected from different scenarios with increasing complexity: Single Digit Bouncing MNIST (SDBM), Two Digits Bouncing MNIST (TDBM) [18], KTH Human

Algorithm 1 Text to Video Generation Training Procedure

Require: Video set V, caption set S, video generator G_v parameters θ_v, video re-embedding module C_r parameters θ_r, batch size m, learning rates α_v and α_r , weighting factors λ_e, λ_g, λ_l

Initialize θ_r by pre-training C_r, randomly initialize θ_v
repeat
  for $n = 1, 2, 3, \ldots$ **do**
    sample m sentence-video pairs $\{\mathcal{S}^{(k)}, \mathcal{V}^{(k)}_{real}\}_{k=1}^{m}$
    extract word-level embeddings s and sentence-level representation $\mathbf{s}$
    encode word embeddings into global and local visual features $\mathbf{f}_g, \mathbf{f}_l$
    generate videos $\mathcal{V}_{syn}$ by the global-local collaborative generator
    compute $\mathcal{L}_{rec}$, $\mathcal{L}_v^g$ and $\mathcal{L}_v^l$ according to Eq (10) and Eq (11)
    re-embedding $\mathcal{V}_{syn}$ and $\mathcal{V}_{real}$ and obtain $\mathcal{L}_{emb}$ by Eq (9)
    compute full objective $\mathcal{L}$ according to Eq (12)
    update $\theta_v \rightarrow \theta_v - \alpha_v \frac{\partial \mathcal{L}}{\partial \theta_v}$
    update $\theta_r \rightarrow \theta_r - \alpha_r \frac{\partial \mathcal{L}}{\partial \theta_r}$
  end for
until convergence

Action Dataset (KTH) [26] and Microsoft Research Video Description Dataset (MSVD) [2]. We present three baseline methods in sentence-to-video generation to provide comparisons of our results with state-of-the-art methods. We give the quantitative results and subjective analysis on four datasets in Sec 4.2. We also conduct a human evaluation for further understanding how reliable and semantically consistent the generated videos are with given sentences.

4.1 Experimental Settings

Evaluation Metrics. Evaluating the quality of generated videos is a difficult problem for video generation due to the multiple possible features of videos. In order to measure the spatial plausibility and temporal smoothness of generated videos, we use Inception Score in [25, 34] for evaluation. Following common practice, we calculate the **frame-level Inception Score** as [36] based on the classification result of each generated frame, which reflects if the video frames are reliable along the sequence. In addition, since videos have the inherent property of spatial-temporal coherence and cohesion, we also measure the **video-level Inception Score** inspired by [1] such that a pre-trained video recognition network can recognize the objects and actions in the generated videos. The frame-level and video-level Inception Scores are calculated accordingly as [1]. In the experiments, an Inception-v3 [30] CNN model trained on ImageNet [5] is used to classify each frame to 1000 categories. The frame-level Inception Score is calculated on each frame and then averaged as the final score. As for the video-level Inception Score, we train an RGB stream based 3D CNN model [35] to classify each video and then compute the inception score.

Compared Approaches. We compare the generation results with the following state-of-the-art methods.

(1) Generative Adversarial Network with Character-Level Sentence Encoder (GAN-CLS) [24] is a method for sentence-to-image generation with additional sentence input to original DC-GAN [23]. We directly replace the 2D convolutions in the generator as 3D convolutions for sentence-to-video synthesis.

(2) Attentive Semantic Video Generation from Captions (Attn-VAE) [16] is a VAE-based method for video generation which utilizes recurrent VAE [8] with Attention to capture long-term and short-term context in video generation.

(3) Temporal GANs conditioning on Captions with Temporal Coherence Adversarial Loss (TGANs-C-A) [20] is proposed to add an adversarial temporal coherence loss between adjacent frames to constrain the smoothness of generated videos.

4.2 Quantitative Evaluation

The performances of our method and other three baselines on the Frame-IS and Video-IS metrics are shown in Table 1. Overall, the results across two metrics with the same input sentences indicate that our proposed CMDL method achieves superior performances against other state-of-the-art techniques on both frame-level Inception Score and video-level Inception Score. Specifically, the Frame-IS of our CMDL can achieve 3.325 on TDBM and 2.682 on SDBM, making the relative improvement over the best competitor TGAN-C-A by 2.044 and 1.594, respectively. The significant progress means the generated videos of CMDL contain more clear objects and have a higher diversity of produced frames, which confirms the effectiveness of the text-to-visual feature encoder in CMDL. Moreover, CMDL by additionally considering global and local visual features for video generation with a soft-attention on input sentence, leads to an improvement over three competitors on Video-IS metric. The results again verify the advantage of our proposed

Table 1: The evaluation results of Frame Inception Score (Frame-IS) and Video Inception Score (Video-IS) by our model and four compared baselines on SDBM, TDBM, KTH Human Action and MSVD datasets.

Metrics	Datasets	GAN-CLS	Attn-VAE	TGAN-C-A	**CMDL**
Frame-IS	SDBM	1.005 ± 0.001	1.212 ± 0.157	1.594 ± 0.187	**2.682 ± 0.168**
	TDBM	1.404 ± 0.681	1.393 ± 0.504	2.044 ± 0.223	**3.325 ± 0.239**
	KTH	1.438 ± 0.097	1.450 ± 0.167	1.937 ± 0.134	**2.077 ± 0.299**
	MSVD	**2.604 ± 0.083**	1.520 ± 0.176	1.749 ± 0.031	2.580 ± 0.125
Video-IS	KTH	1.015 ± 0.003	1.007 ± 0.003	1.005 ± 0.002	**1.280 ± 0.024**
	MSVD	1.018 ± 0.001	1.001 ± 0.001	1.003 ± 0.001	**1.141 ± 0.013**

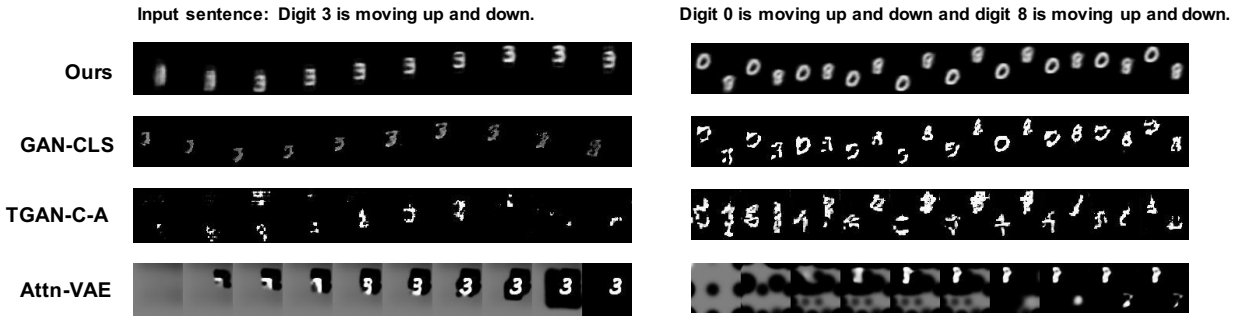

Figure 4: Examples of generated videos by our CMDL model and compared three approaches on SDBM and TDBM datasets.

text-to-visual feature encoder. As expected, GAN-CLS performs worse than CMDL, since it is directly extended from text-to-image generation model such that it ignores the motion changes over time. As a VAE-based method, Attn-VAE tends to generate blurry results, which is illustrated in Sec. 4.3, resulting in worse inception scores. Note that our CMDL method performs similarly to GAN-CLS on MSVD dataset, whereas the qualitative results in Sec. 4.3 indicate that our generated videos are more visually reliable. The overall quantitative results indicate that our CMDL model is capable of producing videos which not only contain realistic frames but also have temporal plausibility over sequence. Our CMDL model outperforms existing state-of-the-art methods.

4.3 Qualitative Evaluation

Subjective Visual Comparisons. We examine the generated videos from four methods by subjective visual comparisons. For each dataset, we randomly sample four instances from the generated results and show these sampled instances in the main paper (Figure 4, 5). For more qualitative results, please refer to our supplemental material. From these exemplar results, it is easy to see all of these automatic methods can generate somewhat video sequences, while our propose CMDL can generate more reliable and semantically consistent videos by exploiting attentive text-to-visual encoding and applying dual learning mechanism to model bidirectional mappings between generated videos and input sentences. In particular, given a sentence "Digit 3 is moving up and down", the video generated by our CMDL is clear and contains the up-and-down motion over frames. The GAN-based methods GAN-CLS and TGAN-C-A are also capable of producing videos in which the digits bounce up and down, but the shapes of digits are sharp and sometimes change in adjacent frames, failing to maintain the coherence among frames. The VAE-based method Attn-VAE generates videos frame by frame and the visual quality is quite good on frame level. However, there are no up-and-down movements of digit 3 in the video, indicating that Attn-VAE is weak at understanding and translating the semantic meaning of input sentence. The comparisons on TDBM dataset are presented on the right part of Figure 4. Similar to the observations on SDBM, our CMDL model generates precisely relevant videos with given sentence, which again verifies the effectiveness of the designed video re-embedding architecture in CMDL. Evidently, our proposed CMDL achieves the best performance on maintaining the semantic consistency and modeling the temporal dynamics as well as the frame-level realness in videos.

Next, we present the qualitative results on KTH and MSVD datasets to evaluate the performances of different approaches on real scenarios. The generated videos are shown in Figure 5. As we can observe, CMDL achieves better results with more details and coherent motions compared to three baselines. Obviously, the video in the first row by CMDL is more realistic than others, indicating the effectiveness of our text-to-visual feature encoder in real scenarios. The VAE-based method Attn-VAE generates blurry video frames, making it hard to recognize the contours of objects. The GAN-based method TGAN-C-A creates more reliable videos but the human motions in adjacent frames are sometimes not coherent. Another GAN-based model GAN-CLS generates contexts which are not as clear as ours distinguished at frame level. More specifically, given a sentence "A person is pouring beans into a pot on a stove", the video generated by CMDL has higher visual quality, where the pot can be recognized as the main object in the video. The subjective visual comparisons indicate that our CMDL model can produce realistic videos and maintain the semantic consistencies with given sentences, and also surpass baselines both at both frame level and video level.

Human Evaluation. To quantitatively evaluate the visual quality and human-likeness of generated videos, we also conduct a human study to compare our CMDL against three approaches, i.e., Attn-VAE, GAN-CLS, TGAN-C-A. A total number of 10 evaluators participate in the study through Amazon Mechanical Turk (AMT). The study includes two tests: the Video Authenticity Test and the Semantic Consistency Test. The Video Authenticity Test aims to evaluate the authenticity of videos generated by different approaches, and the Semantic Consistency Test aims to determine how relevant the generated videos are with given sentences. We randomly sample 100 videos from four methods and divide them into 25 groups, with each group containing four videos with same input sentence. In the Video Authenticity Test, evaluators are provided 25 video groups without sentence descriptions and asked to mark the most realistic one in each group. Similarly, the Semantic Consistency Test provides 25 video groups with one corresponding sentence to evaluators and ask them to determine the most semantically consistent one with the given sentence.

From evaluators' responses, we calculate two metrics: 1) Authenticity: percentage of videos that are visually evaluated better than others; 2) Semantic Consistency: percentage of videos that are more semantically consistent with given sentences than others. Table 2 presents the human evaluation results. Our CMDL method

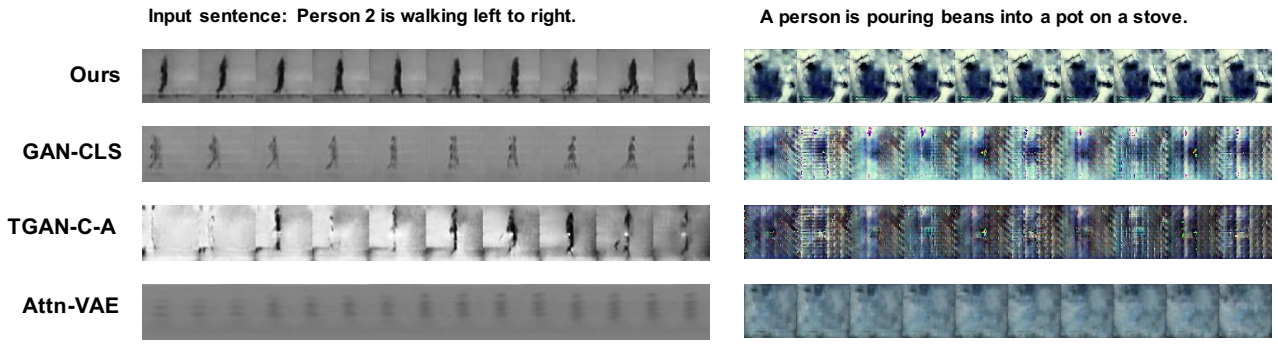

Figure 5: Examples of generated videos by our CMDL model and compared three approaches on KTH and MSVD datasets.

Table 2: Human evaluation results of our CMDL and compared baselines.

Methods	Authenticity	Semantic Consistency
GAN-CLS	31.2%	17.3%
Attn-VAE	12.8%	18.7%
TGAN-C-A	10.4%	28.0%
CMDL	**45.6%**	**36.0%**

evidently outperforms the competing three methods on both visual authenticity and semantic consistency. 45.6% of our generated videos rank first in the Video Authenticity Test and 36.0% of ours win other three methods in the Semantic Consistency Test, which indicates that our model has the capability of generating spatial-temporal plausible videos, and at the same time good at maintaining semantic consistencies with given sentences.

4.4 Ablation Studies

Video re-embedding module. To investigate the effectiveness of our proposed video re-embedding module, we conduct two comparative experiments on KTH dataset by first removing the re-embedding module and then replacing the re-embedding module with a video captioning module, which generates sentences word by word instead of semantic embeddings. As the video re-embedding module performs video-to-sentence mapping, it requires that the objects and actions in generated videos can be identified and at the same time the semantic meanings of videos are aligned with given sentences. When removing this module from our model, the Frame-IS decreases from 2.077 to 1.513, and the Video-IS decreases from 1.280 to 1.010 as presented in Table 3, which verifies that the re-embedding module is crucial in CMDL. Additionally, we replace the re-embedding module with a video captioning module and jointly train it with the video generator. However, it does not converge from the experimental observation. A possible reason is that the vocabulary corpus is limited and discrete such that the process of learning to generate sentences is unstable and difficult to converge. As the re-embedding module maps videos to a continuous semantic space, it considers the synonymy of sentences and is more stable in training.

Text-to-visual feature encoder. We additionally conduct an experiment on KTH dataset to verify the benefit of our text-to-visual feature encoder. The experimental results are shown in Table 3. The Frame-IS and Video-IS both decrease without the text-to-visual feature encoder, which illustrates the advantage of text-to-visual encoding module.

Table 3: Ablation studies results on KTH dataset. "w/o T2V" stands for the model without text-to-visual feature encoder and "w/o V2T" stands for the model without video re-embedding module.

	Frame-IS	Video-IS
w/o T2V	1.009 ± 0.002	1.002 ± 0.001
w/o V2T	1.513 ± 0.117	1.010 ± 0.002
CMDL	**2.077 ± 0.299**	**1.280 ± 0.024**

5 CONCLUSIONS

In this paper, we propose a cross-modal dual learning (CMDL) algorithm for sentence-to-video generation, where the primary task is sentence-to-video generation and the dual task is video-to-sentence re-embedding. The proposed CMDL model tackles the existing issues in temporal coherence and semantic consistency in sentence-to-video generation problem by simultaneously learning the bidirectional mappings between sentences and videos. Experimental results suggest that the dual mechanism can significantly improve the relevance between the generated videos and given sentences. In addition, we employ a multi-scale text-to-visual feature encoder to obtain both global and local representations in the video generation component. The text-to-visual feature encoder provides visual features for generation such that the generated videos can be temporally consistent and sequentially plausible. Extensive experiments conducted on four datasets validate our proposed method. The quantitative and qualitative analysis demonstrate that our CMDL model surpasses current methods both visually and quantitatively.

ACKNOWLEDGMENTS

This work was support by National Program on Key Basic Research Project No. 2015CB352300, National Natural Science Foundation of China Major Project No. U1611461, China Postdoctoral Science Foundation No. BX201700136 and Shenzhen Nanshan District Ling-Hang Team Grant under No. LHTD20170005.

REFERENCES

[1] Haoye Cai, Chunyan Bai, Yuwing Tai, and Chikeung Tang. 2018. Deep Video Generation, Prediction and Completion of Human Action Sequences. In *European Conference on Computer Vision*. 374–390.

[2] David L Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 190–200.

[3] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine learning*. 160–167.

[4] Alexis Conneau, Douwe Kiela, Holger Schwenk, LoÃŕc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[6] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly Supervised Dense Event Captioning in Videos. In *Advances in Neural Information Processing Systems*. 3059–3069.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[8] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In *Proceedings of the 32th International Conference on Machine Learning*. 1462–1471.

[9] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. 2018. Probabilistic video generation using holistic attribute control. In *European Conference on Computer Vision*. 452–467.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[11] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4193–4202.

[12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 221–231.

[13] Shuiwang Ji, Ming Yang, Kai Yu, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 221–231.

[14] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. (2013). arXiv:1312.6114.

[15] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[16] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. 2017. Attentive Semantic Video Generation Using Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1435–1443.

[17] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which Training Methods for GANs Do Actually Converge? (2018). arXiv:1801.04406.

[18] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. 2017. SyncDraw: Automatic Video Generation using Deep Recurrent Attentive Architectures. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1096–1104.

[19] Katsunori Ohnishi. 2018. Hierarchical Video Generation from Orthogonal Information : Optical Flow and Texture. (2018). arXiv:1711.09618v2.

[20] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To Create What You Tell: Generating Videos from Captions. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1789–1798.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[22] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning Text-to-image Generation by Redescription. (2019). arXiv:1903.05854.

[23] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Proceedings of International Conference on Learning Representations*.

[24] Scott Reed, Zeynep Akata, Xinchen Yan, and Lajanugen Logeswaran. 2016. Generative Adversarial Text to Image Synthesis. In *Proceedings of the 33th International Conference on Machine Learning*. 1060–1069.

[25] Tim Salimans, Ian Goodfellow, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*. 1–10.

[26] Christian Schuldt, Ivan Laptev, and Barbara Caputo. 2004. Recognizing Human Actions: A Local SVM Approach. In *Proceedings of International Conference on Pattern Recognition*. 32–36.

[27] Guangyao Shen, Wenbing Huang, Chuang Gan, Mingkui Tan, Junzhou Huang, Wenwu Zhu, and Boqing Gong. 2019. Facial Image-to-Video Translation by a Hidden Affine Transformation. In *Proceedings of the 27th ACM international conference on Multimedia*.

[28] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of International Conference on Learning Representations*.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*. 3104–3112.

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

[31] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3d Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4489–4497.

[32] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. (2014). arXiv:1412.4729.

[33] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*. 613–621.

[34] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. 2017. The Pose Knows: Video Forecasting by Generating Pose Futures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3352–3361.

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. 2015. Towards Good Practices for Very Deep Two-Stream ConvNets. (2015). arXiv:1507.02159.

[36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained Text to Image Generation with Attentional Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1316–1324.

[37] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2868–2876.

[38] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. 2010. Deconvolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[39] Han Zhang, Tao Xu, and Hongsheng Li. 2017. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5908–5916.

[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), 1–1.

[41] Junyan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2242–2251.