DTVNet: Dynamic Time-lapse Video Generation via Single Still Image

Jiangning Zhang[1][⋆], Chao Xu[1][⋆], Liang Liu[1], Mengmeng Wang[1],
Xia Wu[1], Yong Liu[1][†], Yunliang Jiang[2]

[1]APRIL Lab, College of Control Science and Engineering,
Zhejiang University, Hangzhou, Zhejiang, China
[2]Huzhou University, Huzhou, Zhejiang, China
{186368, 21832066, leonliuz, mengmengwang, xiawu}@zju.edu.cn
yongliu@iipc.zju.edu.cn jyl@zjhu.edu.cn

Abstract. This paper presents a novel end-to-end dynamic time-lapse video generation framework, named DTVNet, to generate diversified time-lapse videos from a single landscape image, which are conditioned on normalized motion vectors. The proposed DTVNet consists of two submodules: *Optical Flow Encoder* (OFE) and *Dynamic Video Generator* (DVG). The OFE maps a sequence of optical flow maps to a *normalized motion vector* that encodes the motion information inside the generated video. The DVG contains motion and content streams that learn from the motion vector and the single image respectively, as well as an encoder and a decoder to learn shared content features and construct video frames with corresponding motion respectively. Specifically, the *motion stream* introduces multiple *adaptive instance normalization* (AdaIN) layers to integrate multi-level motion information that are processed by linear layers. In the testing stage, videos with the same content but various motion information can be generated by different *normalized motion vectors* based on only one input image. We further conduct experiments on Sky Time-lapse dataset, and the results demonstrate the superiority of our approach over the state-of-the-art methods for generating high-quality and dynamic videos, as well as the variety for generating videos with various motion information.[1]

Keywords: Generative Adversarial Network, Optical Flow Encoding, Time-lapse Video Generation

1 Introduction

Video generation is a task to generate video sequences from the noise or special conditions such as images and masks, which has promising application capabilities, *e.g.* video dataset expansion, texture material generation, and film production. However, this is a very challenging task where the model has to learn

---

[⋆] indicates equal contributions.
[†] indicates corresponding author.
[1] https://github.com/zhangzjn/DTVNet

the content, motion, and relationship between objects simultaneously, especially when objects have no special shapes such as cloud and fog. In this work, we focus on a more challenging task that generates realistic and dynamic time-lapse videos based on a single natural landscape image.

Thanks to the development of *generative adversarial network* [4, 10, 35], many excellent models have been proposed that achieve generating realistic and dynamic videos. Vondrick *et al.* [28] explore how to leverage a random noise to learn the dynamic video in line with a moving foreground and a static background pathways from large amounts of unlabeled video by capitalizing on adversarial learning methods. However, this kind of noise-based method usually suffers from a low-quality video generation and a hard training, because of its mapping from a vector to a high-resolution feature map. Villegas *et al.* [27] first adopt the LSTM [6] to predict future high-level poses in a sequence to sequence manner, and then use an analogy-based encoder-decoder model to generate future images. The method can well generate high-quality and reasonable images, but it acquires extra label information such as human pose and several past reference images that is unpractical. Subsequent MoCoGAN [26] apply LSTM model to generate a sequence of random vectors that consist of content and motion parts, and then map them to a sequence of video frames. This method can generate diversified dynamic videos that contain the same content in theory, because we can fix the content part while change the motion part of the random vectors. Nevertheless, the resolution of the generated video is still low.

To obtain high-resolution and high-quality time-lapse videos, Aigner *et al.* [1] propose the FutureGAN that uses spatio-temporal 3d convolutions in all encoder and decoder modules, aiming at capturing both the spatial and temporal components of a video sequence. But FutureGAN is time-consuming during the training stage for utilizing concepts of the progressively growing GAN [11]. Xiong *et al.* [32] recently propose a two-stage 3DGAN-based model named MDGAN, which learns to generate long-term realistic time-lapse videos of high resolution given the first frame. Though MDGAN is capable of generating vivid motion and realistic video, it is hard to train such a two stage model and lacks of diverse video generation capabilities, which limits the practicality of the method. We follow the same time-lapse video generation task with MDGAN and design a practical one stage model.

Recently, some researches [14, 13] introduce optical flow maps that contain the motion information during the training and testing stages, aiming at explicitly supplying motion signals to the network. Liang *et al.* [14] design a dual motion GAN that simultaneously solves the primal future-frame prediction and future-flow prediction tasks. Though it can generate high-quality video frames, but it is time-consuming for taking more computational effort on future-flow predictions. Besides the quality and the resolution, the diversification of the generated time-lapse videos is likewise important for practical application. Li *et al.* [13] first map a sampled noise to consecutive flows, and then use proposed video prediction algorithm to synthesize a set of likely future frames in multiple time steps from one single still image. During the testing, the method can
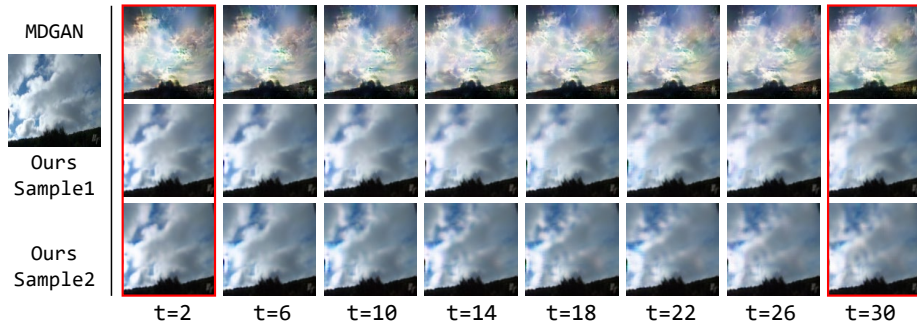
Fig. 1. The first row shows generated frames by MDGAN [32], and other two rows are from our method with random motion vectors. The left image is in-the-wild start landscape frame, and other frames are generation results at different times. Our method could generate more photorealistic and diversified video frames. Please zoom in the red rectangles for a more clear comparison.

directly use sample points from the distribution for predictions. Different from aforementioned methods, our method is designed in an end-to-end manner and we introduce a normalized motion vector to control the generation process, which can generate high-quality and diversified time-lapse videos.

In this work, we propose an end-to-end dynamic time-lapse video generation framework, *i.e.* DTVNet, to generate diversified time-lapse videos from a single landscape image. The time-lapse landscape videos generally contain still objects, *e.g.* house, earth, and tree, as well as unspecific objects, *e.g.* cloud and fog, which is challenging to understand the motion relationship between objects. Specifically, the proposed DTVNet consists of two submodules: *Optical Flow Encoder* (OFE) and *Dynamic Video Generator* (DVG). OFE introduces unsupervised optical flow estimation method to get the motion maps among consecutive images and encoders them to a normalized motion vector. During the testing stage, we exclude the process of flow estimation as well as flow encoding, and directly sample from a normalization distribution as the motion vector, which reduces the network computing overhead and supply diversified motion information simultaneously. DVG contains motion and content streams that learn from the motion vector and the single image respectively, as well as an encoder and a decoder to learn shared content features and construct video frames respectively. In detail, the normalized motion vector is integrated into the motion stream by multiple *adaptive instance normalization* (AdaIN) layers [7]. During the training stage, we apply content loss, motion loss, and adversarial loss that ensures high-quality, dynamic, and diversified video generation, as shown in Figure 1.

Specifically, we make the following four contributions:

- An *optical flow encoder* (OFE) is designed to supply normalized motion information in the training stage that are used to guide diversified video frames generation.

- A new *dynamic video generator* (DVG) is proposed to first learn disentangling content and motion features separately, and then use integrated features to generate target video.
- We apply content loss, motion loss, and adversarial loss during the training stage, which ensures high-quality, dynamic, and diversified video generation.
- Experimental results on Sky Time-lapse dataset indicate that the proposed DTVNet can generate high-quality and dynamic video frames in an end-to-end one stage network.

2   Related Work

Optical Flow Estimation. Optical flow is a reliable representation to characterize motion between frames. Starting from Flownet [3], many supervised methods for optical flow estimation are proposed, *e.g.* FlowNet2 [9], PWC-Net [24], IRR-PWC [8], etc. Though these methods are in high accuracy and efficiency, they heavily depend on the labeled dataset, while it it hard to acquire the ground truth in reality, which reduces the practicality of these methods.

As an alternative, some researchers focus on studying unsupervised methods [22, 34] and have achieved great success. Liu *et al.* [17] propose the SelFlow that distills reliable flow estimations from non-occluded pixels, and uses these predictions as ground truth to learn optical flow for hallucinated occlusions. DDFlow [16] further improve the model performance by distilling unlabeled data. In this paper, we first apply unsupervised method [15] to estimate the optical flow map, and then encoder the flow information to a motion vector that is used as a condition when generating the video.

Generative Adversarial Networks. Since Goodfellow *et al.* [4] first introduces the generative adversarial network (GAN) that contains a generator and a discriminator, many GAN-based approaches are proposed and have achieved impressive results in various aspects, *e.g.* image inpainting, style translation, super-resolution, etc. Mehdi *et al.* [19] propose the cGAN that controls the mode of generated samples by adding extra conditional variable to the network. Pix2Pix [10] uses ℓ_1 and adversarial loss for paired image translation tasks, and Zhu *et al.* [35] further introduces a cycle consistency loss to deal with unpaired image-to-image translation tasks. ProGAN [11] describes a new training methodology that grows both the generator and discriminator progressively, which is capable of generating up to 1024 resolution images. Besides 2D-based GAN methods, Wu *et al.* [31] apply 3D convolution to generate 3D objects from a probabilistic space. MDGAN [32] presents a 3D convolutional based two-stage approach to generate realistic time-lapse videos of high resolution. Our model follows the GAN idea and takes extra motion information into consideration when generating videos.

Video Generation. Video generation aims at generating image sequences from a noise, image(s), or with extra condition such as human pose, semantic label map, and optical flow. Mathieu *et al.* [18] first adopt the GAN idea to mitigate

the inherently blurry predictions obtained from the standard mean squared error loss function. Subsequently, VGAN [28] is proposed to untangle the foreground from the background of the scene with a spatio-temporal convolutional architecture, and many follow-up works borrowed the idea of disentangling. Saito *et al.* [23] exploits two different types of generators, *i.e.* a temporal generator and an image generator, to generate videos and achieve good performance, while MoCoGAN [26] maps a sequence of random vectors that consists of content and motion parts to a sequence of video frames.

However, these noise-input and progressively growing methods generally suffer from a low-quality generation or a difficult training process, so following methods that use a single image as input are proposed. MDGAN [32] adopts a two-stage network to generate long-term future frames. It generates videos of realistic content for each frame in the first stage and then refines the generated video from the first stage. Nam *et al.* [20] learns the correlation between the illumination change of an outdoor scene and the time of the day by a multi-frame joint conditional generation framework. Yang *et al.* [33] propose a pose guided method to synthesize human videos in a disentangled way: plausible motion prediction and coherent appearance generation. Similarly, Cai *et al.* [2] design a skeleton-to-image network to generate human action videos. Researches [29, 21] take one semantic label map as input to synthesize a sequence of photo-realistic video frames. Recently, some flow based methods have made great success. Liang *et al.* [14] design a dual motion GAN that simultaneously solves the primal future-frame prediction and future-flow prediction tasks. Li *et al.* [13] propose a video prediction algorithm that synthesizes a set of likely future frames in multiple time steps from one single still image. Considering the resolution and the motion of generated frames, we design our model as two streams: motion and content stream for solving the motion vector and the image information, respectively, and then fuse features from both streams to generate time-lapse video frames. Our model neither has to map random vectors to a sequence of video frames from scratch, nor generates video sequences frame by frame.

3   Method

In this paper, a novel end-to-end dynamic time-lapse video generation framework named DTVNet is proposed to generate diversified time-lapse videos from a single landscape image. Formally, given a single landscape image I_0, the model learns to generate a time-lapse video sequences $\hat{V} = \{\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_T\}$. As depicted in Figure 2, DTVNet consists of two submodules: *Optical Flow Encoder* ψ and *Dynamic Video Generator* ϕ, and we will explain our approach as follows.

3.1   Optical Flow Encoder

Considering the motion and diversity of the generated video, we design an *optical flow encoder* (OFE) module to encode the motion information to a *normalized motion vector*, thus not only can it supply motion information but also we could
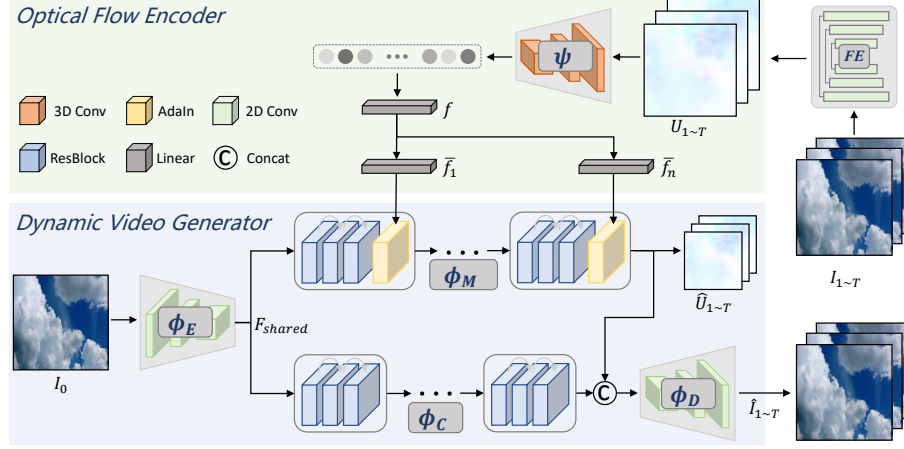
Fig. 2. Overview of the proposed DTVNet that consists of a *Optical Flow Encoder* ψ and a *Dynamic Video Generator* ϕ. Given the first landscape image I_0 and subsequent landscape images $I_{1\sim32}$, the *flow estimator* (FE) first estimates the consecutive flows $U_{1\sim T}$ and then ψ encodes the flows to the *normalized motion vector* $\bar{f}$. Dynamic video predictor ϕ successively apply encoder ϕ_E to learn shared content feature, motion stream ϕ_M and content stream ϕ_C to learn motion and content information from $\bar{f}$ and I_0 respectively, and a decoder ϕ_D to construct consecutive video frames $\hat{I}_{1\sim32}$.

sample various motion vectors from a normalization distribution to generate diversified videos in the testing stage.

As shown in Figure 2, we first apply unsupervised optical flow estimator which is more practical for not requiring label information (we use ARFlow [15] in this paper) to estimate consecutive flows $U_{1\sim T} = \left\{ U_{0\rightarrow 1}, U_{1\rightarrow 2}, \ldots, U_{(T-1)\rightarrow T} \right\}$ from real video frames $V_{0\sim T} = \{I_0, I_1, \ldots, I_T\}$, where T indicates maximum frame number. We formulate this process as:

$$U_{1\sim T} = FE(V_{0\sim T}). \tag{1}$$

Then OFE module ψ encodes consecutive flows $U_{1\sim T}$ to a normalized motion vector f that contains the motion information, denoted as:

$$f = \psi(U_{1\sim T}). \tag{2}$$

Specifically, OFE employs a 3D encoder architecture, which is proven to be more suitable for learning spatial-temporal features than 2D convolution [25]. With the reduction of the time dimension, the 3D encoder can not only model the motion information of local adjacent frames, but also the global motion information of the sequence.

As a result, the output 512-dimensional normalized vector contains continuous motion information for generating future images, and will be integrated into DVG module. Detailed structure and parameters can be found in the supplementary material.

3.2 Dynamic Video Generator

To generate photo-realistic and vivid video that has consistent content with the reference image and dynamic movement, we propose a novel *dynamic video generator* (DVG) that is designed in a disentangling manner. As shown in Figure 2, DVG contains an encoder ϕ_E, a motion stream ϕ_M, a content stream ϕ_C, and a decoder ϕ_D.

In detail, the encoder ϕ_E learns shared content feature F_{shared} from the first landscape image I_0, denoted as:

$$F_{shared} = \phi_E(I_0). \tag{3}$$

Considering the mismatch between the normalized motion vector f and motion stream features in semantic level that could inevitably increase the training difficulty, we introduce multiple linear layers to learn the adaptive motion vectors $\{\bar{f}_1, \ldots, \bar{f}_n\}$. Then adaptive motion vectors are integrated into the motion stream ϕ_M by multiple adaptive instance normalization (AdaIN) layers. Specifically, adaptive motion vector $\bar{f}_i$ is first specialized to motion styles $z_i = (z_i^{scale}, z_i^{shift})$ for i_{th} AdaIn layer with the input feature map F_i^{in}. Then we can calculate the output feature map F_i^{out} in the following formula:

$$F_i^{out} = z_i^{scale} \frac{F_i^{in} - \mu(F_i^{in})}{\sigma(F_i^{in})} + z_i^{shift}, \tag{4}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ calculate mean and variance respectively. Complete formula for motion stream is as follows:

$$\hat{U}_{1\sim T} = \phi_M(F_{shared}, \bar{f}_1, \ldots, \bar{f}_n), \tag{5}$$

where $\hat{U}_{1\sim T}$ indicates adapted low-resolution flows. During the training stage, $\hat{U}_{1\sim T}$ is supervised by real optical flows, aiming at adapting the motion to the input landscape image.

Analogously, the content stream ϕ_C also use F_{shared} as input to further learn deeper features. Subsequent decoder ϕ_D synthesize target video by combined motion and content information:

$$\hat{I}_{1\sim T} = \phi_D(\hat{U}_{1\sim T}, \phi_C(F_{shared})), \tag{6}$$

3.3 Objective Function

During the training stage of the DTVNet, we adopt content loss to monitor image quality at the pixel level, motion loss to ensure reasonable movements of the generated video, and adversarial loss to further boost video quality and authenticity. The full loss function $\mathcal{L}_{all}$ is defined as follow:

$$\mathcal{L}_{all} = \lambda_C \mathcal{L}_C + \lambda_M \mathcal{L}_M + \lambda_{adv} \mathcal{L}_{adv}, \tag{7}$$

where λ_C, λ_M, and λ_{adv} represent weight parameters to balance different terms.

Content Loss. The first term $\mathcal{L}_C$ calculates ℓ_1 errors between generated images $\hat{I}_{1\sim T}$ and real images $I_{1\sim T}$.

$$\mathcal{L}_C = \sum_{i=1}^{T} ||\hat{I}_i - I_i||_1. \tag{8}$$

Motion Loss. The second term $\mathcal{L}_M$ calculates ℓ_1 errors between adapted low-resolution flows $\hat{U}_{1\sim T}$ and real optical flows $U_{1\sim T}^{LR}$. Note that $U_{1\sim T}^{LR}$ are reconstructed low-resolution optical flow maps from $U_{1\sim T}$.

$$\mathcal{L}_M = \sum_{i=1}^{T} ||\hat{U}_i - U_i^{LR}||_1. \tag{9}$$

Adversarial Loss. The third term $\mathcal{L}_{adv}$ employ the improved WGAN with a gradient penalty for adversarial training [5]. Specifically, the discriminator D consists of six (3D-Conv)-(3D-InNorm)-(LeakyReLU) blocks that can capture discriminative spacial and temporal features.

$$\begin{aligned}\mathcal{L}_{GAN} =& \mathbb{E}_{\tilde{V}\sim p_g}[D(\tilde{V})] - \mathbb{E}_{V\sim p_r}[D(V)] + \\ & \lambda \mathbb{E}_{\hat{V}\sim p_{\hat{V}}}[(||\nabla_{\hat{V}} D(\hat{V})||_2 - 1)^2],\end{aligned} \tag{10}$$

where p_r and p_g are real and generated video distribution respectively, and $p_{\hat{x}}$ is implicitly defined by sampling uniformly along straight lines between pairs of points sampled from p_r and p_g.

3.4   Training Scheme

We first train the unsupervised optical flow estimator (FE) under the instruction of ARFlow [15], and fix its parameters once the training is complete in all experiments. When training the DTVNet, we set loss weights λ_C, λ_M, and λ_{adv} to 100, 1, and 1 respectively. The layer number n of AdaIN in the DVG module is set to 6 in the paper.

4   Experiments

In this section, many experiments are conducted to evaluate the effectiveness of the method in various aspects on the Sky Time-lapse dataset. We first qualitatively and quantitatively compare our approach with two state-of-the-art methods, *i.e.* MoCoGAN [26] and MDGAN [32], and then conduct ablation studies to illustrate the effects of the structure and loss functions of our approach. Furthermore, we make a human study to demonstrate that our method can generate high-quality and dynamic video frames. Finally, we analyze the diversified generation ability of the proposed DTVNet.

4.1 Datasets and Implementations Details

Sky Time-lapse. The Sky Time-lapse [32] dataset includes over 5000 time-lapse videos that are cut into short clips from Youtube, which contain dynamic sky scenes such as the cloudy sky with moving clouds and the starry sky with moving stars. Across the entire dataset, there are 35,392 training video clips and 2,815 testing video clips, each containing 32 frames. The original size of each frame is $3 \times 640 \times 360$, and we resize it into a square image of size $3 \times 128 \times 128$ as well as normalize the color values to [-1, 1].

Evaluation Metric. We use *Peak Signal-to-Noise Ratio* (PSNR) to evaluate the frame quality in pixel level, and *Structural Similarity* (SSIM) [30] to measure the structural similarity between synthesized and real video frames. However, these two metrics could not well evaluate the motion information of video sequences, so we introduce another metric, named Flow-MSE [13], to calculate difference of the optical flow between generated video sequences and ground truth sequences. Furthermore, we conduct a *Human Study* (HS) to evaluate the visual quality of generated video frames by real persons.

Implementation Details. We follow the training scheme described in 3.4. The OFE module inputs the flow sequence $U_{1\sim T}$ ($\in \mathbb{R}^{2\times32\times128\times128}$) and produces the normalized motion vector $\bar{f}$ ($\in \mathbb{R}^{512}$). The DVG module inputs the start frame I_0 ($\in \mathbb{R}^{3\times128\times128}$) and $\bar{f}$ to generate the flow sequence $\hat{U}_{1\sim T}$ ($\in \mathbb{R}^{2\times32\times64\times64}$) and synthesized video frames $\hat{I}_{1\sim T}$ ($\in \mathbb{R}^{3\times32\times128\times128}$). During the training stage, we use Adam [12] optimizer for all modules and set β_1=0.99, β_2=0.999. The initial learning rate is set to $3e^{-4}$, and it decays by ten every 150 epochs. We train the DTVNet for 200 epochs and the batch size is 12. We further test the inference speed of the DTVNet that can run 31 FPS with a single 1080 Ti GPU. Our model volume is 8.42M that is nearly a tenth of MDGAN [32] (78.14M). The details of the network architectures are given in the supplementary material.

4.2 Comparison with State-of-the-arts

Qualitative Results. We conduct and discuss a series of qualitative results, compared with MoCoGAN and MDGAN, on Sky Time-Lapse dataset. As shown in Figure 3, we randomly sample two videos from the test dataset with different dynamic speed (The cloud in the top half moves much faster than the bottom), and the first column shows the start frames of two videos while the second to ninth columns are generated video frames by different methods at different times. Note that the first and fifth rows are ground truth frames, and the test models of other methods are supplied by official codes.

Results show that our method can not only keep better content information than other SOTA methods (Comparing the generated results in column at a special time), but also capture the dynamic motion (Viewing the generated results in row). In detail, the generated sequences produced by MoCoGAN (second and sixth rows) become more and more distorted over time, thus the quality and motion cannot be well identified. The results produced by MDGAN
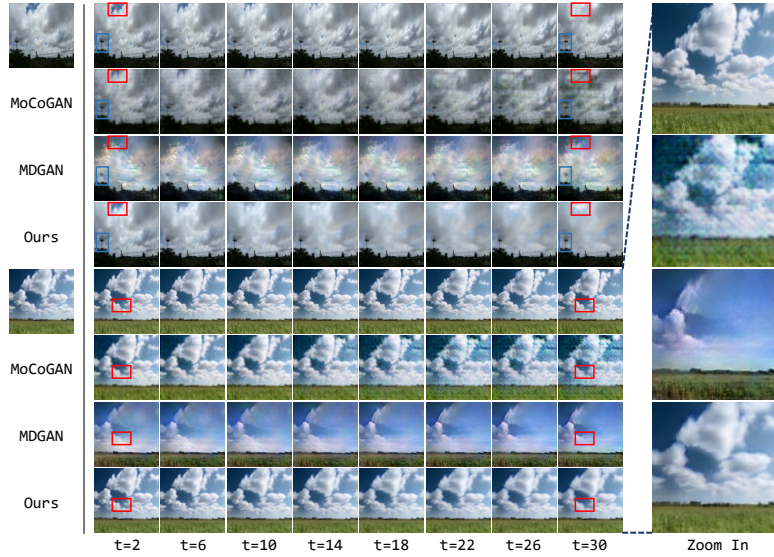
Fig. 3. Generative results compared with MoCoGAN [26] and MDGAN [32] on the Sky Time-lapse dataset. The first column lists two different landscape images as the start frames, the middle eight columns are generated video frames by different methods at different times. Right four enlarged images are long-term results at t=30 for a better visual comparison. Please zoom in red and blue rectangles for a more clear comparison.

suffer from the distortion in color and cannot well keep the content. Specifically, we mark some dynamic and still details in red and blue rectangles respectively, and results show that our method can well keep the content of the still objects while predict reasonable dynamic details, which obviously outperforms all other state-of-the-art methods.

Quantitative Results. We choose PSNR, SSIM, and Flow-MSE metrics to quantitatively evaluate the effectiveness of our proposed method on Sky Time-lapse dataset. In detail, we use all start frames in the test dataset to generate corresponding videos by different methods, and then calculate metric scores with ground truth videos.

As shown in Table 1, our approach gains +6.05 and +6.875 improvements for PSNR as well as 0.067 and 0.094 for SSIM compared to MoCoGAN and MDGAN, respectively. For the Flow-MSE metric, our method achieves the lowest value, which means that our generated video sequences are the closest to the ground truth videos in terms of the motion. On the whole, evaluation results indicate that our approach outperforms other two baselines in all three metrics, which illustrates that our model can generate more high-quality and dynamic videos than other methods.

Human Study. We further perform a human study for artificially evaluating the video frames quality. In detail, we first random choose 100 start images from

Table 1. Metric evaluation results of MoCoGAN [26], MDGAN [32], and out approach on the Sky Time-lapse dataset. The up arrow indicates that the larger the value, the better the model performance, and vice versa.

Method	PSNR ↑	SSIM ↑	Flow-MSE ↓
MoCoGAN [26]	23.867	0.849	1.365
MDGAN [32]	23.042	0.822	1.406
Ours	**29.917**	**0.916**	**1.275**

Table 2. Human study about video frames and video quality evaluation on Sky Time-lapse dataset.

| Comparison Methods	Frame Quality Score	Video Quality Score
MoCoGAN *vs.* GT	3 *vs.* 97	2 *vs.* 98
MDGAN *vs.* GT	2 *vs.* 98	1 *vs.* 99
Ours *vs.* GT	**12** *vs.* **88**	**7** *vs.* **93**
Ours *vs.* MoCoGAN	94 *vs.* 6	96 *vs.* 4
Ours *vs.* MDGAN	97 *vs.* 3	99 *vs.* 1

the test dataset and generate corresponding videos by different methods. Then the generated videos by two different methods are simultaneously shown to a real person and we can get a result indicating "which one is better". Totally, the aforementioned experiment is conduct 30 times by 30 real workers, and we take the average of all results as the final result. Specifically, we split video quality evaluation to two aspects: the single frame quality evaluation (selecting frame 16 for each video sequence) and the dynamic video evaluation, which evaluates content and motion information separately. Note that the image is shown for 2 seconds and the video is played only once for a worker.

From the comparison results shown in Table 2, our method and other two baselines have a lower quality score than ground truth images, which means that there are still many challenges for the video generation task. Nevertheless, our method has a better quality score than others: 12*vs.* 3/2 in frame and 7*vs.* 2/1 in video. To make a more intuitive comparison between our method and the others, we conduct another experiments in the bottom two rows. Results demonstrate that our approach outperforms other two SOTA methods.

4.3   Ablation Study

In this section, we conduct several ablation studies on the Sky Time-Lapse dataset to analyze the contribution of the *optical flow encoder* submodule and the effectiveness of different loss terms.

Influence of OFE. To evaluate the effectiveness of the optical flow encoder module, we conduct an ablation experiment that with or without OFE module of our proposed DTVNet. As show in Figure 4, images in the first row are ground
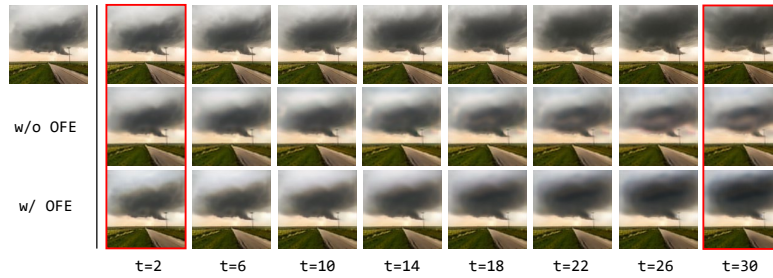
Fig. 4. A toy experiment for testing the effect of the *flow stream* of our approach. Images in the first row are ground truth frames at different times. Second and third rows are generated video frames without and with flow stream of our approach from the same start frame (the image in the first column). Please zoom in the red rectangles for a more clear comparison.

truth frames, and the second and third rows illustrate generation results without and with OFE module respectively.

By analyzing the results, we find that the generated frames loss the motion information and could not capture the clear movement and texture of the cloud without OFE module (as shown in the second row of Figure 4). In detail, the generated frames remains almost stationary and becomes more obscure as time goes on. When adding the OFE module, the model can generate high-quality and long-term dynamic video (as shown in the third row of Figure 4), which demonstrates that the OFE is critical for synthesizing photorealistic and dynamic videos. Please zoom in the red rectangles to compare long-term generation results of different structures.

Table 3. Quantitative comparisons of our approach with different loss terms on Sky Time-lapse dataset.

Method	PSNR ↑	SSIM ↑	Flow-MSE ↓
$\mathcal{L}_C$	28.768	0.897	1.624
$\mathcal{L}_C + \mathcal{L}_M$	29.364	0.906	1.509
$\mathcal{L}_C + \mathcal{L}_M + \mathcal{L}_{adv}$	**29.917**	**0.916**	**1.275**

Influence of Loss Functions. To further illustrate the effectiveness of different loss functions, *i.e.* content loss, motion loss, and adversarial loss, we conduct qualitative and quantitative experiments with different loss functions on Sky Time-lapse dataset. As shown in Figure 5, the first row indicates the ground truth of the video frames, and the other three rows are generation results under the supervision of different loss functions. Generated video frames in odd columns indicate that *motion loss* and *adversarial loss* can greatly improve the model
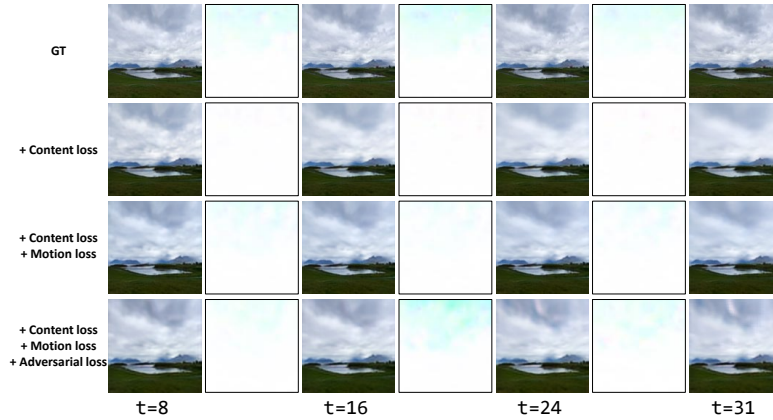
Fig. 5. Qualitative comparisons of our approach with different loss terms on Sky Time-lapse dataset. GT indicates the ground truth of the video frames. The images in even columns are the visualization results of the optical flow between two adjacent images.

performance, where the generated video frames have more clear details and a higher quality when gradually adding loss terms. *i.e.* the results in the third row are better than the second row and the results in the fourth row are better than the third row.

We further visualize the optical flow between two adjacent images in even columns, and the result shows that our approach can well learn the motion information when generating video frames. Also, we quantitatively evaluate the effectiveness of different loss functions, and obtain a similar result. As shown in Table 3, the model obtains the best metric results, *i.e.* PSNR=29.917, SSIM=0.916, and Flow-MSE=1.275, when all the three loss functions are used.

4.4 Diversified Video Generation

Diversity is a key factor in content creation. Besides generating a fixed target video from a single start frame, diversified video generation from one still landscape image is critical for practical application. In this section, we conduct an additional diversified video generation experiment to illustrate the advantage of our approach for using *normalized motion vector* to provide motion information.

In detail, different video frames are generated by DTVNet under the different motion vectors, *e.g.* encoded from the real video (denoted as **Real**) and sampled from the normalization distribution (denoted as **Random1** and **Random2**), as shown in Figure 6. We can obviously observe that our approach can generated diversified video frames by different motion vectors, which have the same content information but different motion information.
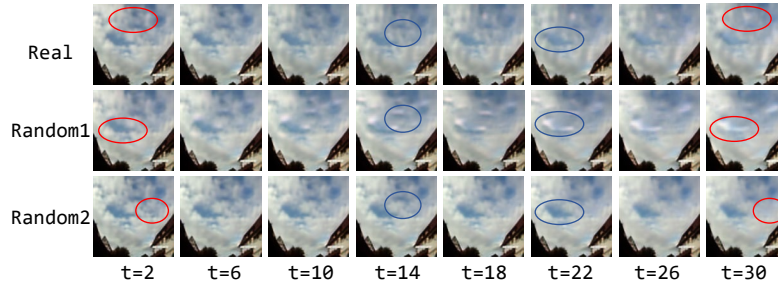
Fig. 6. Diversified video generation experiment. Given the same start frame, our approach can generate different videos under the different motion vectors, *e.g.* encoded from the real video (denoted as **Real**) and sampled from the normalization distribution (denoted as **Random1** and **Random2**). Please zoom in the red rectangles to compare each generated video in temporal and blue ellipses to compare different video frames generated by different motion vectors.

5   Conclusions

In this paper, we propose a novel end-to-end one-stage dynamic time-lapse video generation framework, *i.e.* DTVNet, to generate diversified time-lapse videos from a single landscape image. The *Optical Flow Encoder* submodule maps a sequence of optical flow maps to a *normalized motion vector* that encodes the motion information inside the generated video. The *Dynamic Video Generator* submodule contains motion and content streams that learn the movement and the texture of the generated video separately, as well as an encoder and a decoder to learn shared content features and construct target video frames respectively. During the training stage, we design three loss functions, *i.e.* content loss, motion loss, and adversarial loss, to the network, in order to generate high-quality and diversified dynamic videos. During the testing stage, we exclude the OFE module and directly sample from normalization distribution as the motion vector, which reduces the network computing overhead and supply diversified motion information simultaneously. Furthermore, extensive experiments demonstrate that our approach is capable of generating high-quality and diversified videos.

We hope our study to help researchers and users to achieve more effective works in the video generation task, and we will explore how to efficiently produce higher resolution and higher quality videos in the future.

Acknowledgements

References

1. Aigner, S., Körner, M.: Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans. arXiv preprint arXiv:1810.01325 (2018)
2. Cai, H., Bai, C., Tai, Y.W., Tang, C.K.: Deep video generation, prediction and completion of human action sequences. In: ECCV. pp. 366–382 (2018)
3. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV. pp. 2758–2766 (2015)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: NeurIPS. pp. 5767–5777 (2017)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
7. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501–1510 (2017)
8. Hur, J., Roth, S.: Iterative residual refinement for joint optical flow and occlusion estimation. In: CVPR. pp. 5754–5763 (2019)
9. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR. pp. 2462–2470 (2017)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)
11. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Flow-grounded spatial-temporal video prediction from still images. In: ECCV. pp. 600–615 (2018)
14. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion gan for future-flow embedded video prediction. In: ICCV. pp. 1744–1752 (2017)
15. Liu, L., Zhang, J., He, R., Liu, Y., Wang, Y., Tai, Y., Luo, D., Wang, C., Li, J., Huang, F.: Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In: CVPR. pp. 6489–6498 (2020)
16. Liu, P., King, I., Lyu, M.R., Xu, J.: Ddflow: Learning optical flow with unlabeled data distillation. In: AAAI. vol. 33, pp. 8770–8777 (2019)
17. Liu, P., Lyu, M., King, I., Xu, J.: Selflow: Self-supervised learning of optical flow. In: CVPR. pp. 4571–4580 (2019)
18. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)
19. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
20. Nam, S., Ma, C., Chai, M., Brendel, W., Xu, N., Kim, S.J.: End-to-end time-lapse video synthesis from a single outdoor image. In: CVPR. pp. 1409–1418 (2019)
21. Pan, J., Wang, C., Jia, X., Shao, J., Sheng, L., Yan, J., Wang, X.: Video generation from single semantic label map. In: CVPR. pp. 3733–3742 (2019)

22. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: CVPR. pp. 12240–12249 (2019)
23. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: ICCV. pp. 2830–2839 (2017)
24. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR. pp. 8934–8943 (2018)
25. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
26. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: CVPR. pp. 1526–1535 (2018)
27. Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., Lee, H.: Learning to generate long-term future via hierarchical prediction. In: ICML. pp. 3560–3569 (2017)
28. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NeurIPS. pp. 613–621 (2016)
29. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. arXiv preprint arXiv:1808.06601 (2018)
30. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
31. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: NeurIPS. pp. 82–90 (2016)
32. Xiong, W., Luo, W., Ma, L., Liu, W., Luo, J.: Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In: CVPR. pp. 2364–2373 (2018)
33. Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., Lin, D.: Pose guided human video generation. In: ECCV. pp. 201–216 (2018)
34. Zhong, Y., Ji, P., Wang, J., Dai, Y., Li, H.: Unsupervised deep epipolar flow for stationary or dynamic scenes. In: CVPR. pp. 12095–12104 (2019)
35. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)