DUAL-MTGAN: STOCHASTIC AND DETERMINISTIC MOTION TRANSFER FOR IMAGE-TO-VIDEO SYNTHESIS

[1]*Fu-En Yang (楊福恩)*, [1]*Yu-Chiang Frank Wang (王鈺強)*, [2]*Chiou-Shann Fuh (傅楸善)*

[1]Graduate Institute of Communication Engineering,

[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

E-mail: {r07942077, ycwang}@ntu.edu.tw, fuh@csie.ntu.edu.tw

ABSTRACT

Generating videos with content and motion variations is a challenging task in computer vision. While the recent development of GAN allows video generation from latent representations, it is not easy to produce videos with particular content of motion patterns of interest. In this paper, we propose Dual Motion Transfer GAN (Dual-MTGAN), which takes image and video data as inputs while learning disentangled content and motion representations. Our Dual-MTGAN is able to perform deterministic motion transfer and stochastic motion generation. Based on a given image, the former preserves the input content and transfers motion patterns observed from another video sequence, and the latter directly produces videos with plausible yet diverse motion patterns based on the input image. The proposed model is trained in an end-to-end manner, without the need to utilize pre-defined motion features like pose or facial landmarks. Our quantitative and qualitative results would confirm the effectiveness and robustness of our model in addressing such conditioned image-to-video tasks.

Index Terms— Generative Adversarial Networks, Representation Disentanglement, Image-to-Video Synthesis.

1. INTRODUCTION

Producing video sequences with desirable information is among active research topics in computer vision and machine learning. In general, a video sequence can be generated based on different corresponding conditioned inputs, such as an input image, another video sequence, or a combination of both, depending on the information of interest. In this work, we address conditional video generation given an input image, which either allows transfer of deterministic motion patterns from another video sequence of interest, or exhibits the ability to produce realistic video outputs with diverse motion patterns.

Several existing works [1, 2, 3] generate video sequences by randomly sampling latent vectors from some prior distributions (e.g., Gaussian prior), which are observed from training videos. To better manipulate video contents, recent works consider a single image [4, 5] or consecutive video frames [6, 7, 8, 9, 10, 11, 12, 13] as inputs to produce videos. Via modeling motion information in a probabilistic manner, diverse motion patterns may be observed in their output videos (e.g., [4, 9, 10]). Despite promising performances, the above methods cannot easily capture and transfer motion information from another video sequence of interest, which would limit the use of their video generation models.

Aiming at transferring motion patterns across videos, some recent works either extend Generative Adversarial Networks (GANs) [14] for video retargeting in a frame-by-frame manner[15], or leverage explicit information (e.g., pose skeleton, 3D facial model, or key-point) from video sequences of interest with input frames to perform video-to-video translation [16, 17, 18, 19, 20, 21, 22, 23, 24]. Nevertheless, the above models are specifically designed to transfer *deterministic* motion patterns from particular video inputs, and thus lack the flexibility in showing motion stochasticity in the output videos.

In this paper, we propose a *Dual Motion Transfer GAN* (Dual-MTGAN), which addresses image-to-video synthesis with joint capabilities of stochastic motion generation and deterministic motion transfer (as depicted in Figure 1). Our Dual-MTGAN disentangles input image and video data into latent representations of content and motion features, which describe visual appearance and model dynamic motion patterns, respectively. While we advance inherent temporal coherence as self-supervision for content feature disentanglement, the motion features are learned via cycle-consistency of motion features during translation. With the decomposed content/motion representations, our model animates the target image with the transferred motion patterns of the source video in a *deterministic* manner by combining the content/motion features extracted from image/video inputs. Moreover, to further exploit the *stochasticity* of motion dynamics in video sequences, the *motion latent space* encoded from source videos would fit a prior distribution. This allows out Dual-MTGAN to synthesize realistic yet diverse output videos by sampling
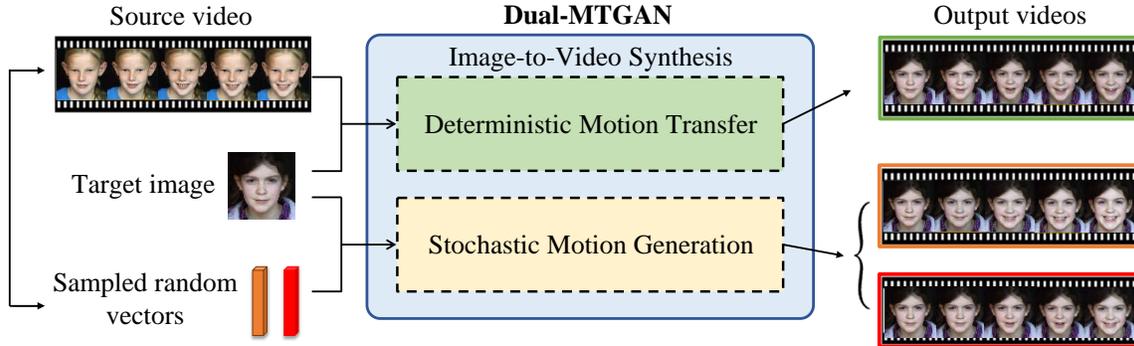
Fig. 1. Illustration of our Dual-MTGAN for image-to-video synthesis. Given a target image of interest, the goal is to either produce video outputs with motion diversity, or to generate a video sequence with motion patterns matching those of the source video. Note that motion stochasticity is modeled from source video data during training.

motion features. We note that, our Dual-MTGAN utilizes adversarial learning strategies, which further preserve the plausibility and continuity of the generated video outputs. In the experiments, we show that our model not only transfers motion patterns across videos, but is also able to synthesize realistic videos with motion diversity given a single input image.

The contributions of this paper are highlighted below:

- We address image-to-video generation with flexibility in controlling motion information. Given an input image, our proposed model allows transfer of motion patterns from video data, or synthesis of video sequences with motion diversity.

- By enforcing appearance coherence and motion consistency, our Dual-MTGAN factorizes visual latent representations into disjoint features describing content and motion features in a self-supervised manner.

- Our proposed model is simply trained by observing a sourced-domain video and a target-domain image in an end-to-end manner. No auxiliary information or supervision like facial landmarks or shape models.

2. RELATED WORKS

Stochastic Video Synthesis. Based on GAN [14] architecture, several works [1, 2, 3] generate video frames from prior distributions observed from training data. For example, Vondrick *et al.* [1] and Saito *et al.* [2] proposed VGAN and TGAN respectively, to learn a mapping between video data and the associated latent spaces. While they can produce video sequences similar to real ones, they cannot synthesize videos conditioned on the content of interest. With the aim of controlling the content of synthesis videos, recent works [4, 5] further perform image-to-video generation, which produce plausible video outputs based on a single image. For example, Li *et al.* [4] proposed a two-stage training framework incorporating generative models with optical flow supervision

to generate video frames with diversity. Nevertheless, these works typically require pre-defined motion priors like optical flow, which hampers the motion diversity and cannot be easily extended to motion transfer.

Deterministic Motion Transfer. The goal of motion transfer is to translate a source video into an output one, in which the motion information is derived from the source video while the visual appearance is preserved in the target output. Video-to-video translation is among the solutions, which directly learns mapping across videos. For example, Bansal *et al.* [15] extended CycleGAN [27] and combined spatiotemporal constraints to translate one video into another in a frame-by-frame manner. Vid2vid [16] aims to translate a sequence of semantic representation to consecutive video frames. While satisfactory results have been reported, the above methods are only applicable to translation between two domains (e.g., John Oliver to Stephen Colbert). Thus, they have limited flexibility in practical scenarios. Another group of works learned to use explicit information as prior information like skeleton or keypoints [20, 21, 22, 23] to address this task. For example, Monkey-Net [23] incorporated keypoints information to realize motion transfer and also applied the extracted keypoints to generate videos from an image. However, the above assumptions with deterministic settings cannot perform video generation with sufficient motion diversity.

Representation Disentanglement for Video Synthesis. Aiming at learning interpretable data representations [28, 29, 30, 31, 32], the idea of representation disentanglement has been applied to video generation [25, 26, 33]. For example, Tulyakov *et al.* [3] presented MoCoGAN to generate videos from decomposed random noise representing motion and content information respectively. Note that, while MoCoGAN present an extension version for image-to-video synthesis, they fail to capture and transfer the motion patterns across videos. Moreover, DRNet [26] disentangled the latent representation into content and pose, and performed motion transfer across videos. Despite promising prediction and motion transfer results, the above methods lack the ability to exhibit motion stochasticity during the transfer process, and

Table 1. Comparisons of recent works on video synthesis.

	Representation Disentanglement	Motion Transfer	Without Pre-defined Prior Information	Generate Video from an Image	Results with Diversity
MoCoGAN [3]	√	-	√	(√)	-
MCNet [25]	√	-	√	-	-
DRNet [26]	√	√	√	-	-
SVG [10]	-	-	√	√	√
Chan *et al.* [21]	-	√	-	-	-
Vid2vid [16]	-	√	-	-	-
Recycle-GAN [15]	-	√	√	-	-
FR-GAN [22]	-	√	-	√	-
Li *et al.* [4]	-	-	-	√	√
seg2vid [5]	-	-	-	√	-
Monkey-Net [23]	-	√	-	√	-
Ours (Dual-MTGAN)	√	√	√	√	√

cannot be directly applied to transfer motion across videos. In Table 1, we compare the state-of-the-art methods with ours, showing the robustness of our model.

3. PROPOSED METHOD

3.1. Problem Definition and Model Overview

With the goal of generating videos with plausible content and motion information, we propose a novel Dual-Motion Transfer GAN (Dual-MTGAN) to produce video sequences via observing a combination of source video $V = [x^t]_1^T$ with T input frames and a target image y, or only a given target image y. With the content observed from either input domain, our output video would directly transfer motion information from V or exhibit sufficiently realistic motion patterns.

As illustrated in Figure 2, our Dual-MTGAN utilizes E_C and E_M as content and motion encoders, producing time-invariant content features z_c and video dynamic motion features z_M, respectively. For dual motion transfer purposes, we generate frame-based motion features $z_{m,d}^t$ or $z_{m,s}^t$ by either directly decoding z_M via a RNN-based motion generator G_T, or utilizing feature sampled from the motion latent space as input to G_T. Moreover, the decomposed content and motion representations are learned with both data recovery and plausibility objectives, which are realized by an image generator G_I, a video-based discriminator D_V, and an image-based discriminator D_I. With such adversarial learning strategies, we are able to produce a video sequence based on the target image y, with motion patterns either encoded from the source video V or sampled from prior distributions.

3.2. Self-Supervised Disentanglement of Content/Motion Features

As illustrated in Figure 2, we have a content encoder E_C to extract content features from either source video or target image data. As the appearance in a video sequence is temporally

coherent across all frames, we leverage this property as a *self-supervision* to ensure that content features extracted from the source video are time-invariant (i.e., $z_{c,x}^t = z_{c,x}$). In other words, we enforce temporal consistency between cross-frame content features $z_{c,x}^t$ and $z_{c,x}^{t+1}$, which correspond to the loss function $\mathcal{L}_C$ defined as follows:

$$\mathcal{L}_C = ||E_C(x^t) - E_C(x^{t+1})||_1. \quad (1)$$

By enforcing the above loss, our E_C extracts time-invariant content features from input frames, and thus can be applied to extract content features from (target) images as well.

In order to encode motion information observed from video data, we utilize a separate motion encoder E_M with spatial-temporal convolutional architectures (i.e., 3D convolution) [34] in our proposed framework. Moreover, to better exploit the motion features and allow stochastic sampling during inference, we train our network module in a generative manner, and establish a motion latent space to model the distribution of inherent motion dynamics. More precisely, this is realized by enforcing the *Kullback-Leibler* divergence to encourage the distribution of extracted video motion feature z_M to fit a prior Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus, the corresponding objective function $\mathcal{L}_{KL}$ is defined as:

$$\mathcal{L}_{KL} = \mathbb{E}[KL(\mathcal{P}(z_M)||\mathcal{N}(\mathbf{0}, \mathbf{I}))]. \quad (2)$$

where $\mathcal{P}(z_M)$ denotes the distribution of z_M. Note that, with the learned motion latent space, we are able to perform dual deterministic and stochastic motion transfer by using motion representations that are either directly encoded or stochastically sampled.

To generate frame-based motion features for each output frame, we then apply recurrent neural network modules of *LSTM* [35] as our frame-based motion generator G_T, which learns the distribution of motion latent feature trajectories and outputs $z_{m,d}^t$ or $z_{m,s}^t$ across frames from encoded or sampled motion as input of G_T, respectively.

By deploying image generator G_I to synthesize the output frames via concatenating $z_{m,d}^t$ or $z_{m,s}^t$, we perform image-
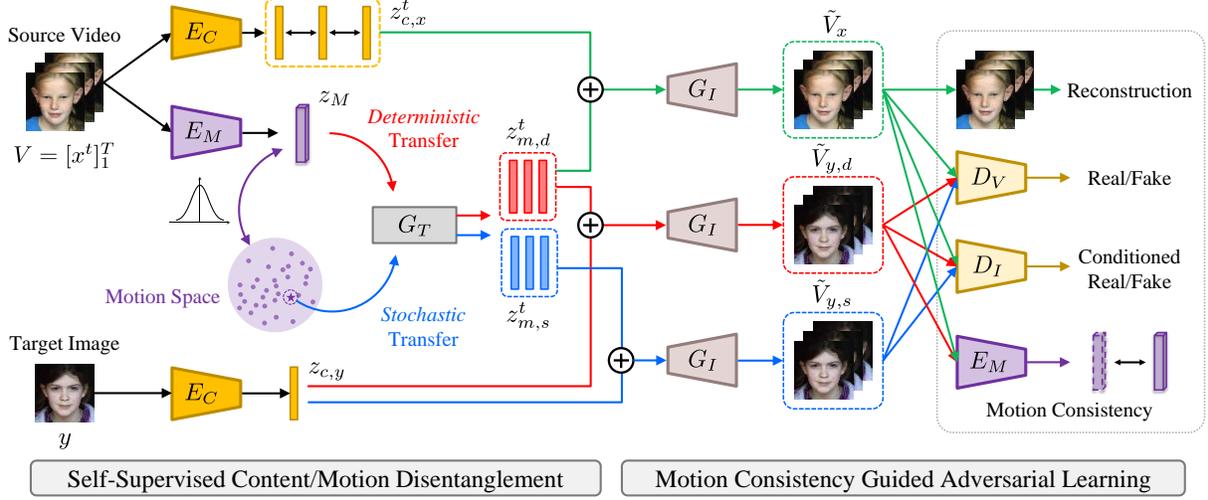
Fig. 2. The network architecture of our proposed Dual-MTGAN, which consists of a content encoder E_C, a motion encoder E_M, a RNN-based motion generator G_T, and an image generator G_I, with video and image-based discriminators D_V and D_I. Note that z_c, z_M, and z_m^t denote time-invariant content, video motion, and frame-based motion features, respectively. Note that $\oplus$ indicates concatenation of content and motion features. Best viewed in color.

to-video generation conditioned on either content vector $z_{c,x}$ (from source video) or $z_{c,y}$ (from target image). We denote the recovered video sequence from x as $\tilde{V}_x$, and the generated videos from y with encoded or sampled motion as $\tilde{V}_{y,d}$ and $\tilde{V}_{y,s}$, respectively. To be more specific, the output videos are expressed as:

$$\tilde{V}_x = G_I([z_{c,x}^t]_1^T, [z_{m,d}^t]_1^T),$$
$$\tilde{V}_{y,d} = G_I(z_{c,y}, [z_{m,d}^t]_1^T), \ \tilde{V}_{y,s} = G_I(z_{c,y}, [z_{m,s}^t]_1^T), \quad (3)$$

The plausibility of the generated output $\tilde{V}_{y,d}$ and $\tilde{V}_{y,s}$ (3) is enforced via adversarial training (detailed in the next subsection), while the recovered output $\tilde{V}_x$ in (3) can be constrained by the reconstruction loss $\mathcal{L}_{rec}$ defined as:

$$\mathcal{L}_{rec} = ||\tilde{V}_x - V||_1. \quad (4)$$

Through $\mathcal{L}_{rec}$, we not only guarantee G_I with the ability of reconstruction, but also enforce the derivation of disjoint time-variant motion information from the input video to achieve disentanglement of latent representations.

3.3. Motion-Consistency Guided Adversarial Learning

With our derived time-invariant content feature z_c and time-variant motion feature z_M, we advance adversarial learning strategies to ensure realistic video outputs with appearance guarantees. Moreover, cycle-consistency for motion representations is enforced to facilitate the preservation of learned motion information. We now describe the details below.

Appearance-Aware Visual Realism. We note that, when generating output videos with the appearance of interest, there is no guarantee that the output video would adequately satisfy the appearance information based on the given target im-

age. Hence, as shown in Figure 2, we deploy in our proposed network an image-based discriminator D_I which takes a pair of images as input, and a video-based discriminator D_V which observes an entire video. Inspired by [36], we design our D_I as a conditional discriminator by introducing an adversarial loss with appearance-aware terms to further confirm that content information can be properly encoded/recovered. This loss function encourages image-based discriminator D_I to not only distinguish the generated video outputs from real ones, but also let D_I identify the mismatch between the generated and the conditioned inputs. In this way, we guarantee the plausibility of generated video frames, while ensuring the content of the output to match the conditioned image (i.e., the first frame of the source video (x^1), or the target image (y)). For the sake of simplicity, we use $\tilde{V}_y$ to denote either one of $\tilde{V}_{y,d}$ or $\tilde{V}_{y,s}$. The objective functions for the above learning process can be defined as follows:

$$\mathcal{L}_{GAN}^I = \mathcal{L}_{GAN,x}^I + \mathcal{L}_{GAN,y}^I, \text{ where} \quad (5)$$

$$\mathcal{L}_{GAN,x}^I = \log(D_I(x^1, S_I(V))) + \frac{1}{2}[\log(1 - D_I(x^1, S_I(\tilde{V}_x)))$$
$$+ \log(1 - D_I(y, S_I(V)))]$$

$$\mathcal{L}_{GAN,y}^I = \log(D_I(y, y)) + \frac{1}{2}[\log(1 - D_I(y, S_I(\tilde{V}_y)))$$
$$+ \log(1 - D_I(x^1, S_I(V_y)))],$$

where S_I is an access function which randomly samples a frame from the output video, and V_y is the real video with the same appearance as y. We note that, in both $\mathcal{L}_{GAN,x}^I$ and $\mathcal{L}_{GAN,y}^I$, the first term indicates the "real" pairs. The second term represents the "fake" pairs, where the first subterm ensures the plausibility and the second subterm avoids content mismatch.

As for video-level adversarial learning, we apply a video-based discriminator D_V to ensure both visual quality and temporal continuity of the entire video output. Thus, the objective function of D_V is calculated as:

$$\mathcal{L}_{GAN}^V = \mathcal{L}_{GAN,x}^V + \mathcal{L}_{GAN,y}^V, \text{ where} \qquad (6)$$

$$\mathcal{L}_{GAN,x}^V = \log(D_V(V)) + \log(1 - D_V(\tilde{V}_x)),$$
$$\mathcal{L}_{GAN,y}^V = \log(D_V(V_y)) + \log(1 - D_V(\tilde{V}_y)).$$

Cycle-Consistency for Motion Features. While the quality and continuity of produced videos are enforced by the above deployment of image and video-based discriminators, they do not explicitly guarantee the disentanglement or modeling of motion information from input videos. To further enrich the derivation of motion space, we additionally advance a motion consistency constraint for preserving motion dynamics between the encoded motion vector z_M and the reconstructed video $\tilde{V}_x$ (or the generated one $\tilde{V}_{y,d}$). We calculate such consistency losses at feature level, which suggests the associated objective function $\mathcal{L}_M$ as:

$$\mathcal{L}_M = ||E_M(\tilde{V}_x) - z_M||_1 + ||E_M(\tilde{V}_{y,d}) - z_M||_1. \qquad (7)$$

3.4. Full Objectives

The full objective function of learning our Dual-MTGAN can be summarized as below:

$$\min_{E_C,E_M,G_T,G_I} \max_{D_I,D_V} \mathcal{L} = \lambda_1(\mathcal{L}_C + \mathcal{L}_{rec} + \mathcal{L}_M)$$
$$+ \lambda_{KL}\mathcal{L}_{KL} + \lambda_I\mathcal{L}_{GAN}^I + \lambda_V\mathcal{L}_{GAN}^V, \qquad (8)$$

In all our experiments, we set the hyperparameters as follows: $\lambda_1 = 10$, $\lambda_{KL} = 10^{-2}$, $\lambda_V = 1$, and $\lambda_I = 10^{-4}$.

Once the learning process is complete, our Dual-MTGAN can be applied to deterministic motion transfer and stochastic motion generation:

1) Given a source video V and a target image y, we utilize E_M and E_C to capture the motion feature z_M from V and visual content $z_{c,y}$ from y, respectively. We then derive frame-based motion features $z_{m,d}^t$ through G_T. After concatenating with $z_{c,y}$, the image generator G_I is applied to output a video with visual appearance of y and motion patterns of V.

2) With the input of a target image y, we extract its content feature z_c by E_C, with a sampled motion feature z_M for producing frame-based motion trajectories from G_T. By combining $z_{m,s}^t$ with z_c, the image generator G_I would output videos with motion diversity.

4. EXPERIMENTS

4.1. Datasets

Facial expression. The UvA-Nemo dataset [37] contains 1240 videos of several identities. Following [23], each frame

is resized to $64 \times 64 \times 3$ pixels and each video is divided into 16-frame length sequences. Also, we use 1110 videos for training and 124 for inference.

Human actions. The Weizmann Human Action dataset [38] consists of 90 videos covering 10 action categories performed by 9 people. We cropped each frame to center on the person and resize the frames to the size of $64 \times 64 \times 3$. Following [3], we split the first $\frac{2}{3}$ frames of each video for training, and the remaining $\frac{1}{3}$ frames of video for inference.

Robot pushing. The BAIR robot pushing dataset [39] is composed of videos presented by a robotic arm pushing various objects over a table. It contains 40960 videos for training and 256 videos for testing. Each frame is $64 \times 64 \times 3$ pixels and each video has 30 frames.

4.2. Deterministic Motion Transfer

4.2.1. Qualitative results and comparisons.

We demonstrate the ability of our Dual-MTGAN in realizing deterministic motion transfer across videos using *facial expression* and *human actions* data. As described in Figure 2, given a target image with a source video, our model is able to synthesize a corresponding video with motion patterns transferred from those observed in the source video. In Figure 3, we see that we were able to generate output videos whose appearances were properly preserved from the input target image, while the motion patterns were consistent to those of the source videos. Take the right part of Figure 3 for example, despite the poses of target images were different from those in the source videos (e.g., the person in the source video is waving both of his hands, while a different person in the target image is raising one hand only), our model was able to preserve the content of the target image and successfully extracted/transferred the motion pattern across videos (e.g., the hand(s) waving up or down).

In the *left* part of Figure 5, we compare our model with Monkey-Net [23] (a state-of-the-art deterministic motion transfer model) to show superior visual realism achieved by our Dual-MTGAN. Since Monkey-Net integrates keypoints information to perform motion transfer, its output may exhibit inferior visual realism when the appearance of the target image is largely different from that of the source video. A number of frames in the output of Monkey-Net contain irrational limbs resulted from keypoints extracted from the source video (highlighted in red blocks), leading to unrealistic and blurry outputs. From the above experiments, we see that our Dual-MTGAN exhibits capabilities in performing image-to-video generation with deterministic motion transfer.
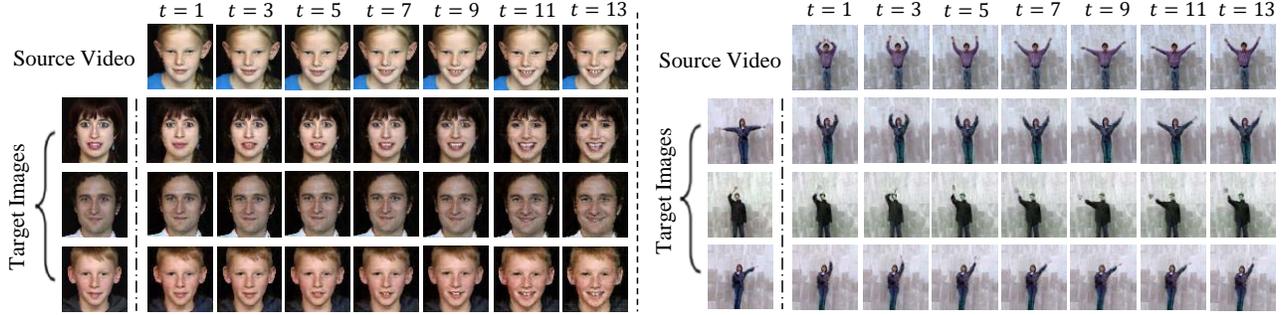
Fig. 3. Example results of deterministic motion transfer using *facial expression* (left) and *human actions* (right) video data, showing that our model produces videos with content based on the target image while the motion patterns matching those in the source video.
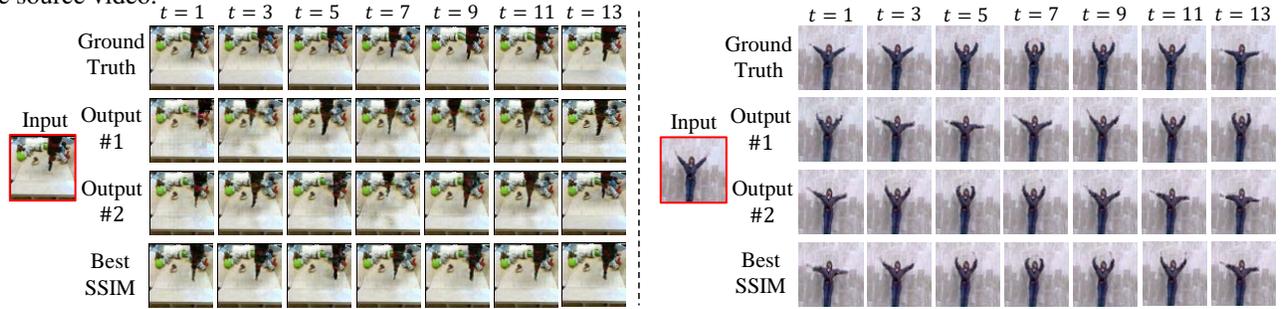
Fig. 4. Example results of stochastic motion generation using *robot pushing* (left) and *human actions* (right) video data. The frame bounded in red denotes the input frame, and each row shows ground truth or video outputs with motion stochasticity. The last row shows the video with the highest SSIM matching the ground truth video.

4.3. Stochastic Motion Generation

4.3.1. Qualitative results and comparisons.

In order to provide motion stochasticity in the produced video outputs, our Dual-MTGAN is able to generate frame-based motion features $z_{m,s}^t$ by sampling the video motion representation z_M from a prior Gaussian distribution, and then synthesizes a corresponding output video based on a single input image. Example results are shown in Figure 4 on *robot pushing* (left) and *human actions* (right) datasets respectively, in which diverse outputs can be produced based on different sampled random vectors. From the above results, we see that our model is capable of synthesizing a set of plausible videos with reasonable motion by sampling different z_M from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

To verify the plausibility of our stochastic motion generation model, we take the first frame of a video as the single input image, and compute *Structural Similarity* (SSIM) between the generated videos and the ground truth video. We show the output videos with the best SSIM in the bottom rows of Figure 4, and observe both visual and motion similarity between these outputs and their associated ground truth ones. Therefore, we confirm that our model is able to produce diverse yet realistic output videos.

As shown in the *right* part of Figure 5, we perform qualitative comparisons on *robot pushing* dataset using SVG [10], Monkey-Net [23], and our model. Our model achieved improved temporal continuity and video quality across frames

than [10] did and presented comparable visual fidelity as [23]. It is worth noting that SVG was not able to produce satisfactory results under the condition that only an input image was provided. In addition, Monkey-Net requires extra information (i.e., keypoints) for video generation, while ours is capable of synthesizing videos without the need to utilize pre-defined motion information. The quantitative comparison is available in supplementary material.

4.4. Ablation Study

We now provide ablation studies (more available in supplementary). For stochastic motion generation, we consider the *human actions* and *robot pushing* datasets, using our model 1) without D_I, 2) without D_V, 3) without $\mathcal{L}_C$, and 4) without $\mathcal{L}_M$. For each model, we produce 25 output videos, then compute the best *Structural Similarity* (SSIM) and *Peak Signal-to-Noise Ratio* (PSNR) scores with respect to the ground truth. As shown in Figure 6, our model surpassed others in terms of both metrics, which confirms the video quality achieved by Dual-MTGAN. Note that, without the presence of the D_I, both scores dropped since there would be no guarantee for preserving image quality and its property. Moreover, when D_V was disabled, both scores dropped drastically. This verifies that our D_V has the ability to ensure that the produced videos exhibit both temporal continuity and visual quality.

Next, we further disable our temporal consistency $\mathcal{L}_C$ and motion consistency $\mathcal{L}_M$ to verify the effectiveness of our de-
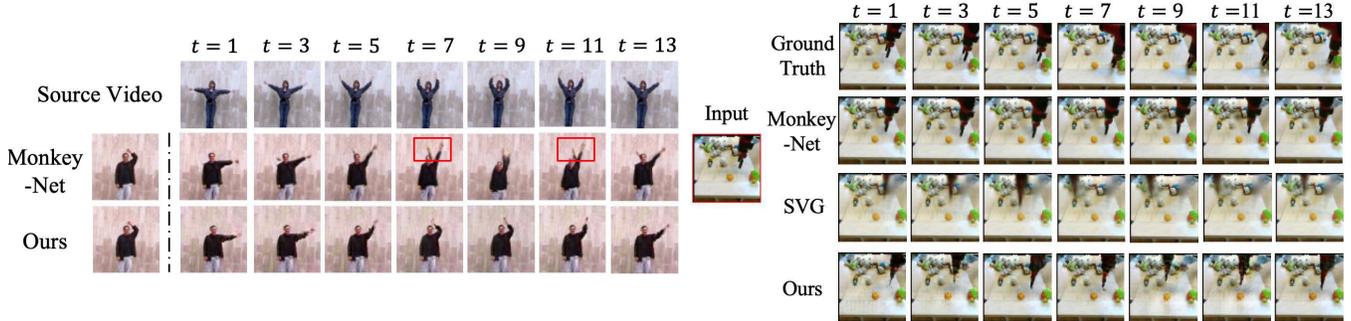
Fig. 5. Qualitative comparisons with SVG and Monkey-Net. Left part: comparisons of deterministic motion transfer; Right part: comparisons of stochastic motion generation.
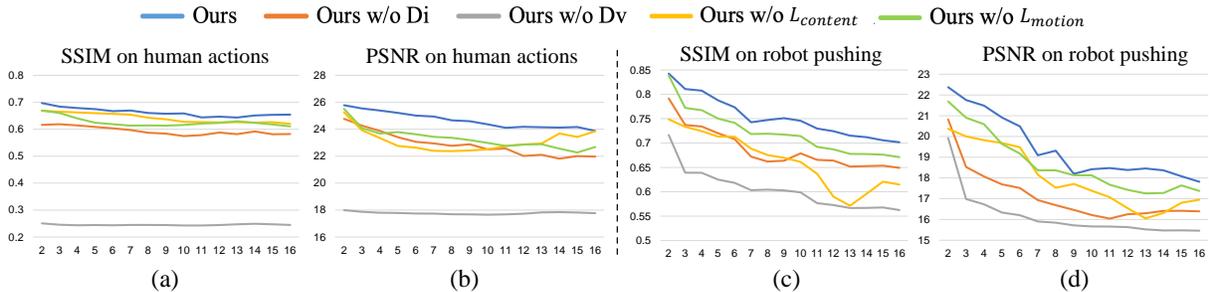
Fig. 6. Ablation studies of stochastic motion generation. Note that x and y axes indicating time step and SSIM (or PSNR) scores. (a) and (b) are evaluated on *human actions*, while (c) and (d) present results on *robot pushing* dataset.

sign for representation disentanglement. If the temporal consistency is disabled, we observe a huge drop on both scores. Without $\mathcal{L}_C$, the extracted content features are not trained to be time-invariant, and the image generator G_I would synthesize frames with the entangled representations and ignore the motion features. Hence, we cannot derive a rich motion space to produce videos with sufficiently realistic dynamics and stochasticity. Finally, if the motion consistency term is not applied, the motion information would be omitted during the generation process, and thus lead to worse visual quality and lower SSIM and PSNR scores. With the above experiments, we confirm the effectiveness and robustness of our proposed Dual-MTGAN in performing image-to-video synthesis.

5. CONCLUSION

In this paper, we proposed a unified deep learning model of Dual Motion Transfer GAN (Dual-MTGAN). This unique network addresses image-to-video synthesis, which is able to perform deterministic motion transfer or stochastic motion generation given an input image. This is realized by the design of encoders which disentangle temporal-coherent content and motion features, while the latter is modeled by a generative recurrent network modules. In our experiments, we successfully verified that our model performed promising deterministic motion transfer and stochastic motion genera-

tion results using facial expression, human actions, and robot pushing datasets with satisfactory visual quality and motion stochasticity.

REFERENCES

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

[2] Masaki Saito, Eiichi Matsumoto, and Shunta Saito, "Temporal generative adversarial nets with singular value clipping," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2830–2839.

[3] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.

[4] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang, "Flow-grounded spatial-temporal video prediction from still images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 600–615.

[5] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang, "Video generation from single semantic label map," *arXiv preprint arXiv:1903.04480*, 2019.

[6] Michael Mathieu, Camille Couprie, and Yann LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[7] Tianfan Xue, Jiajun Wu, Katherine L Bouman, and William T Freeman, "Visual dynamics: Stochastic future generation via layered cross convolutional networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[8] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee, "Learning to generate long-term future via hierarchical prediction," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3560–3569.

[9] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine, "Stochastic variational video prediction," *arXiv preprint arXiv:1710.11252*, 2017.

[10] Emily Denton and Rob Fergus, "Stochastic video generation with a learned prior," *arXiv preprint arXiv:1802.07687*, 2018.

[11] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine, "Stochastic adversarial video prediction," *arXiv preprint arXiv:1804.01523*, 2018.

[12] Yunseok Jang, Gunhee Kim, and Yale Song, "Video prediction with appearance and motion conditions," *arXiv preprint arXiv:1807.02635*, 2018.

[13] Tsun-Hsuan Wang, Yen-Chi Cheng, Chieh Hubert Lin, Hwann-Tzong Chen, and Min Sun, "Point-to-point video generation," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[15] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh, "Recycle-gan: Unsupervised video retargeting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–135.

[16] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[17] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, and Ziwei Liu, "One-shot face reenactment," *arXiv preprint arXiv:1908.03251*, 2019.

[18] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro, "Few-shot video-to-video synthesis," *arXiv preprint arXiv:1910.12713*, 2019.

[19] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9459–9468.

[20] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee, "Neural kinematic networks for unsupervised motion retargetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8639–8648.

[21] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros, "Everybody dance now," *arXiv preprint arXiv:1808.07371*, 2018.

[22] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas, "Learning to forecast and refine residual motion for image-to-video generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 387–403.

[23] Aliaksandr Siarohin, Stéphane Lathuilière, Segey Tulyakov, Elisa Ricci, and Nicu Sebe, "Animating arbitrary objects via deep motion transfer," *arXiv preprint arXiv:1812.08861*, 2018.

[24] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang, "Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis," *arXiv preprint arXiv:1911.09224*, 2019.

[25] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee, "Decomposing motion and content for natural video sequence prediction," *arXiv preprint arXiv:1706.08033*, 2017.

[26] Emily L Denton et al., "Unsupervised learning of disentangled representations from video," in *Advances in neural information processing systems*, 2017, pp. 4414–4423.

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[28] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[29] Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[30] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, "Diverse image-to-image translation via disentangled representations," *arXiv preprint arXiv:1808.00948*, 2018.

[31] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," *arXiv preprint arXiv:1804.04732*, 2018.

[32] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang, "Drit++: Diverse image-to-image translation viadisentangled representations," *arXiv preprint arXiv:1905.01270*, 2019.

[33] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles, "Learning to decompose and disentangle representations for video prediction," in *Advances in Neural Information Processing Systems*, 2018, pp. 517–526.

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[35] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[37] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *European Conference on Computer Vision*. Springer, 2012, pp. 525–538.

[38] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. IEEE, 2005, vol. 2, pp. 1395–1402.

[39] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine, "Self-supervised visual planning with temporal skip connections," *arXiv preprint arXiv:1710.05268*, 2017.