Dual-MTGAN: Stochastic and Deterministic Motion Transfer for Image-to-Video Synthesis

Fu-En Yang\(^1\), Jing-Cheng Chang\(^1\), Yuan-Hao Lee\(^1\), and Yu-Chiang Frank Wang\(^{1,2}\)
\(^1\)Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
\(^2\)ASUS Intelligent Cloud Services, Taiwan
Email: \{f07942077, b04901138, r07942074, ycwang\}@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.

I. 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\(^{[1]}\)–\(^{[3]}\) 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\(^{[6]}\)–\(^{[13]}\) 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\(^{[16]}\)–\(^{[24]}\). 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 our Dual-MTGAN to synthesize realistic yet diverse output videos by sampling motion features. We note that, our Dual-MTGAN...
works typically require pre-defined motion priors like optical flow to generate video frames with diversity. Nevertheless, these approaches incorporate generative models with optical flow supervision to generate plausible video outputs based on a single image. For example, Li et al. [4] proposed a two-stage training framework that further performs image-to-video generation, which produces 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 is required, unlike facial landmarks or shape models.

## II. Related Work

### Stochastic Video Synthesis

Based on GAN [14] architecture, several works [1]–[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 [15], [16], [20]–[23], [27]–[29]. Video-to-video translation is among the solutions, which directly learns mapping across videos. For example, Bansal et al. [15] extended CycleGAN [30] 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]–[23], [27]–[29]. Video-to-video translation is among the solutions, which directly learns mapping across videos. 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 [31]–[36], the idea of representation disentanglement has been applied to video generation [25], [26], [37]. 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 presents 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 cannot be directly applied to transfer motion across videos.

<table>
<thead>
<tr>
<th>Representation Disentanglement</th>
<th>Motion Transfer</th>
<th>Without Pre-defined Prior Information</th>
<th>Generate from an Image</th>
<th>Results with Motion Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoCoGAN [3]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MCNet [25]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DRNet [26]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SVG [10]</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Chan et al. [21]</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Vid2vid [16]</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Recycle-GAN [15]</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FR-GAN [22]</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Li et al. [4]</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>seg2vid [5]</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>X2Face [27]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Monkey-Net [23]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Ours (Dual-MTGAN)</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**TABLE I**

COMPARISONS OF RECENT WORKS ON VIDEO SYNTHESIS.
III. PROPOSED METHOD

A. 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 and motion information, we propose a novel Dual-Motion Transfer GAN (Dual-MTGAN) to produce video sequences and motion consistency guided adversarial learning.

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 $z_{m,d}$ or $z_{m,s}$ 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.

B. Self-Supervised Disentanglement of Content/Motion Representations

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., $z_{c,x} = z_{c,x}$). In other words, we enforce temporal consistency between cross-frame content features $z_{c,x}$ and $z_{c,x}^{t+1}$, which correspond to the loss function $L_C$ defined as follows:

$$L_C = ||E_C(x^t) - E_C(x^{t+1})||_1.$$  (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) [38] 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 $\mathcal{N}(0, I)$. Thus, the corresponding objective function $L_{KL}$ is defined as:

$$L_{KL} = E[KL(P(z_M)||\mathcal{N}(0, I))].$$  (2)

where $P(z_M)$ 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 [39] as our frame-based motion generator $G_T$, which learns the distribution of motion latent feature trajectories and outputs $z_{m,d}$ or $z_{m,s}$ 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 $z_{m,d}$ or $z_{m,s}$, we perform image-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 \( \hat{V}_x \), and the generated videos from \( y \) with encoded or sampled motion as \( \hat{V}_{y,d} \) and \( \hat{V}_{y,s} \), respectively. To be more specific, the output videos are expressed as:

\[
\begin{align*}
\hat{V}_x &= G_I([z_{c,x}]^T, \{z_{m,d}^t\}^T), \\
\hat{V}_{y,d} &= G_I(z_{c,y}, [z_{m,d}^t]^T), \\
\hat{V}_{y,s} &= G_I(z_{c,y}, [z_{m,s}^t]^T),
\end{align*}
\]

(3)

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

\[
L_{\text{rec}} = ||\hat{V}_x - V||_1.
\]

(4)

Through \( L_{\text{rec}} \), we not only guarantee \( G_I \) with the ability of reconstruction, but also enforce the derivation of disjoint time-vanariant motion information from the input video to achieve disentanglement of latent representations.

C. 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 image. 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 [40], 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 appearance 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 \( \bar{V}_y \) to denote either one of \( \bar{V}_{y,d} \) or \( \bar{V}_{y,s} \). The objective functions for the above learning process can be defined as follows:

\[
L_{\text{GAN}}^t = L_{\text{GAN},x}^t + L_{\text{GAN},y}^t, \quad \text{where} \quad (5)
\]

\[
L_{\text{GAN},x}^t = \log(D_I(x^1, S_I(V))) + \frac{1}{2} \log(1 - D_I(x^1, S_I(\bar{V}_y))) + \log(1 - D_I(y, S_I(V)))
\]

\[
L_{\text{GAN},y}^t = \log(D_I(y, y)) + \frac{1}{2} \log(1 - D_I(y, S_I(\bar{V}_y))) + \log(1 - D_I(\bar{x}^1, S_I(V)))]
\]

where \( S_I \) is an access function which randomly samples a frame from the output video, and \( \bar{V}_y \) is the real video with the same appearance as \( y \). We note that, in both \( L_{\text{GAN},x}^t \) and \( L_{\text{GAN},y}^t \), 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:

\[
L_{\text{GAN}}^V = L_{\text{GAN},x}^V + L_{\text{GAN},y}^V, \quad \text{where} \quad (6)
\]

\[
L_{\text{GAN},x}^V = \log(D_V(V)) + \log(1 - D_V(\bar{V}_x)),
\]

\[
L_{\text{GAN},y}^V = \log(D_V(\bar{V}_y)) + \log(1 - D_V(V)).
\]

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 \( \hat{V}_x \) (or the generated one \( \bar{V}_{y,d} \)). We calculate such consistency losses at feature level, which suggests the associated objective function \( L_M \) as:

\[
L_M = ||E_M(\hat{V}_x) - z_M||_1 + ||E_M(\bar{V}_{y,d}) - z_M||_1. \quad (7)
\]

D. Full Objectives

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

\[
\min_{E_C, E_M, G_T, G_I, D_I, D_V} \max_{\lambda_C, \lambda_M} \mathcal{L} = \lambda_1(\mathcal{L}_C + \mathcal{L}_{\text{rec}} + \mathcal{L}_M) + \lambda_{KL} \mathcal{L}_{KL} + \lambda_M L_{\text{GAN}}^t + \lambda_V L_{\text{GAN}}^V.
\]

(8)

In all our experiments, we set the hyperparameters as follows: \( \lambda_1 = 10, \lambda_{KL} = 10^{-2}, \lambda_V = 1, \) and \( \lambda_M = 10^{-4} \). Please refer to the supplementary materials for implementation details.

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 \( z_c \) 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.
IV. Experiments

A.Datasets

Facial expression. The UvA-Nemo dataset [41] 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 [42] 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 $\frac{3}{4}$ frames of each video for training, and the remaining $\frac{1}{4}$ frames of video for inference.

Robot pushing. The BAIR robot pushing dataset [43] 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.

B. Deterministic Motion Transfer

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.

C. Stochastic Motion Generation

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 $z_M^{t,s}$ 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 \( \mathcal{N}(0, 1) \).

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.

**D. Quantitative comparisons**

We now conduct quantitative comparisons on robot pushing dataset to measure both the visual realism and diversity using SSIM and LPIPS [44], respectively. We compute the best SSIM among all samples to measure the video quality, followed by [13], and evaluate the diversity of generated videos by calculating average LPIPS distance between videos generated from the same target image, where higher LPIPS distance indicates outputs with more diversity. In Table II, we show that our model achieved satisfactory SSIM score comparing with SVG and Monkey-Net, and also surpassed SVG in terms of LPIPS distance. Note that, despite Monkey-Net is able to produce videos with fidelity, it only performs deterministic video generation. Thus, the LPIPS distance for Monkey-Net is not considered in Table II. With the above comparisons, we verify that our model is capable of synthesizing results with both plausibility and stochasticity.

**E. 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 design 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.

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

ACKNOWLEDGMENT

This work is supported in part by the Ministry of Science and Technology of Taiwan under grant MOST 109-2634-F-002-037.

REFERENCES


