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

Fu-En Yang*, Jing-Cheng Chang*, Yuan-Hao Lee, Yu-Chiang Frank Wang

國立臺灣大學 National Taiwan University

ASUS | AICS

ICPR 2020

Motivation

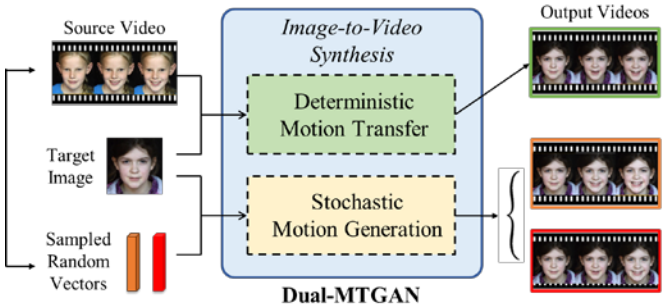
- **Image-to-Video Synthesis**

Synthesize videos from an input image with the motion of interest.

- **Contributions**

(1) Given an input image, our proposed model allows transfer of motion patterns from video data, or synthesis of video sequences with motion diversity.

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



Approach

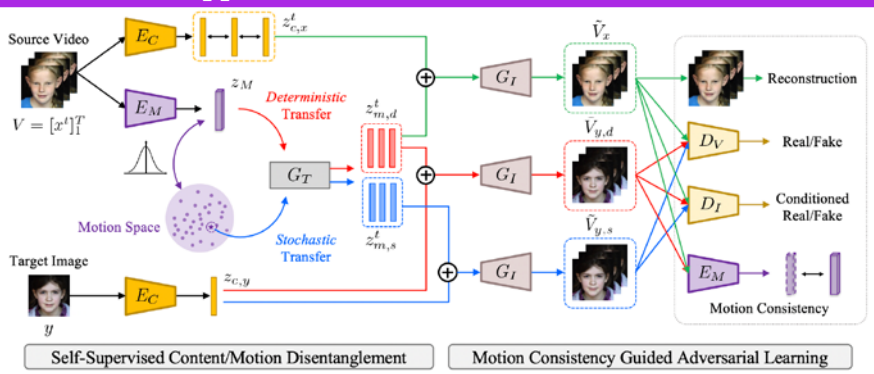
- **Self-Supervised Content/Motion Disentanglement**

- Content encoder E_C aims to extract time-invariant content features z_c by enforcing the **temporal consistency** across frames.

- Video motion feature z_M is derived from motion encoder E_M and fits Gaussian prior for generating diverse outputs via sampling during testing.

- **Motion Consistency Guided Adversarial Learning**

- *Video-level adversarial learning* ensures both video quality and temporal continuity.

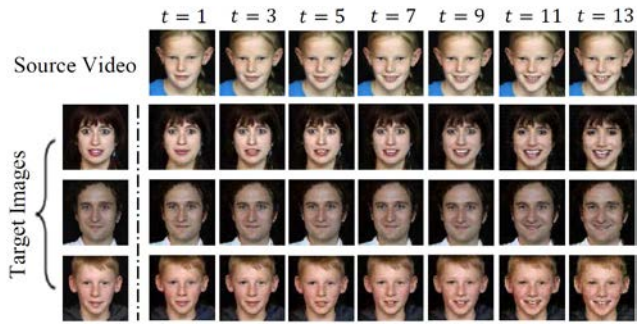


- *Image-level adversarial learning* guarantees the plausibility of synthesized frames, while ensures the appearance of output to match the conditioned image.

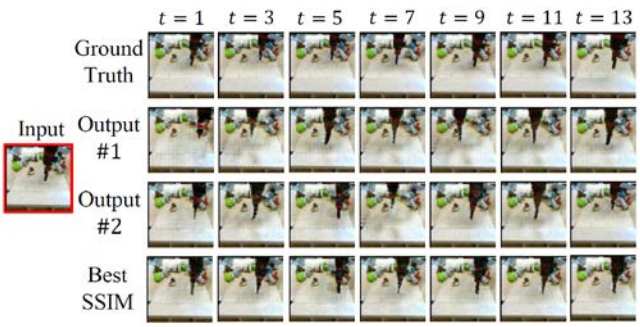
- *Motion consistency* preserves motion information z_M during training process.

Experiment Results

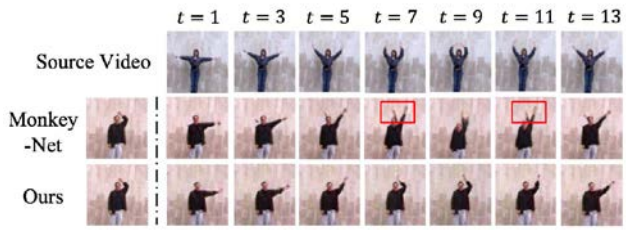
Deterministic Motion Transfer



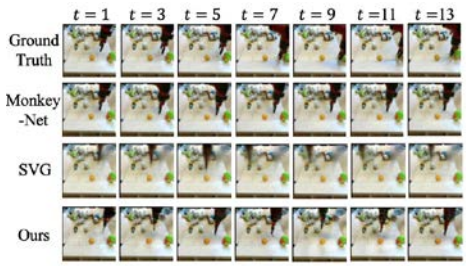
Stochastic Motion Generation



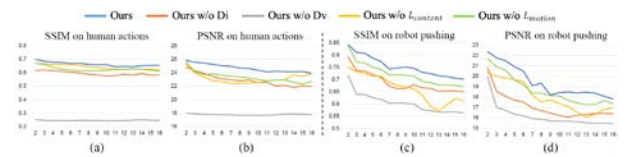
Comparisons with SOTAs



Comparisons with SOTAs



Quantitative Ablation Studies



Method	*Robot pushing*	
	SSIM (↑)	LPIPS (↑)
SVG	0.815 ± 0.006	0.0398 ± 0.0005
Monkey-Net	0.783 ± 0.008	N/A
Ours	$\mathbf{0.827} \pm 0.007$	$\mathbf{0.0422} \pm 0.0003$

- **SSIM**: measure the visual realism
- **LPIPS**: evaluate the visual diversity