國立臺灣大學
National Taiwan University

ASUS | AICS

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

ICPR 2020

[1,2]Fu-En Yang*, [1]Jing-Cheng Chang*, [1]Yuan-Hao Lee, [1,2]Yu-Chiang Frank Wang

[1]Graduate Institute of Communication Engineering, National Taiwan University

[2]ASUS Intelligent Cloud Services (AICS)

What is image-to-video synthesis?

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

Deterministic Motion Transfer

- Transfer motion pattern across different videos

Source Video

Synthesized Videos with Different Identities (Preserved Facial Expressions/Motions)

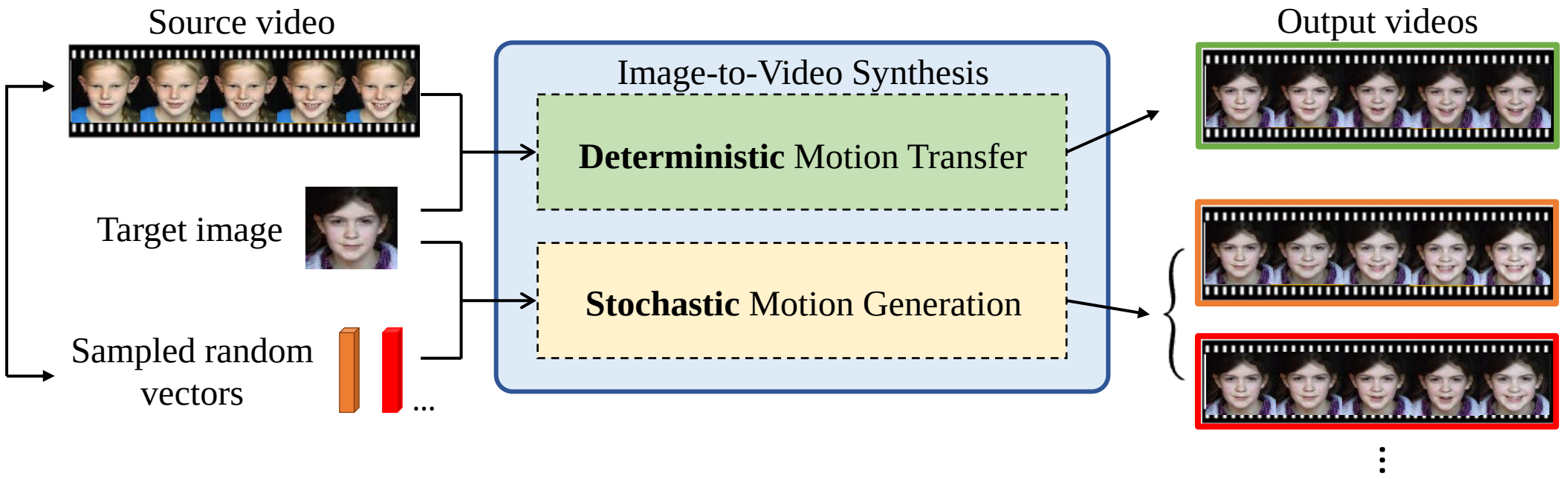
Huang *et al.* " Learning Identity-Invariant Motion Representations for Cross-ID Face Reenactment. " *CVPR* 2020

Stochastic Motion Generation

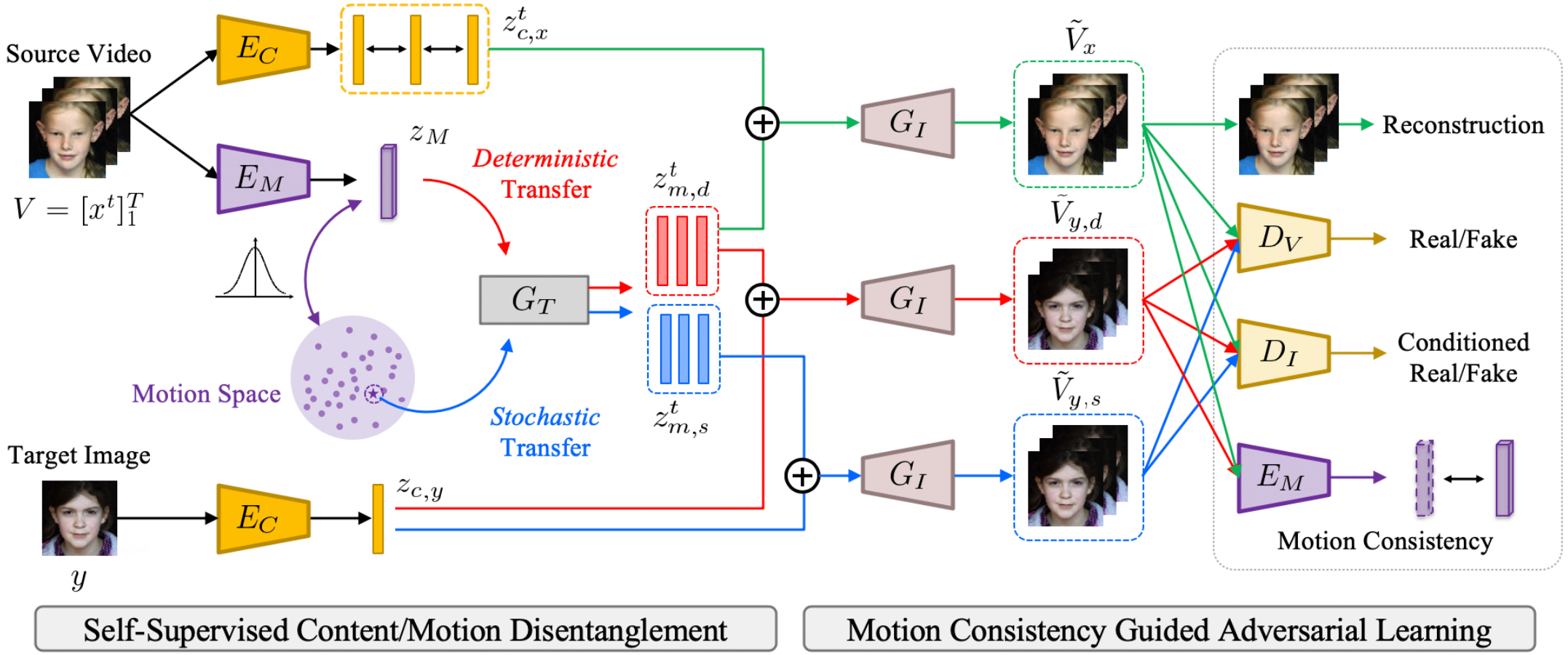
- Generate videos from an image or few frames

Pan *et al.* Video Generation from Single Semantic Label Map. ” *CVPR* 2019

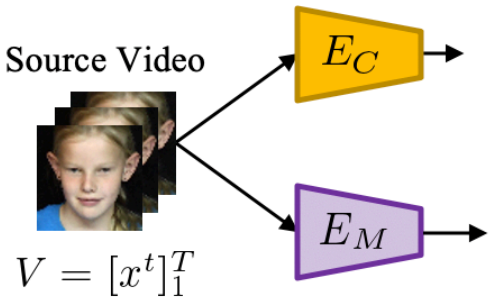
*Can we jointly perform **deterministic** and **stochastic** motion transfer in a **unified** framework?*

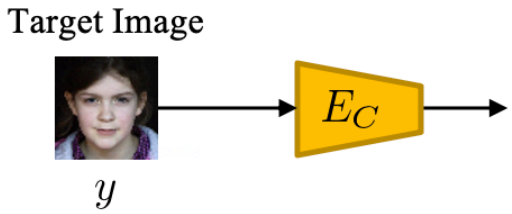


Method – Dual-Motion Transfer GAN (Dual-MTGAN)



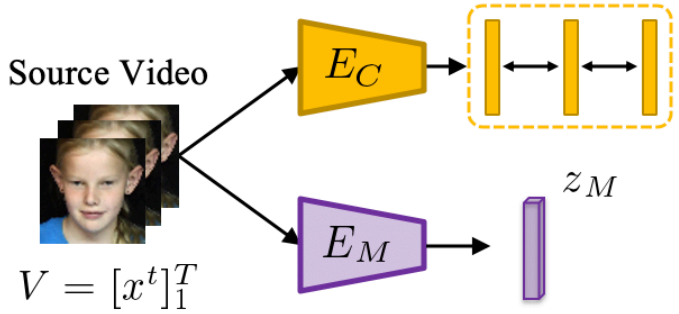
Method





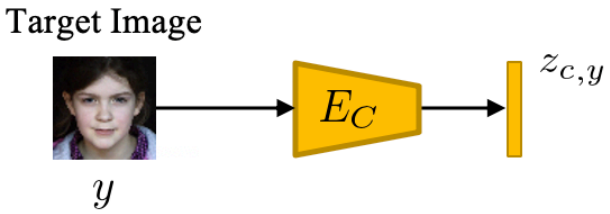
Method

– Self-Supervised Content/Motion Disentanglement



Temporal coherence across frames:

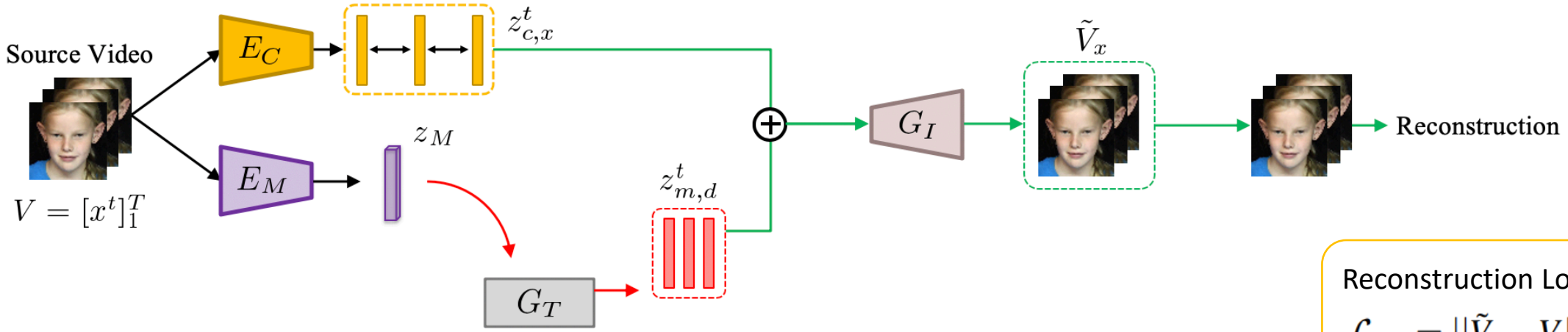
$$\mathcal{L}_C = ||E_C(x^t) - E_C(x^{t+1})||_1$$



Self-Supervised Content/Motion Disentanglement

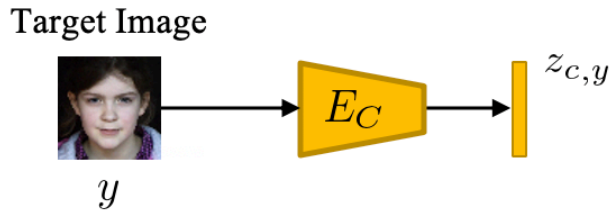
Method

– Source Video Reconstruction



Reconstruction Loss:

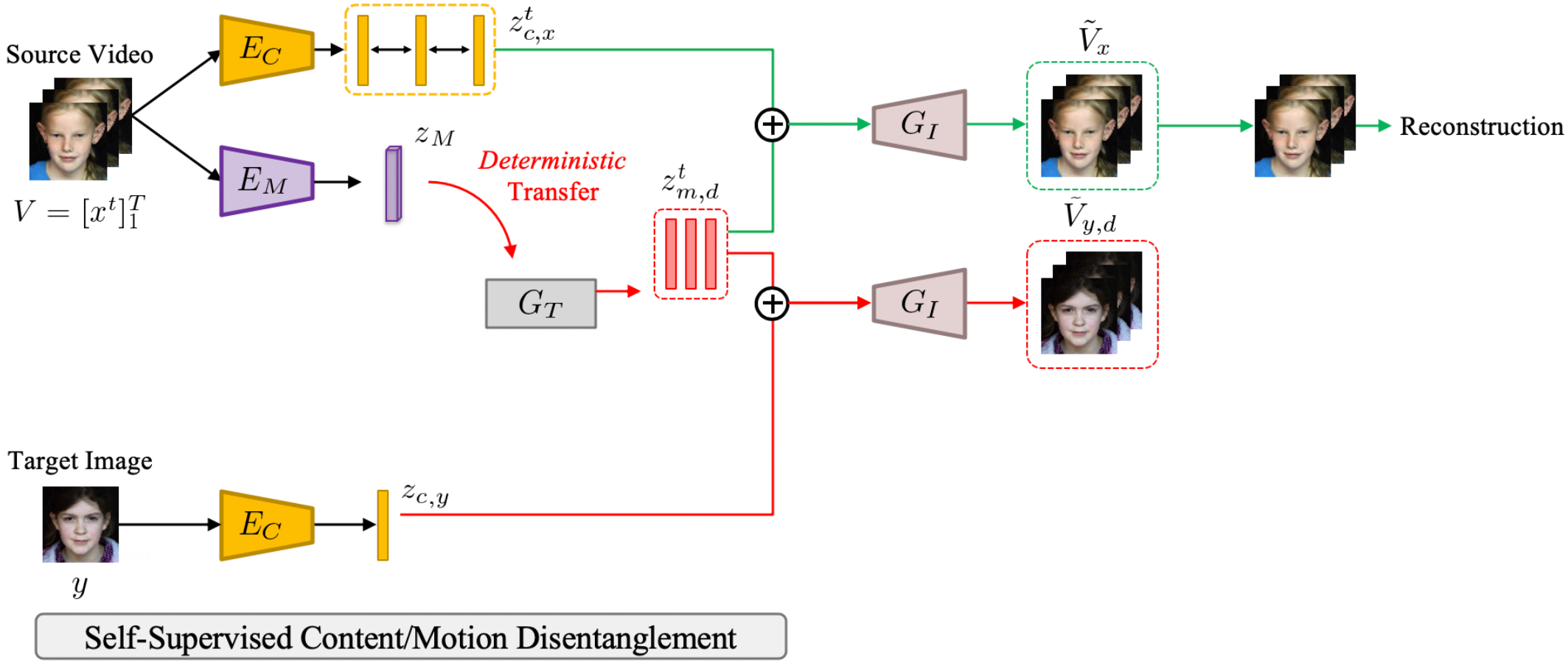
$$\mathcal{L}_{rec} = ||\tilde{V}_x - V||_1$$



Self-Supervised Content/Motion Disentanglement

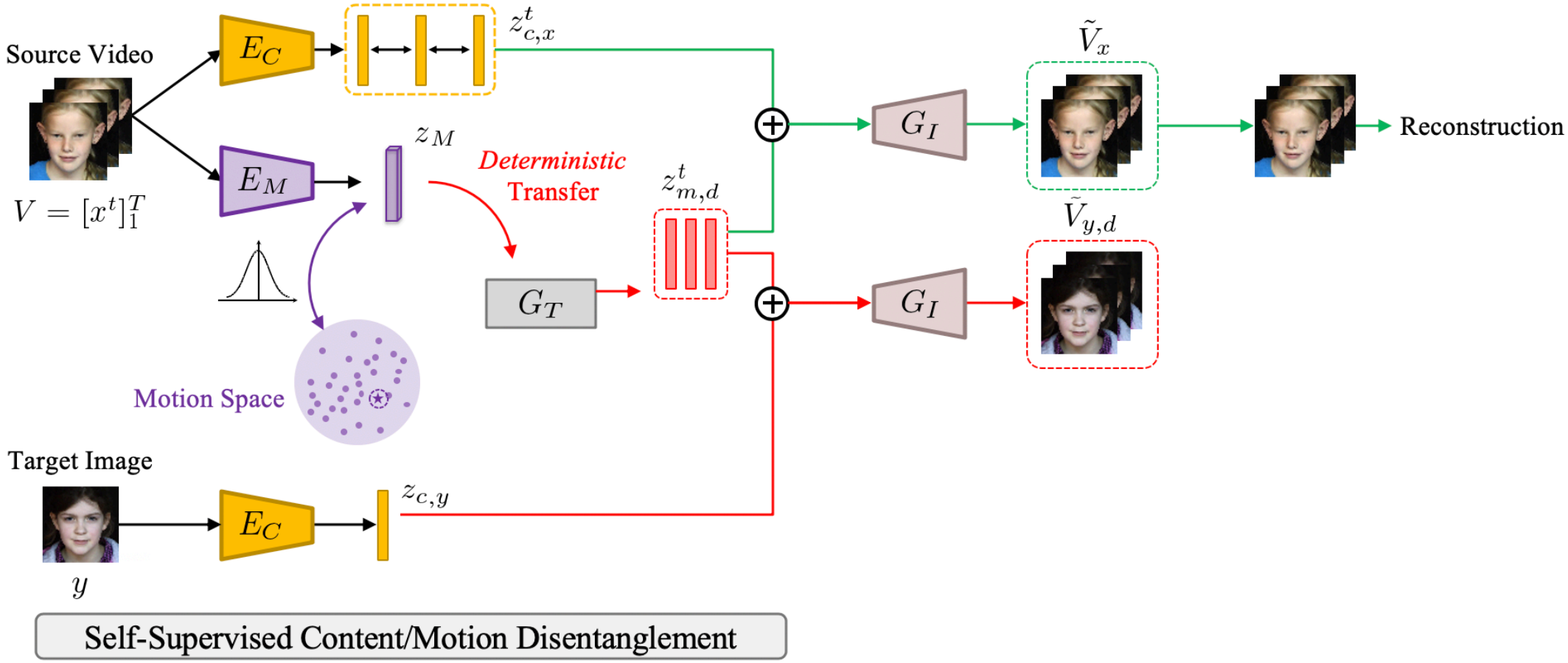
Method

– Deterministic Motion Transfer



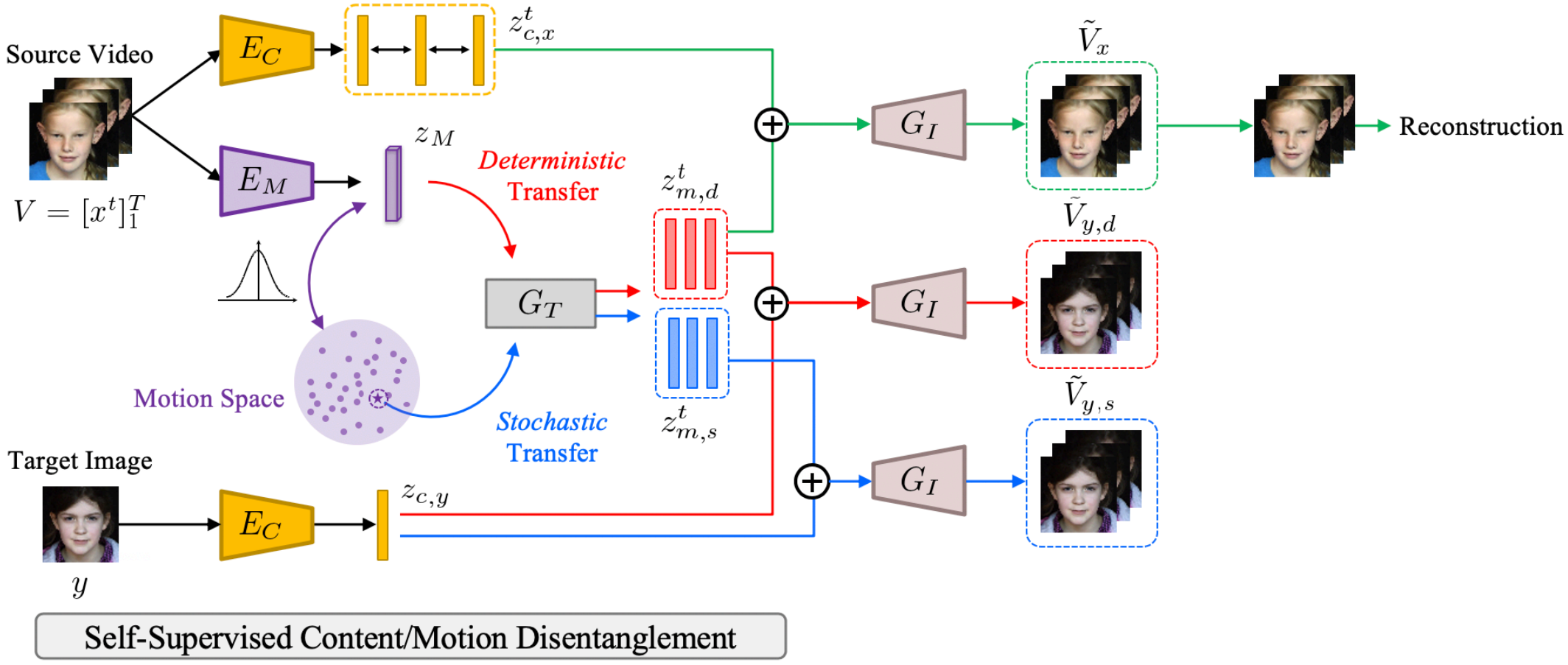
Method

– Learning Motion Latent Space



Method

– Stochastic Motion Transfer



Method

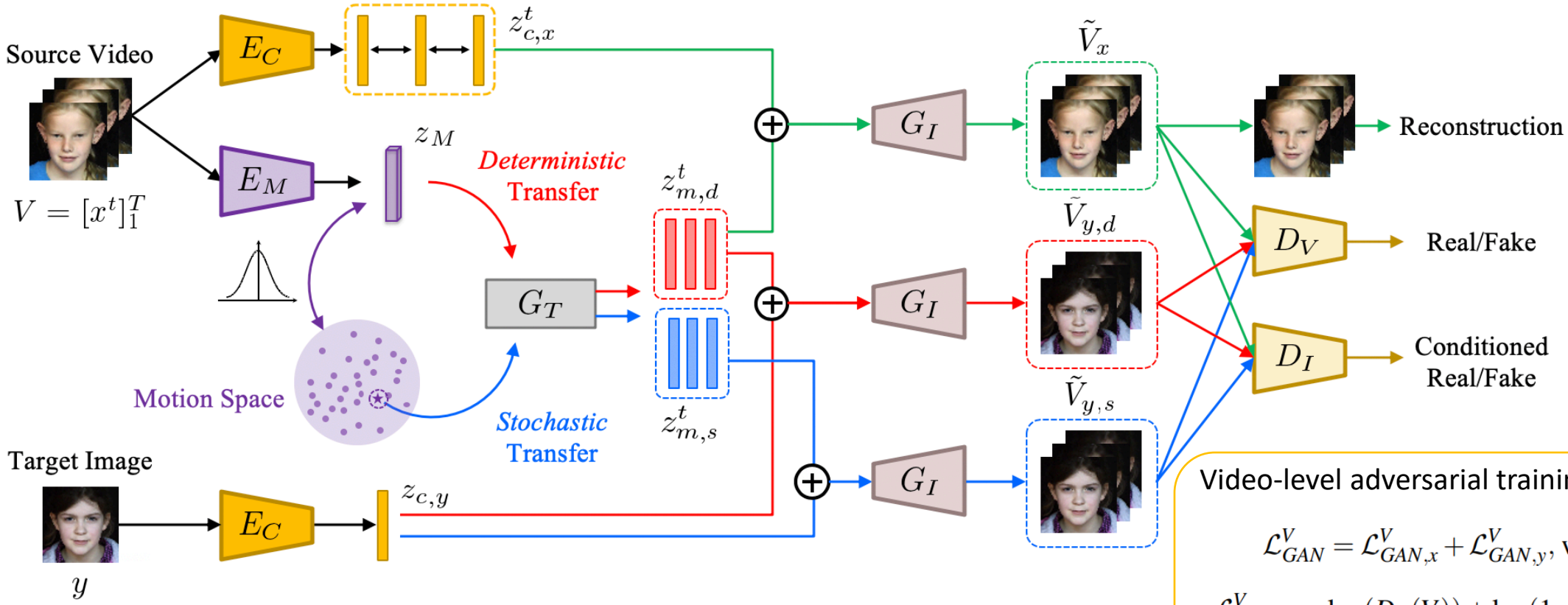
– Adversarial Learning

Image-level conditional adversarial training:

$$\mathcal{L}^I_{GAN} = \mathcal{L}^I_{GAN,x} + \mathcal{L}^I_{GAN,y}, \text{ where}$$

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

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



Video-level adversarial training:

$$\mathcal{L}^V_{GAN} = \mathcal{L}^V_{GAN,x} + \mathcal{L}^V_{GAN,y}, \text{ where}$$

$$\mathcal{L}^V_{GAN,x} = \log(D_V(V)) + \log(1 - D_V(\tilde{V}_x)),$$

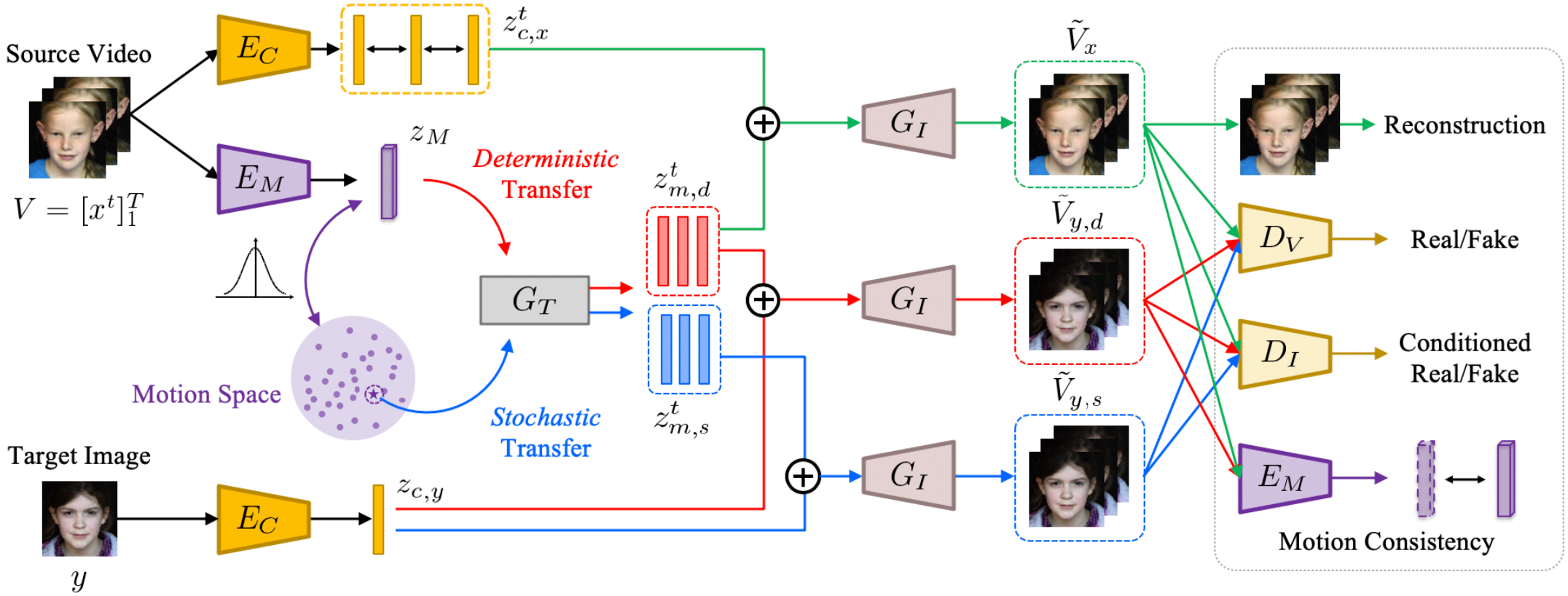
$$\mathcal{L}^V_{GAN,y} = \log(D_V(V_y)) + \log(1 - D_V(\tilde{V}_y)).$$

Method

– Motion Consistency

Motion consistency:

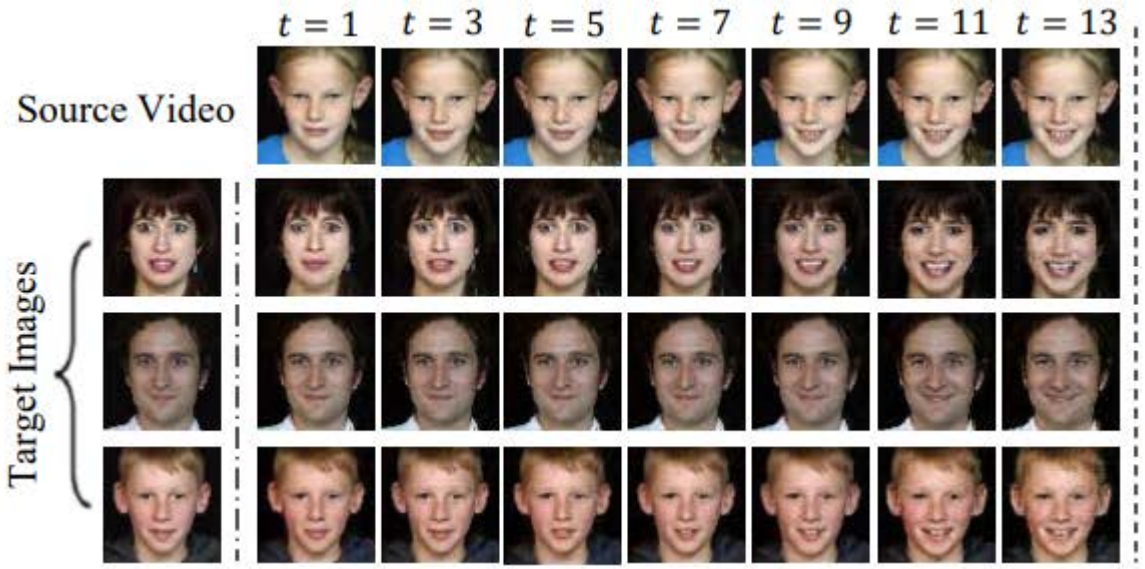
$$\mathcal{L}_M = ||E_M(\tilde{V}_x) - z_M||_1 + ||E_M(\tilde{V}_{y,d}) - z_M||_1$$



Result

– **Deterministic** Face Reenactment and Motion Retargeting

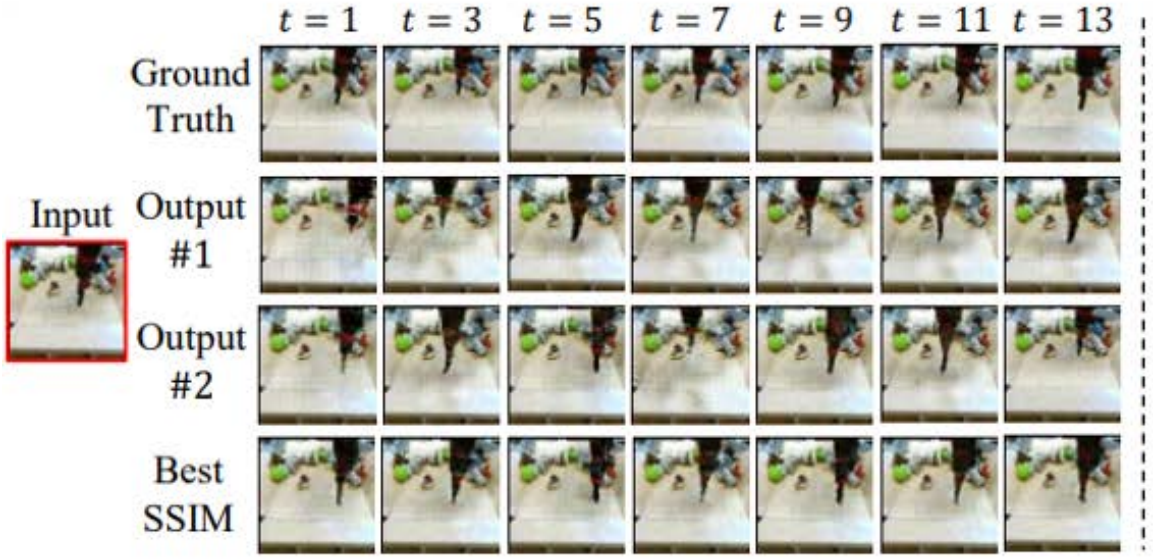
- Facial expression & Human actions

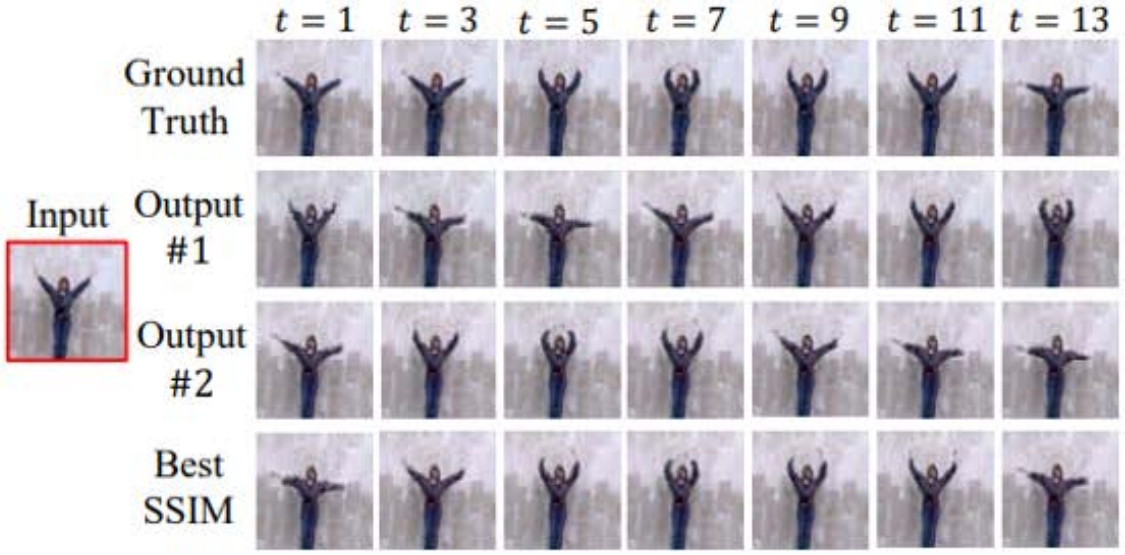


Result

– **Stochastic** Robot Movement and Action Generation

- Robot pushing & Human actions





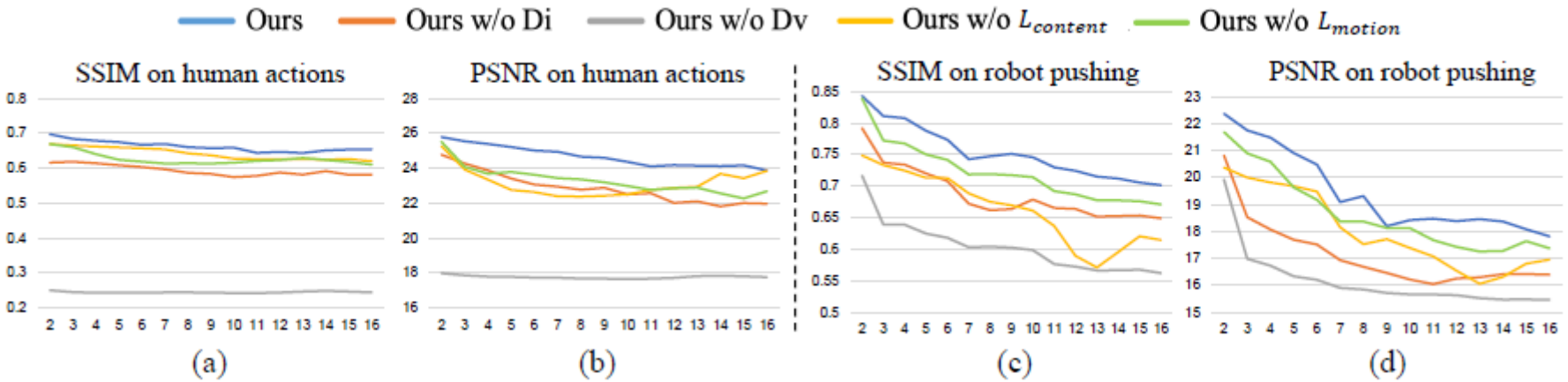
Result

– Quantitative Comparison & Ablation Study

- Quantitative Comparison

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

- Ablation Study



Conclusion

- 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 information in a self-supervised manner.

Thanks for listening!