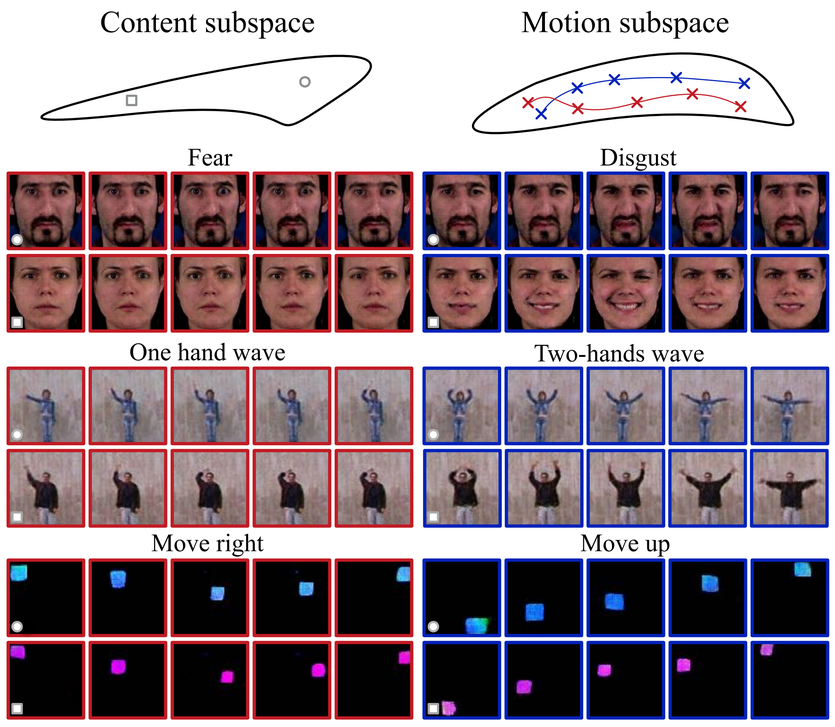
On Decomposing Motion and Content for Video Generation

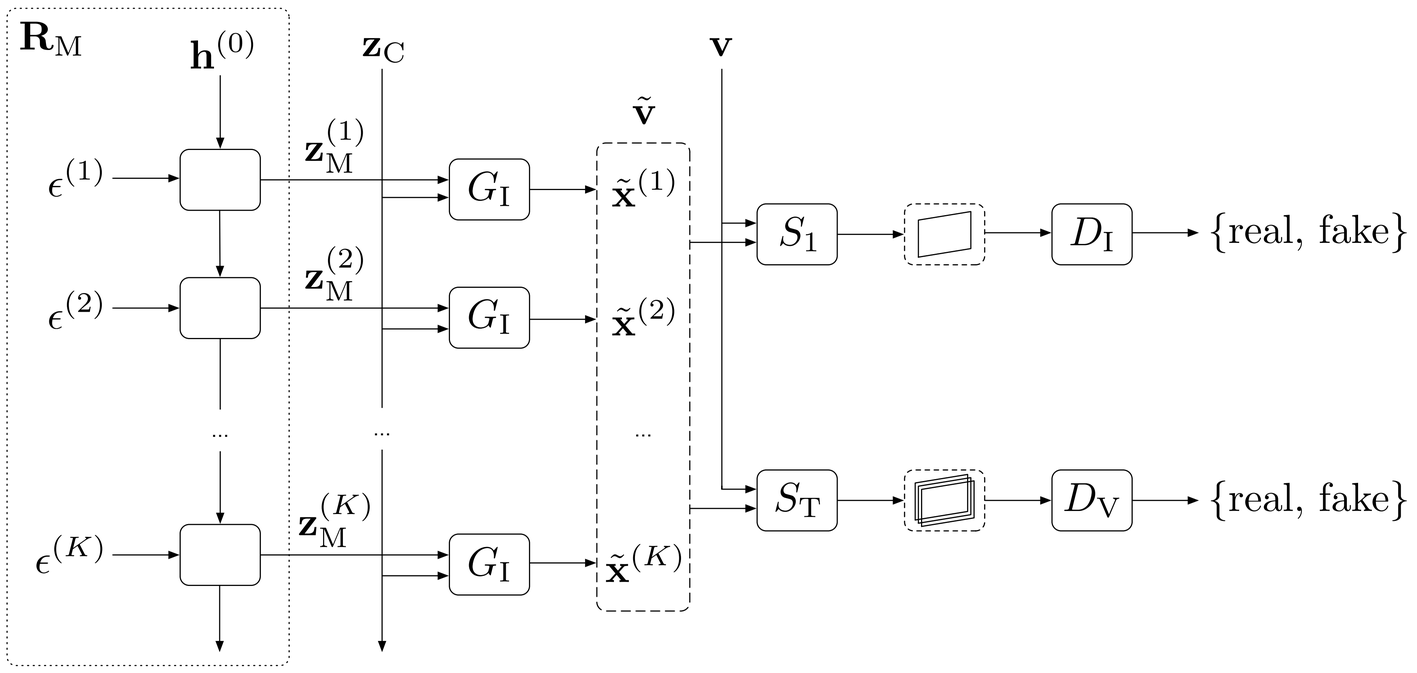
Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz

stulyakov@snap.com {mingyul,xiaodongy,jkautz}@nvidia.com

Motivation



Framework



Quantitative results

Average Content Distance

ACD	Shape Motion	Facial Expression
Reference	0	0.116
VGAN	5.02	0.322
MoCoGAN	1.79	0.201

User Study

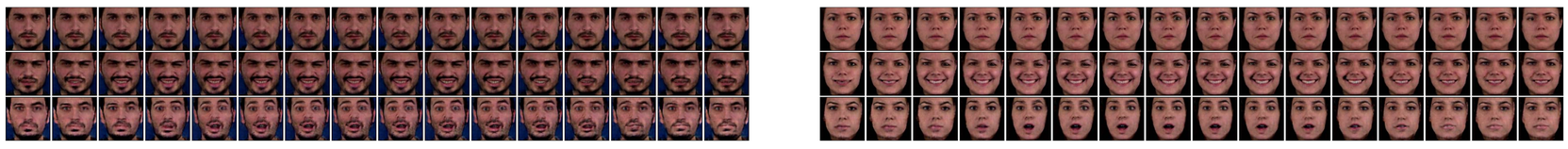
User preference, %	Facial Expression	Tai-Chi
MoCoGAN / VGAN	**84.2** / 15.8	**75.4** / 24.6
MoCoGAN / TGAN	**54.7** / 45.3	**68.0** / 32.0

Performance on Categorical Video Generation

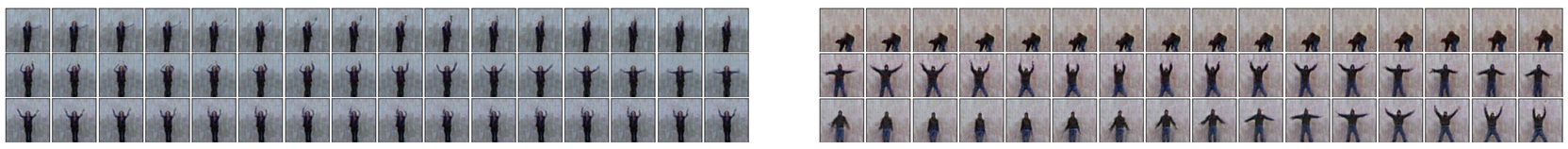
Setting	MCS	ACD
~~D_I~~ $\mathbf{z}_\mathrm{A} \to G_\mathrm{I}$	0.472	1.115
~~D_I~~ $\mathbf{z}_\mathrm{A} \to R_\mathrm{M}$	0.491	1.073
D_I $\mathbf{z}_\mathrm{A} \to G_\mathrm{I}$	0.355	0.738
D_I $\mathbf{z}_\mathrm{A} \to G_\mathrm{I}$	**0.581**	**0.606**

Categorical Video Generation Results

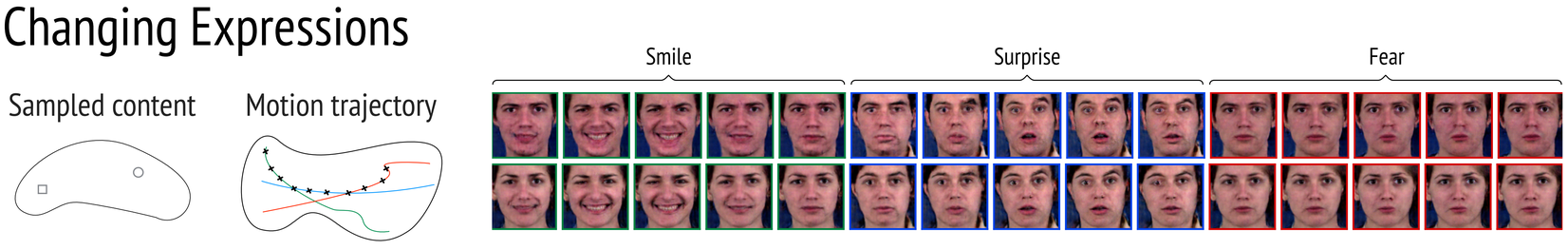
Facial Expression



Human Actions



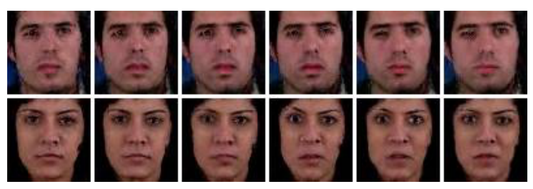
Changing Expressions



Video Generation Results

MoCoGAN

TGAN



VGAN

Content and Motion Separation

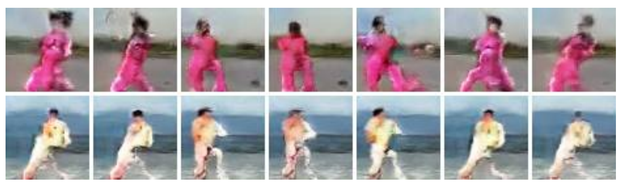
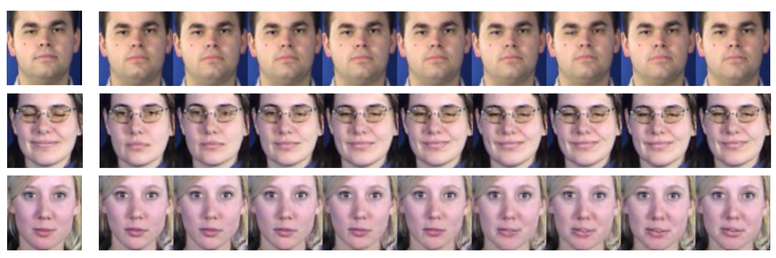


Image to Video Translation



References:

[TGAN] Saito, Masaki, Eiichi Matsumoto, and Shunta Saito. "Temporal generative adversarial nets with singular value clipping." ICCV'2017

[VGAN] Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." NIPS'2016