MoCoGAN: Decomposing Motion and Content for Video Generation

CVPR 2018

Sergey Tulyakov, Snap Research
stulyakov@snap.com

Ming-Yu Liu, Xiaodong Yang, Jan Kautz
NVIDIA
{mingyul,xiaodongy,jkautz}@nvidia.com

Motivation

- deep feature representation of data in a unsupervised manner
- generate novel data for various applications
- very good image generation models
- weather prediction, autonomous driving

https://thispersondoesnotexist.com/

Challenges

- both the appearance model and the motion model
- the time dimension brings in a huge amount of variations
- human beings are more sensitive to motion

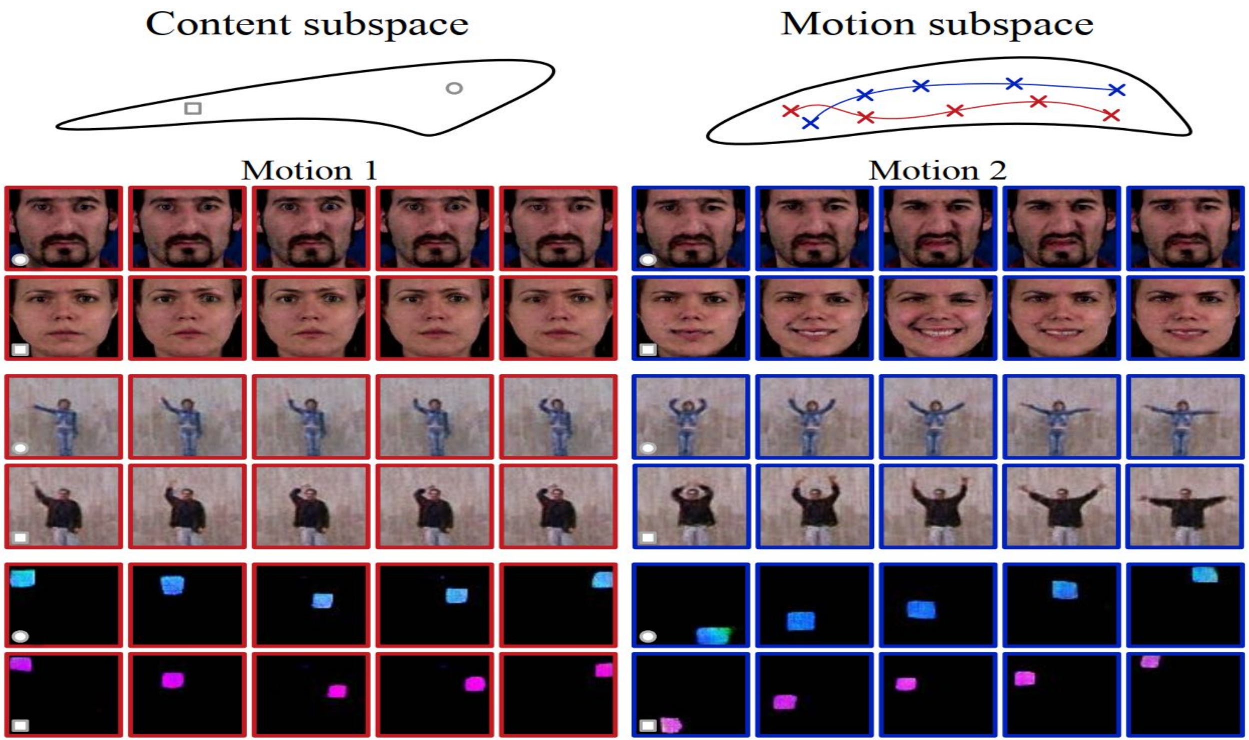
Previous approach

VGAN, TGAN: video as a point in latent space

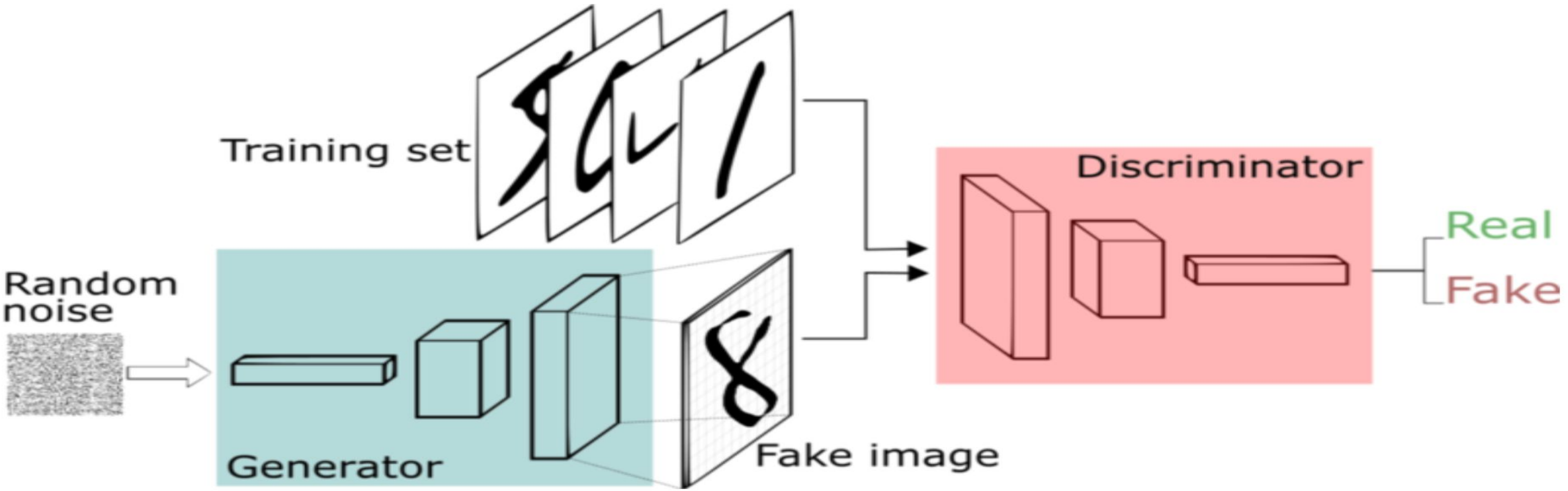
Cons:

- Complexity
- Fixed length video

Approach



GAN (Image)



$$\max_{G_I} \min_{D_I} \mathcal{F}_I(D_I, G_I)$$

$$\mathcal{F}_I(D_I, G_I) = \mathbb{E}_{\mathbf{x} \sim p_X}[-\log D_I(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{Z_I}}[-\log(1 - D_I(G_I(\mathbf{z})))]$$

MoCoGAN

Latent space: $Z_{\rm I} = Z_{\rm C} \times Z_{\rm M}$ $Z_{\rm C} = \mathbb{R}^{d_{\rm C}}, Z_{\rm M} = \mathbb{R}^{d_{\rm M}}$ $d = d_{\rm C} + d_{\rm M}$

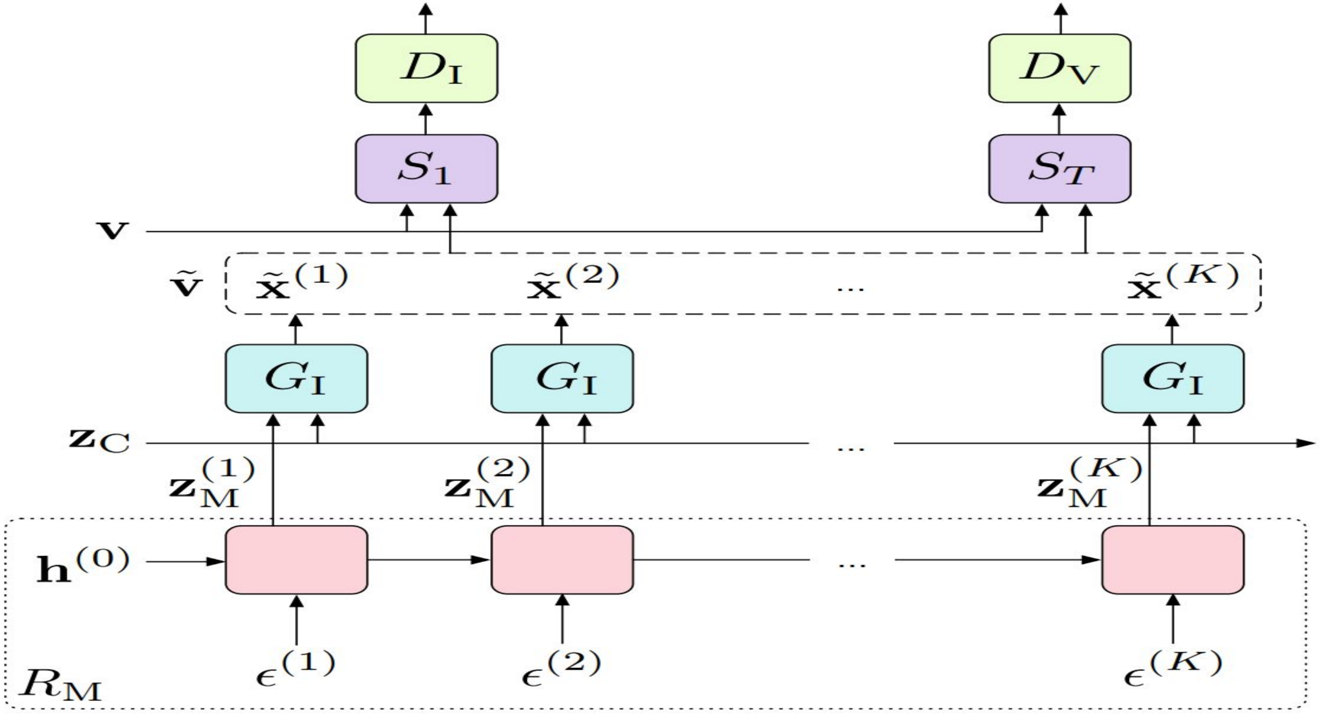
$$[\mathbf{z}^{(1)}, ..., \mathbf{z}^{(K)}] = \left[\left[\begin{array}{c} \mathbf{z}_{\rm C} \\ \mathbf{z}_{\rm M}^{(1)} \end{array} \right], ..., \left[\begin{array}{c} \mathbf{z}_{\rm C} \\ \mathbf{z}_{\rm M}^{(K)} \end{array} \right] \right]$$

Content subspace: $\mathbf{z}_{\rm C} \sim p_{Z_{\rm C}} \equiv \mathcal{N}(\mathbf{z}|0, I_{d_{\rm C}})$

Motion subspace: $[\epsilon^{(1)}, ..., \epsilon^{(K)}]$ —RNN→ $Z_M(1), \ldots, Z_M(K)$

$$\epsilon^{(k)} \sim p_{\rm E} \equiv \mathcal{N}(\epsilon|0, I_{d_{\rm E}})$$

MoCoGAN



D_I
D_V
S_1
S_T
$\mathbf{v}$
$\tilde{\mathbf{v}}$
$\tilde{\mathbf{x}}^{(1)}$
$\tilde{\mathbf{x}}^{(2)}$
...
$\tilde{\mathbf{x}}^{(K)}$
G_I
G_I
G_I
$\mathbf{z}_C$
...
$\mathbf{z}_M^{(1)}$
$\mathbf{z}_M^{(2)}$
$\mathbf{z}_M^{(K)}$
$\mathbf{h}^{(0)}$
...
$\epsilon^{(1)}$
$\epsilon^{(2)}$
$\epsilon^{(K)}$
R_M

MoCoGAN

G_I	Configuration
Input	$[\mathbf{z}_a \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{z}_m \sim R_M]$
0	DCONV-(N512, K6, S0, P0), BN, LeakyReLU
1	DCONV-(N256, K4, S2, P1), BN, LeakyReLU
2	DCONV-(N128, K4, S2, P1), BN, LeakyReLU
3	DCONV-(N64, K4, S2, P1), BN, LeakyReLU
4	DCONV-(N3, K4, S2, P1), BN, LeakyReLU

D_I	Configuration
Input	height × width × 3
0	CONV-(N64, K4, S2, P1), BN, LeakyReLU
1	CONV-(N128, K4, S2, P1), BN, LeakyReLU
2	CONV-(N256, K4, S2, P1), BN, LeakyReLU
3	CONV-(N1, K4, S2, P1), Sigmoid

D_V	Configuration
Input	16 × height × width × 3
0	CONV3D-(N64, K4, S1, P0), BN, LeakyReLU
1	CONV3D-(N128, K4, S1, P0), BN, LeakyReLU
2	CONV3D-(N256, K4, S1, P0), BN, LeakyReLU
3	CONV3D-(N1, K4, S1, P0), Sigmoid

MoCoGAN

$$\max_{G_\mathrm{I}, R_\mathrm{M}} \min_{D_\mathrm{I}, D_\mathrm{V}} \mathcal{F}_\mathrm{V}(D_\mathrm{I}, D_\mathrm{V}, G_\mathrm{I}, R_\mathrm{M})$$

$$\mathbb{E}_{\mathbf{v}}[-\log D_\mathrm{I}(S_1(\mathbf{v}))] + \mathbb{E}_{\tilde{\mathbf{v}}}[-\log(1 - D_\mathrm{I}(S_1(\tilde{\mathbf{v}})))] +$$
$$\mathbb{E}_{\mathbf{v}}[-\log D_\mathrm{V}(S_\mathrm{T}(\mathbf{v}))] + \mathbb{E}_{\tilde{\mathbf{v}}}[-\log(1 - D_\mathrm{V}(S_\mathrm{T}(\tilde{\mathbf{v}})))],$$

Categorical dynamics: $$\left[\begin{bmatrix} \mathbf{z}_\mathrm{A} \\ \epsilon^{(1)} \end{bmatrix}, \ldots, \begin{bmatrix} \mathbf{z}_\mathrm{A} \\ \epsilon^{(K)} \end{bmatrix}\right]$$

$$\mathcal{F}_\mathrm{V}(D_\mathrm{I}, D_\mathrm{V}, G_\mathrm{I}, R_\mathrm{M}) + \lambda L_I(G_\mathrm{I}, Q)$$

Dataset

Name	Number of videos
Shape motion	4000
MUG Facial Expression	1254
Tai-Chi	4500
Weizmann Human Actions	81
UCF101	13 220

Video Generation Performance

ACD	Shape Motion	Facial Expressions
Reference	0	0.116
VGAN [40]	5.02	0.322
TGAN [30]	2.08	0.305
MoCoGAN	**1.79**	**0.201**

ACD: Average Content Distance

L2 distance between average color vectors (Shape motion)

L2 distance between OpenFace feature vector (Facial expression)

Video Generation Performance

Inception score

	VGAN	TGAN	MoCoGAN
UCF101	$8.18 \pm .05$	$11.85 \pm .07$	$\mathbf{12.42 \pm .03}$

Facial expression and Tai-Chi datasets

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

Inception score:

- Images have variety
- Each image distinctly looks like something

Qualitative evaluation

Categorical Video Generation

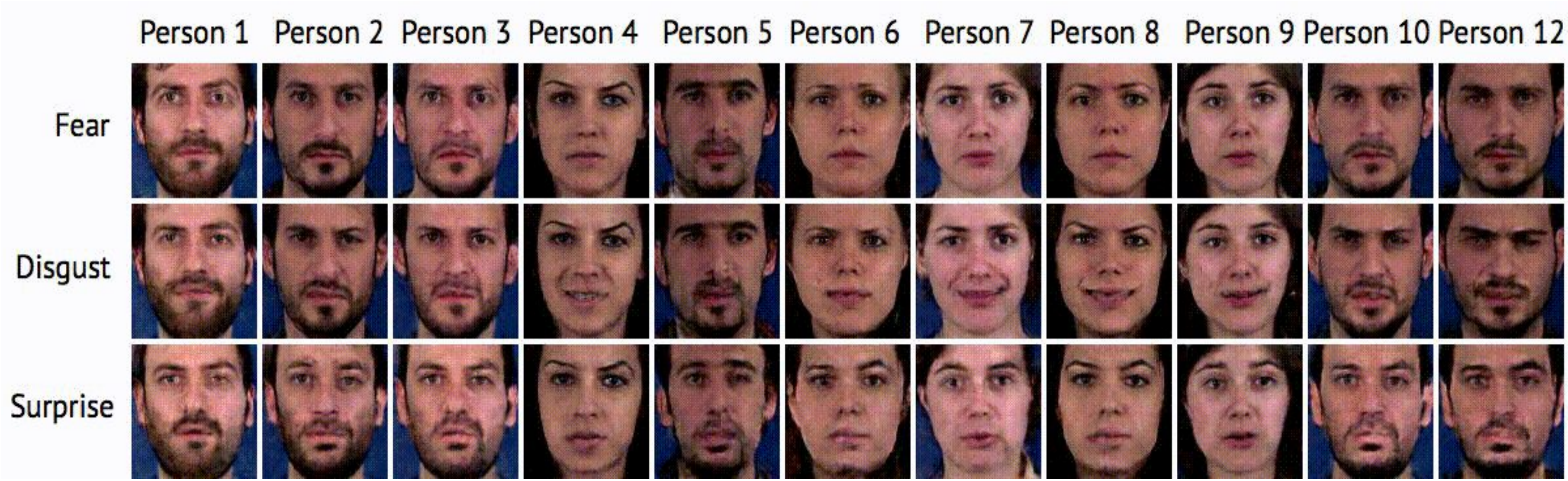


Image-to-video Translation

User preference, %	Tai-Chi
MoCoGAN / C-VGAN	**66.9** / 33.1
MoCoGAN / MCNET	**65.6** / 34.4

Strengths

- A Novel GAN framework for video generation.
- Can control content and motion in video generation.
- Several experiments with multiple datasets.

Weaknesses

- Small and medium sized dataset.
- Assuming there is a fixed content for the whole video.
- Does not work well with bigger dataset (Kinetics or even UCF101).

Discussion questions

- Why do we need image discriminator?
- Do you think the assumption of fixed content for the video is reasonable? Can we learn everything from data?

Thank you
Questions?