Video Synthesis

Sergey Tulyakov

Motivation

Imagination is more important than knowledge…

Face

Plant

Room

Karras et al. "Progressive growing of GANs for improved quality, stability, and variation", 2018

Motivation

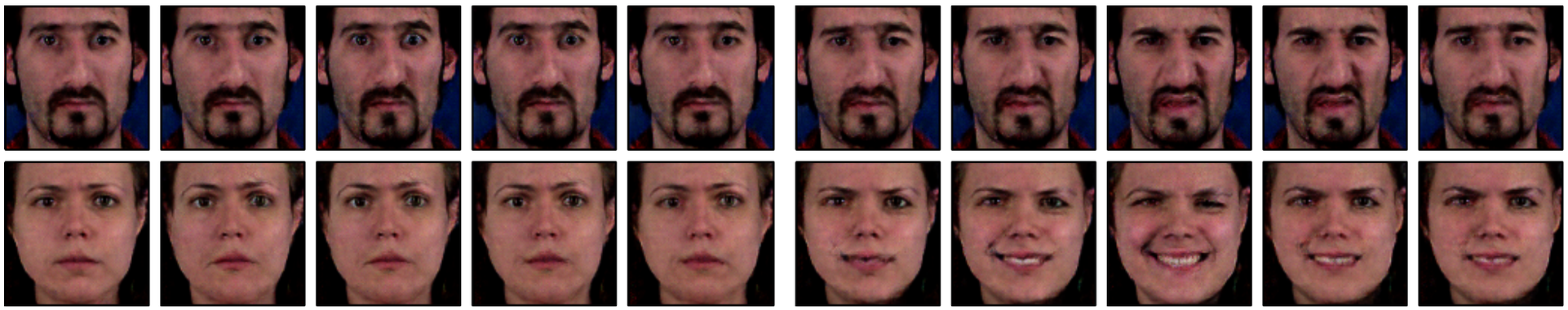
Likewise we can imagine actions

Video Generation

$$x \sim p(x)$$

Fear

Disgust



One hand wave

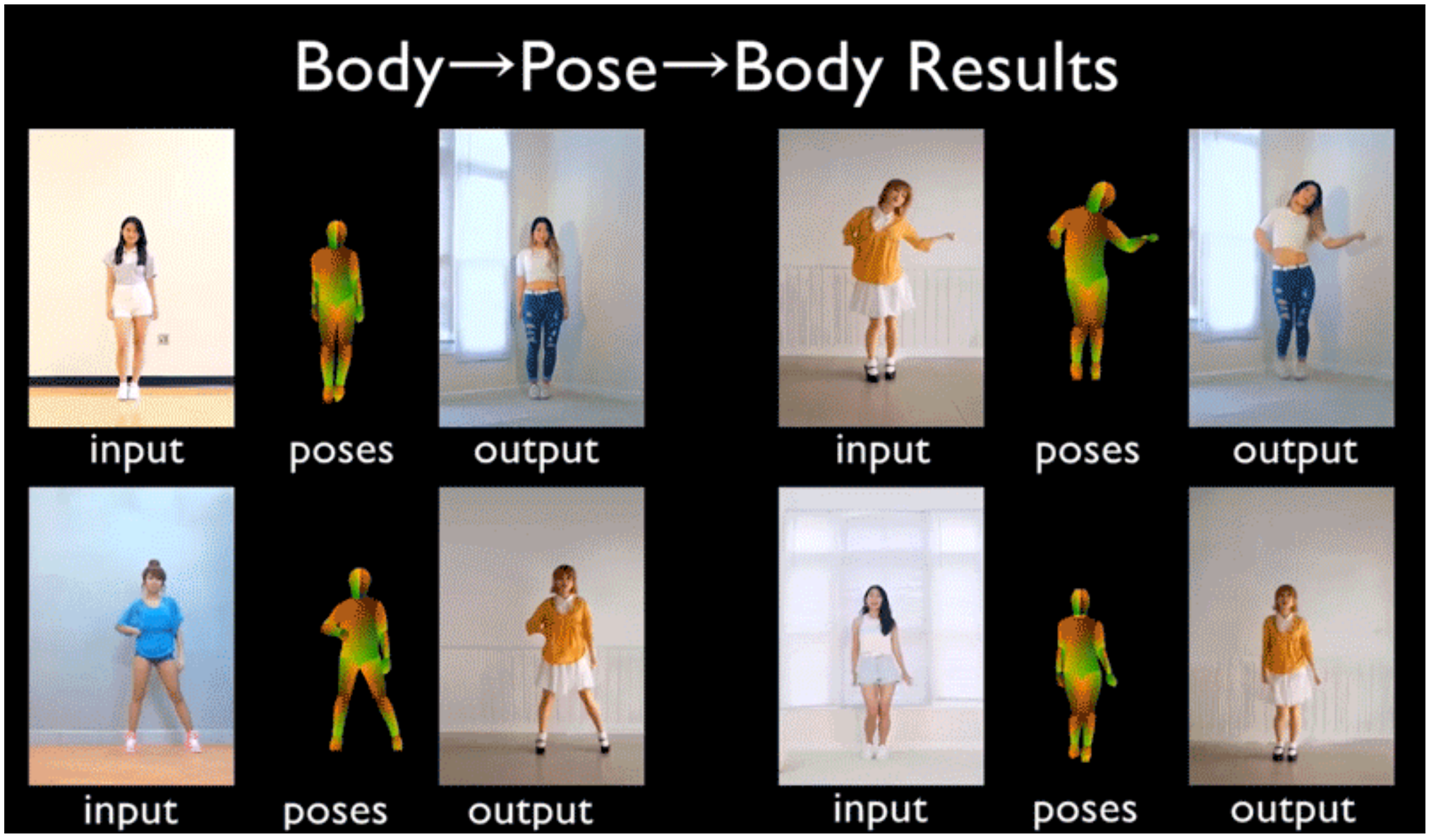
Two-hands wave

Video Prediction

$$x \sim p(x|x_0)$$

Video Translation

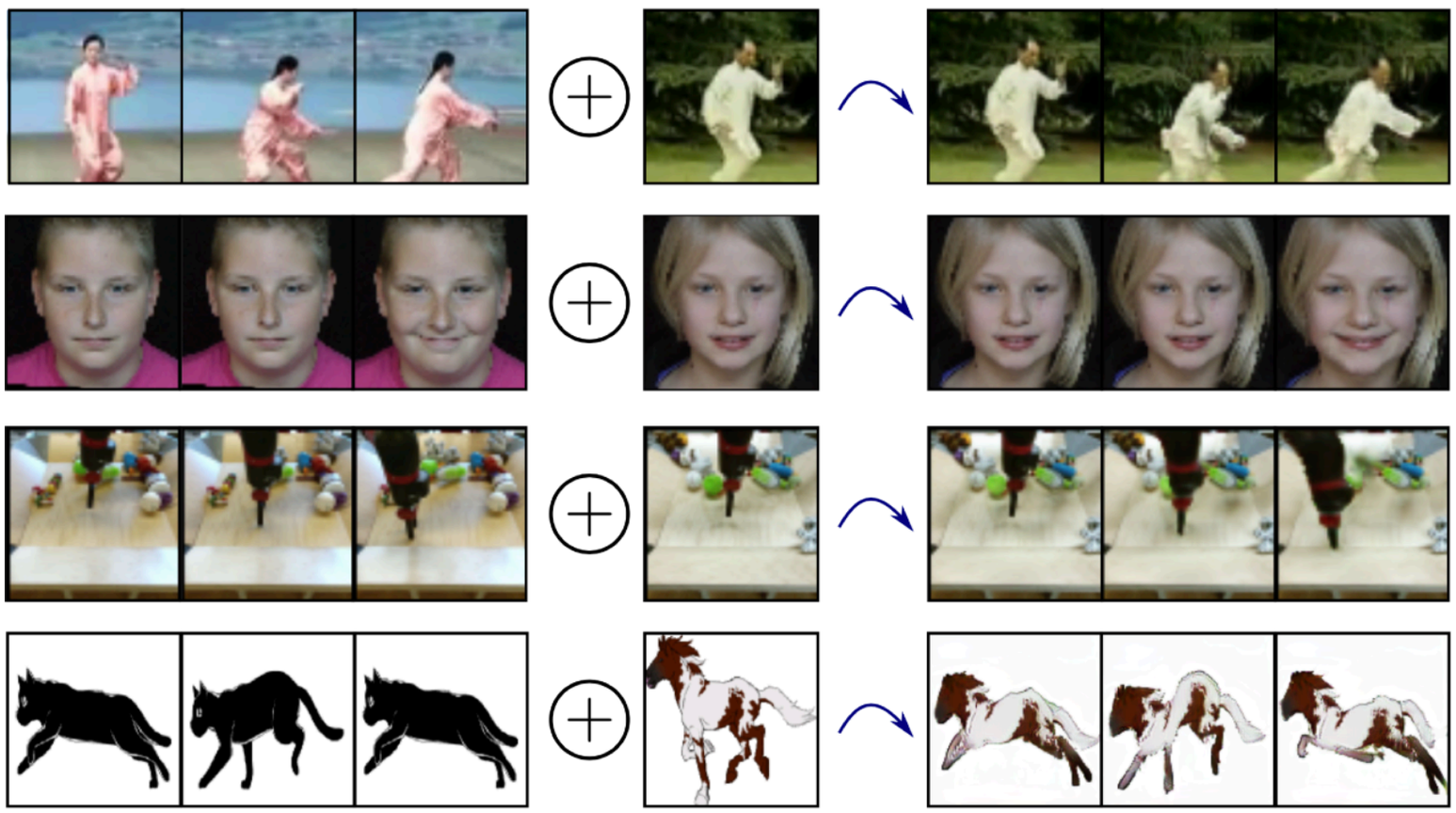
$$x_t \sim p(x_t | x_{t-1}, \ldots, x_2, x_1)$$



Video Retargeting

$$x \sim p(x|x_0, y)$$

Driving video y | Source image x_0 | Generated video x



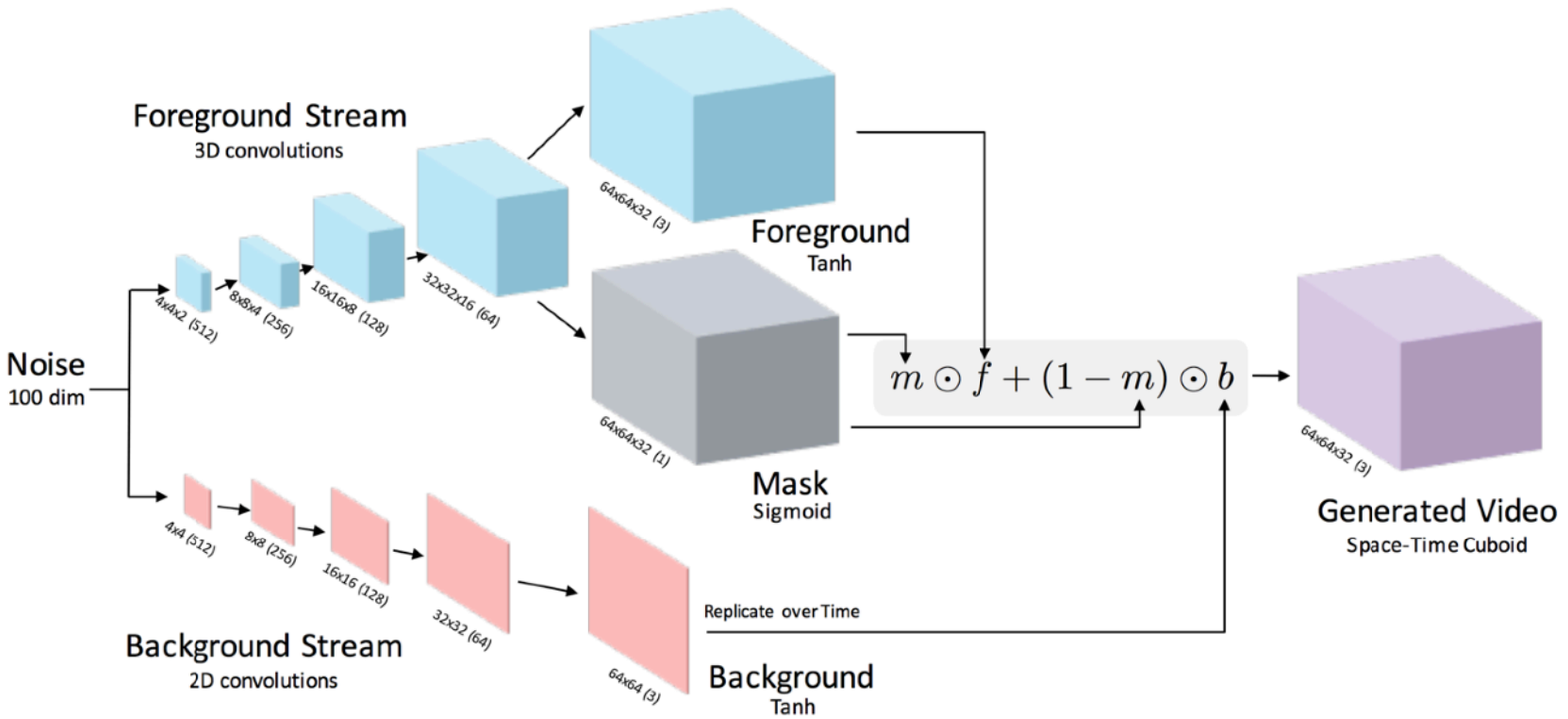
Video Generation

Video generation starts with random noise vector or a series of vectors

- Video-based models
  - VideoGAN
- Image-based models
  - MoCoGAN
  - TGAN
  - DVD-GAN

Video representation

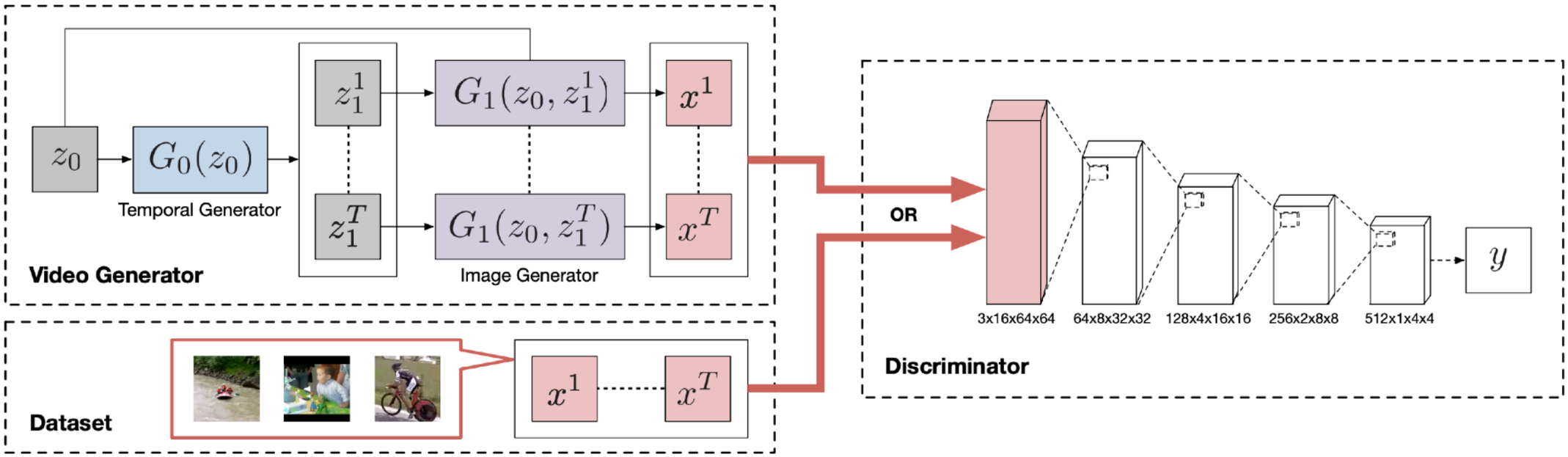
Hidden space: every points is a video



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

Video representation

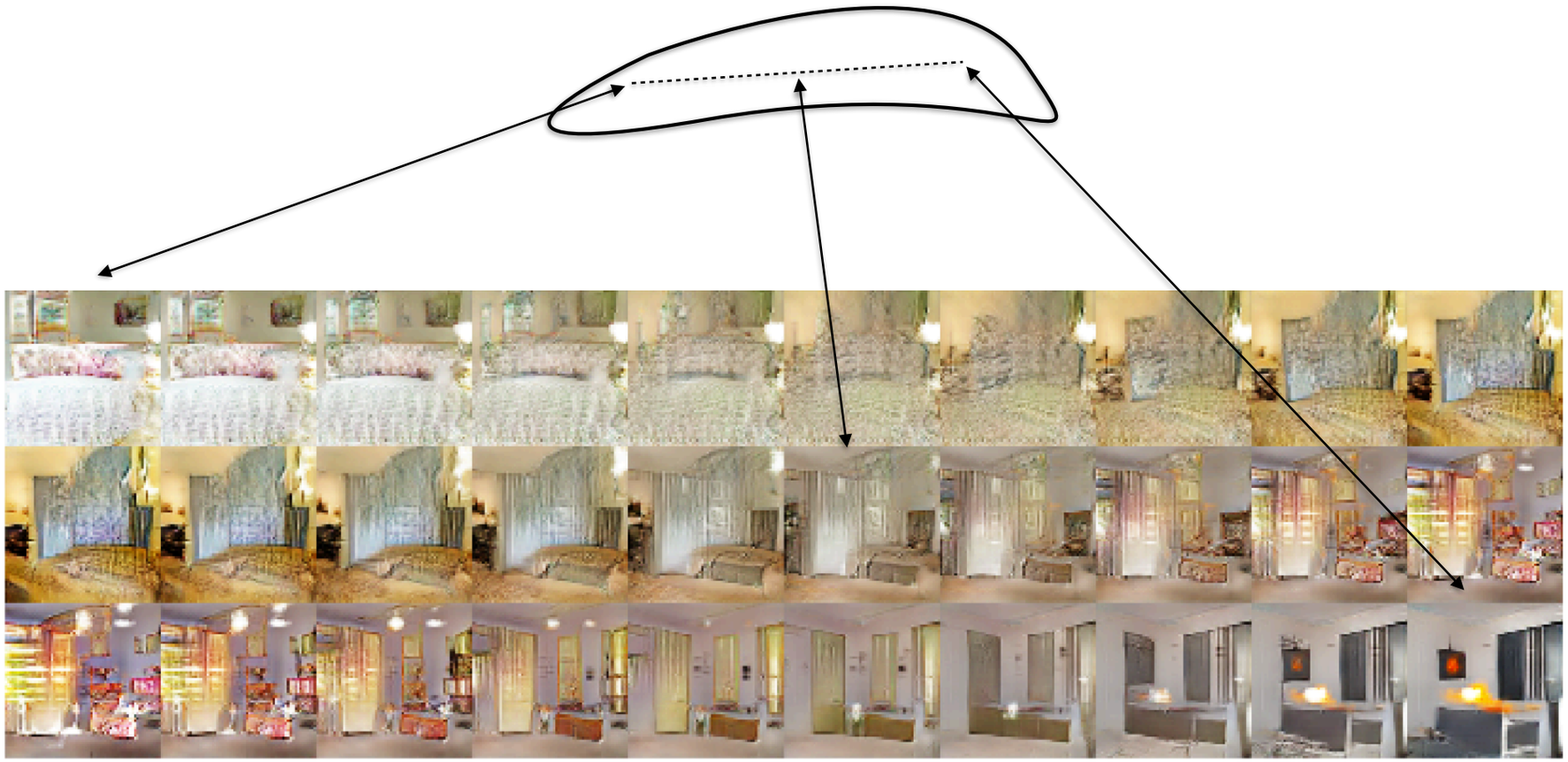
Hidden space: every points is a video



Saito, Matsumoto, and Saito. "Temporal generative adversarial nets with singular value clipping." ICCV'2017.

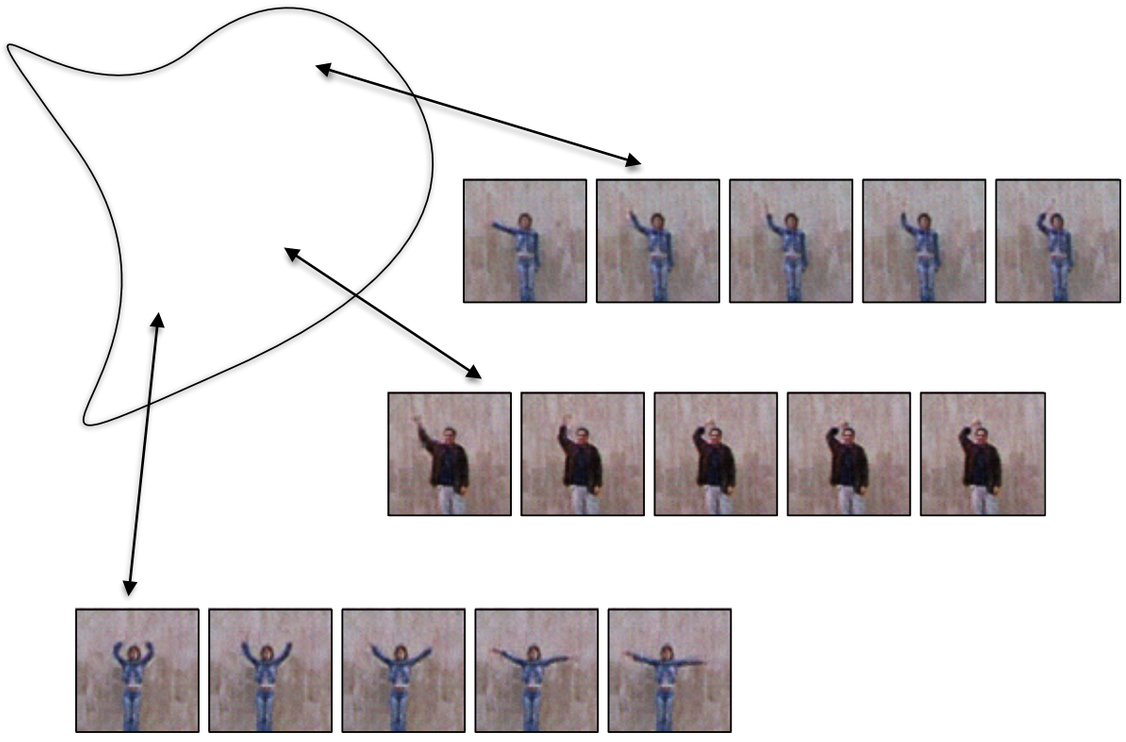
Image-based GANs

Hidden space: every point is an image



Video representation

Hidden space: every point is an **video**

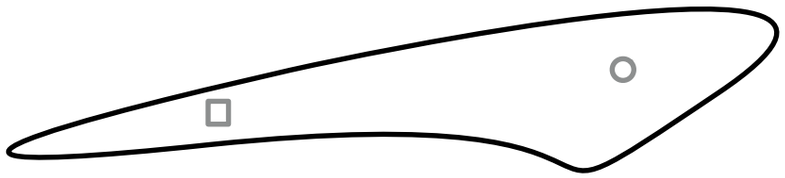


Limitations:

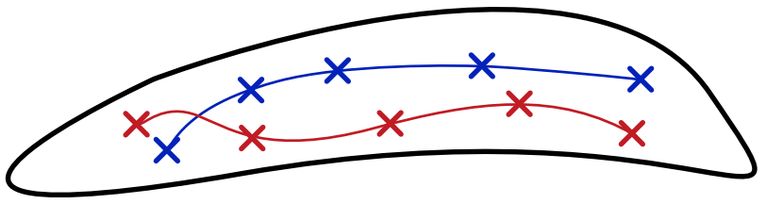
- Fixed length videos only
- No control over motion and content

MoCoGAN representation

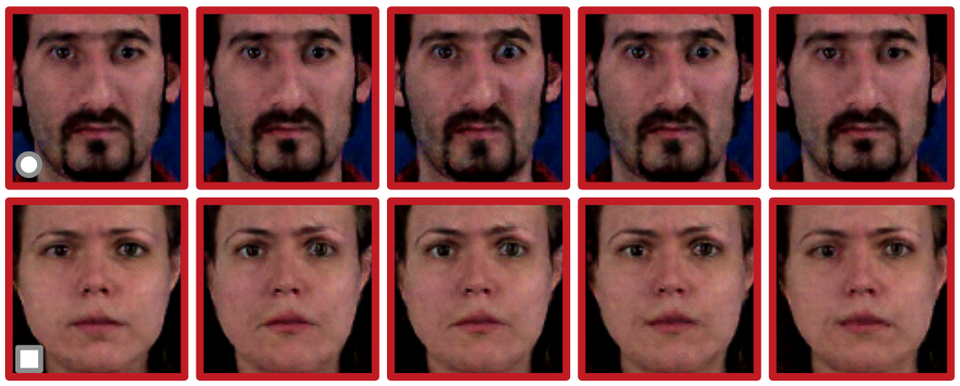
Content subspace



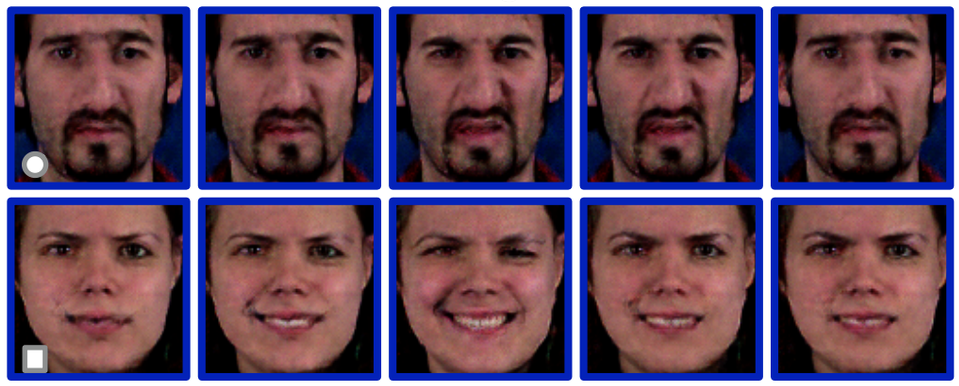
Motion subspace



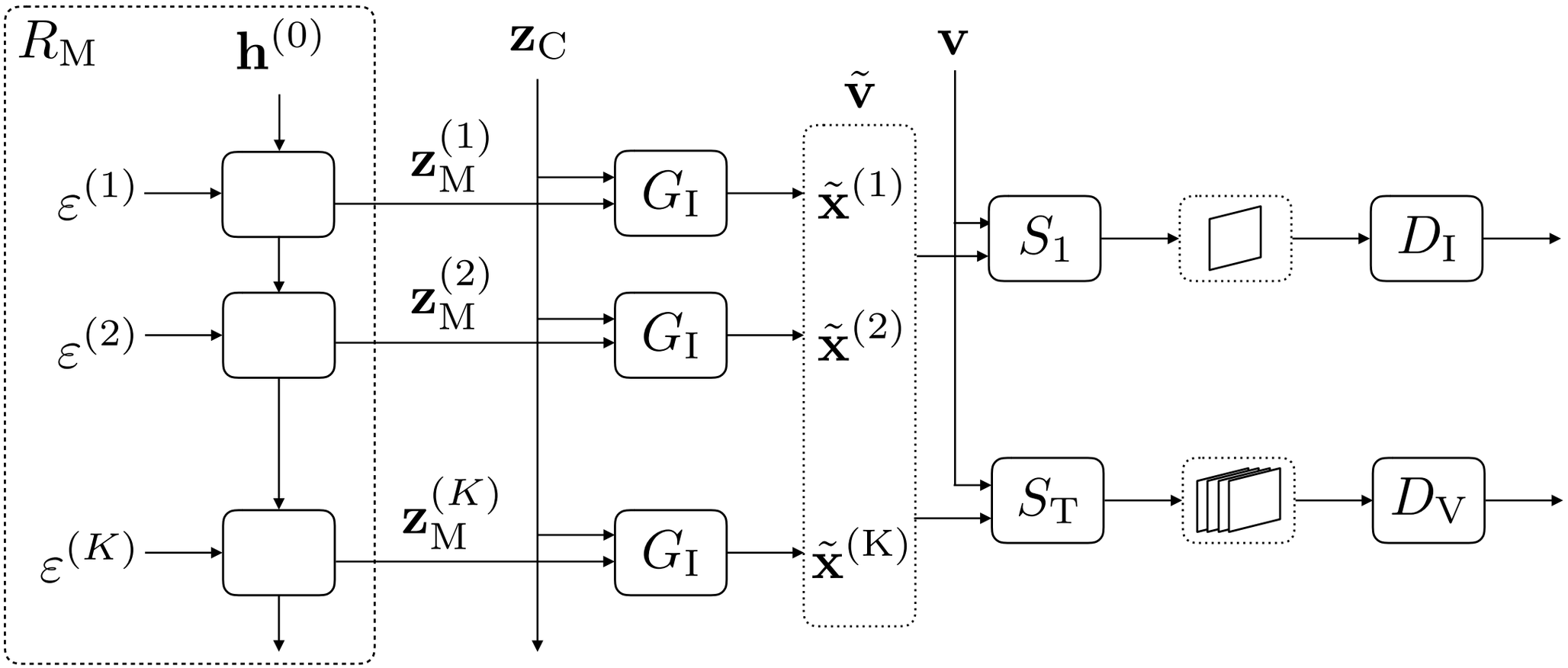
Fear



Disgust



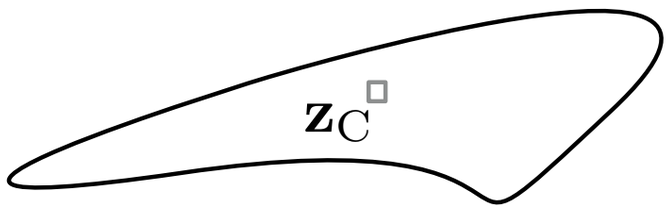
MoCoGAN framework



MoCoGAN framework

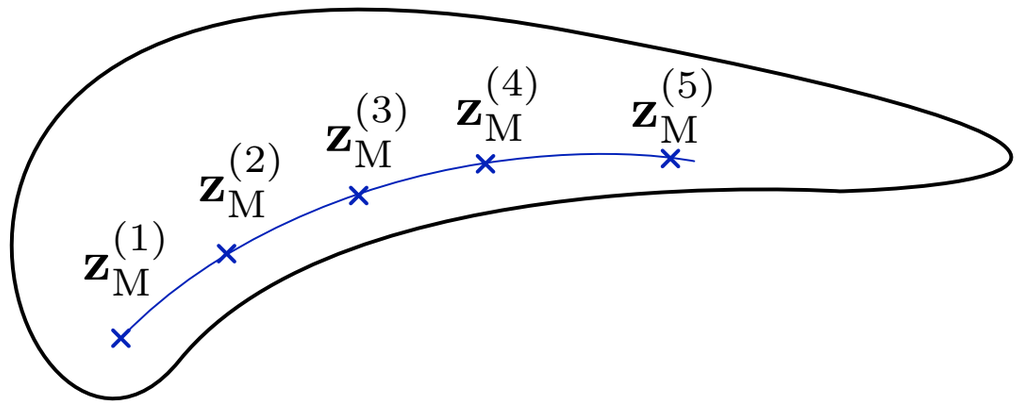
Sampled content

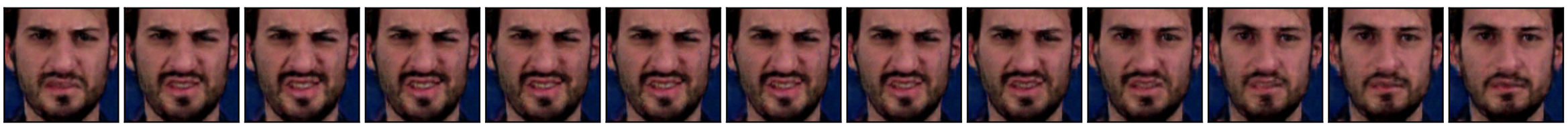
$$\mathbf{Z}_\mathrm{C} = [\mathbf{z}_\mathrm{C}, \mathbf{z}_\mathrm{C}, ..., \mathbf{z}_\mathrm{C}]$$



Motion trajectory

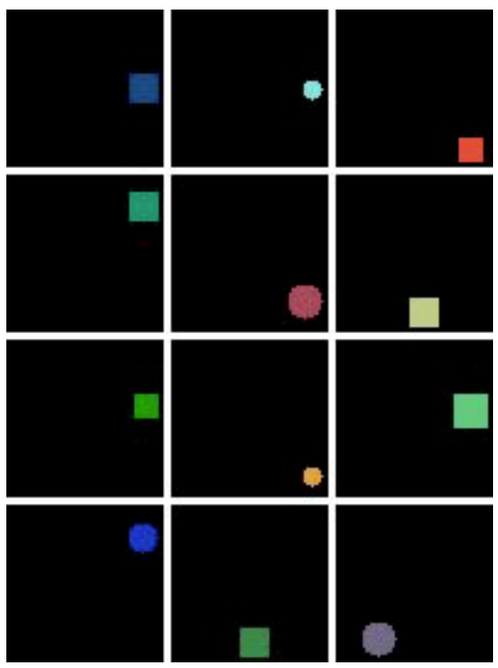
$$\mathbf{Z}_\mathrm{M} = [\mathbf{z}_\mathrm{M}^{(1)}, \mathbf{z}_\mathrm{M}^{(2)}, ..., \mathbf{z}_\mathrm{M}^{(K)}]$$





Datasets

Shape motion



Synthetic

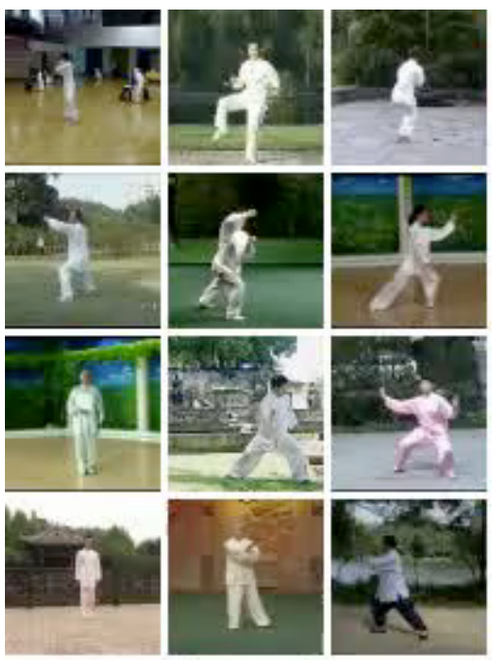
Human actions

[Gorelick et al]

Expressions

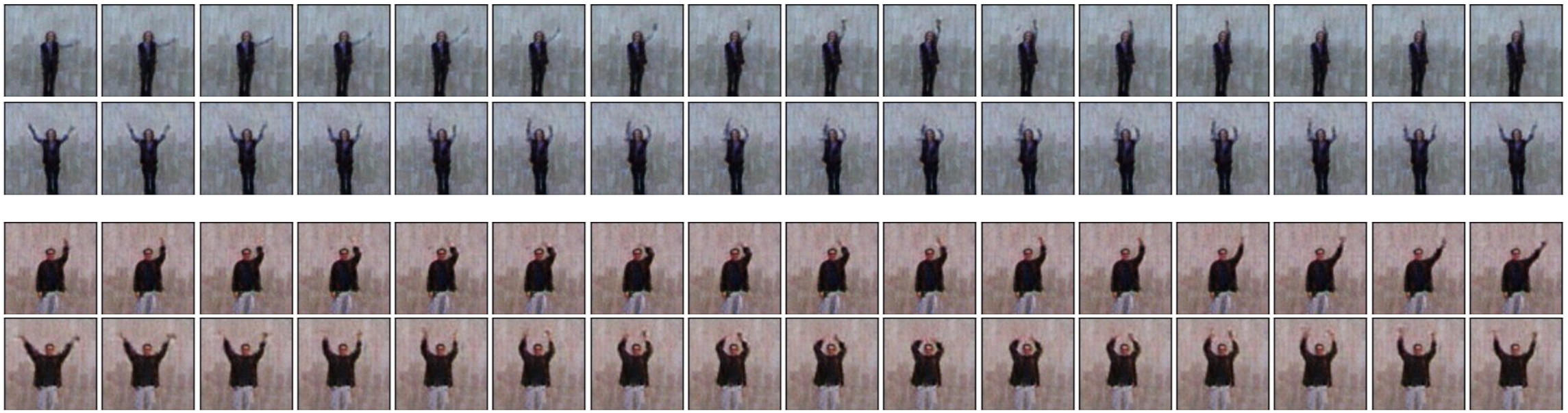
[Aifanti et al]

TaiChi



From YouTube

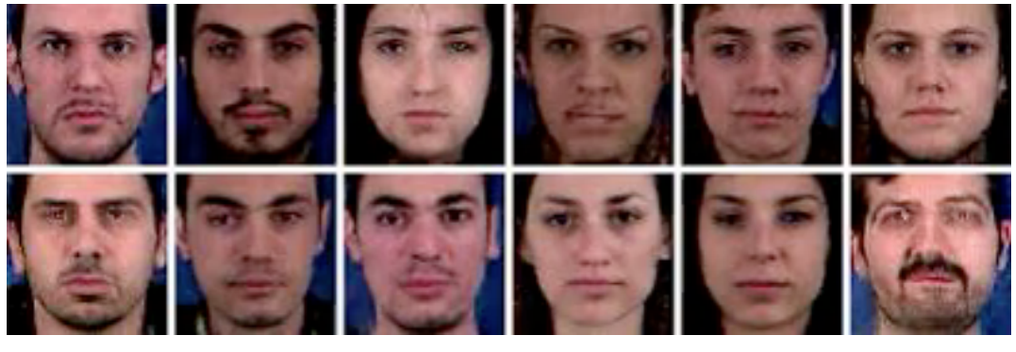
Generating human actions



MoCoGAN vs VGAN

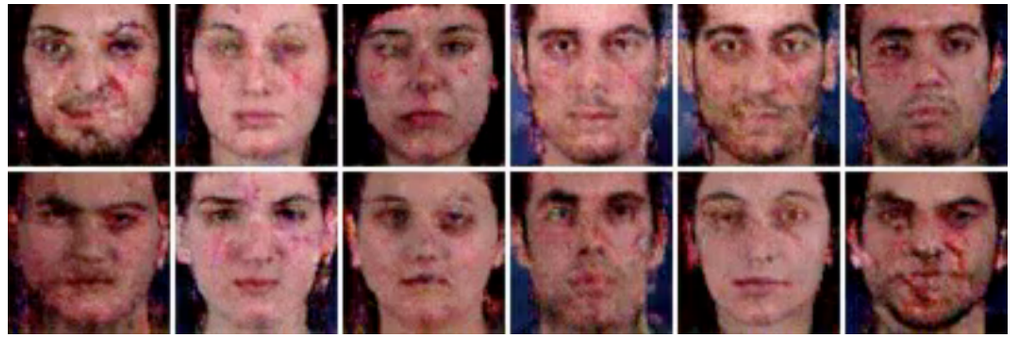
MoCoGAN

Preferred by 84.2%



VGAN

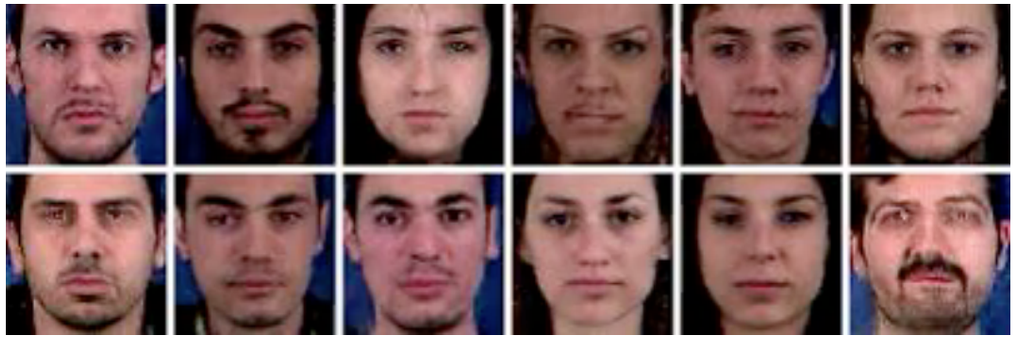
Preferred by 15.8%



MoCoGAN vs TGAN

MoCoGAN

Preferred by 54.7%

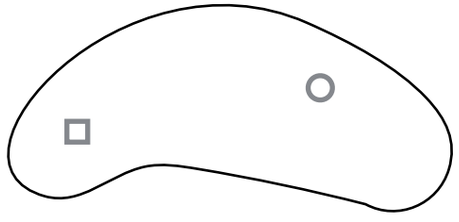


TGAN

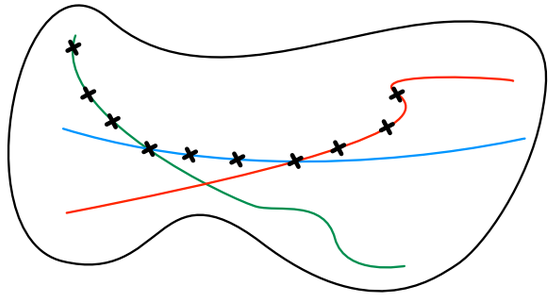
Preferred by 45.3%

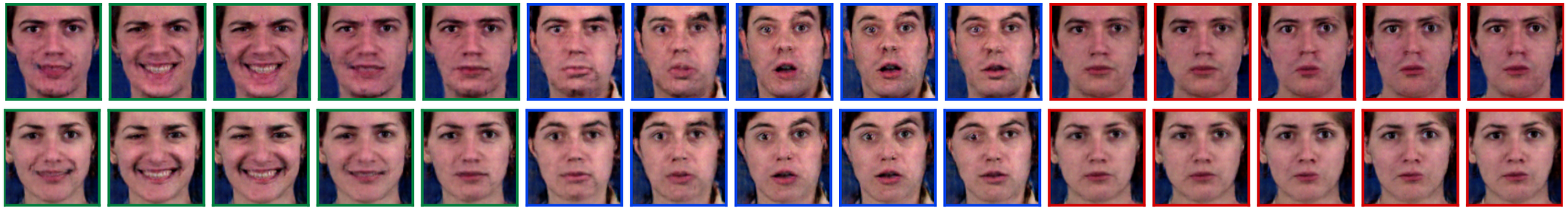
Changing content and motion

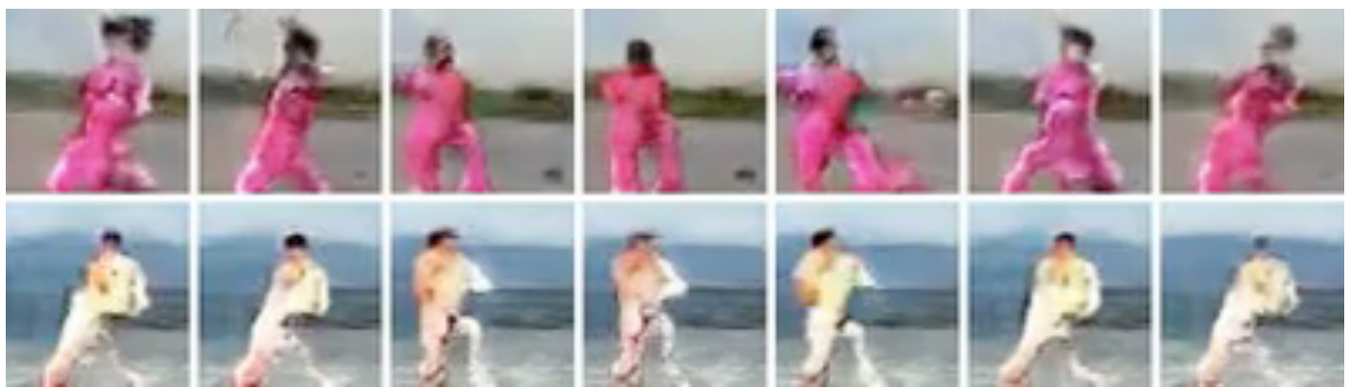
Sampled content



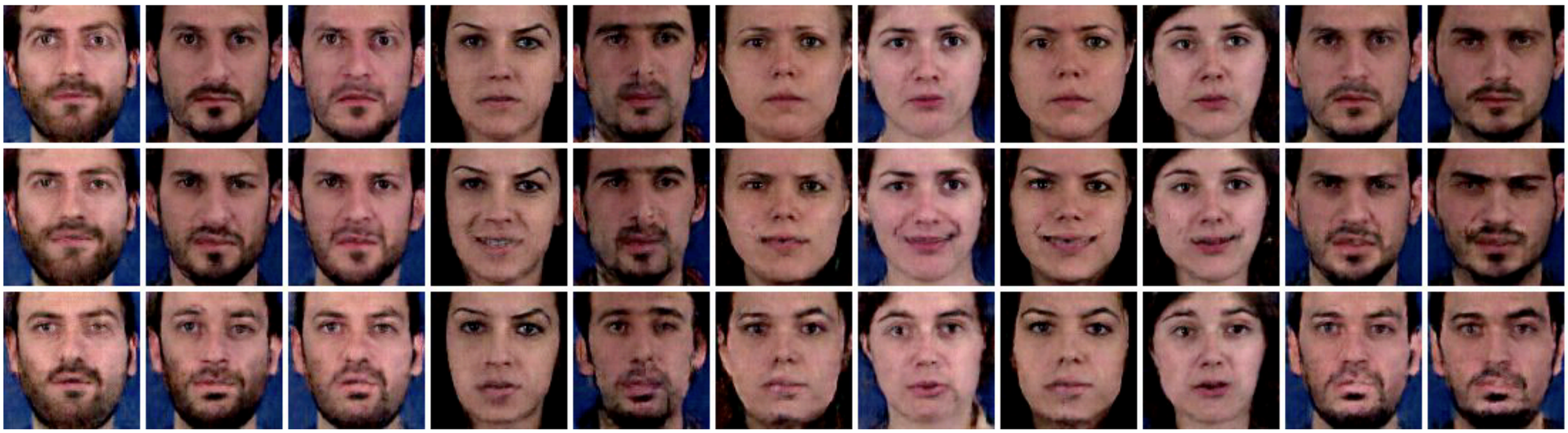
Motion trajectory







Generating facial expressions

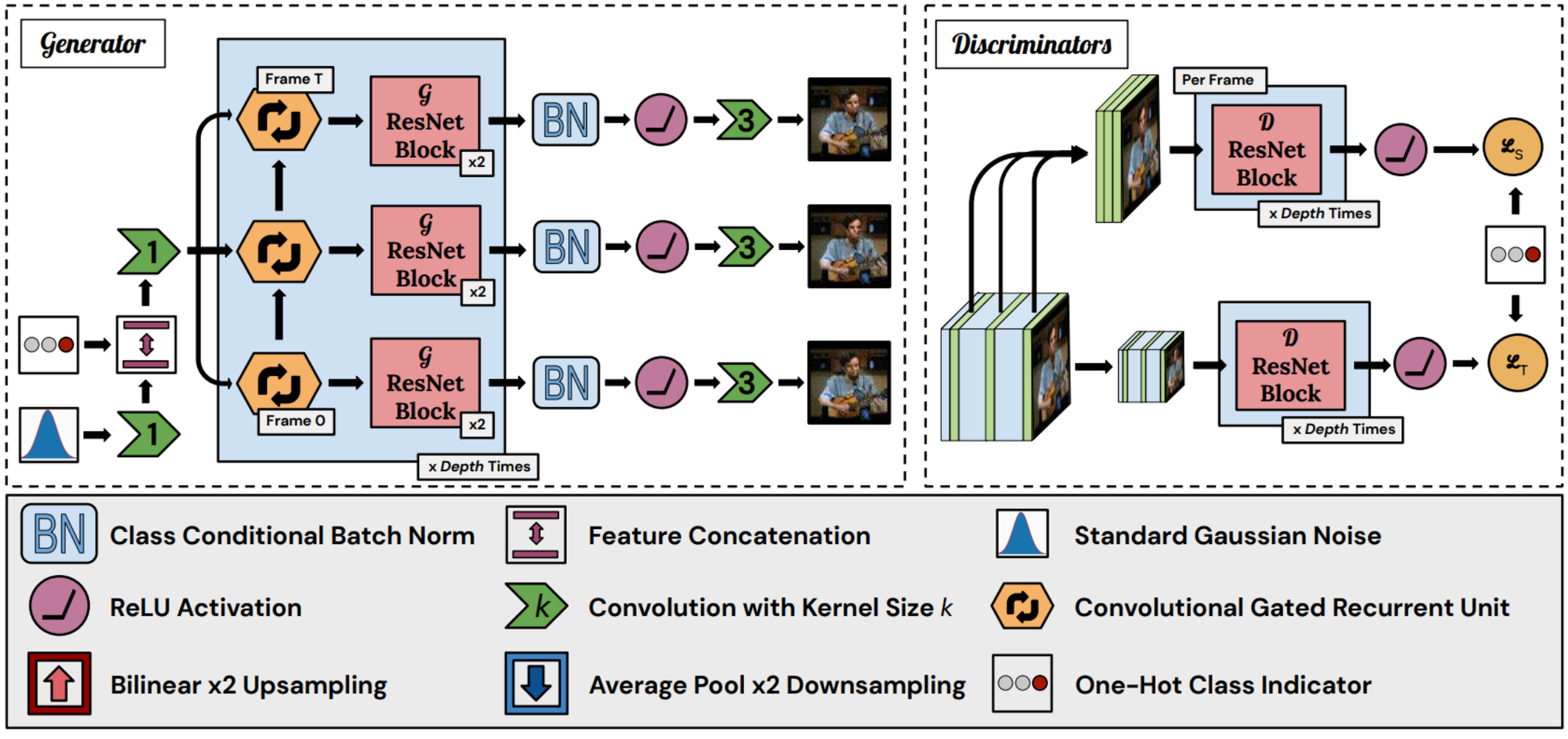


Recent Extension: DVD-GAN

High-fidelity, supports variety of objects and scenes

Recent Extension: DVD-GAN

High-fidelity, supports variety of objects and scenes



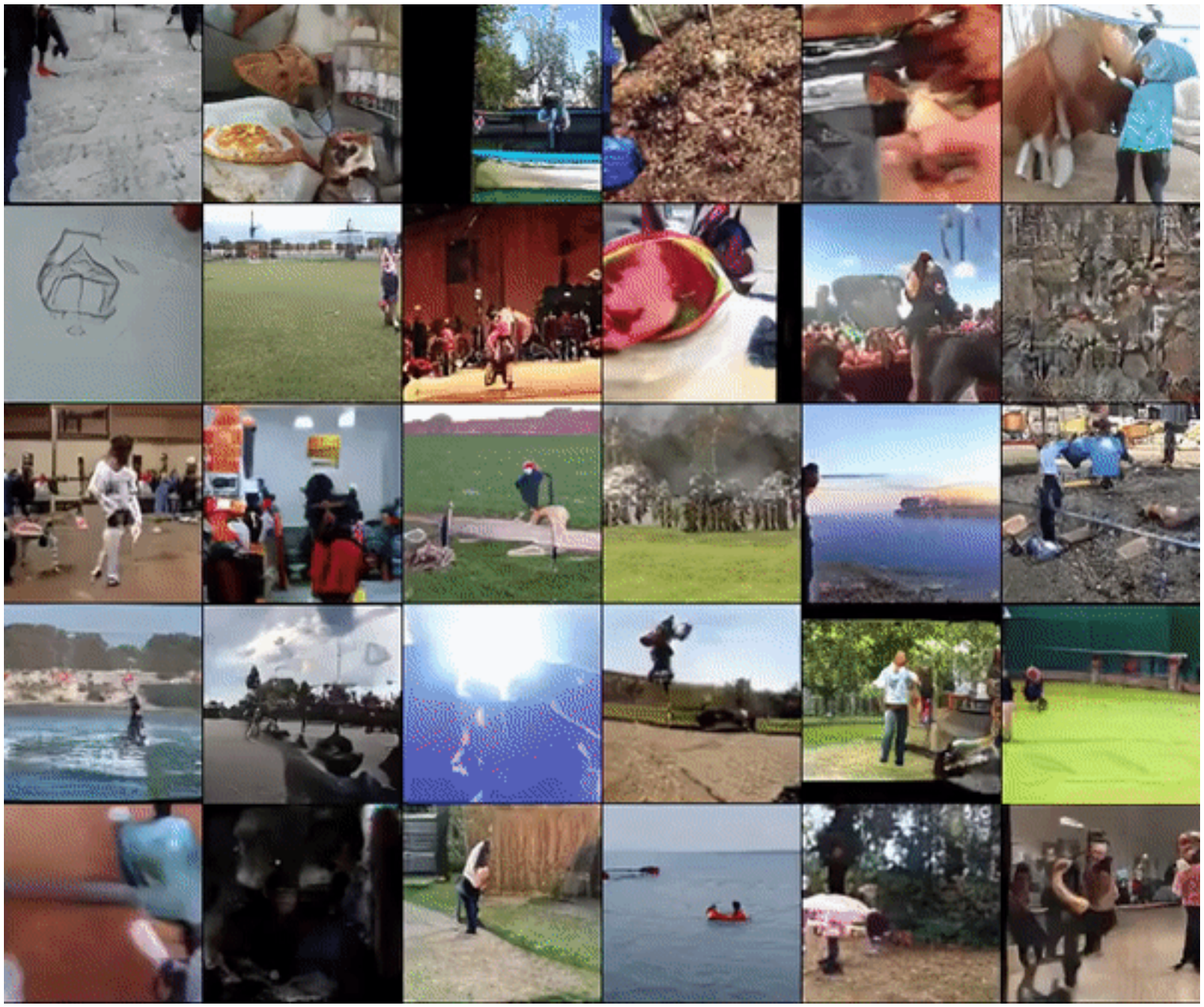
Clark, Aidan, Jeff Donahue, and Karen Simonyan. "Efficient video generation on complex datasets." ArXiv preprint

DVD-GAN: Results

Videos with people and faces

DVD-GAN: Results

Scenes

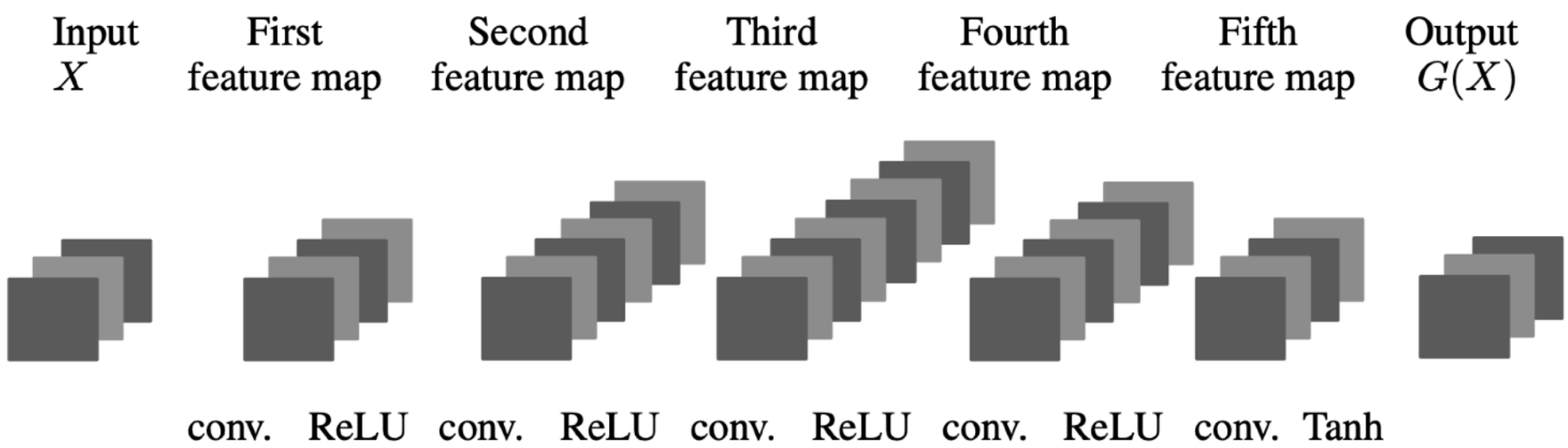


Video Prediction

Given an input image or a sequence, generate future frames

- ICLR'2016
- MCNET
- MoCoGAN
- Timelapse:
  - CVPR'2018
  - CVPR'2019
- SV2P

Initial Works



Objective:

$$\mathcal{L}_p(X, Y) = \ell_p(G(X), Y) = \|G(X) - Y\|_p^p$$

L2 and L1 losses produce blurry predictions, increasingly worse when predicting further

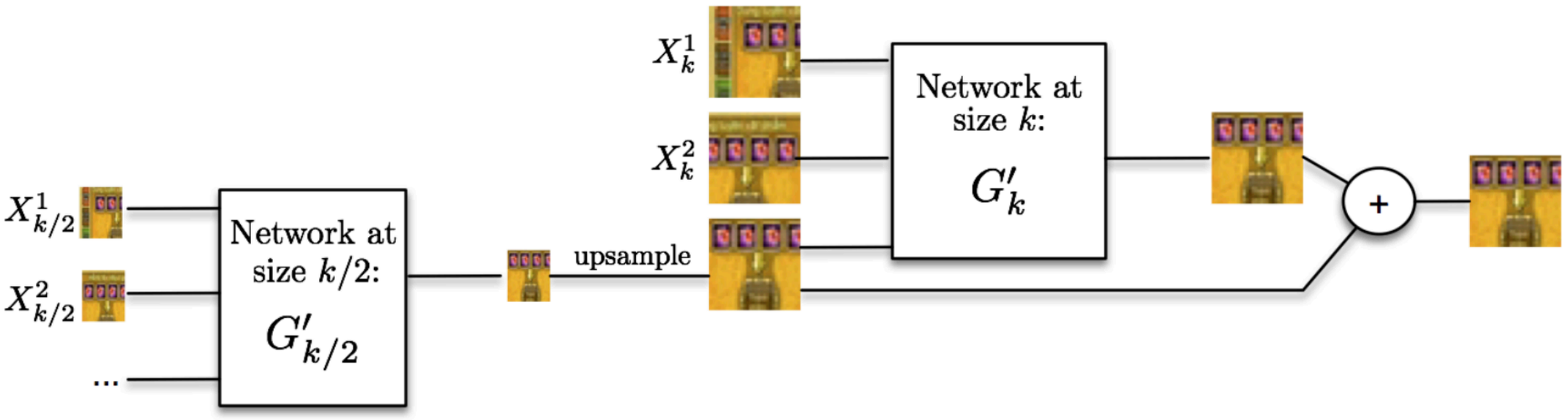
Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR'2016

Three Ideas to Improve Sharpness

- Multiscale network
- Adversarial training – you just add the discriminator
- Image gradient difference loss

Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR'2016

Multiscale Network



Recursive definition:

$$\hat{Y}_k = G_k(X) = u_k(\hat{Y}_{k-1}) + G'_k\left(X_k, u_k(\hat{Y}_{k-1})\right)$$

Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR'2016

Image Gradient Difference Loss

Compute differences in image pixels:

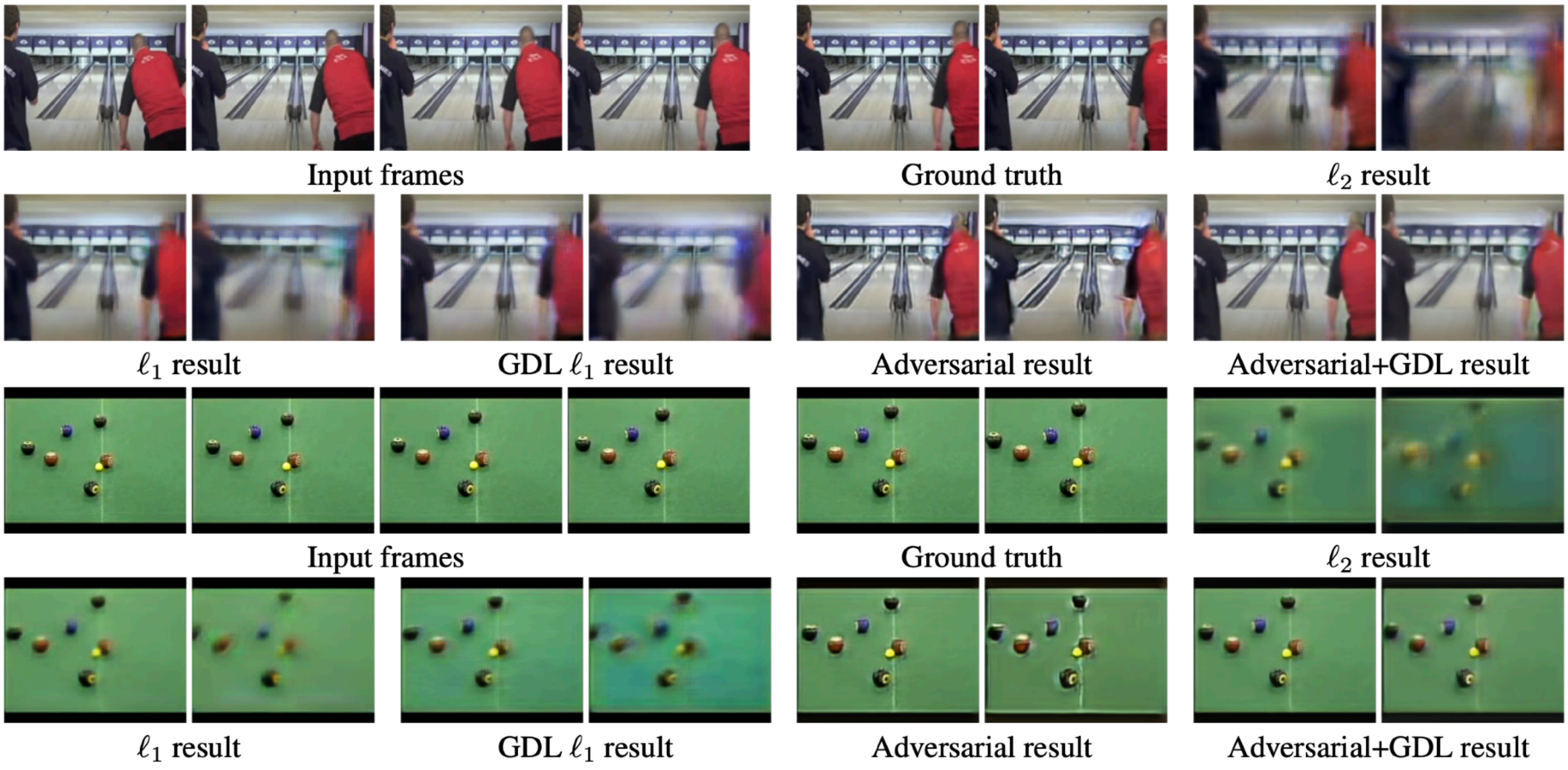
$$\mathcal{L}_{gdl}(X, Y) = L_{gdl}(\hat{Y}, Y) =$$
$$\sum_{i,j} \left| |Y_{i,j} - Y_{i-1,j}| - |\hat{Y}_{i,j} - \hat{Y}_{i-1,j}| \right|^{\alpha} + \left| |Y_{i,j-1} - Y_{i,j}| - |\hat{Y}_{i,j-1} - \hat{Y}_{i,j}| \right|^{\alpha},$$

Final loss:

$$\mathcal{L}(X, Y) = \lambda_{adv}\mathcal{L}^{G}_{adv}(X, Y) + \lambda_{\ell_p}\mathcal{L}_{p}(X, Y) + \lambda_{gdl}\mathcal{L}_{gdl}(X, Y)$$

Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR'2016

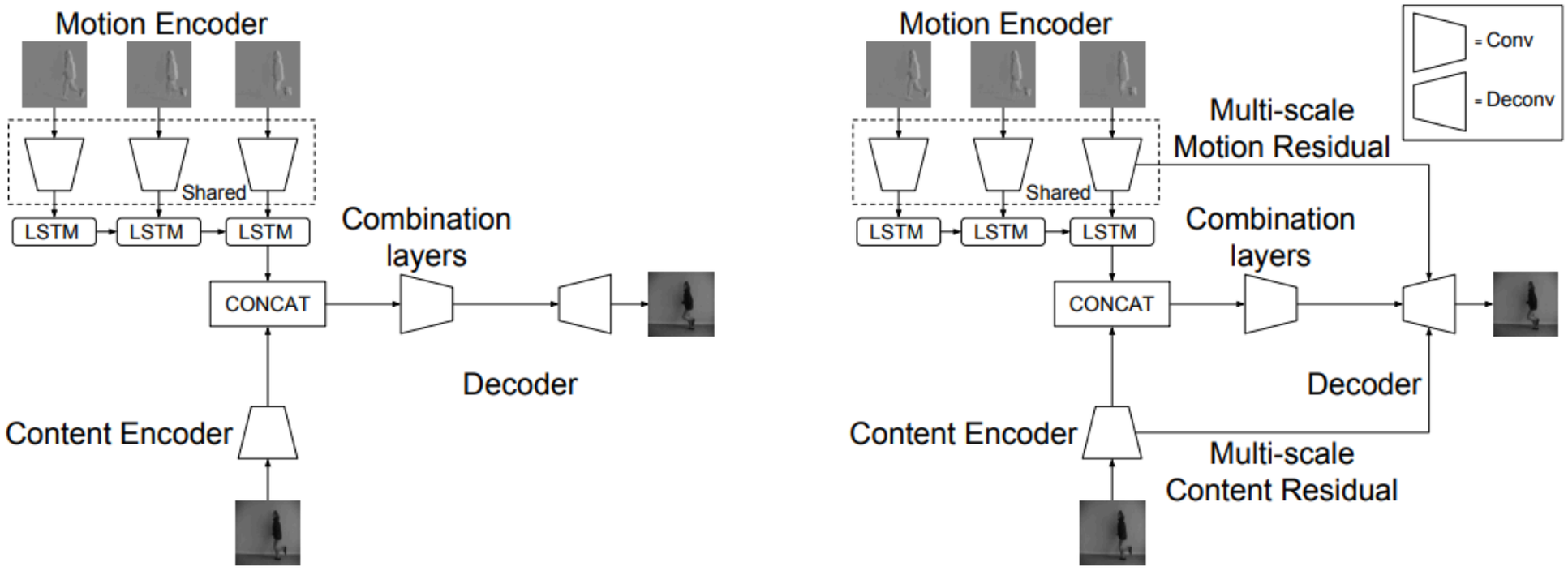
Results



Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." ICLR'2016

MCNET: Decomposing Content and Motion

- Motion encoder uses image differences to encode motion
- Content encoder uses the last frame to produce content code
- Multi-scale motion-content residual uses features from both encoders

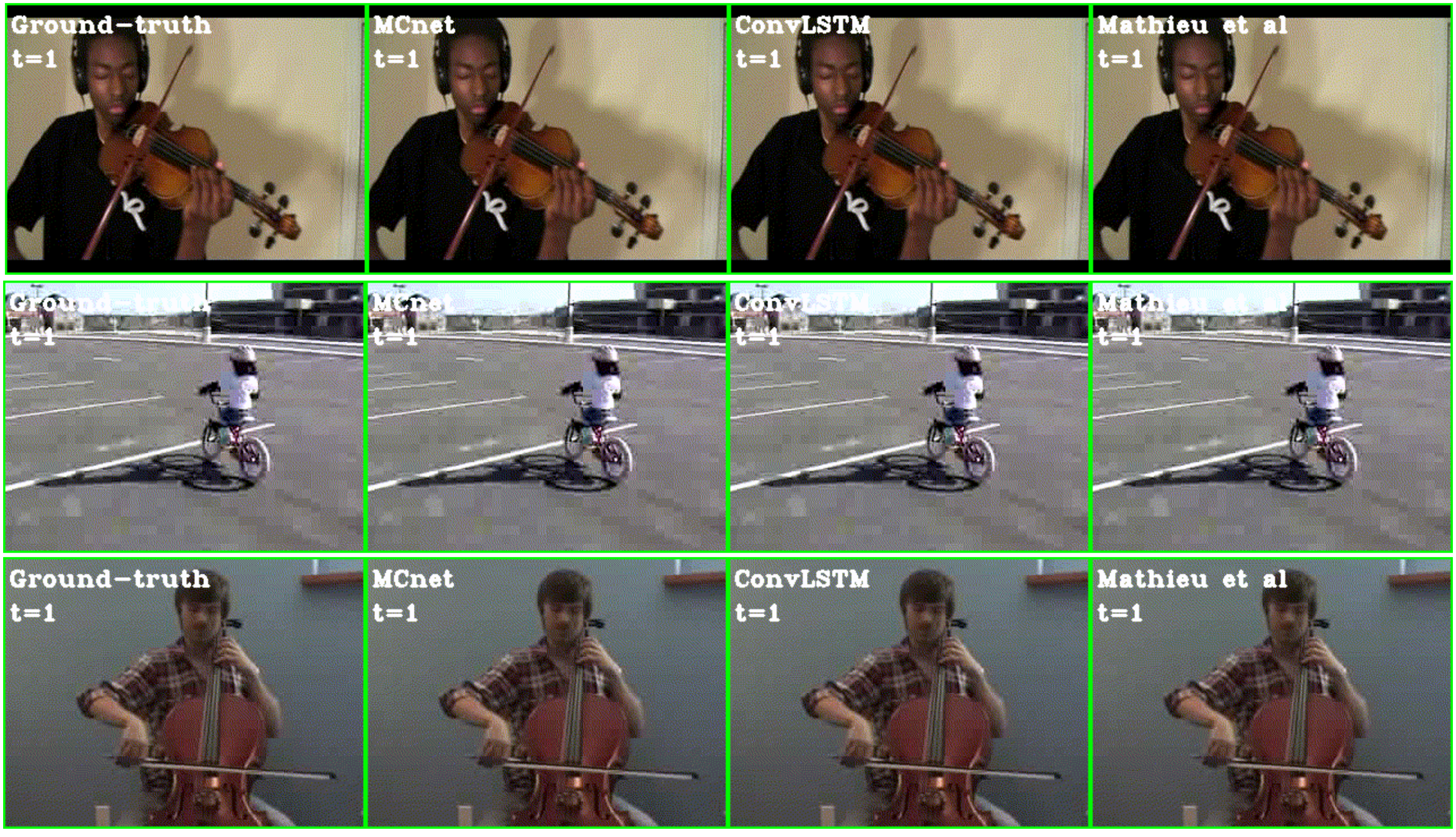


(a) Base MCnet

(b) MCnet with Multi-scale Motion-Content Residuals

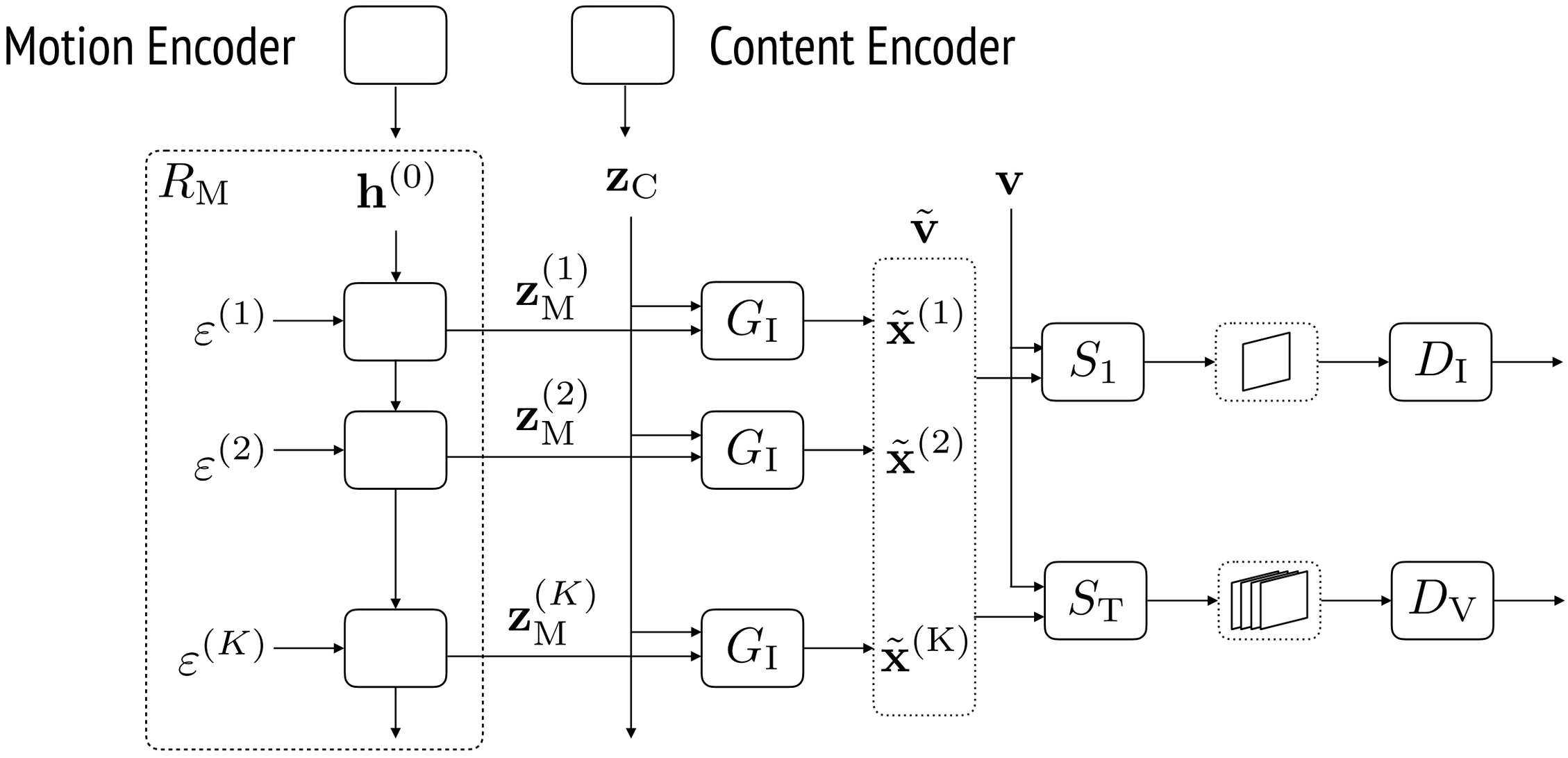
Villegas, Ruben, et al. "Decomposing motion and content for natural video sequence prediction." ICLR'2017

MCNET: Results



Villegas, Ruben, et al. "Decomposing motion and content for natural video sequence prediction." ICLR'2017

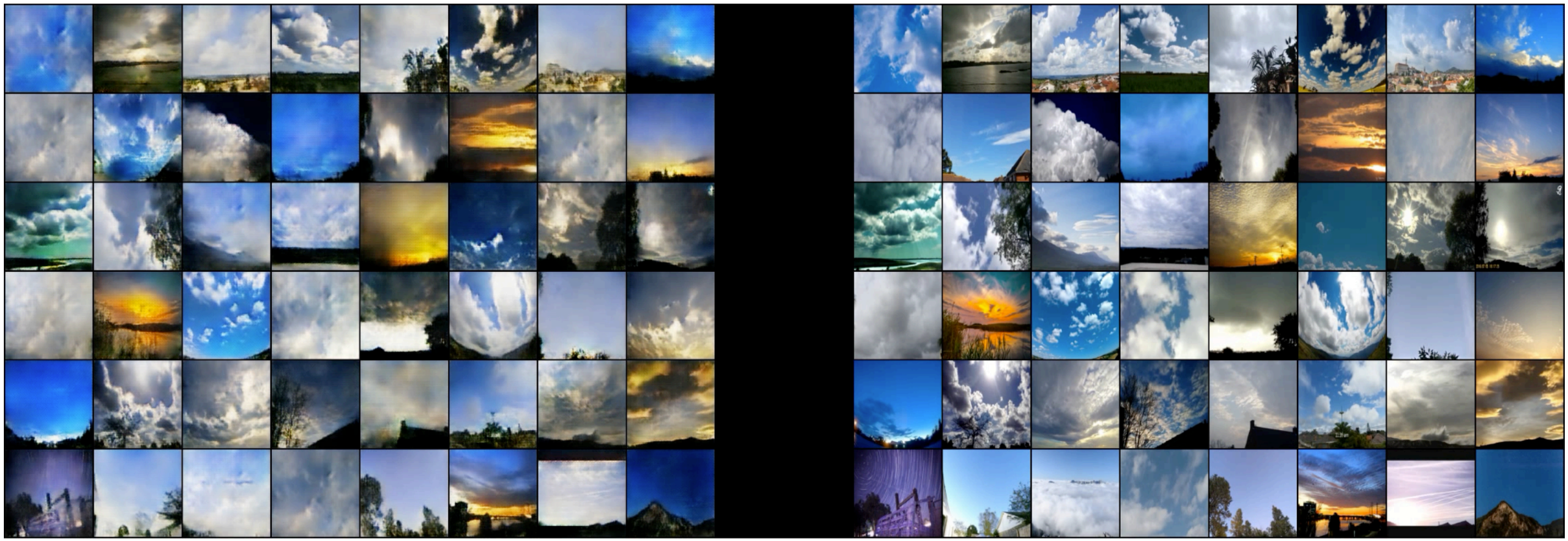
MoCoGAN framework



MoCoGAN: Results

Cloud Animation Generation

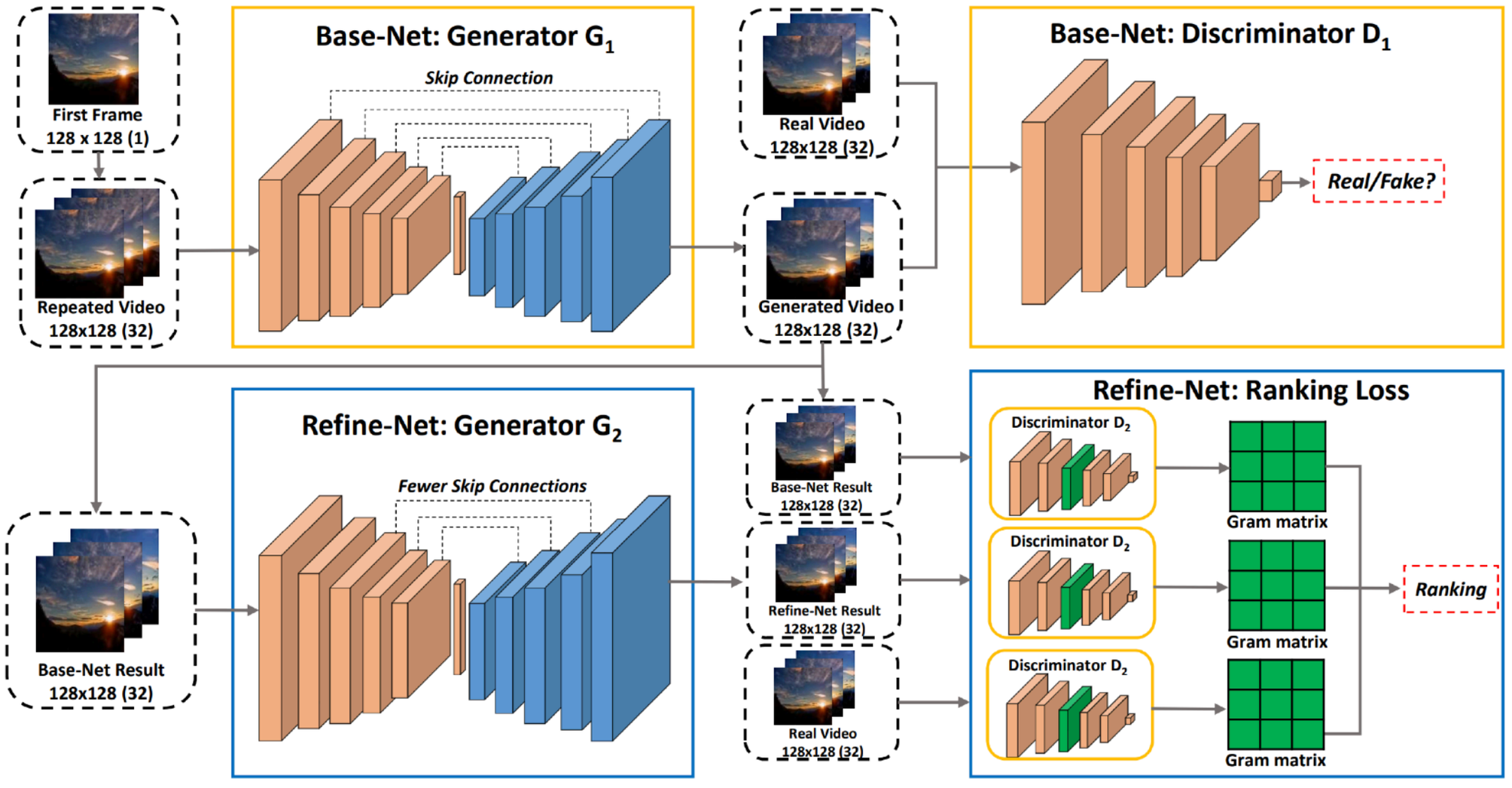
Predicting near future



Xiong, Wei, et al. "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks." CVPR'2018

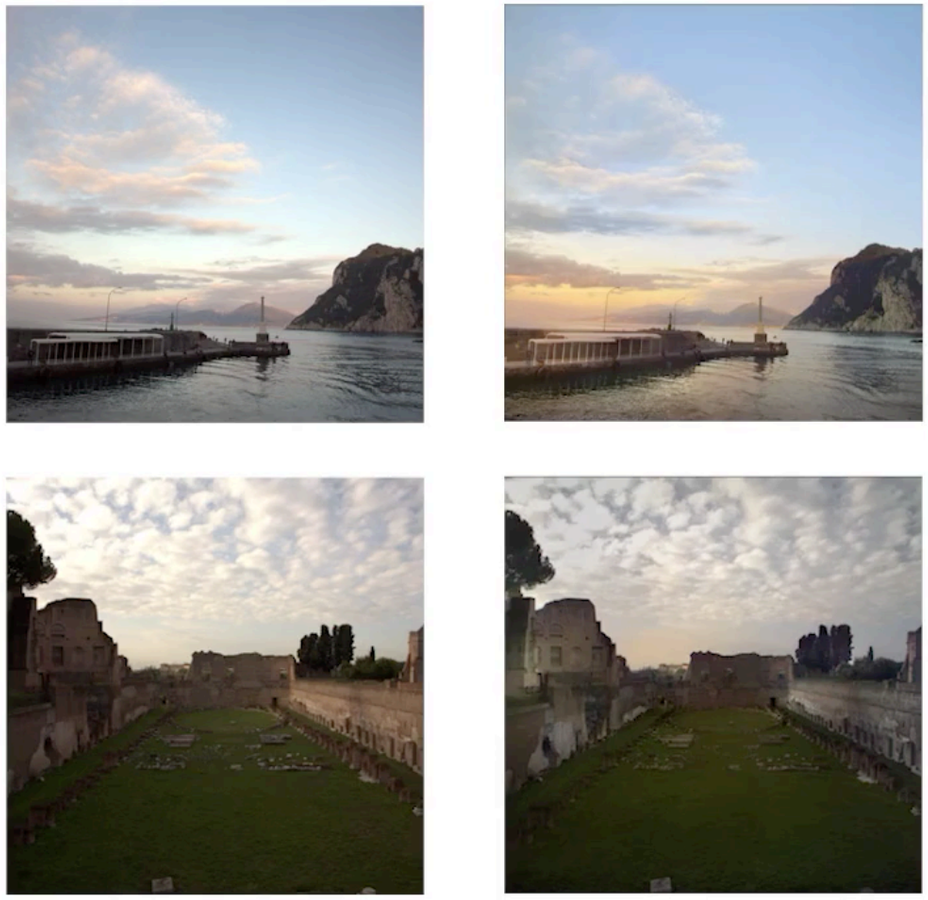
Cloud Animation Generation

2 Stage Generation



Xiong, Wei, et al. "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks." CVPR'2018

Time of the Day Animation



- Single input image (day-time)
- Uses a timestamp to generate outputs
- Continuous and photorealistic illumination changes
- Models change of color, not moving objects

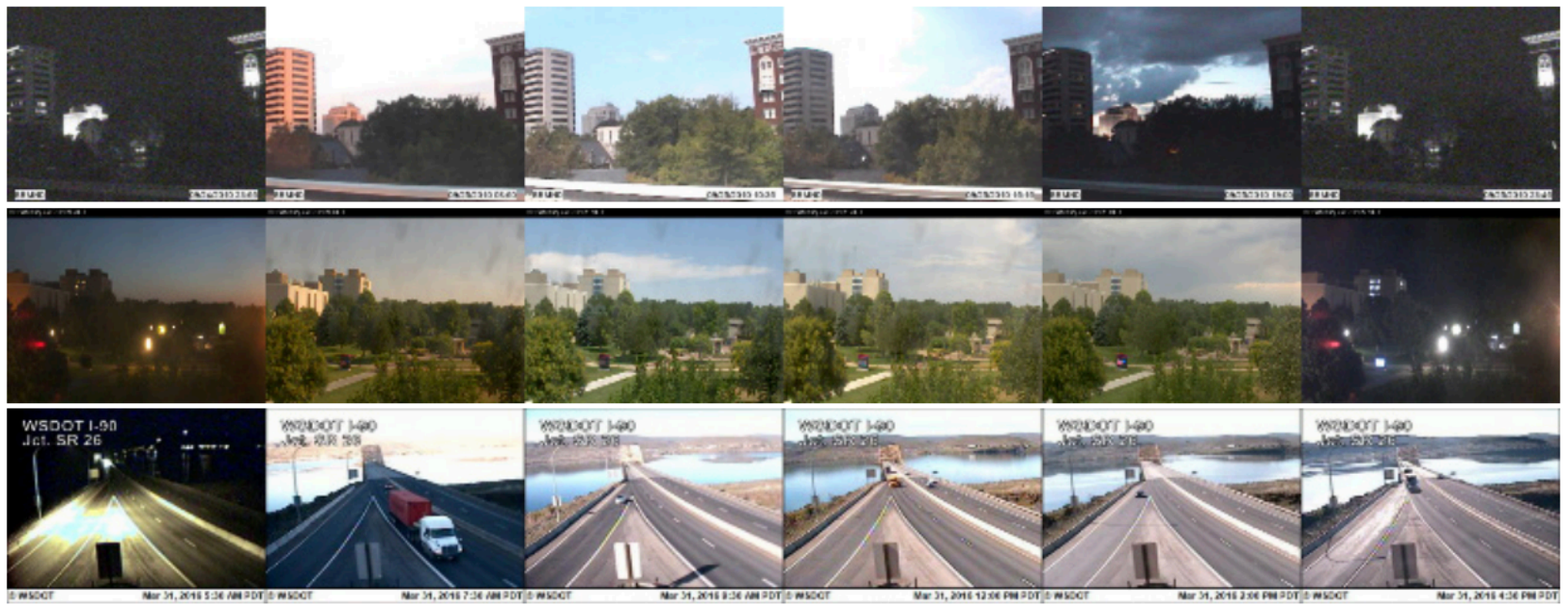
Nam, Seonghyeon, et al. "End-to-end time-lapse video synthesis from a single outdoor image." CVPR'2019

Time of the Day Animation

Training on two datasets

AMOS

WebCam
35K videos
Timestamps
Couple of frames per 24h



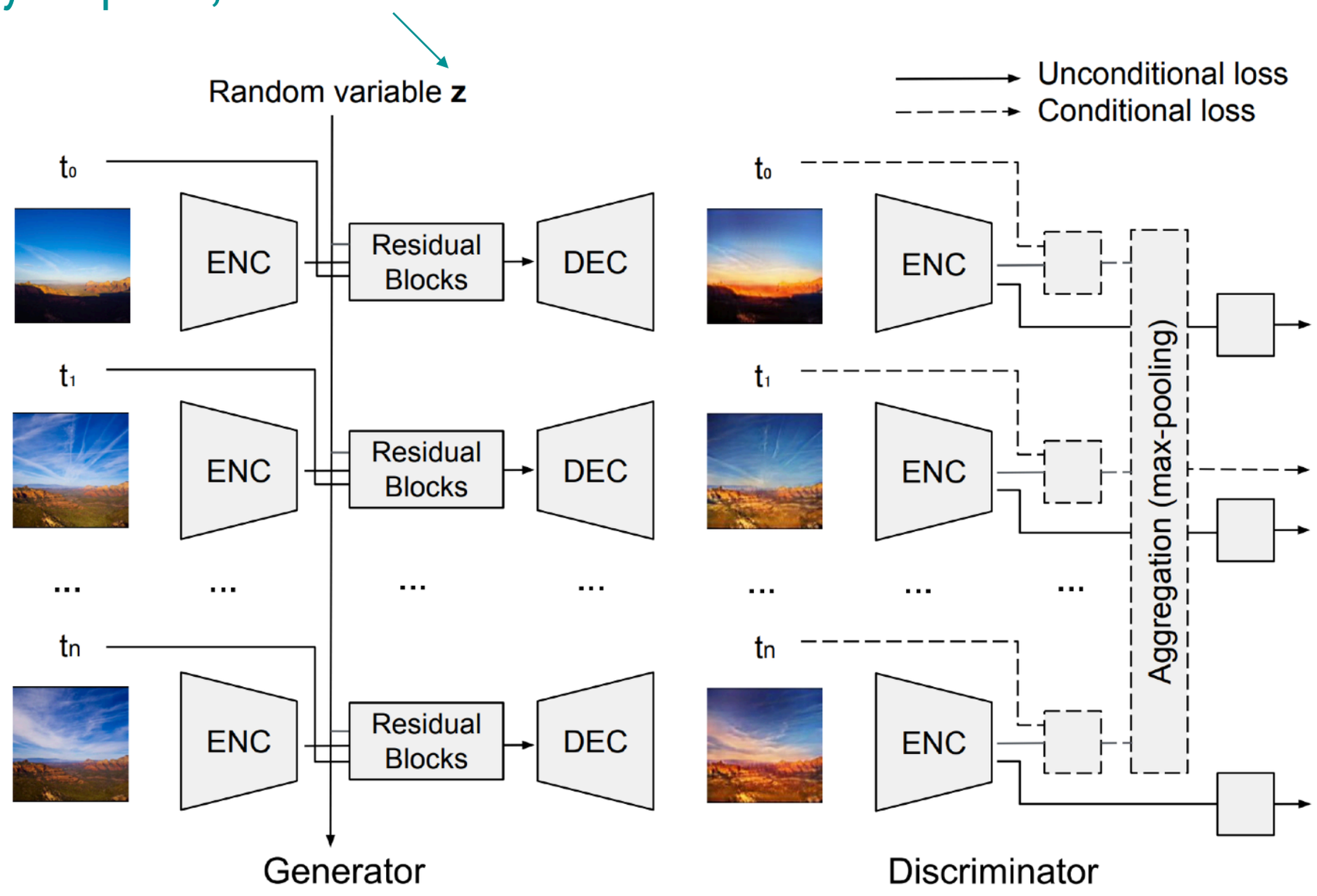
TLVDB

Landmarks scenes
Sometimes w/o time stamps
Thousands of frames per 24h

Nam, Seonghyeon, et al. "End-to-end time-lapse video synthesis from a single outdoor image." CVPR'2019

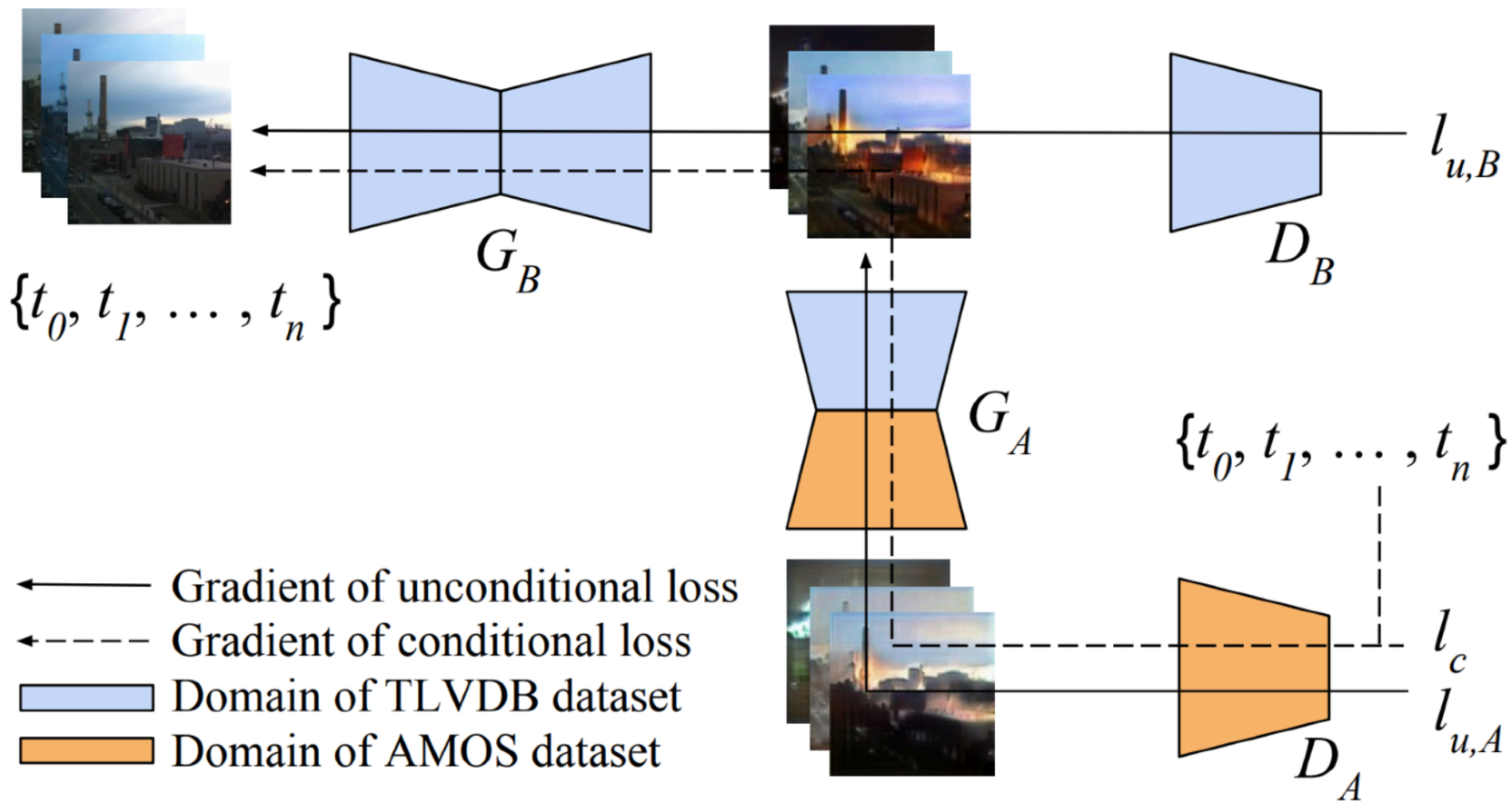
Time of the Day Animation

Dependency on place, season etc



Nam, Seonghyeon, et al. "End-to-end time-lapse video synthesis from a single outdoor image." CVPR'2019

Multi Domain Training



Nam, Seonghyeon, et al. "End-to-end time-lapse video synthesis from a single outdoor image." CVPR'2019

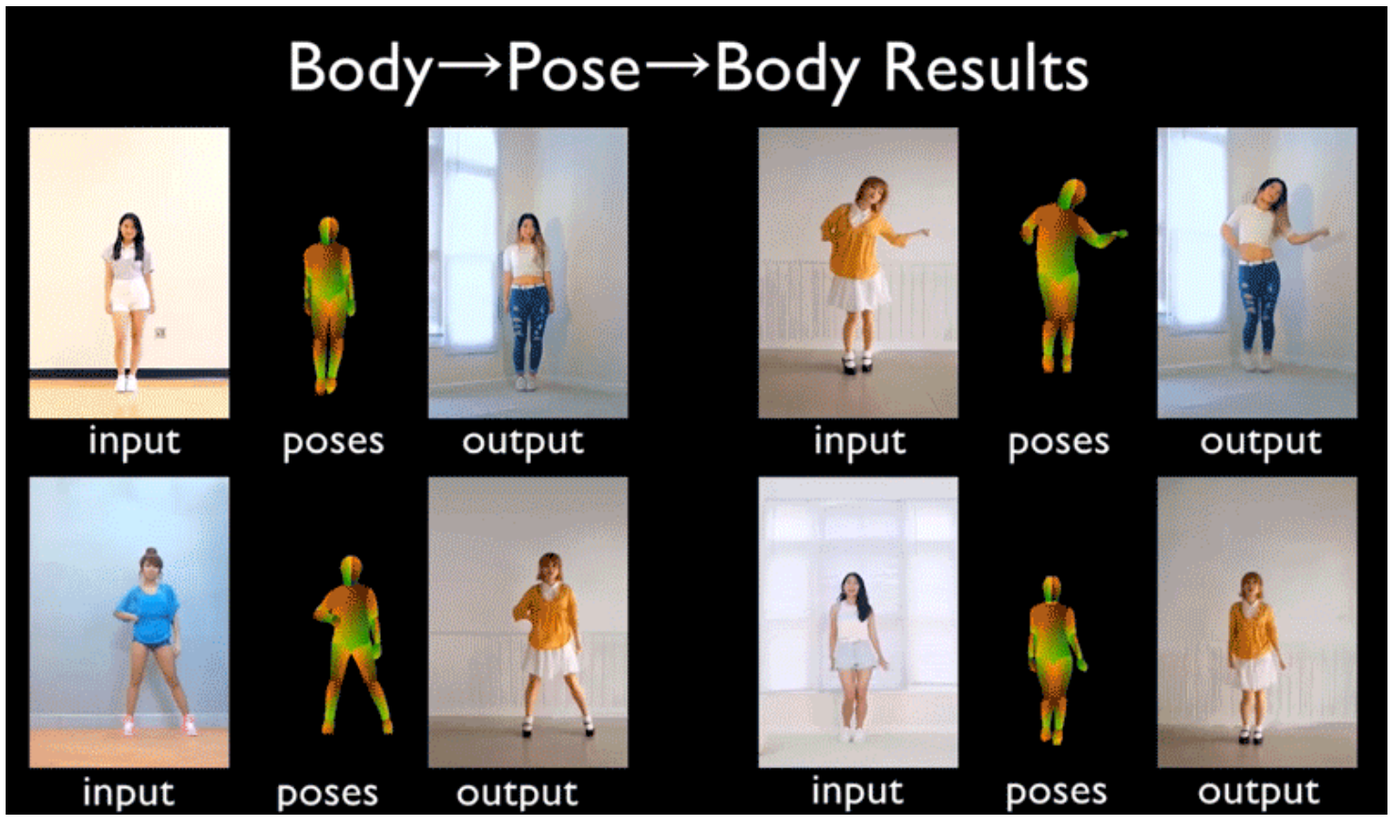
Results

Nam, Seonghyeon, et al. "End-to-end time-lapse video synthesis from a single outdoor image." CVPR'2019

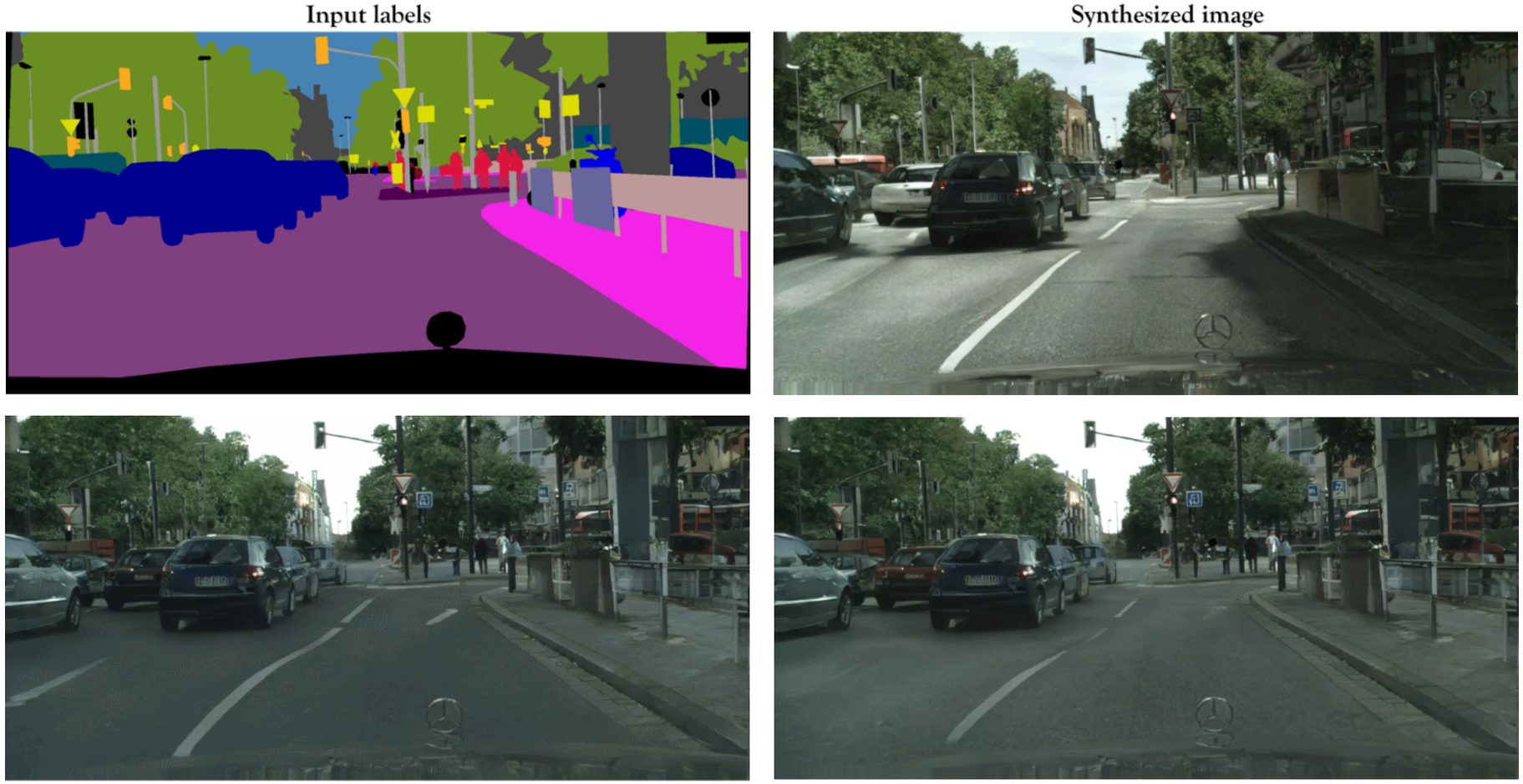
Video Translation

Translate a video between domains

- Pix2PixHD
- Everybody dance now
- Vid2Vid
- Deep video portraits
- Few shot video to video



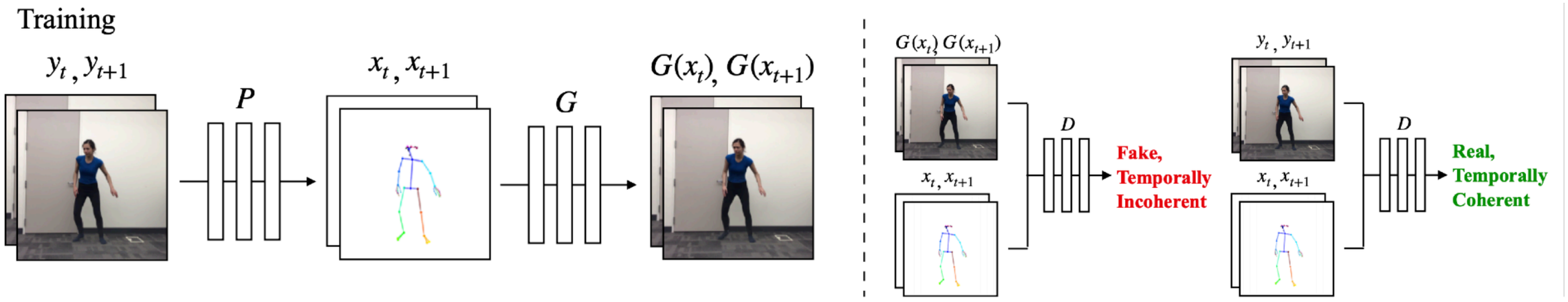
Pix2PixHD: Results

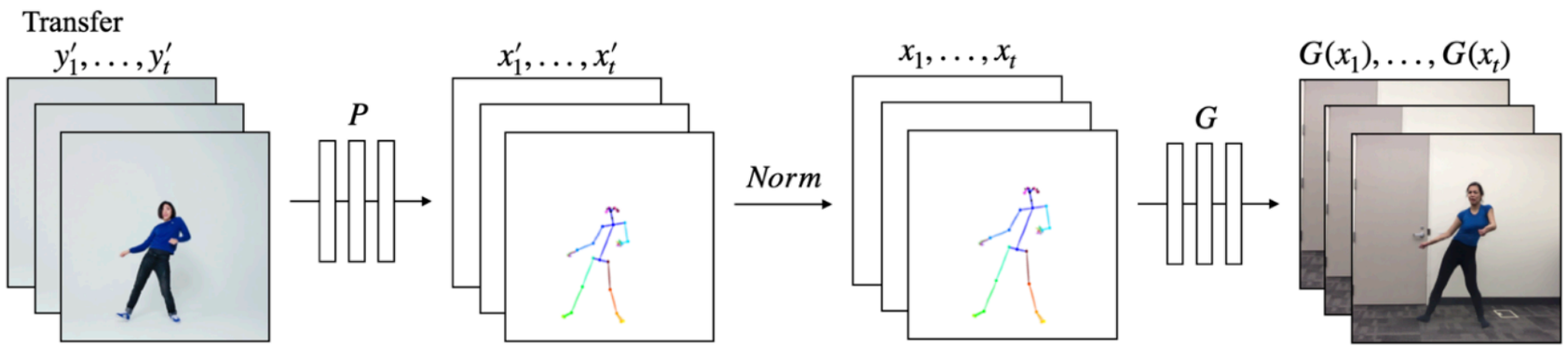


Wang, Ting-Chun, et al. "Pix2pixHD: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs." CVPR'2018

Animating Single Subject

Paired motion retargeting—"Do as I do"





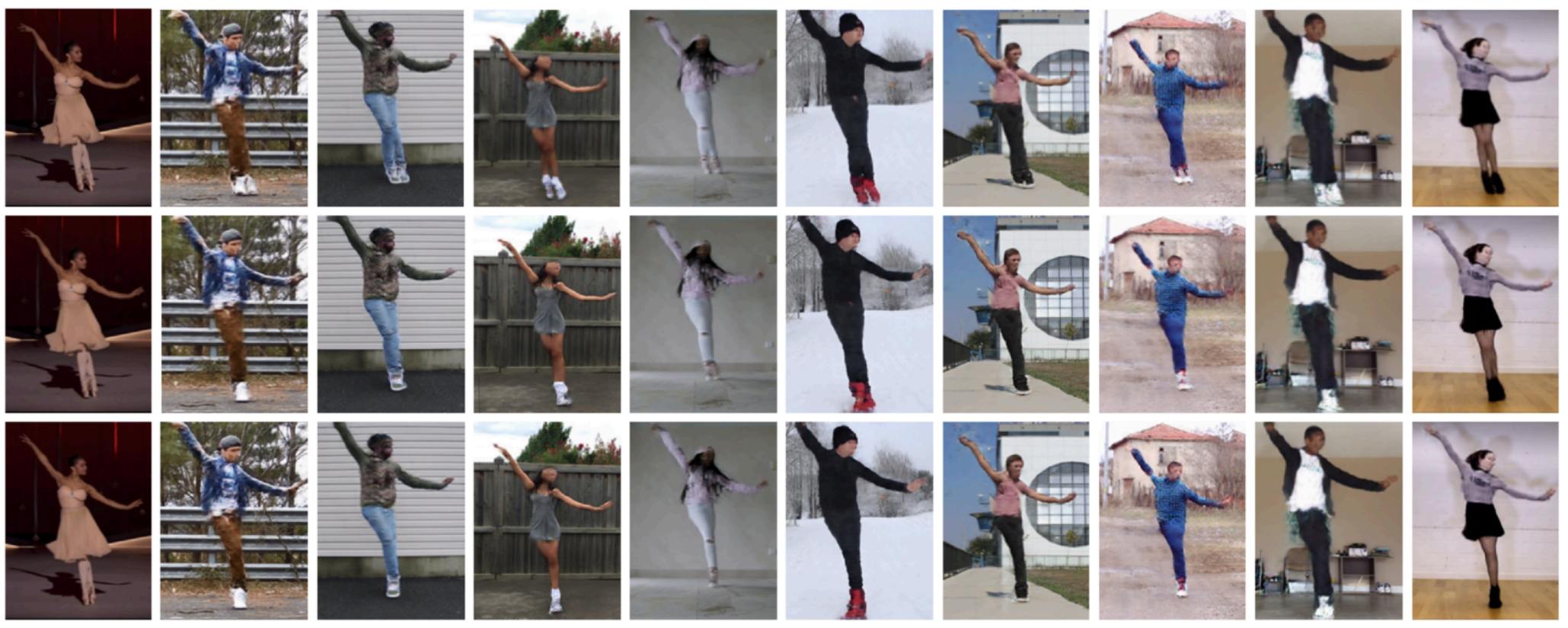
Chan, Caroline, et al. "Everybody dance now." ICCV'2019.

Retargeting Appearance vs Motion

In fact, everybody-dance now is about appearance retargeting

Motion: how would you do what I'm doing?

Appearance: how would I do what I'm doing in your jeans and jacket?



Chan, Caroline, et al. "Everybody dance now." ICCV'2019.

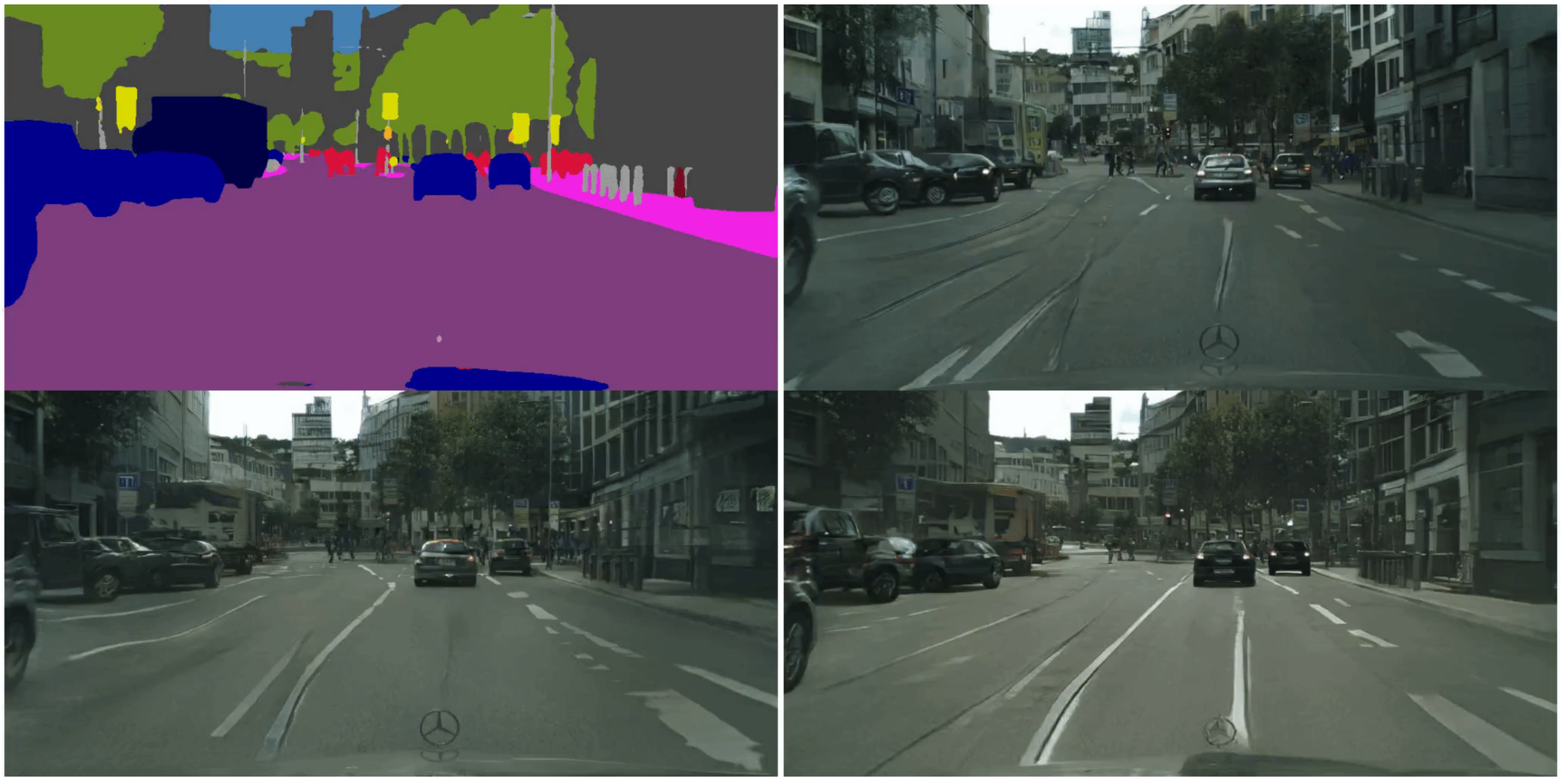
Results

Chan, Caroline, et al. "Everybody dance now." ICCV'2019.

Video-to-video Synthesis

Input

Pix2PixHD



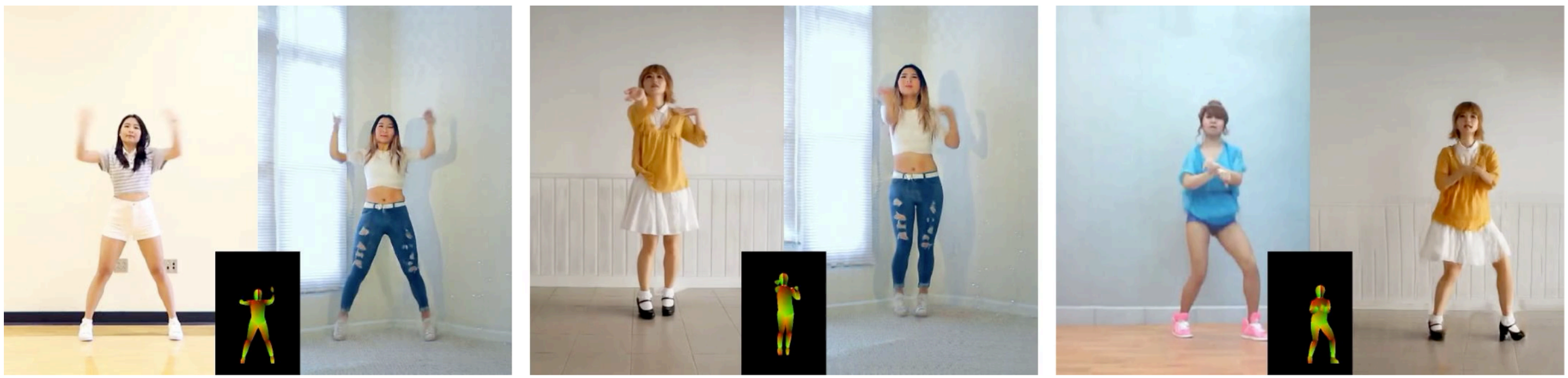
COVST-warping based

vid2vid

Wang, Ting-Chun, et al. "Video-to-video synthesis." NIPS'2018

Video-to-video Synthesis

Again, appearance transfer



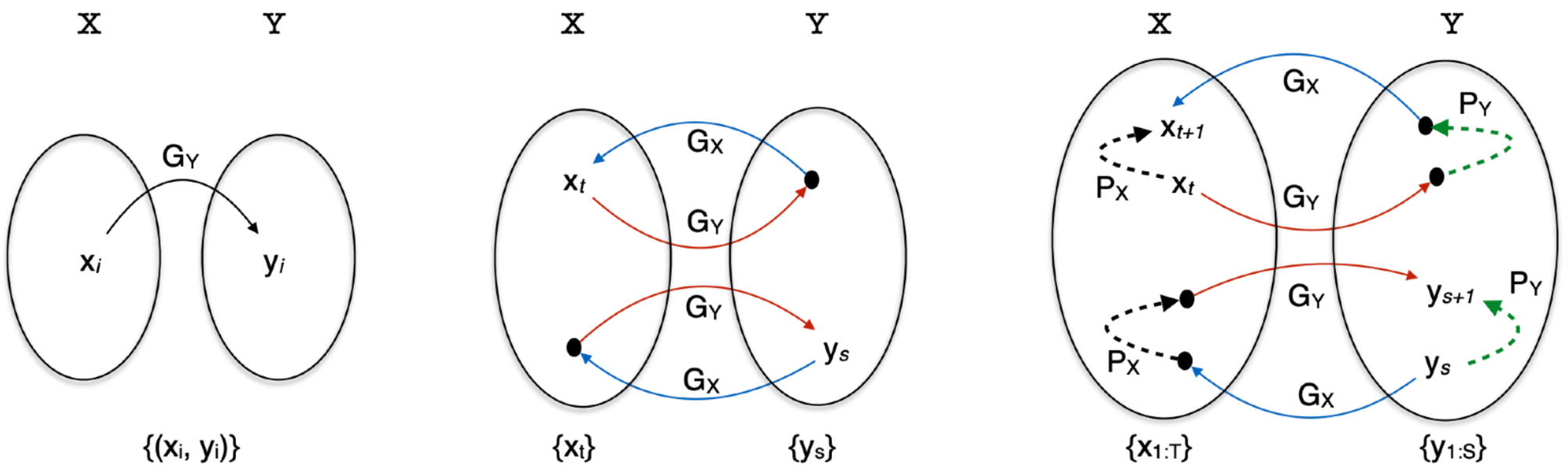
Wang, Ting-Chun, et al. "Video-to-video synthesis." NIPS'2018

RecycleGAN: Unpaired Vid2Vid

Pix2Pix: use **pairs** of images for **paired** i2i translation

CycleGAN: use cycle consistency to do **unpaired** i2i translation

RecycleGAN: use **temporal** cycle consistency to do **unpaired** vid2vid translation



Pix2Pix CycleGAN RecycleGAN

Bansal, Aayush, et al. "Recycle-gan: Unsupervised video retargeting." ECCV'2018

RecycleGAN: Unpaired Vid2Vid

RecycleGAN does video retargeting, but since it only can retarget a single video, we consider it as a video2video translation method

Bansal, Aayush, et al. "Recycle-gan: Unsupervised video retargeting." ECCV'2018

RecycleGAN: Unpaired Vid2Vid

RecycleGAN does video retargeting, but since it only can retarget a single video, we consider it as a video2video translation method

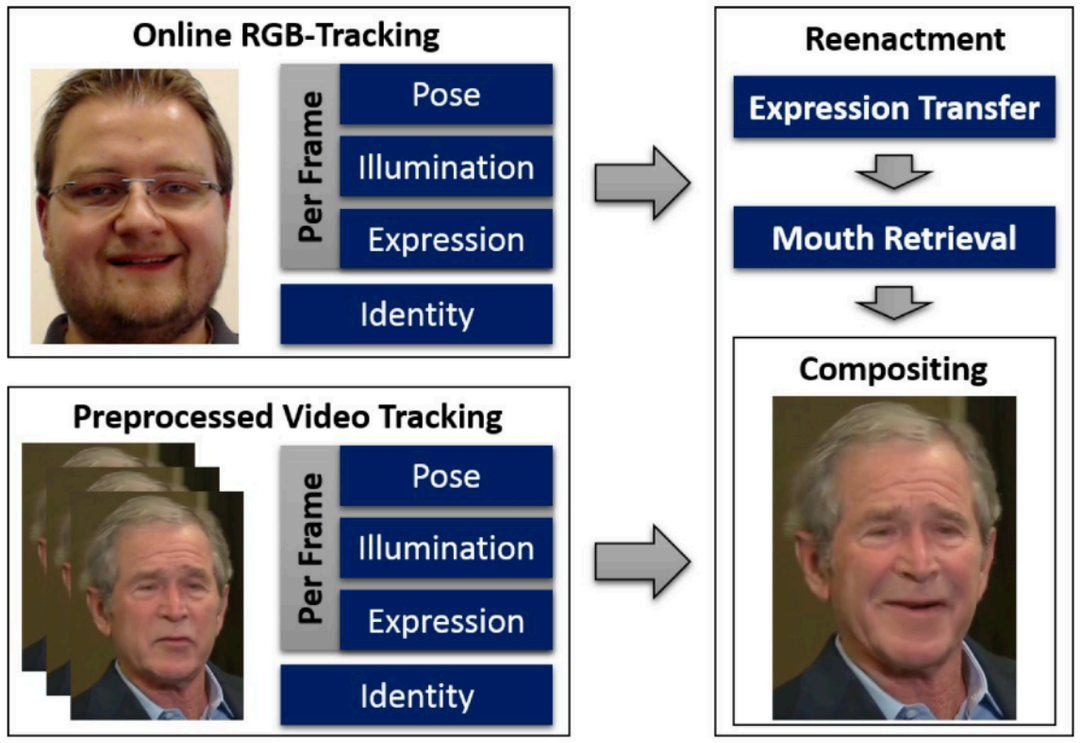
Bansal, Aayush, et al. "Recycle-gan: Unsupervised video retargeting." ECCV'2018

Face2Face

- Person-independent
- Operates only on faces
- Uses a very strong prior for re-enacting
- Works on high quality videos
- But it is one of the first works to photo-realistically retarget motions from faces

Thies, Justus, et al. "Face2face: Real-time face capture and reenactment of rgb videos." CVPR'2016

Face2Face



- Fit a 3DMM model to each subject
- Re-enact target
- Render

Thies, Justus, et al. "Face2face: Real-time face capture and reenactment of rgb videos." CVPR'2016

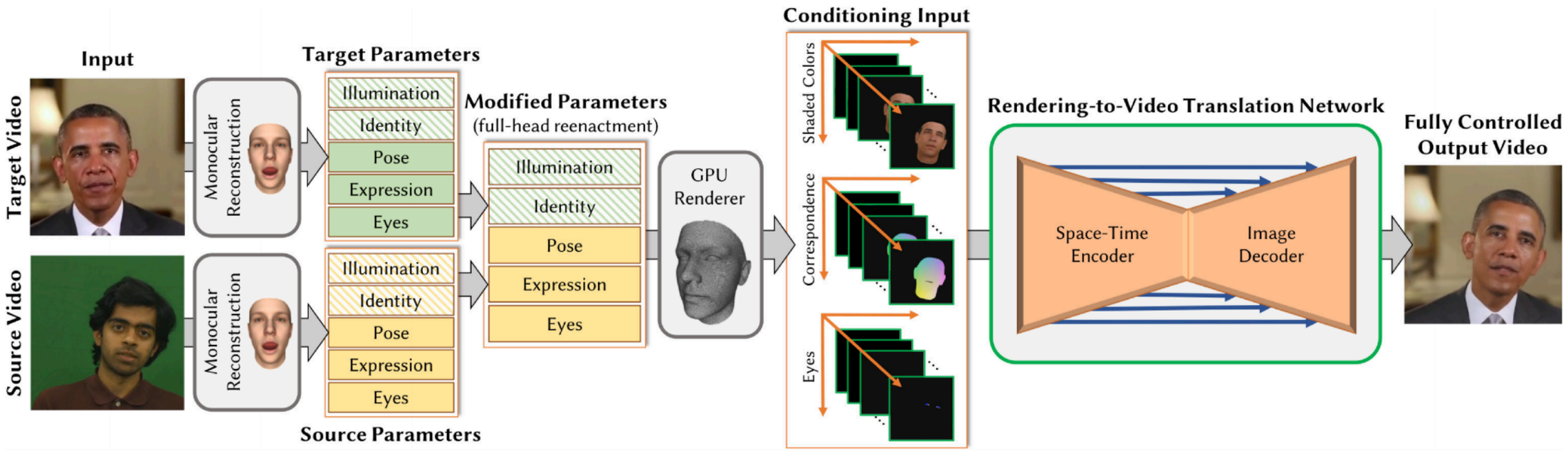
Face2Face

Thies, Justus, et al. "Face2face: Real-time face capture and reenactment of rgb videos." CVPR'2016

Face2Face


Real-time Facial Reenactment
"THE REMOVAL OF SADDAM HUSSEIN WAS RIGHT DECISION"
Pres. Bush reflects on the Iraq War
LIVE
CNN
ARRIS LIVED IN WEST; WAS HELPING BATTLE FIRE WHEN THE EXPLO
10:41 PM ET
DELL
Live capture using a commodity webcam

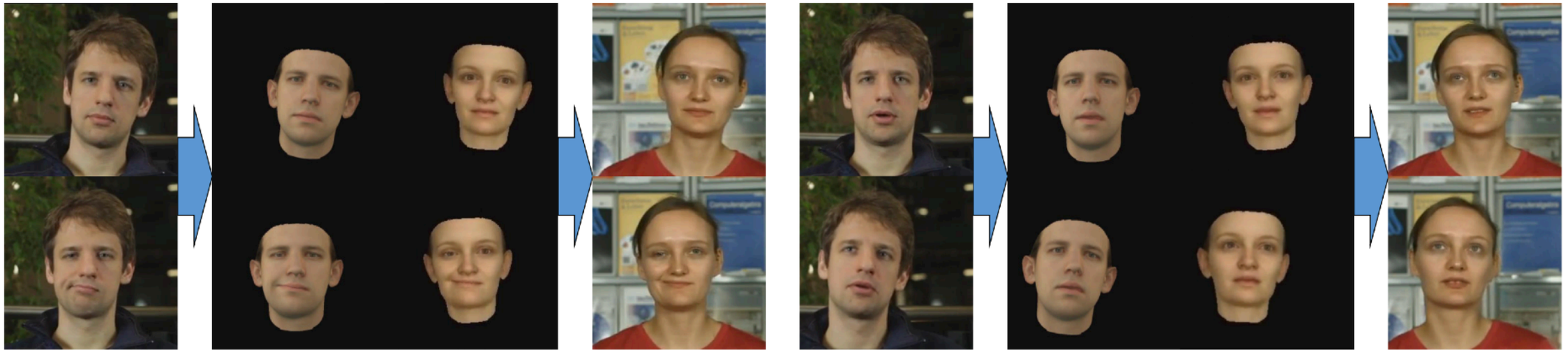
Face2Face -> Deep Video Portraits



- Person-**dependent**
- Operates only on faces
- Uses a very strong prior for re-enacting
- Works on high quality videos
- Uses graphics and neural networks to re-enact or translate

Kim, Hyeongwoo, et al. "Deep video portraits." TOG'2018

Face2Face -> Deep Video Portraits



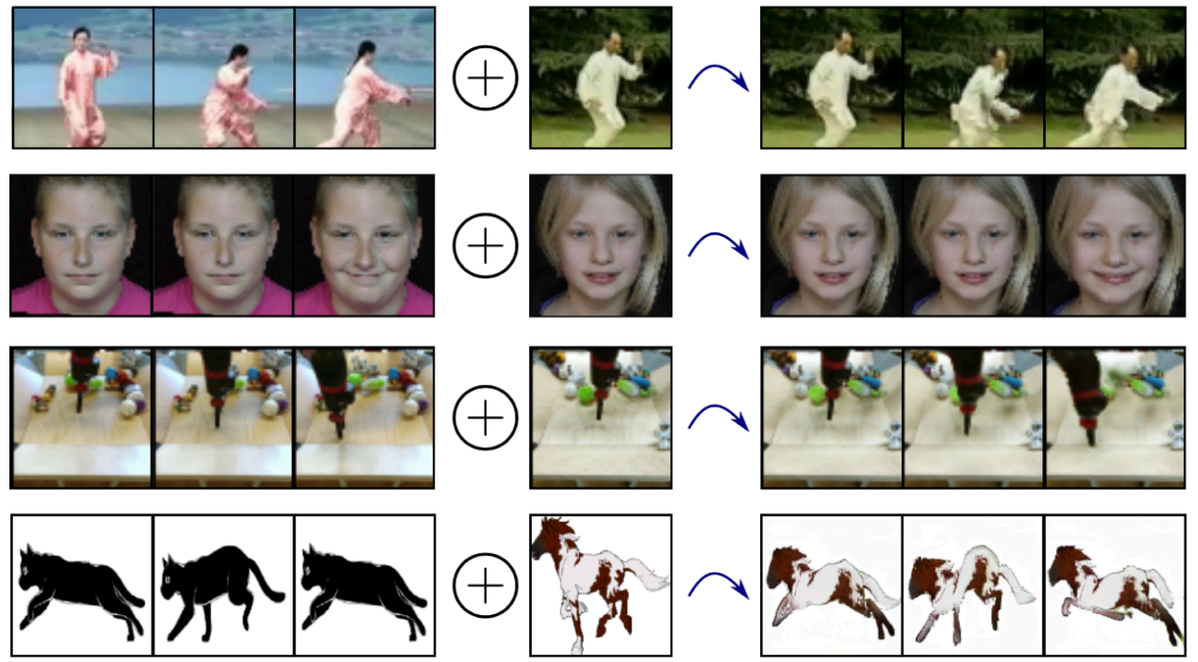
- Person-**dependent**
- Operates only on faces
- Uses a very strong prior for re-enacting
- Works on high quality videos
- Uses graphics and neural networks to re-enact or translate

Kim, Hyeongwoo, et al. "Deep video portraits." TOG'2018

Video Retargeting

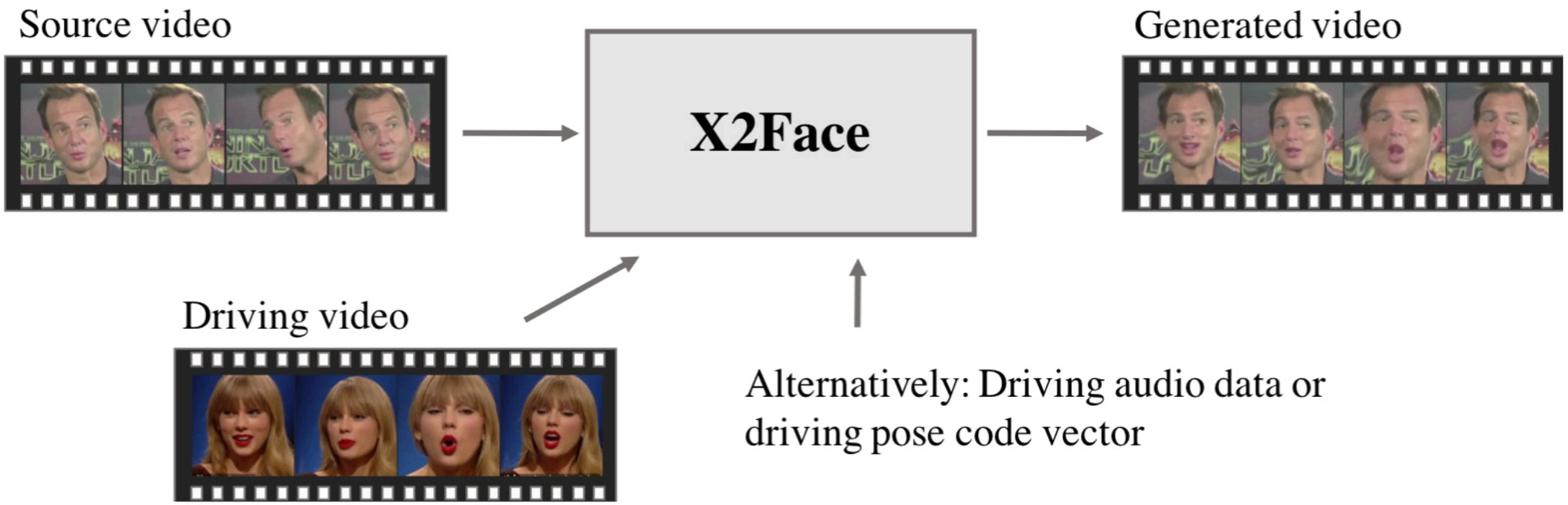
Extract and retarget motion from the driving and apply it to source

- Face2Face
- X2Face
- Talking face generation
- MonkeyNet and FOMM



X2Face

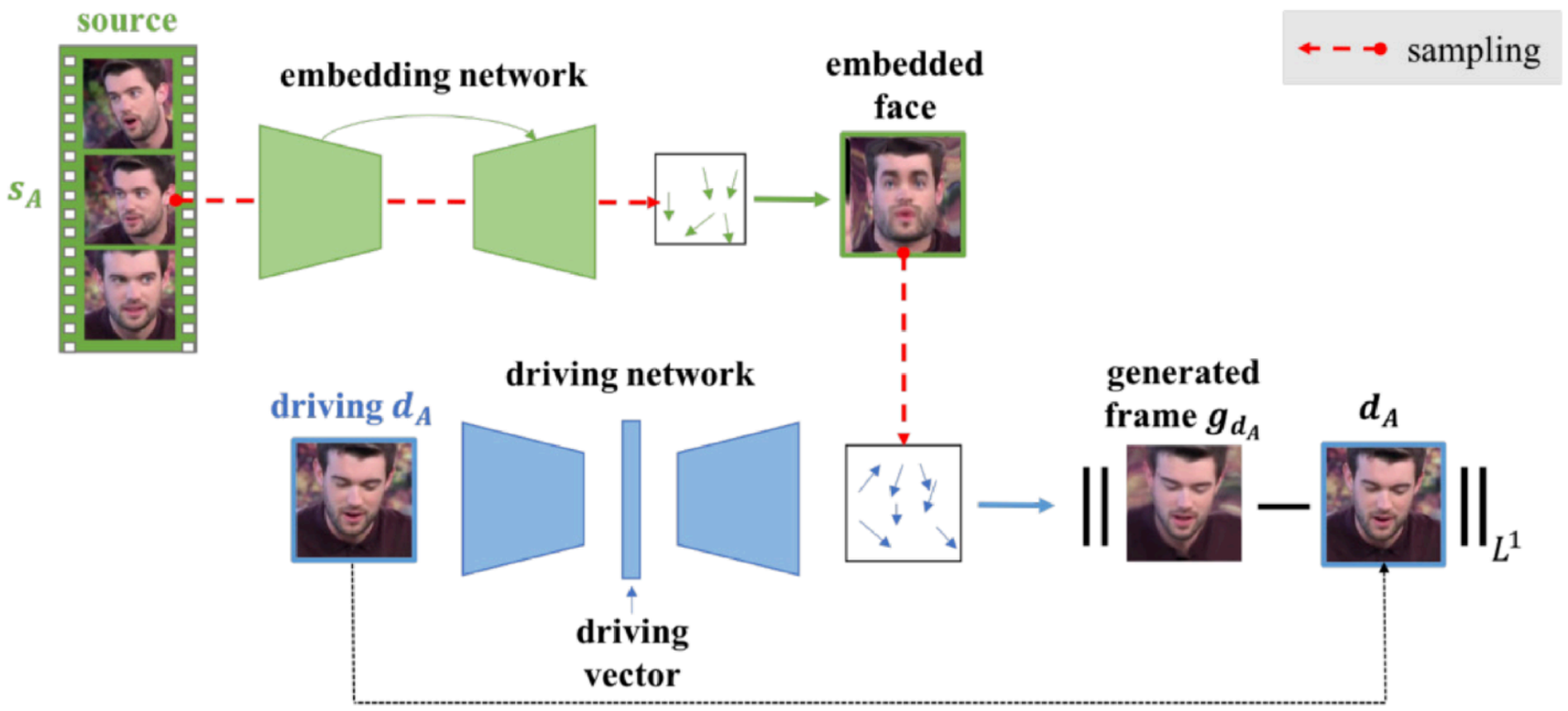
Given an input sequence animate it using driving video or **audio**



- Face only model
- Presented also a new dataset with thousands talking people

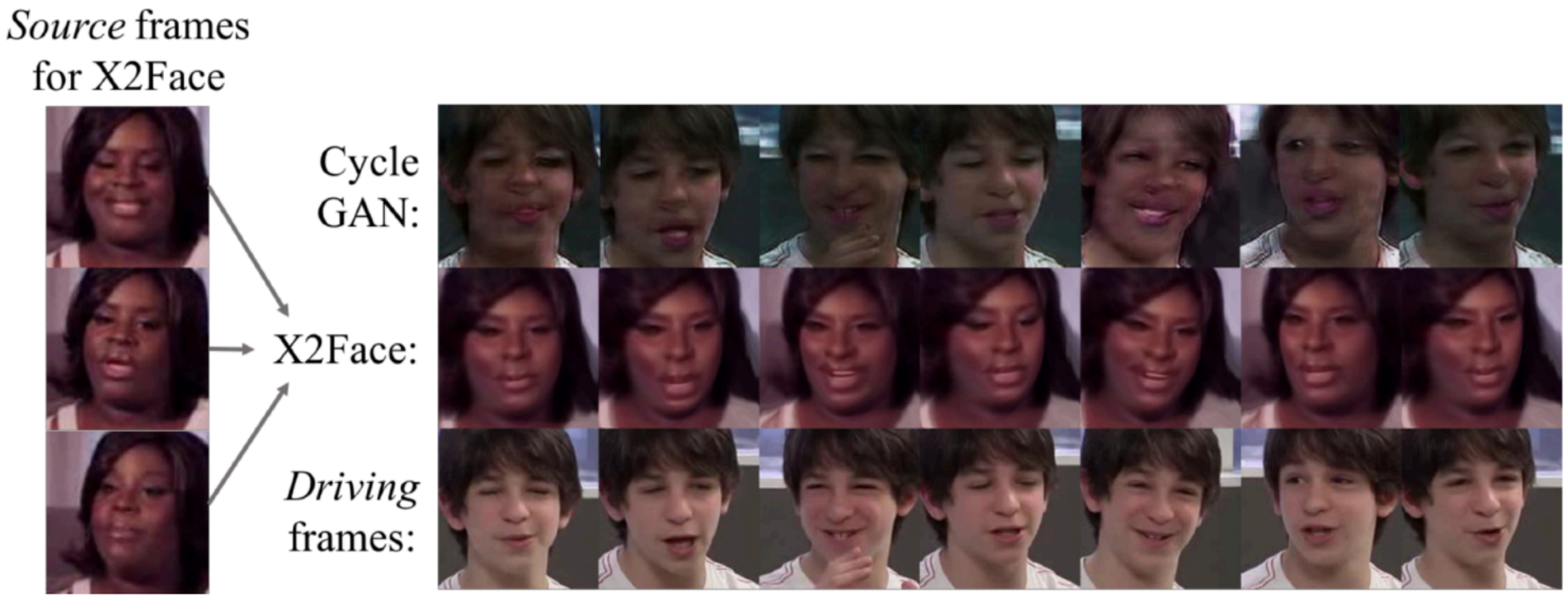
Content and Motion Disentanglement

Content is an image, motion is a vector field

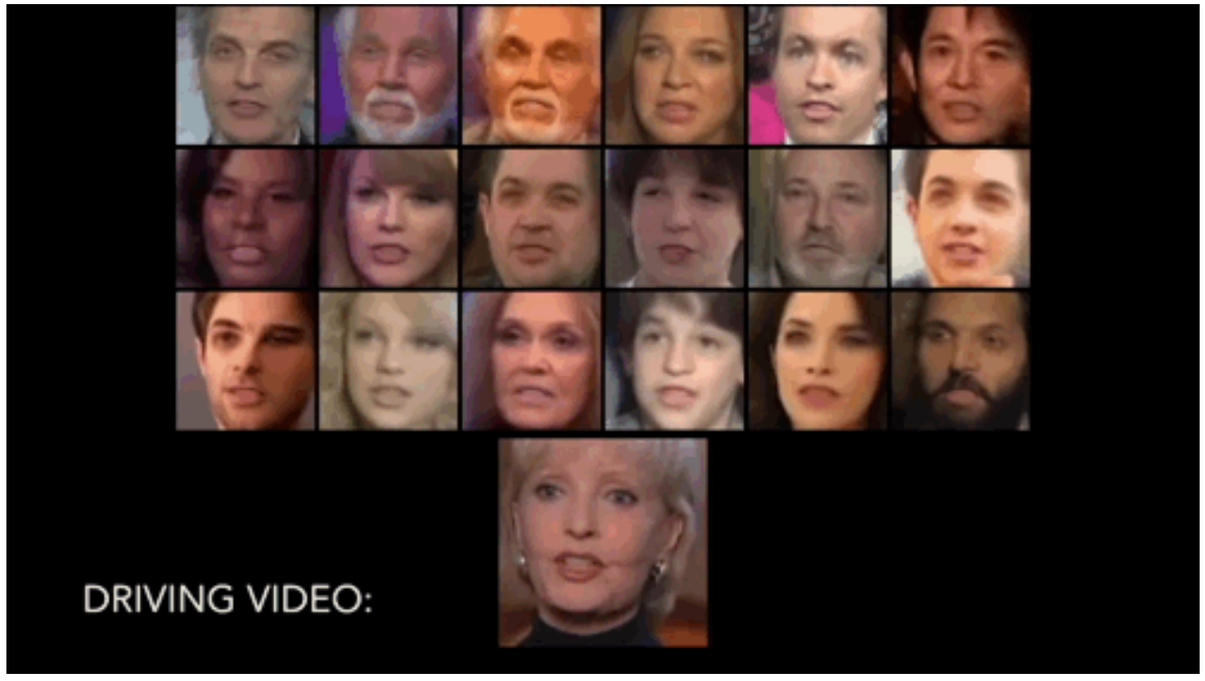


- Embedding network combines input images into a unified embedded face
- Driving network estimates flow field either from the driving image or other inputs, such as audio or pose
- Cannot generate parts not present in the embedding, for example teeth

Comparisons

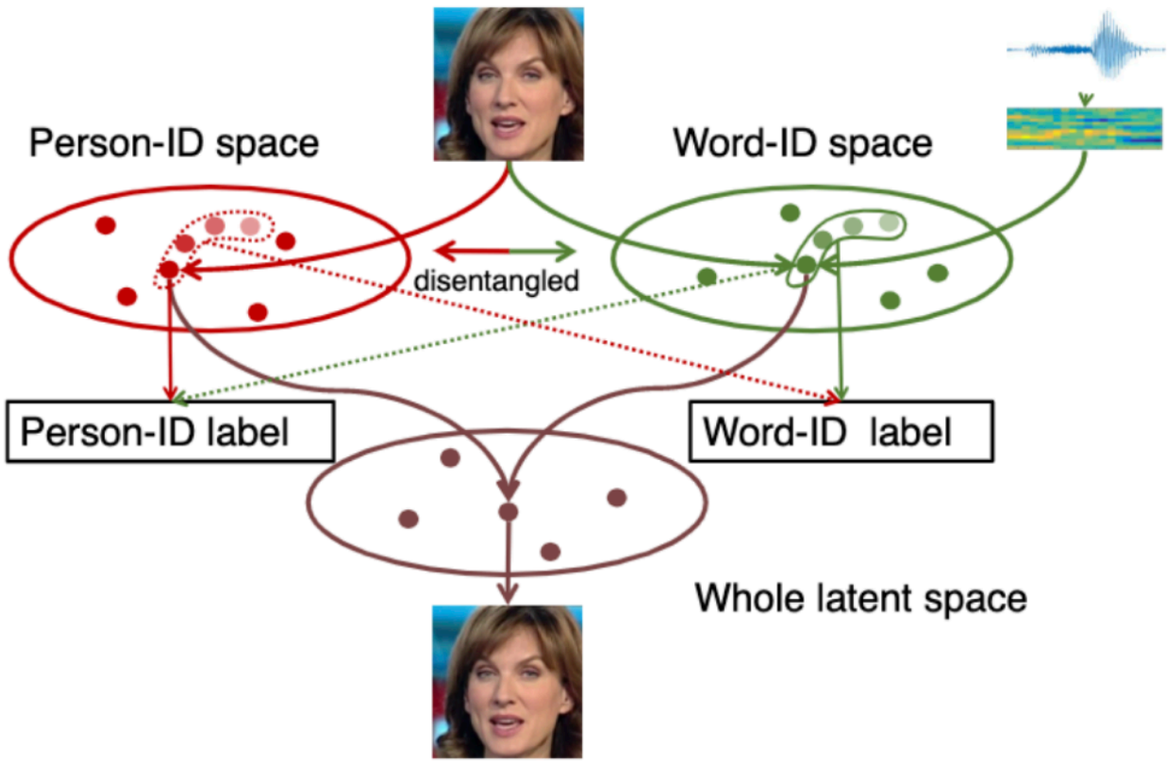


More Results



Talking Face Generation

Animate faces using audios or videos by disentangling information

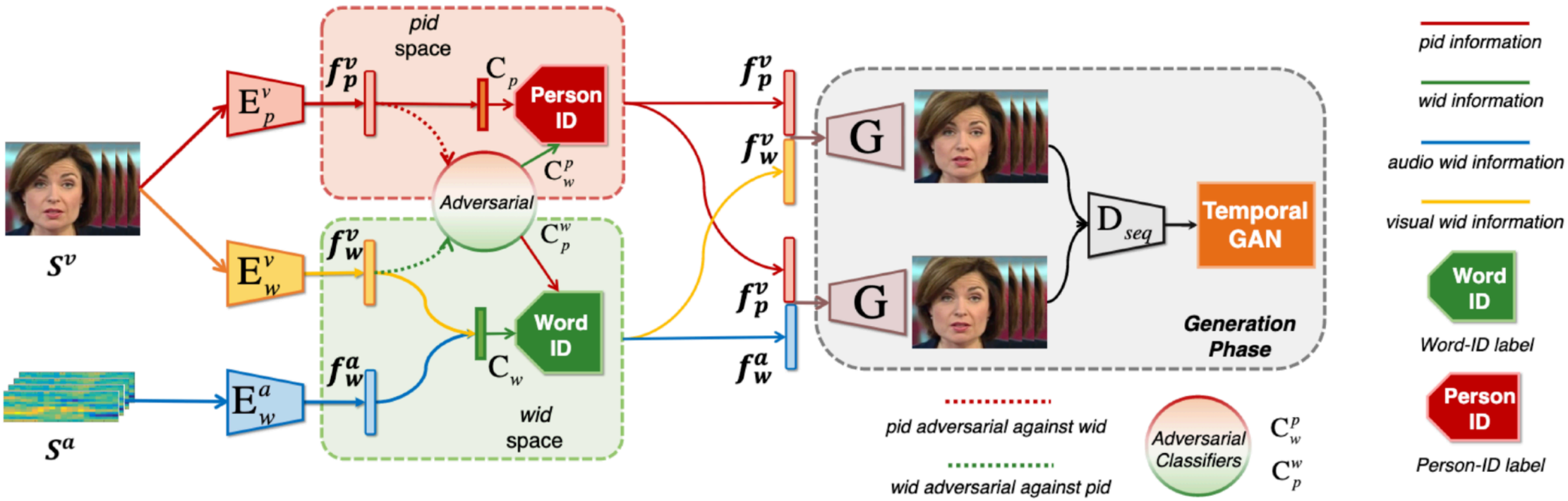


- Extracts appearance information (or person-id) from the video
- Extracts word-id either from video or from audio
- Head pose information is not modeled

Zhou, Hang, et al. "Talking face generation by adversarially disentangled audio-visual representation." AAAI'2019

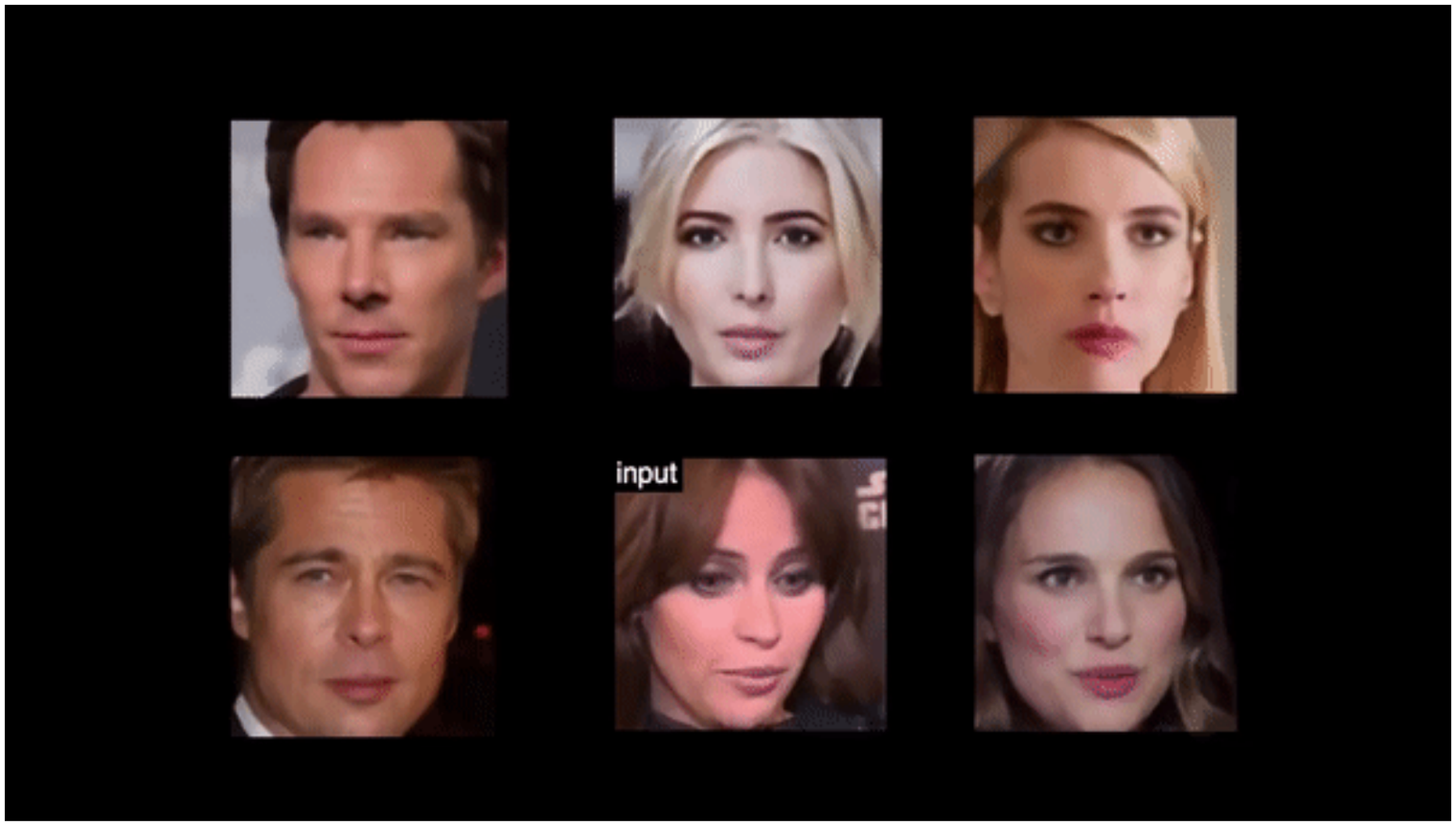
Training

Animate faces using audios or videos by disentangling information



Zhou, Hang, et al. "Talking face generation by adversarially disentangled audio-visual representation." AAAI'2019

Results

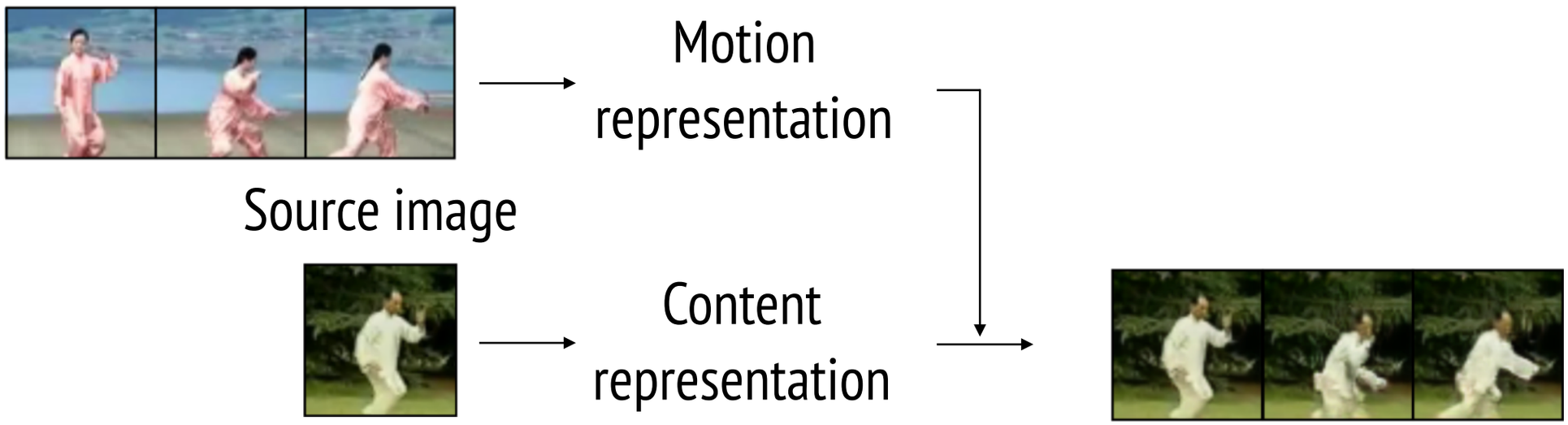


Zhou, Hang, et al. "Talking face generation by adversarially disentangled audio-visual representation." AAAI'2019

Retargeting Video for Arbitrary Objects

Previous methods considered only one type of motions: **talking humans**

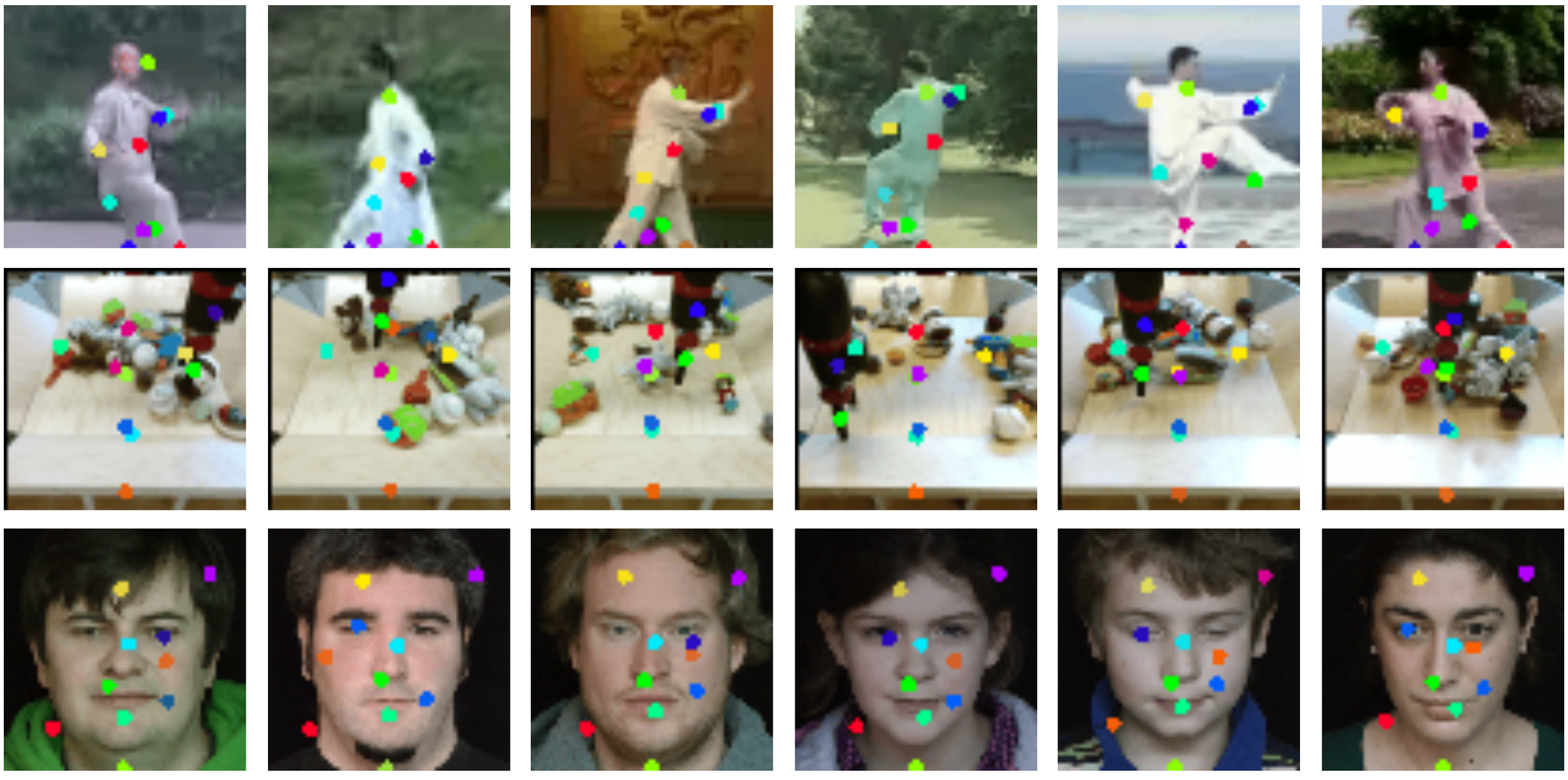
Can we animate any object by extracting motion from objects of the same class?



Siarohin, Aliaksandr, et al. "Animating arbitrary objects via deep motion transfer." CVPR'2019

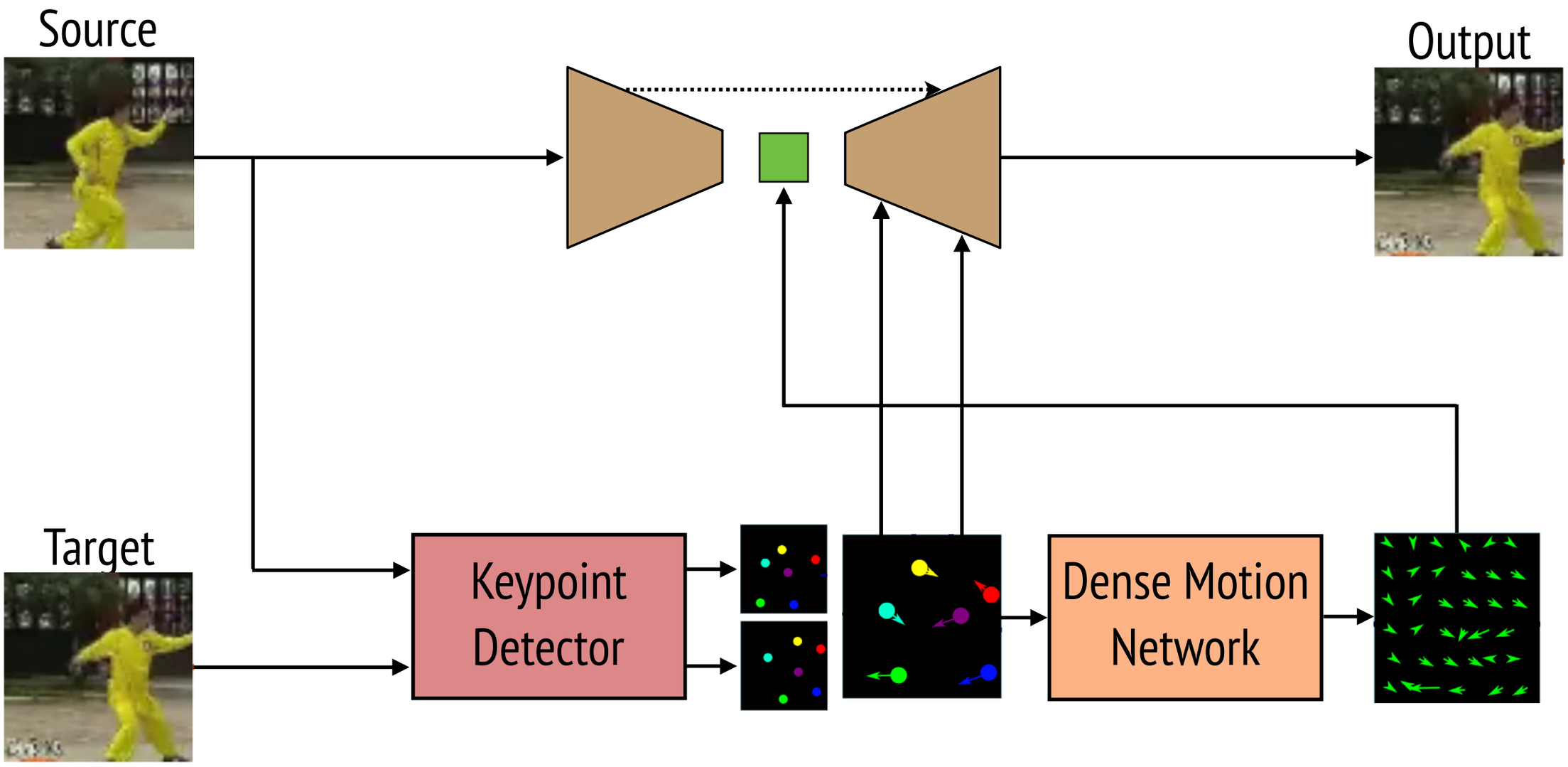
Retargeting Video for Arbitrary Objects

What is the appropriate motion representation?

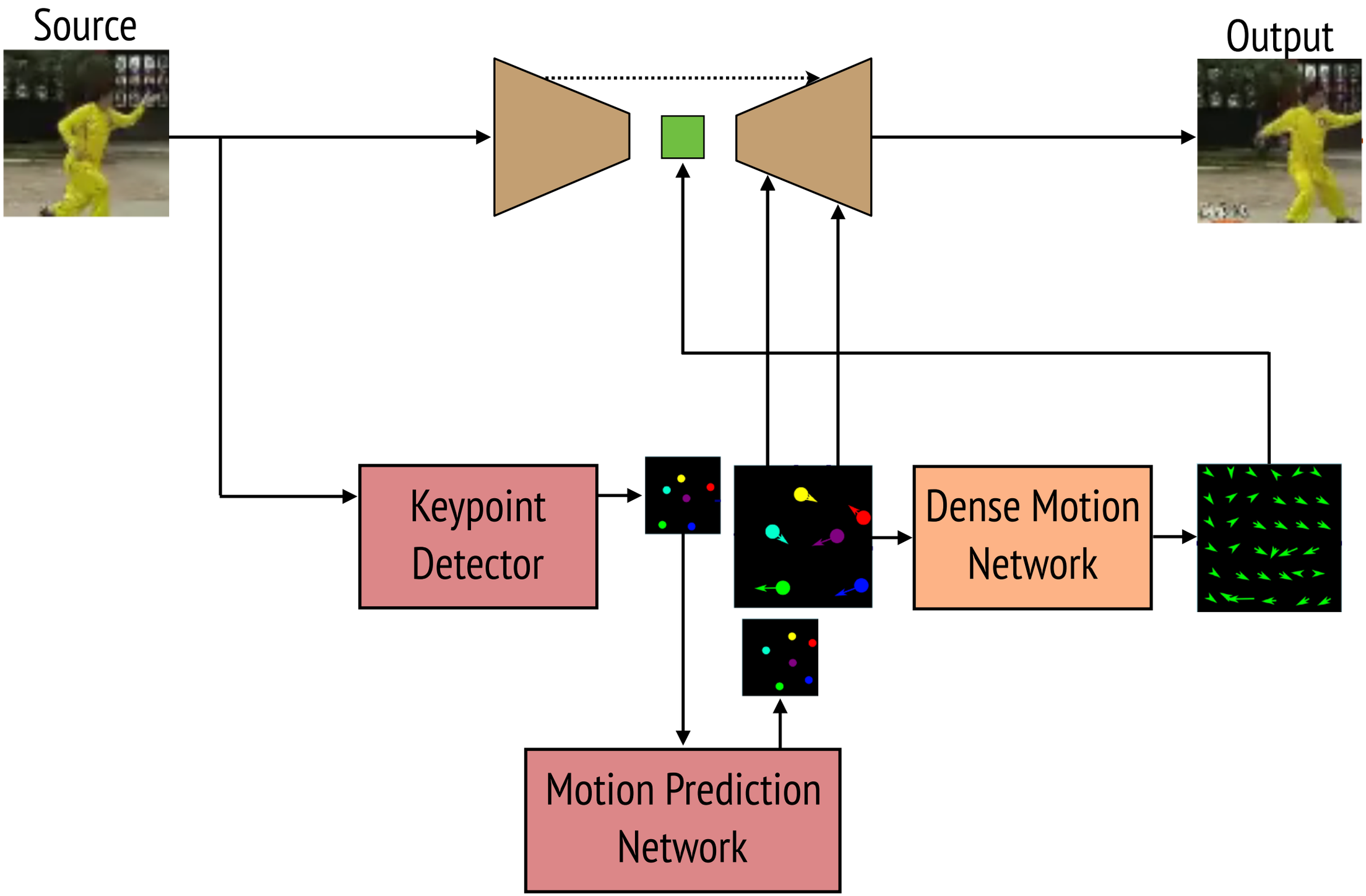


Siarohin, Aliaksandr, et al. "Animating arbitrary objects via deep motion transfer." CVPR'2019

End-to-end self-supervised training

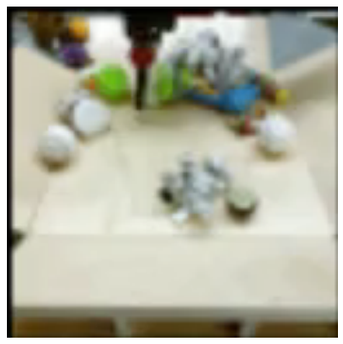


Video prediction



BAIR robot dataset

Driving video



Source image

Generated video

Source image

Generated video

NEMO face dataset

Driving video

Source image

Generated video

Source image

Generated video

Taichi dataset

Driving video

Source image

Generated video

Source image

Generated video

MGIF dataset

Driving video

Source image

Generated video

Source image

Generated video

More Recent Version

A single model animates all images given only a single source image

Driving video