Cross-modal understanding and generation of multimodal content

NICU SEBE
Univ. of Trento
niculae.sebe@unitn.it

Collaborators: Xavier Alameda-Pineda, Stephane Lathuiliere, Willi Menapace, Elisa Ricci, Subhankar Roy, Aliaksandr Siarohin, Hao Tang, Sergey Tulyakov, etc.

Deep Fakes: Driving Video, Static Input

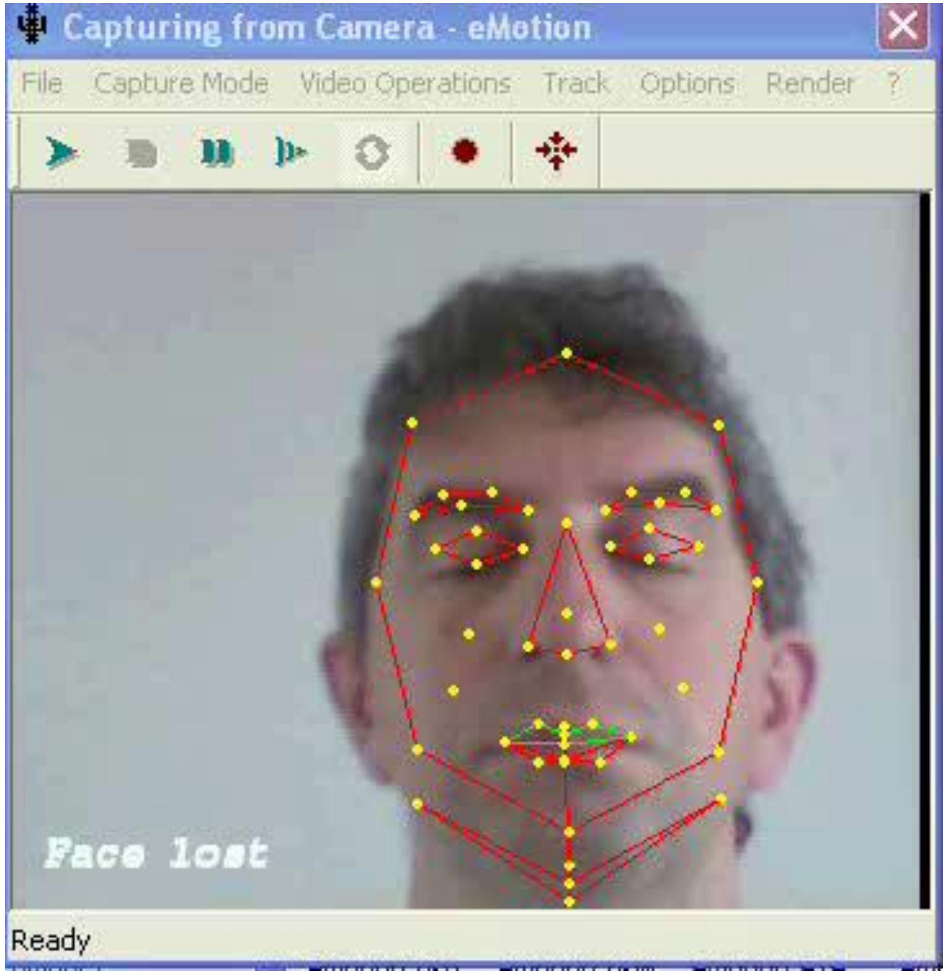
Deep Fakes: Video/Voice Inpainting

Creating Games with Real Footage

The player moves to the left corner waiting for the serve

The player serves the ball to the left corner of the field

A Bit of History



… about 2008

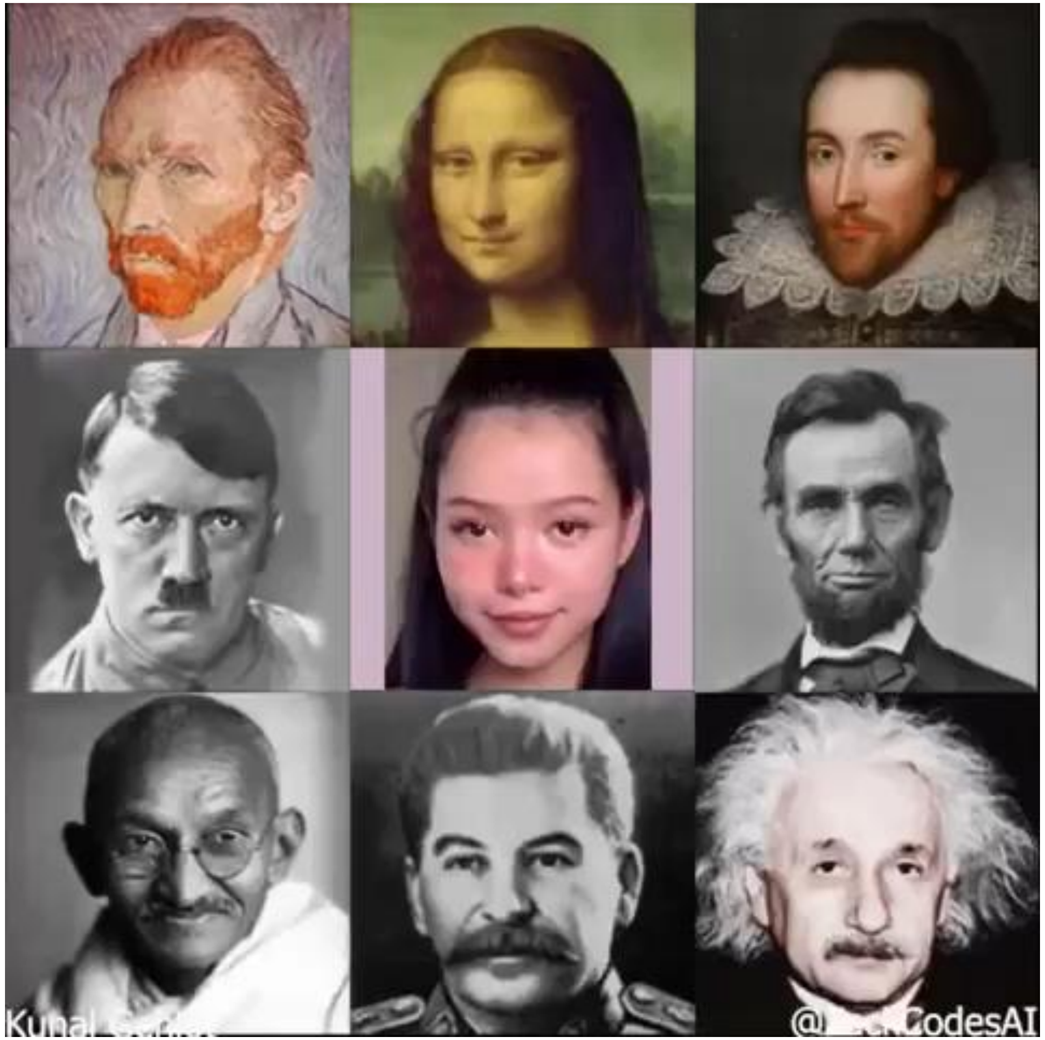
… about 2018

A Bit of History

… about 2019

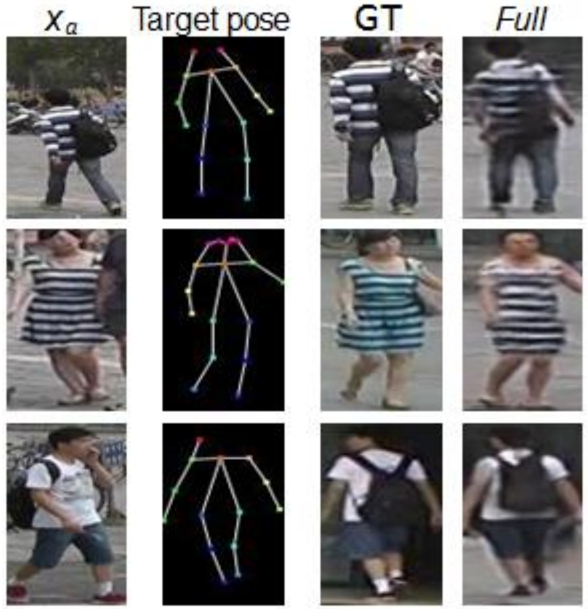
A Bit of History

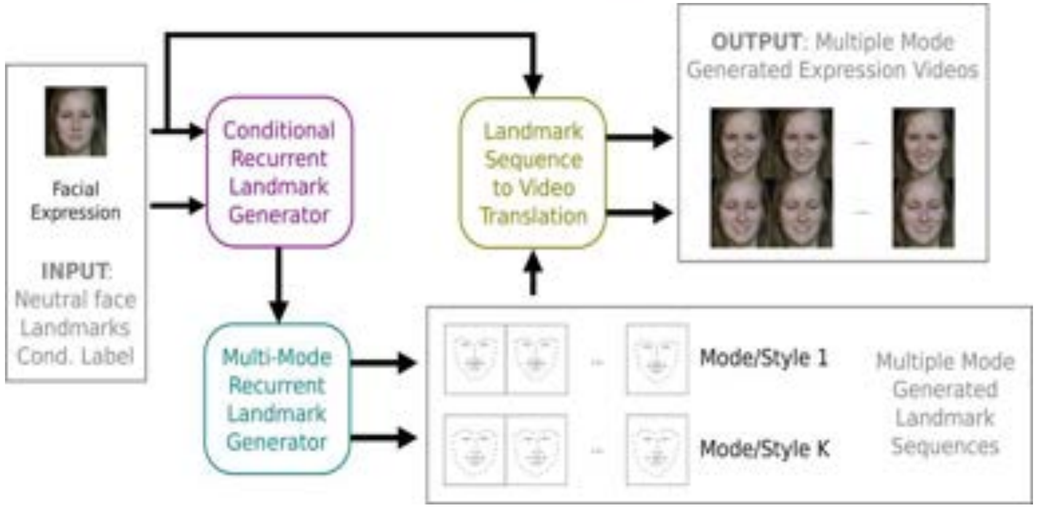


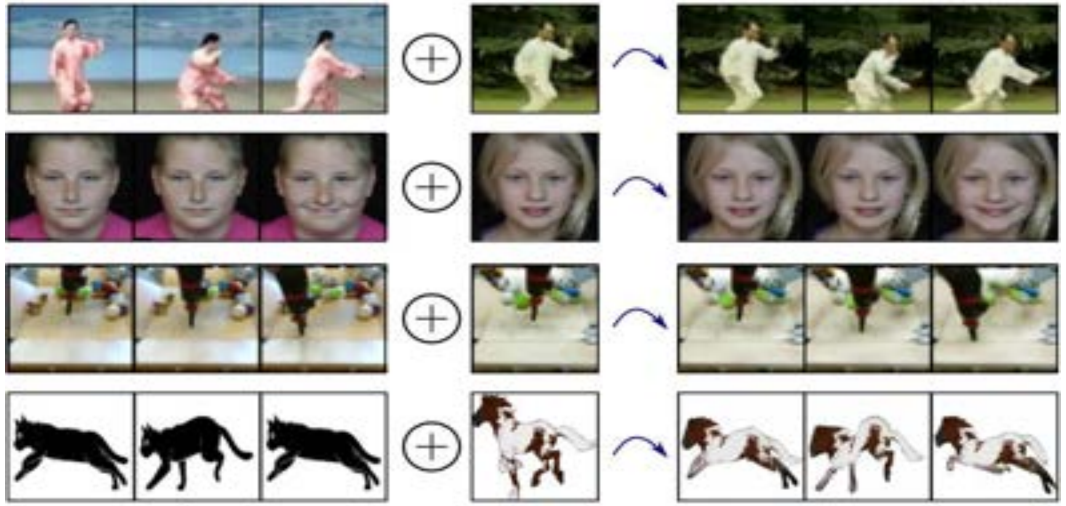
… nowadays

Image and Video Generation

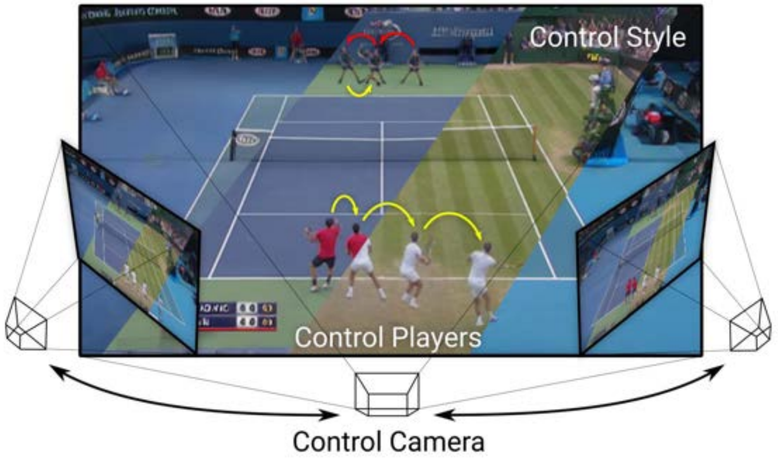
Deep Generative Models for Image/Video Generation & Animation







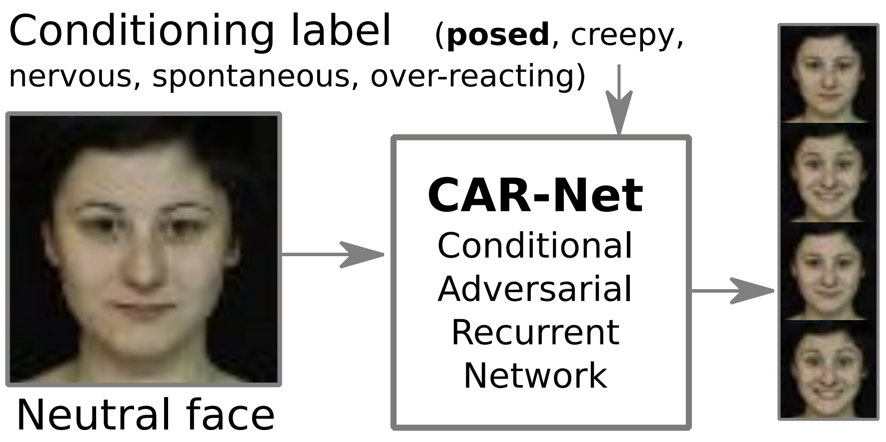
Arbitrary Object Animation with/without 3D Modeling



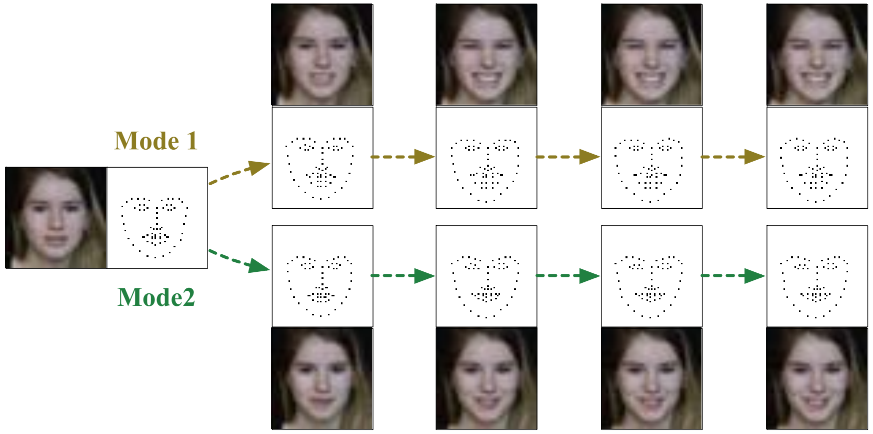
Diverse Smile Video Generation

- Wang, et al., “Every Smile is Unique: Landmark-Guided Diverse Smile Generation”, in CVPR 2018
- Wang, et al., “Learning How to Smile: Expression Video Generation with Conditional Adversarial Recurrent Nets”, in IEEE Transactions on Multimedia, 22(11):2808-2819, Nov. 2020

Landmark-Guided Diverse Smile Generation



(a) Generate sequence of smiles conditioned on labels

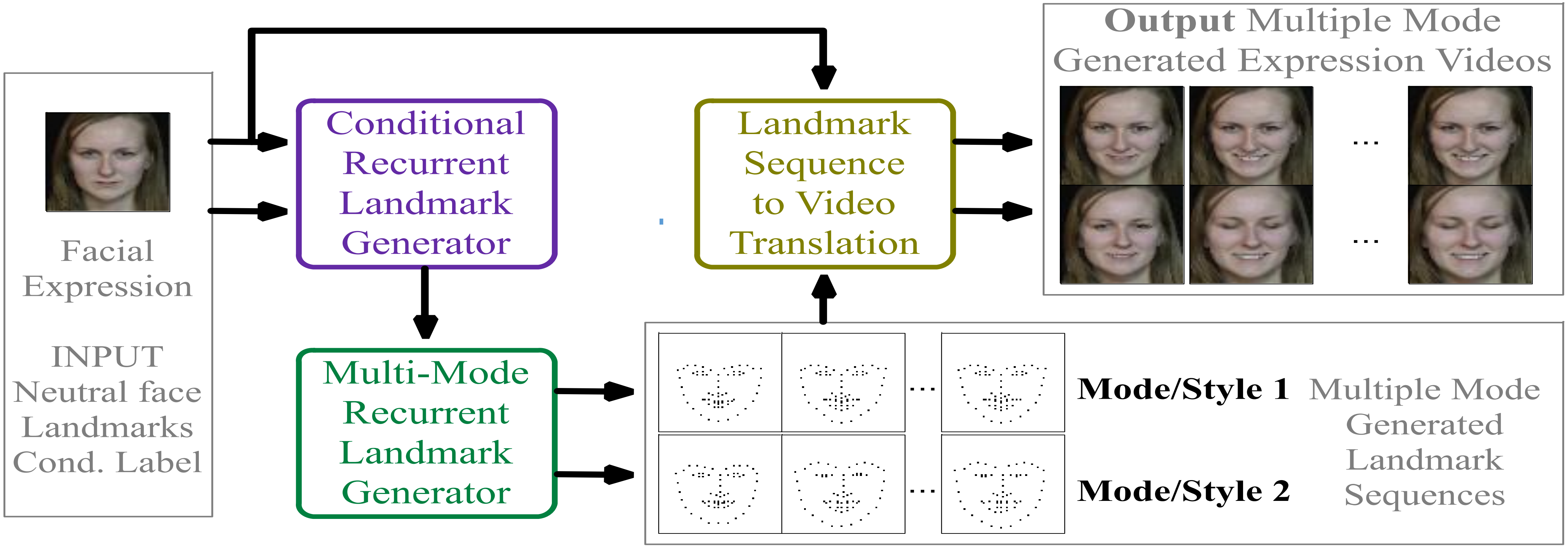


(b) Generate K different sequences of smiles

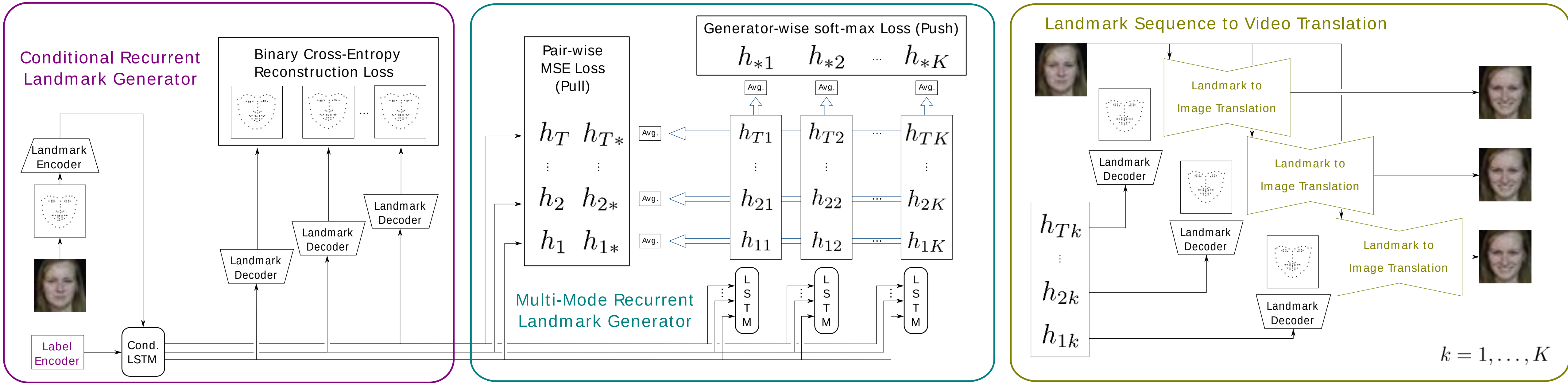
Challenges

- Sequence Generation conditioned on priors (i.e., input neutral face and smile label)
  - Conditional Recurrent Neural Network
- One-to-Many
  - Push-Pull Loss
- Preserve the identity
  - Landmark Sequence → Real Face via U-Net

Landmark-Guided Diverse Smile Generation

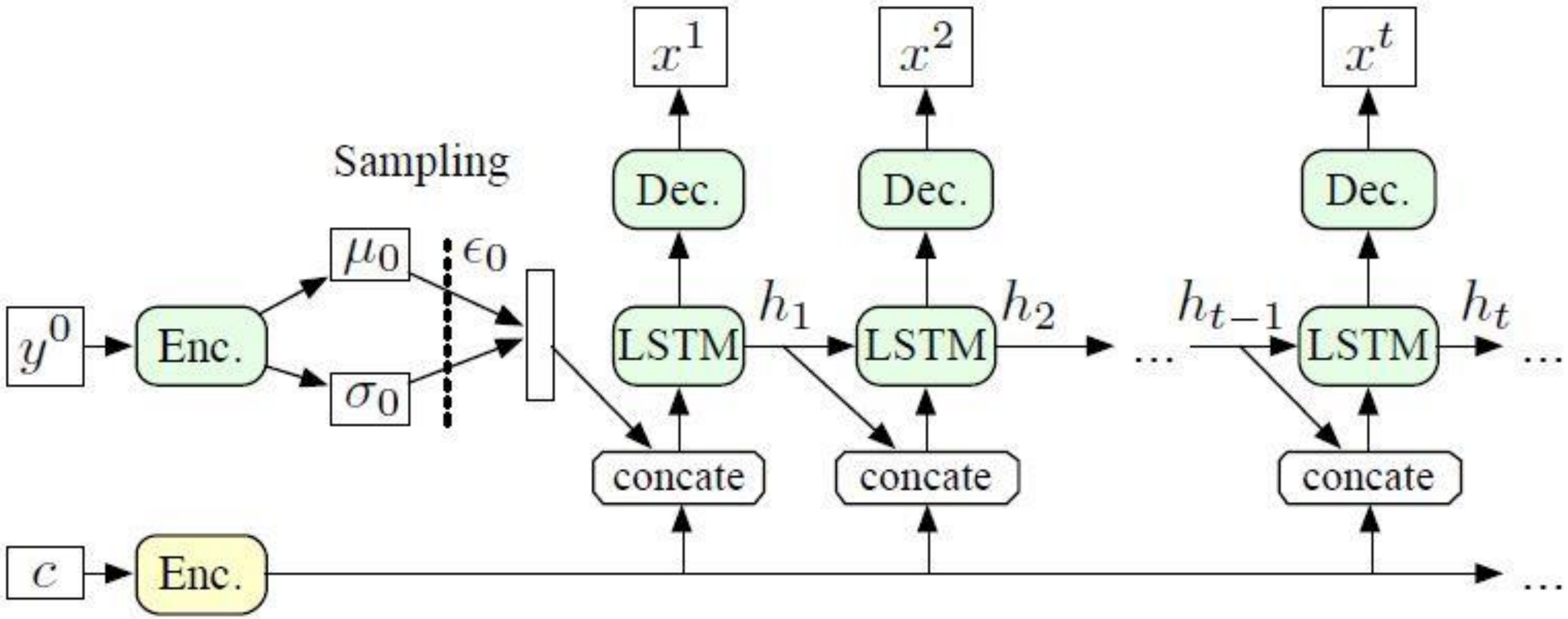


Landmark-Guided Diverse Smile Generation



- (left) encode the landmark image and generates a sequence of landmark embeddings according to the conditioning label
- (middle) generates K different landmark embedding sequences
- (right) translate each of the sequences into a face video

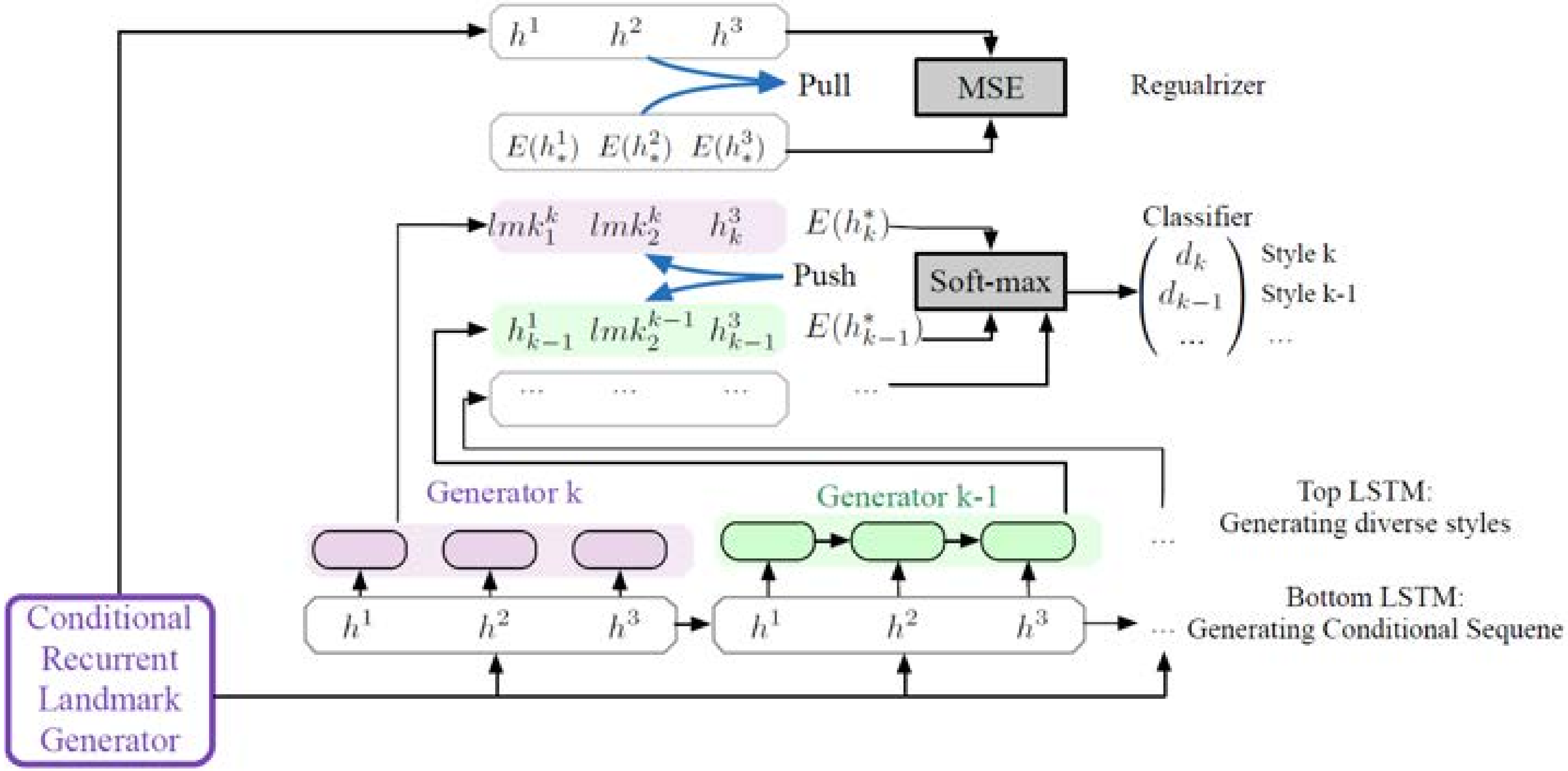
Landmark-Guided Diverse Smile Generation



(1) Conditional Recurrent Neural Network

- y^0 => initial input neutral face landmark image
- x^i => generated face landmark images
- LSTM is the recurrent unit receiving as input the concatenation of h_{t-1} and the embedding of conditioning label c

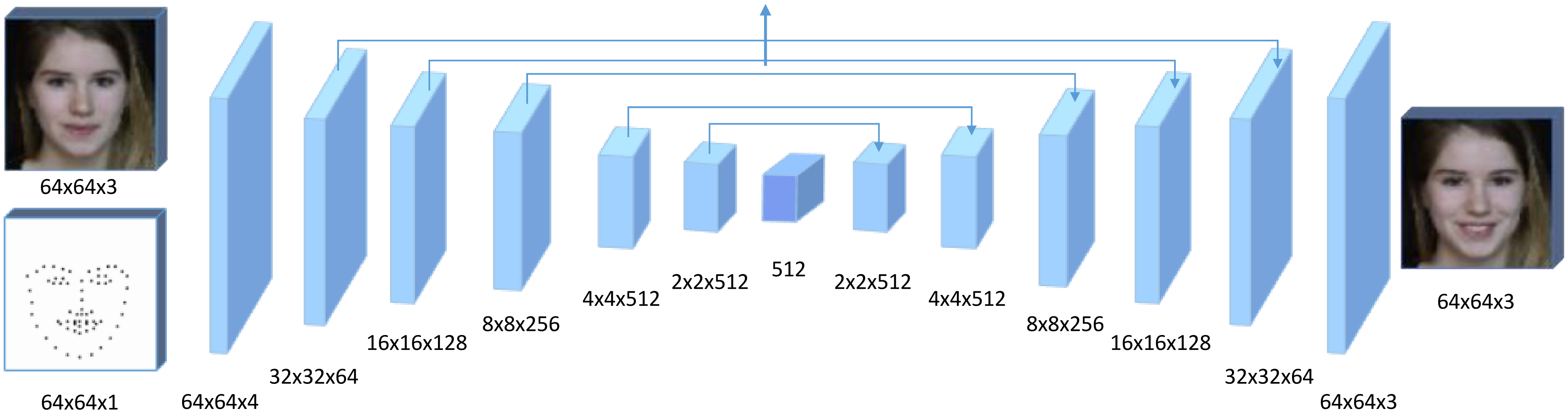
Landmark-Guided Diverse Smile Generation



(2) One-to-Many Mapping: Push & Pull loss

Landmark-Guided Diverse Smile Generation

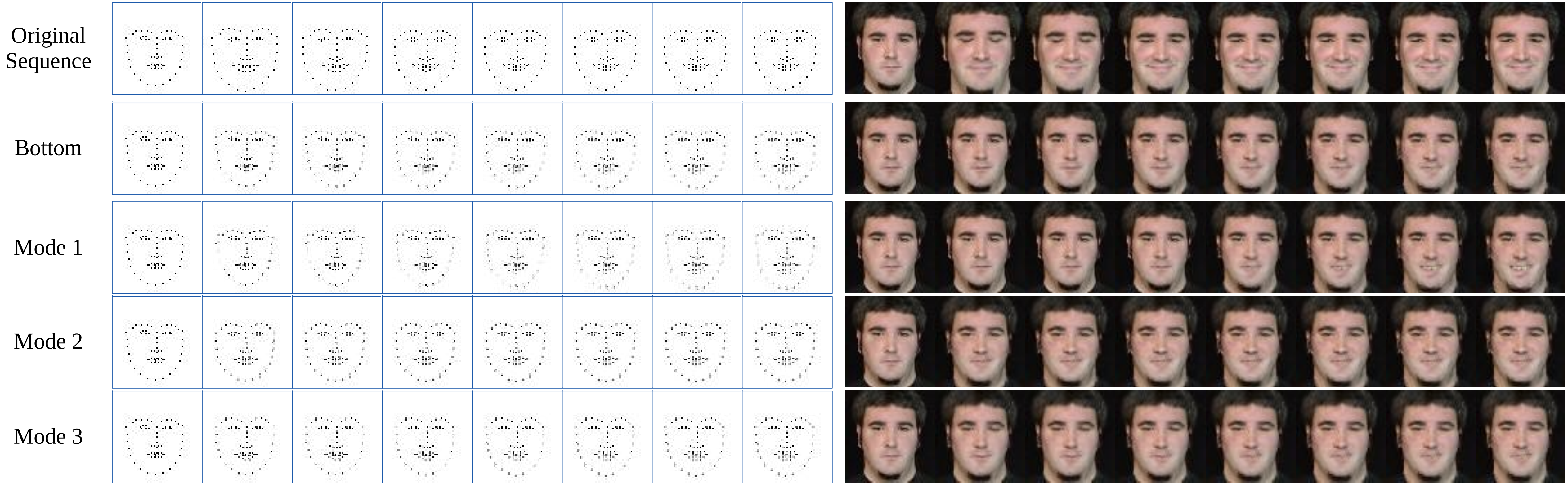
Skip Connection allows texture passing from source to target to preserve the identity



(3) Landmark Sequence to Video Generation via U-Net

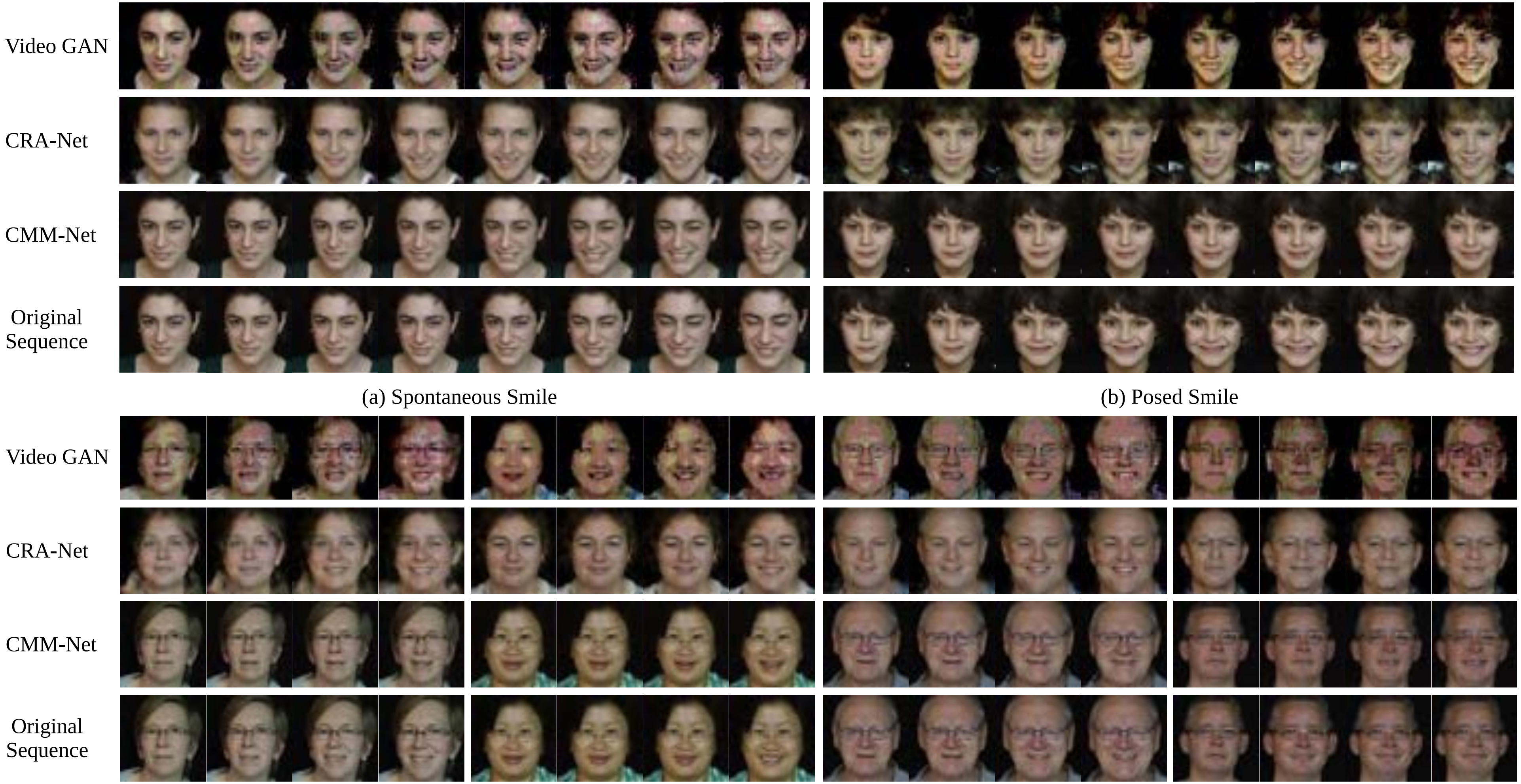
Landmark-Guided Diverse Smile Generation

Multi-Mode



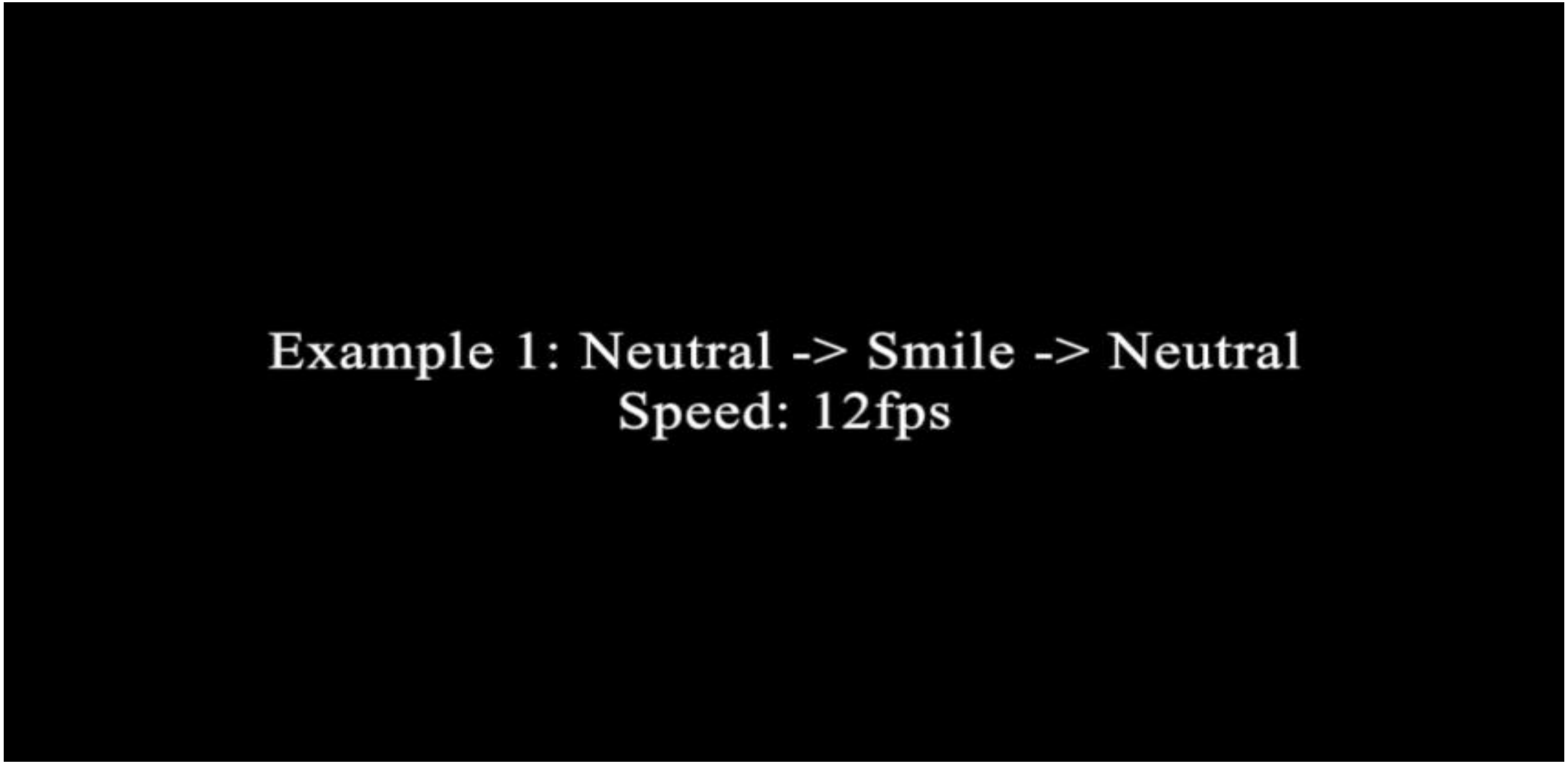
Landmark-Guided Diverse Smile Generation

Comparison with the state-of-the-art



(c) Spontaneous Smile with Glasses

(d) Posed Smile with Glasses

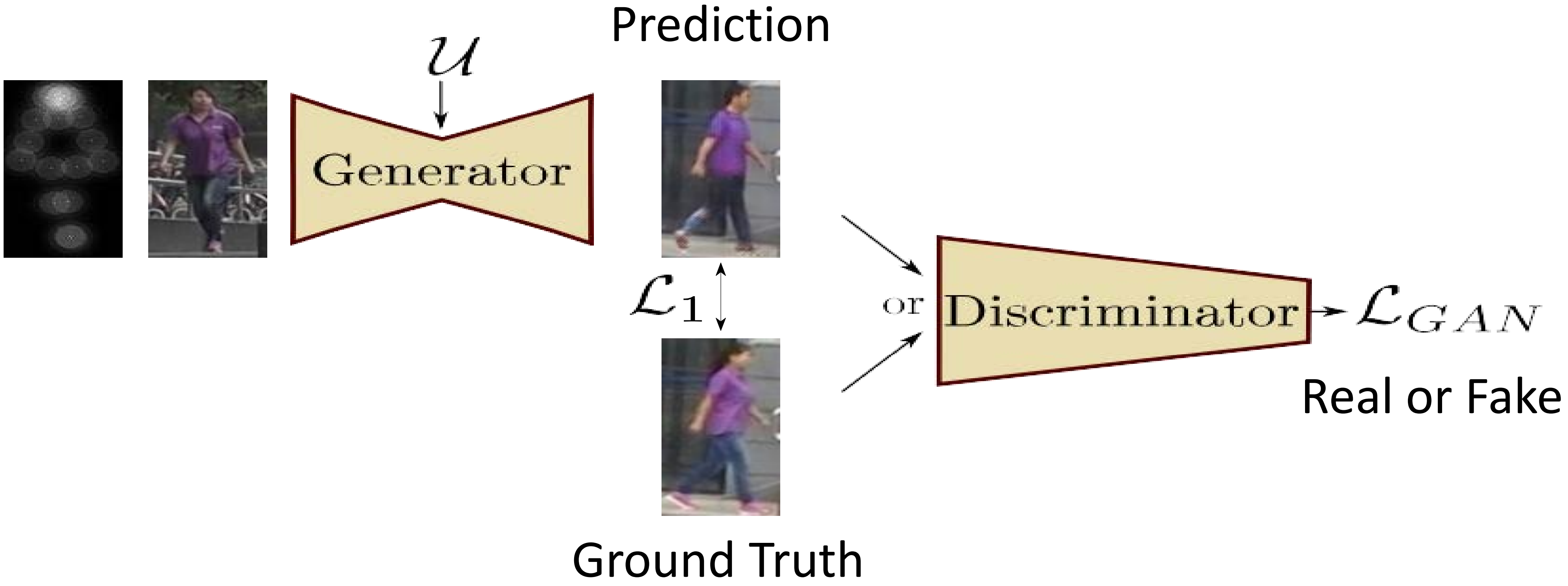

Example 1: Neutral -> Smile -> Neutral
Speed: 12fps

Pose-based Human Image Generation

- Siarohin, et al., "Appearance and Pose-Conditioned Human Image Generation using Deformable GANs", PAMI, 43(4):1156-1171, April 2021

https://github.com/AliaksandrSiarohin/pose-gan

Pose-based Human Image Generation

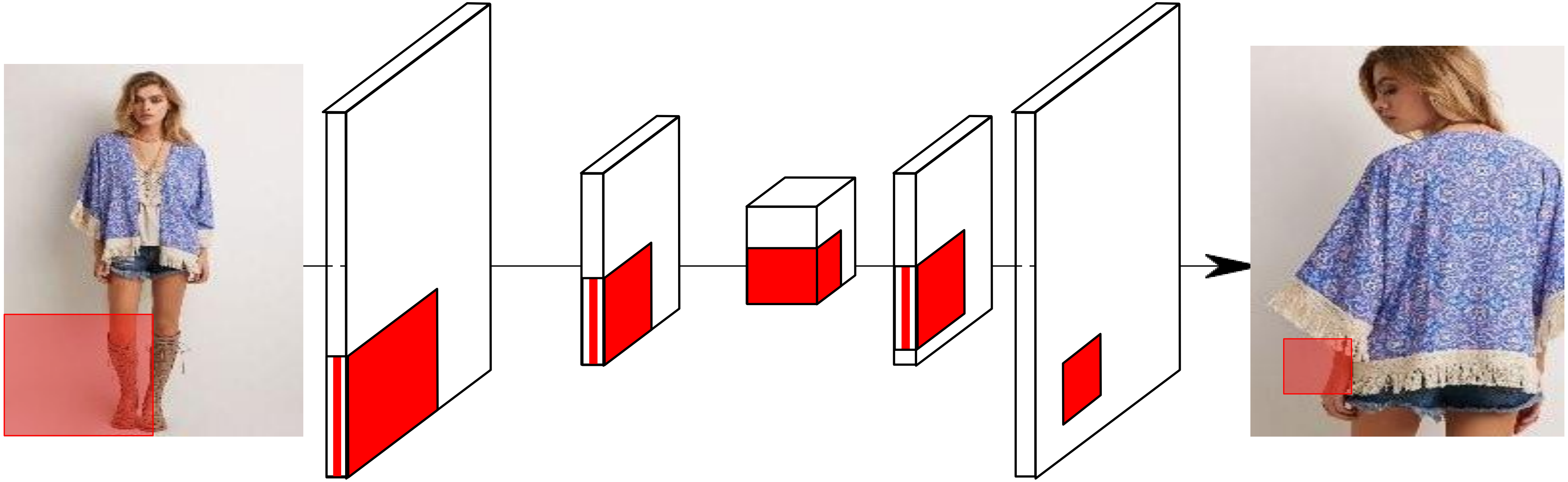


Pose-based Human Image Generation

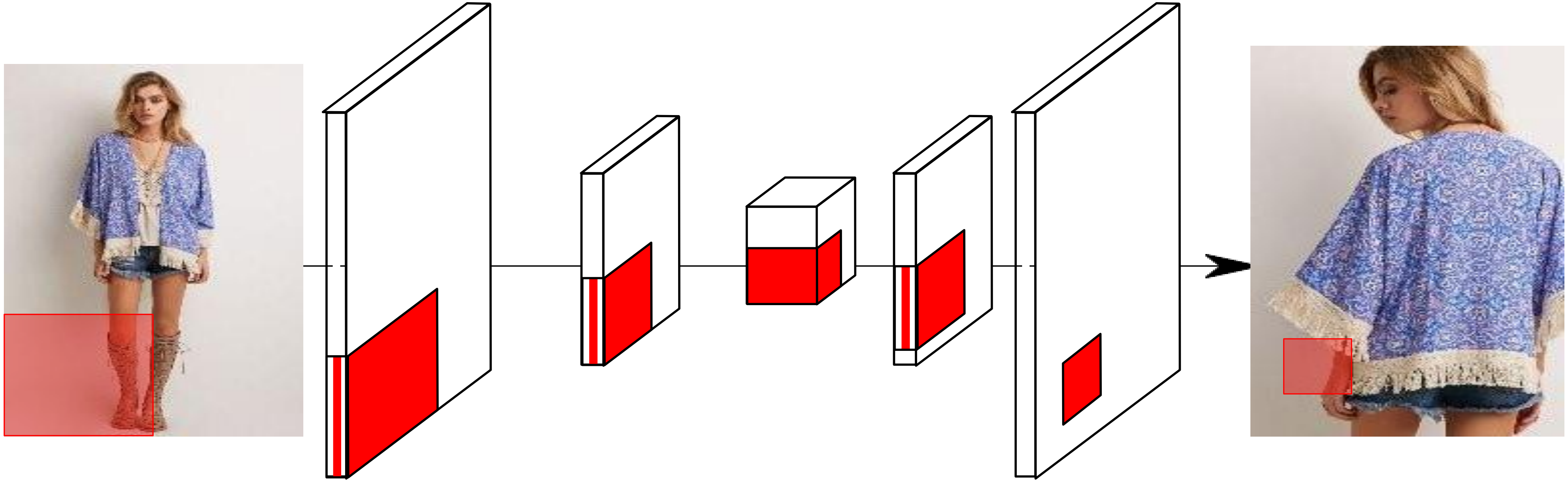
(a) Aligned task (b) Unaligned task

(a) typical “rigid” scene generation task: the local structures of conditioning and output image local structures are well aligned
(b) deformable-object generation task: the input and output are not spatially aligned

Pose-based Human Image Generation

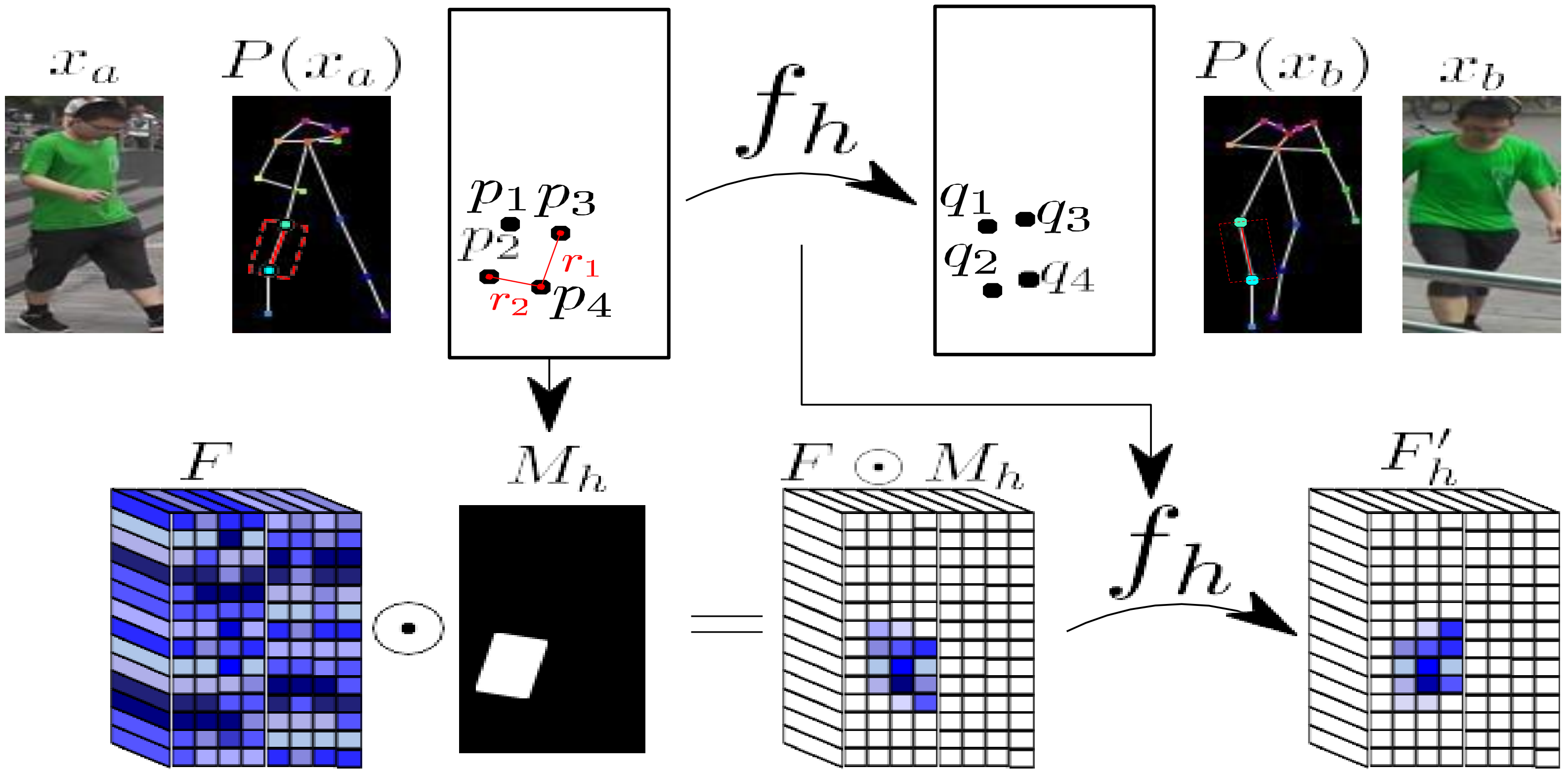


Pose-based Human Image Generation



We need a deformation model

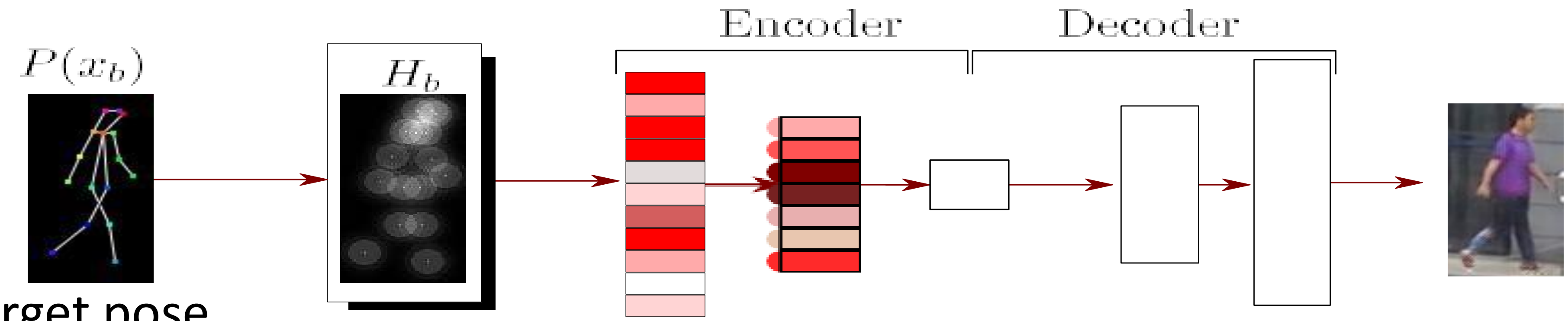
Pose-based Human Image Generation



- For each specific body part, compute an affine transformation f_h
- Use f_h to "move" the corresponding feature-map content

Pose-based Human Image Generation

Target Stream



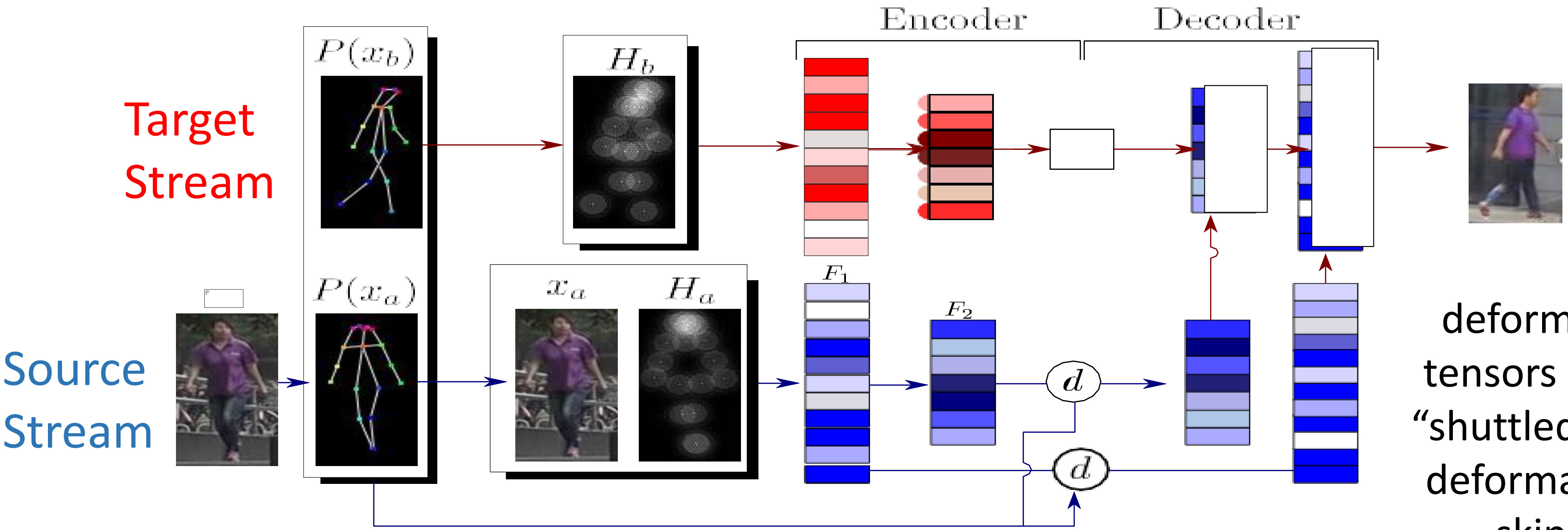
target pose

heat map using gaussian blurring

feature maps "shuttled" by skip connections from the H_b stream

feature maps directly obtained using up-convolutional filters applied to the previous-layer maps

Pose-based Human Image Generation



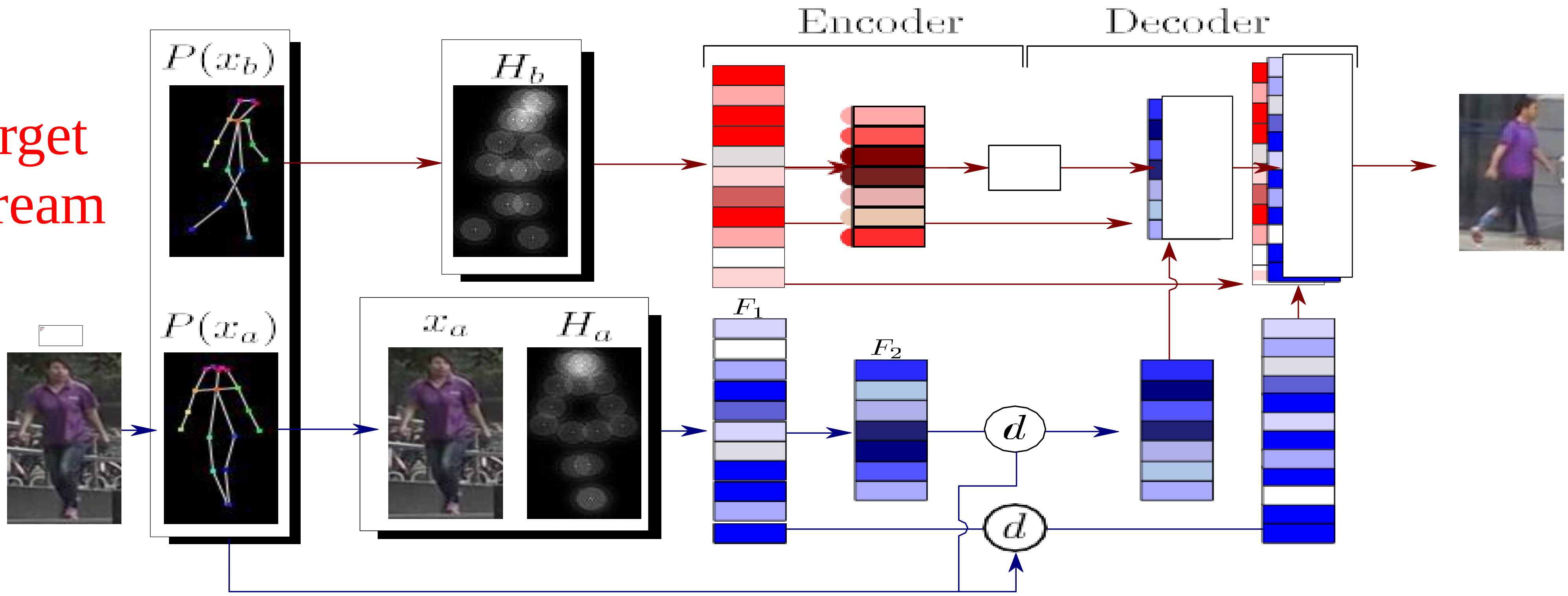
deformed tensors d(F) "shuttled" by deformable skip connections from (x_a, H_a) stream

- joint locations in x_a and H_a are spatially aligned (by construction)
- in H_b the joint locations may be far apart from x_a
- Hence, H_b is not concatenated with the other input tensors

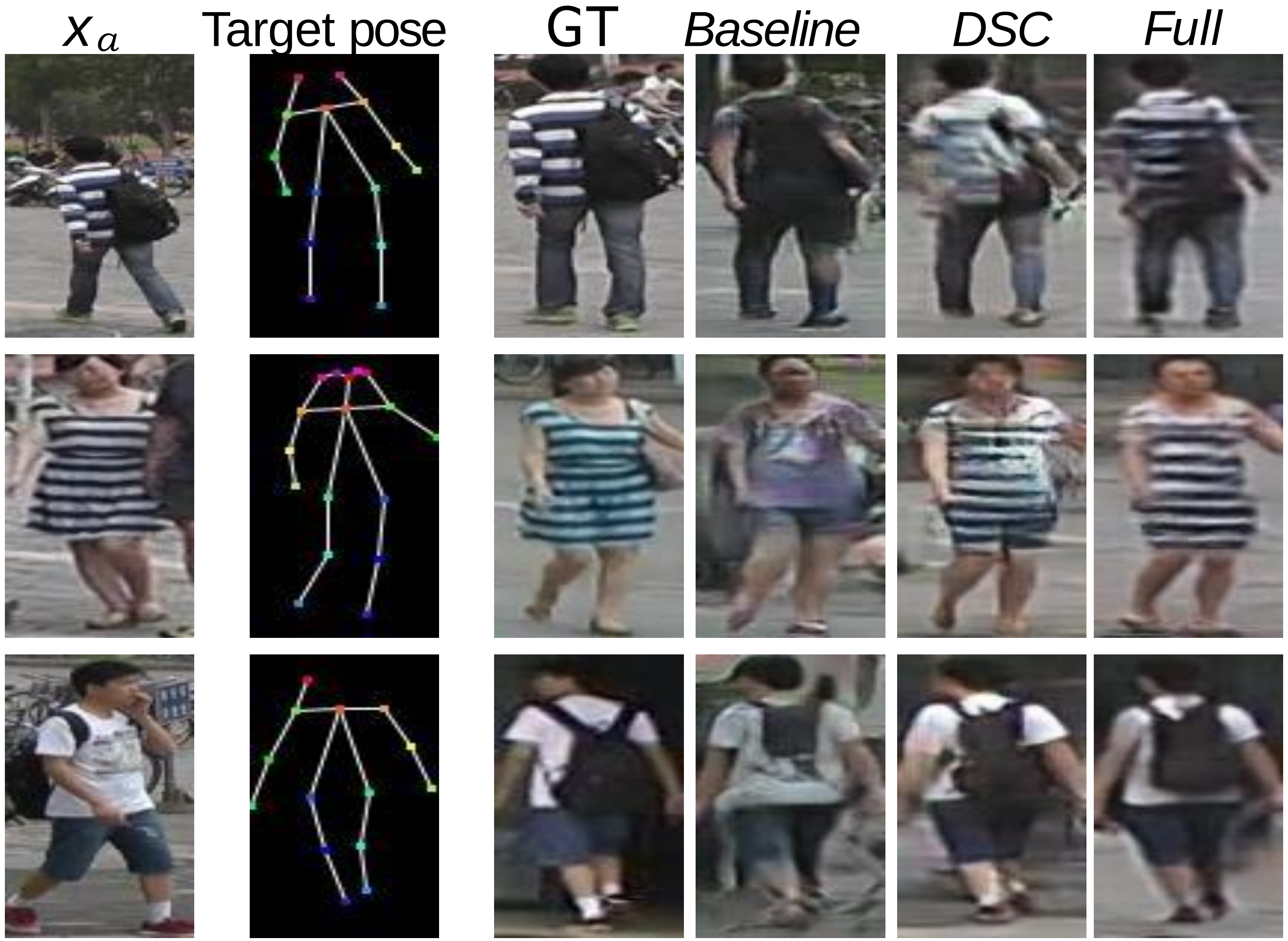
Pose-based Human Image Generation

Target Stream

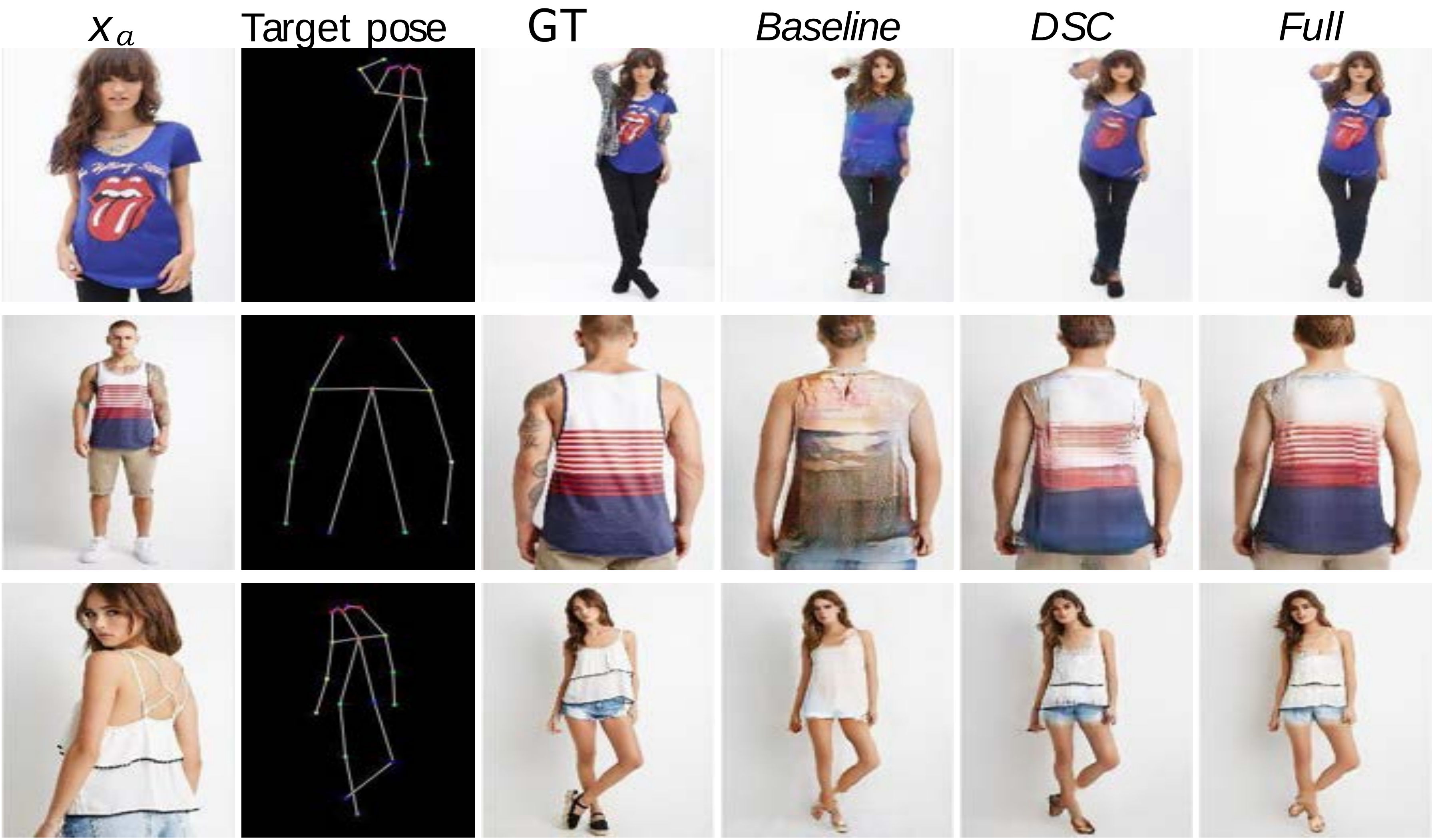
Source Stream



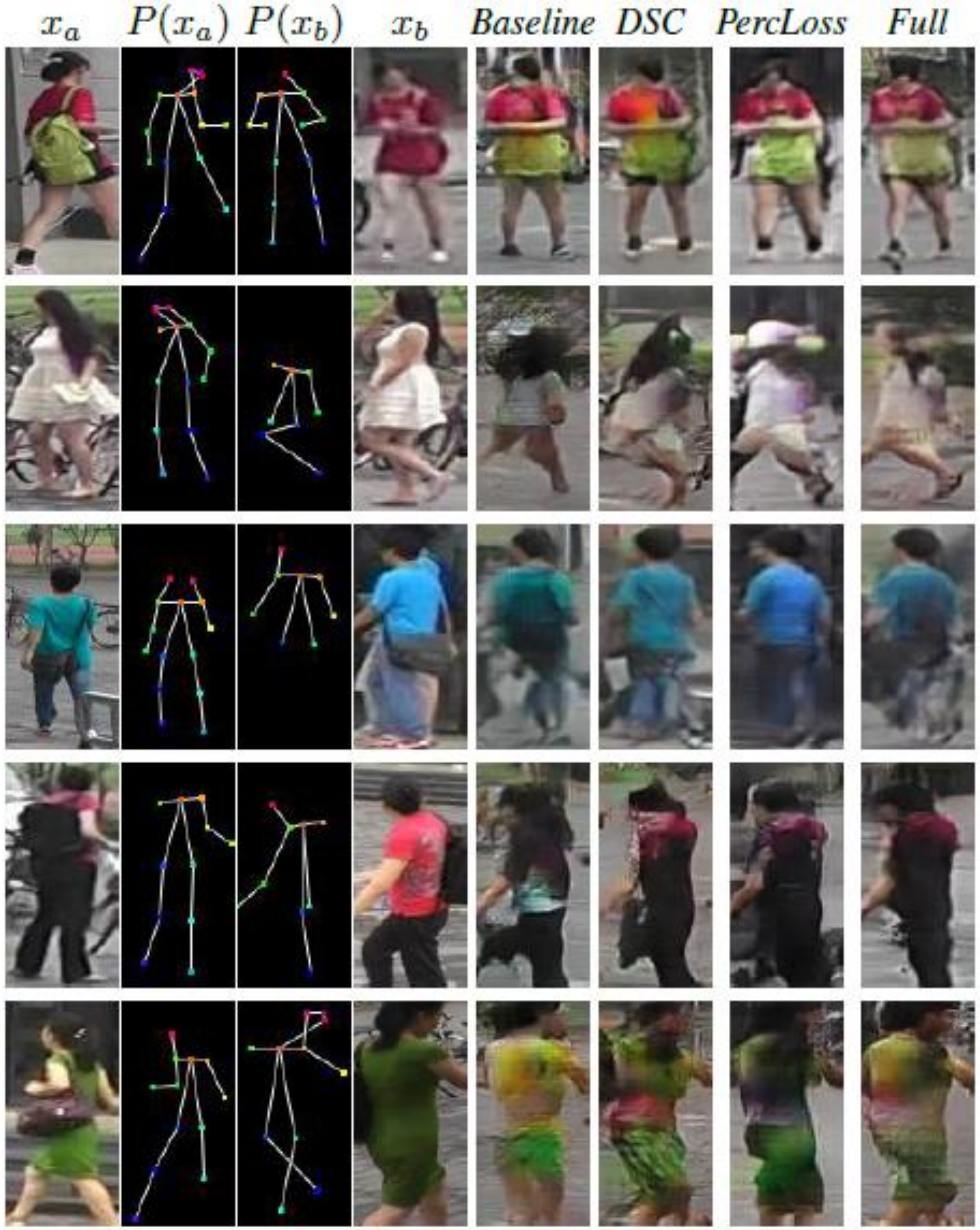
Conditional Image Generation



Qualitative results on the Market-1501 dataset



Qualitative results on the DeepFashion dataset



Badly generated images

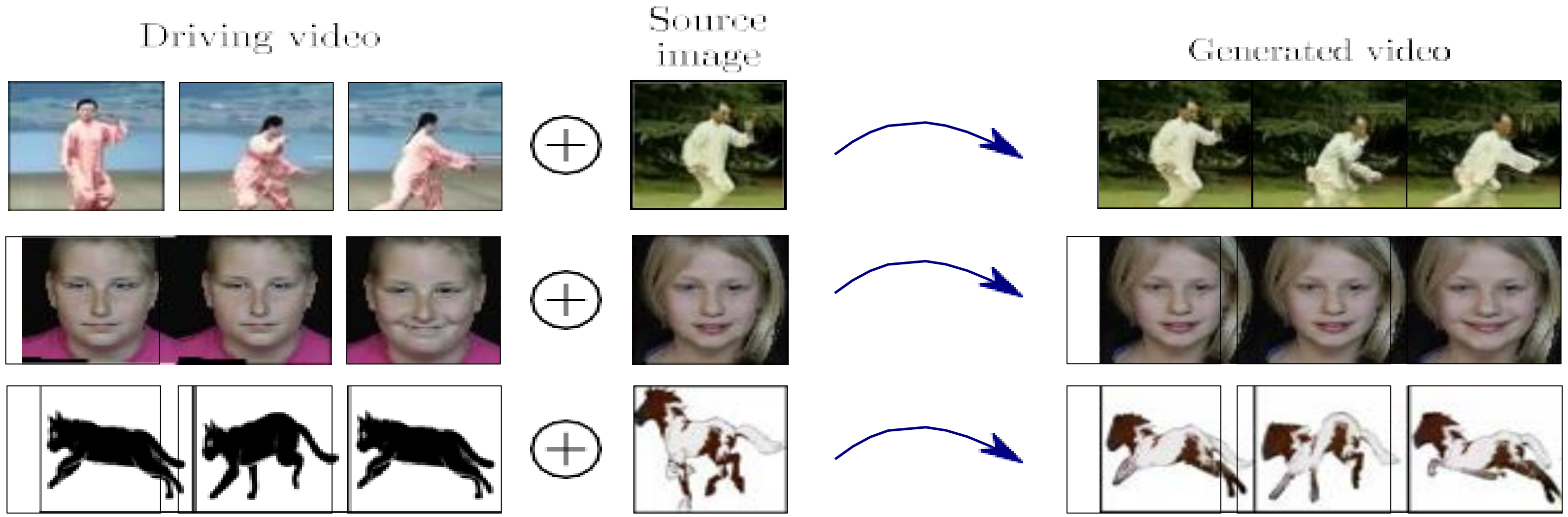
- errors of the pose estimation
- ambiguity of the pose estimation
- rare object appearance
- rare poses

Image Animation

- Siarohin, et al., "Animating Arbitrary Objects via Deep Motion Transfer", CVPR19
- Siarohin, et al., "First Order Motion Model for Image Animation", NeurIPS19

https://github.com/AliaksandrSiarohin/first-order-model

Image Animation: Appearance or Motion Transfer?



Appearance transfer

Detect pose in each frame of the driving video

Apply our pose-base image generator with the source image and each detected pose

Problems: requires a detector, does not work when the shapes of the object are different (ie. short to tall persons) => **Use Unsupervised Transfer Motion**

Image Animation with MOviNg KEYpoints

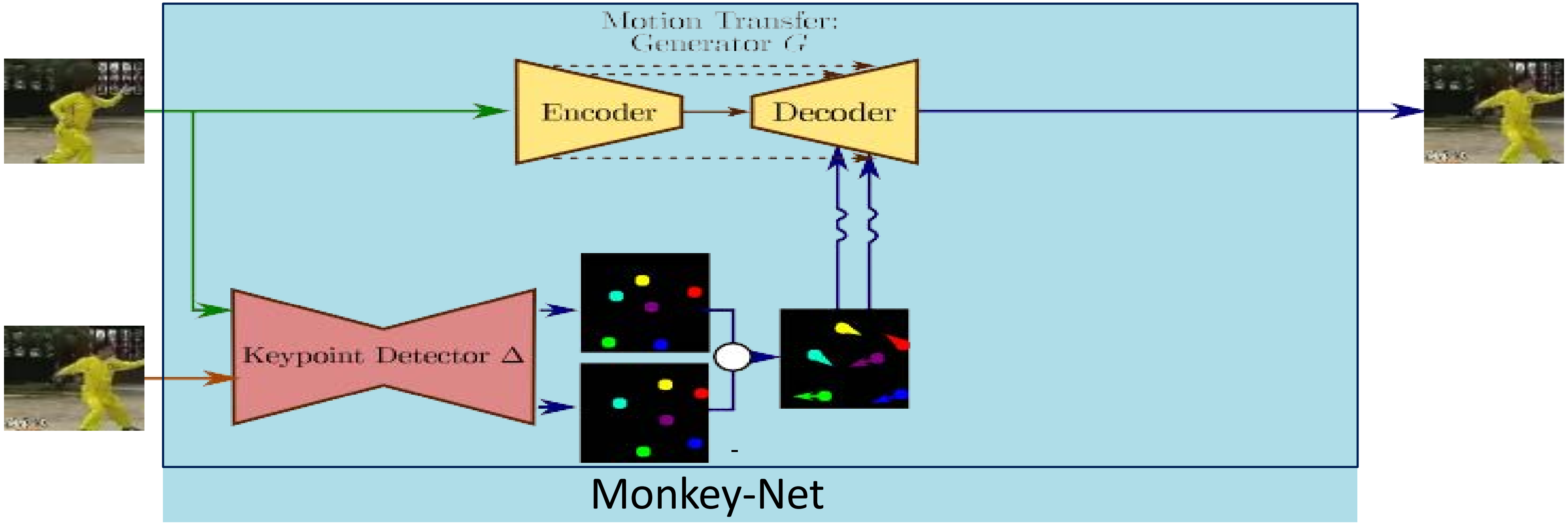
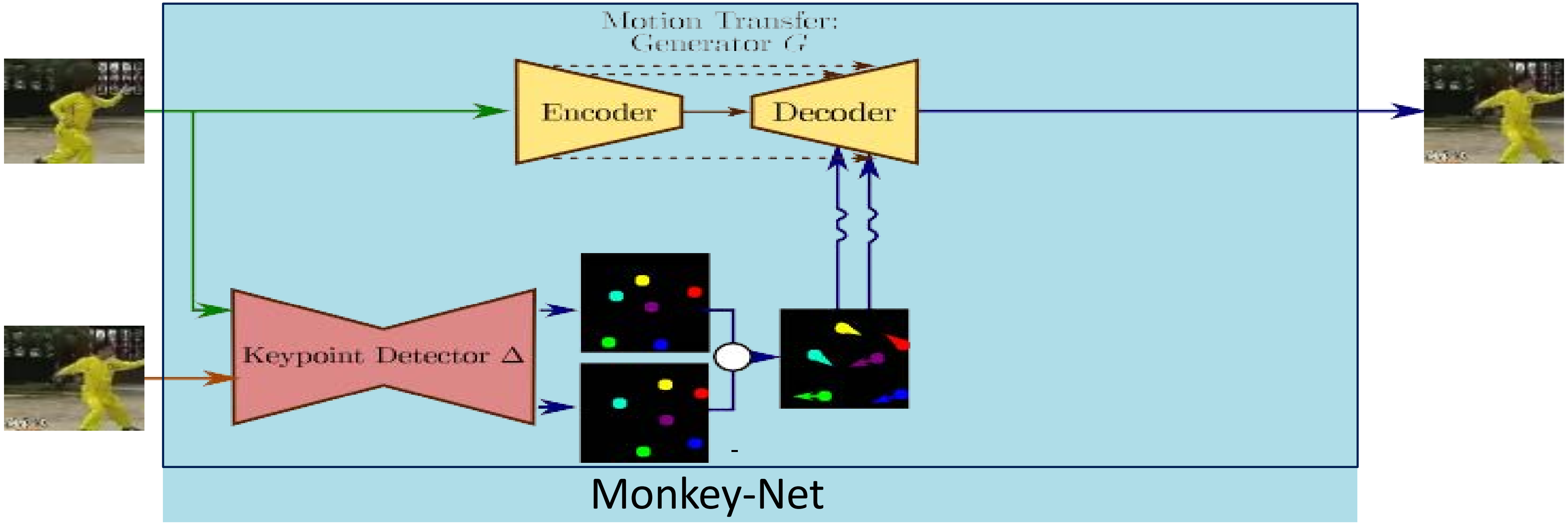
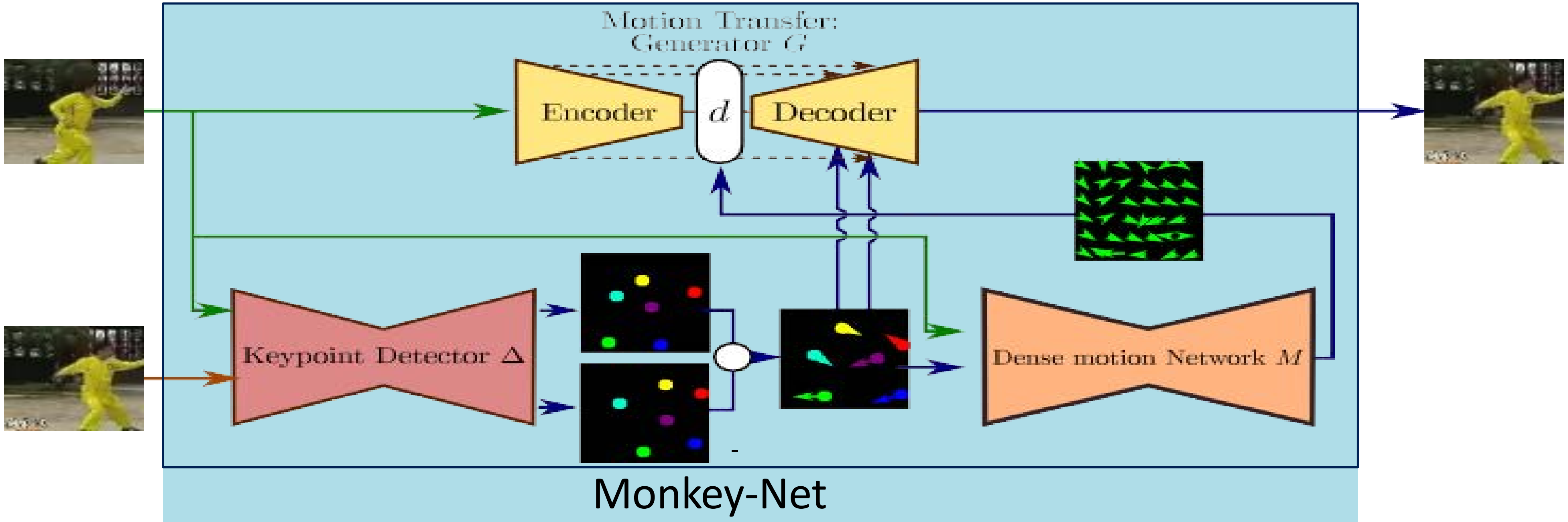


Image Animation with MOviNg KEYpoints



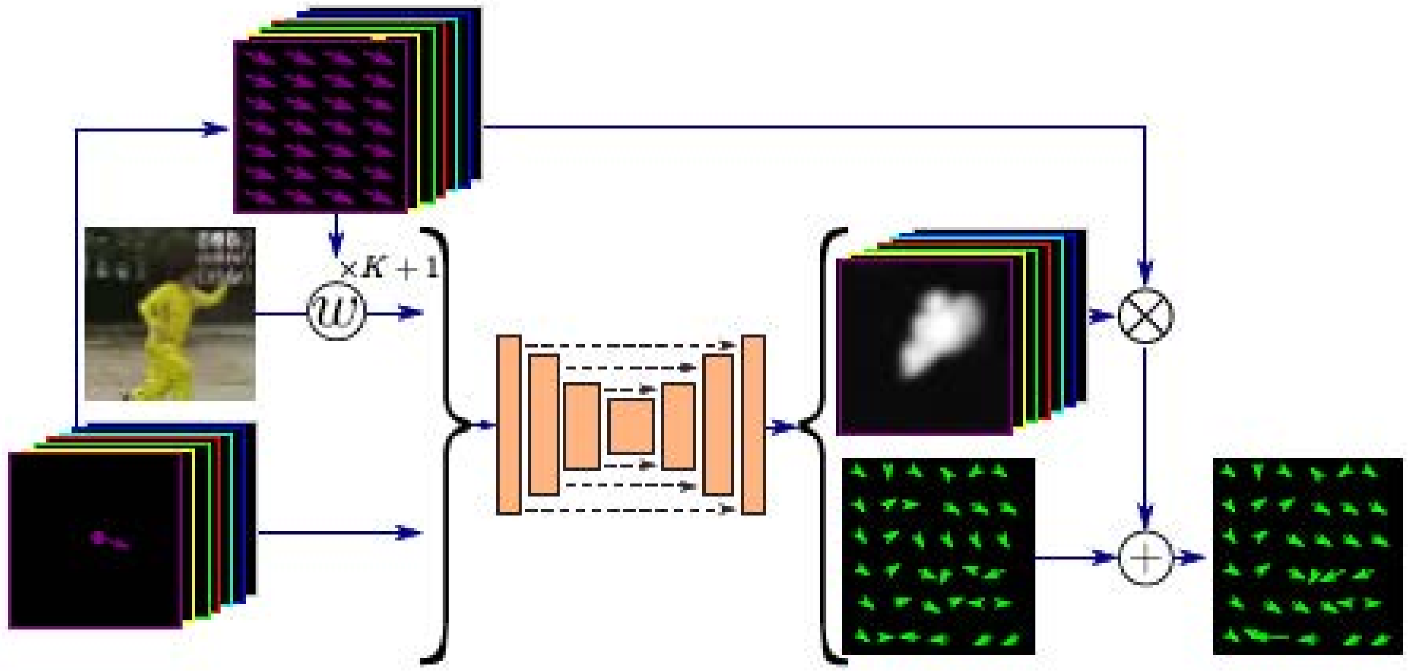
Again, we have an alignment problem

Image Animation with MOviNg KEYpoints



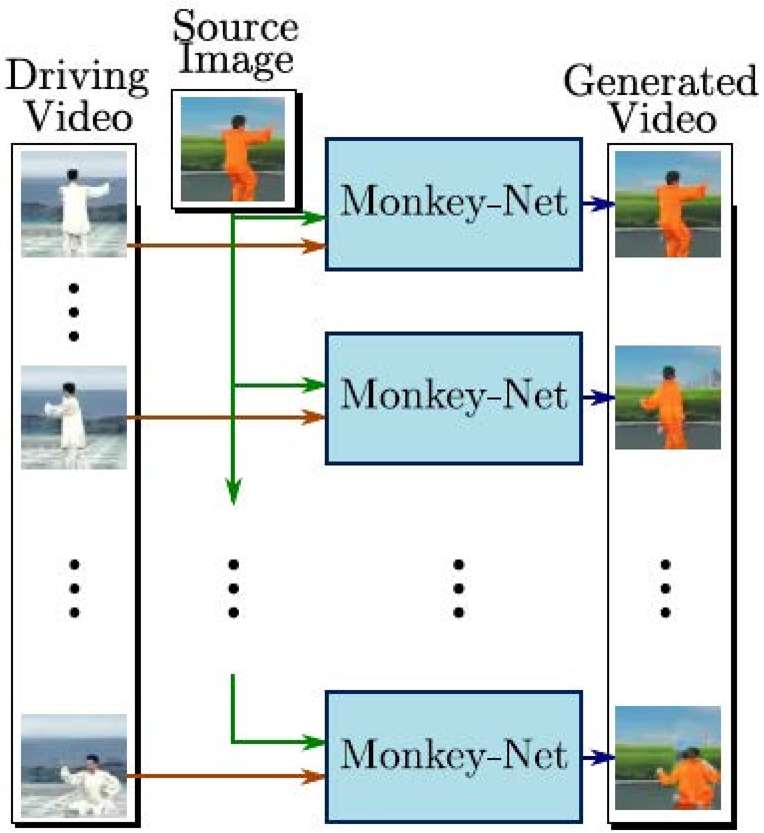
- Monkey-Net has a motion-specific keypoint detector Δ, a motion prediction network M, and an image generator G (reconstructs the image x′ from the keypoint positions Δ(x) and Δ(x′)); Optical flow computed by M is used by G to handle misalignments between x and x′
- The model is learned with a self-supervised learning scheme

Image Animation: Motion Prediction



From the appearance of the first frame and the keypoints motion, the network M predicts a mask for each keypoint and the residual motion

Image Animation Generation



At testing time the model generates a video with the object appearance of the source image but with motion from driving video:

- transfer the motion between the source image and each driving frame
- provide the generator the relative difference between keypoints

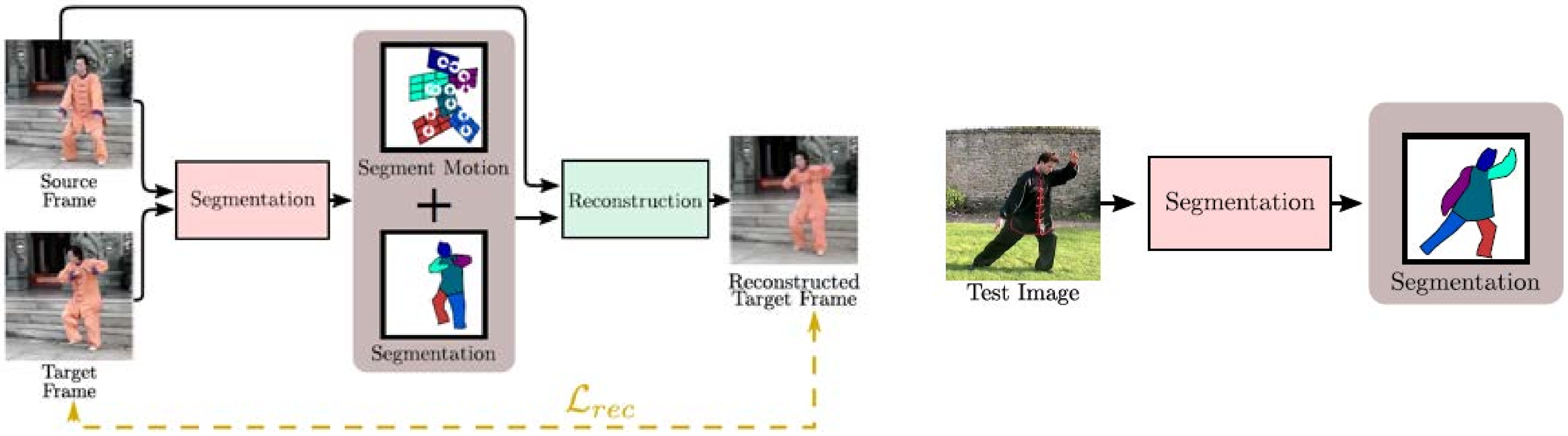
Learned Keypoints

Motion-supervised Co-Part Segmentation

- Siarohin, et al., "Motion Supervised Co-Part Segmentation", ICPR20

https://github.com/AliaksandrSiarohin/motion-cosegmentation

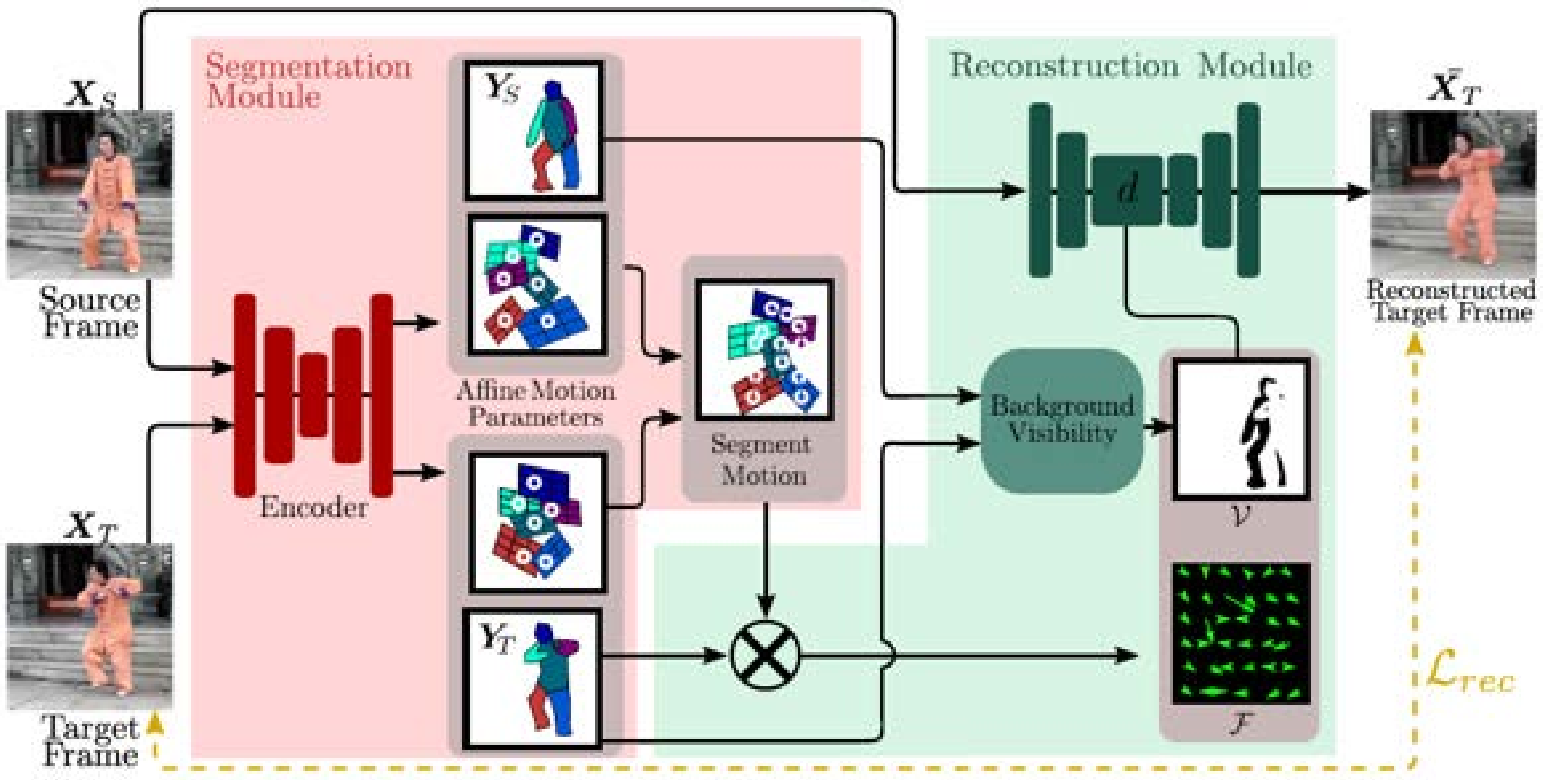
Self-supervised Co-Part Segmentation



Leverage motion info to train a segmentation network without annotation

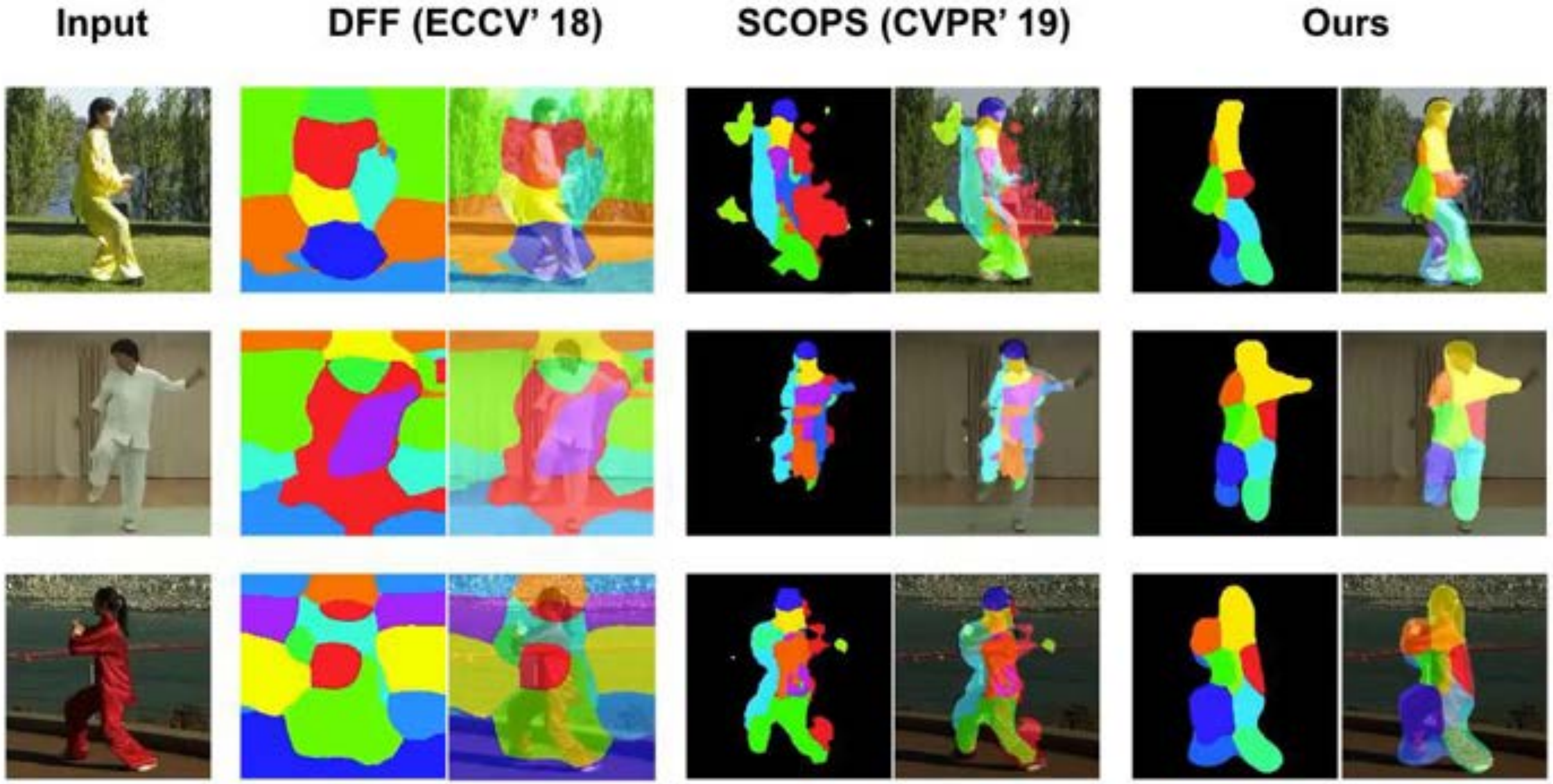
- At training, use frame pairs (source and target) extracted from the same video => predict segments in target that can be combined with a motion representation between the two frames to reconstruct the target frame
- At inference, use the trained segmentation model to predict object parts segments

Self-supervised Co-Part Segmentation



- **Segmentation Module** predicts the segmentation maps Y_S and Y_T, and the affine motion parameters
- **Reconstruction Module:** (1) computes a background visibility mask V and an optical flow F; (2) reconstructs the target frame X_T by warping the features of the source frame X_S and masking occluded features

Tai-Chi-HD

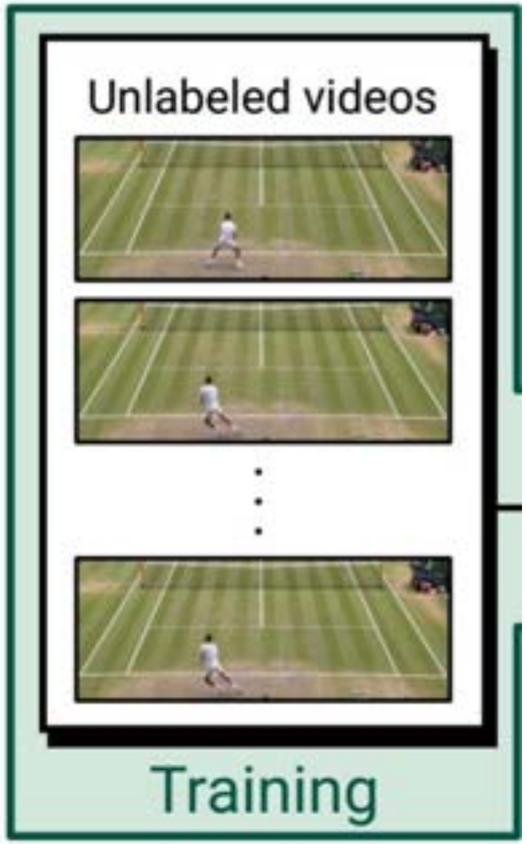


Playable Video Generation

- Menapace, et al., “Playable Video Generation”, CVPR21

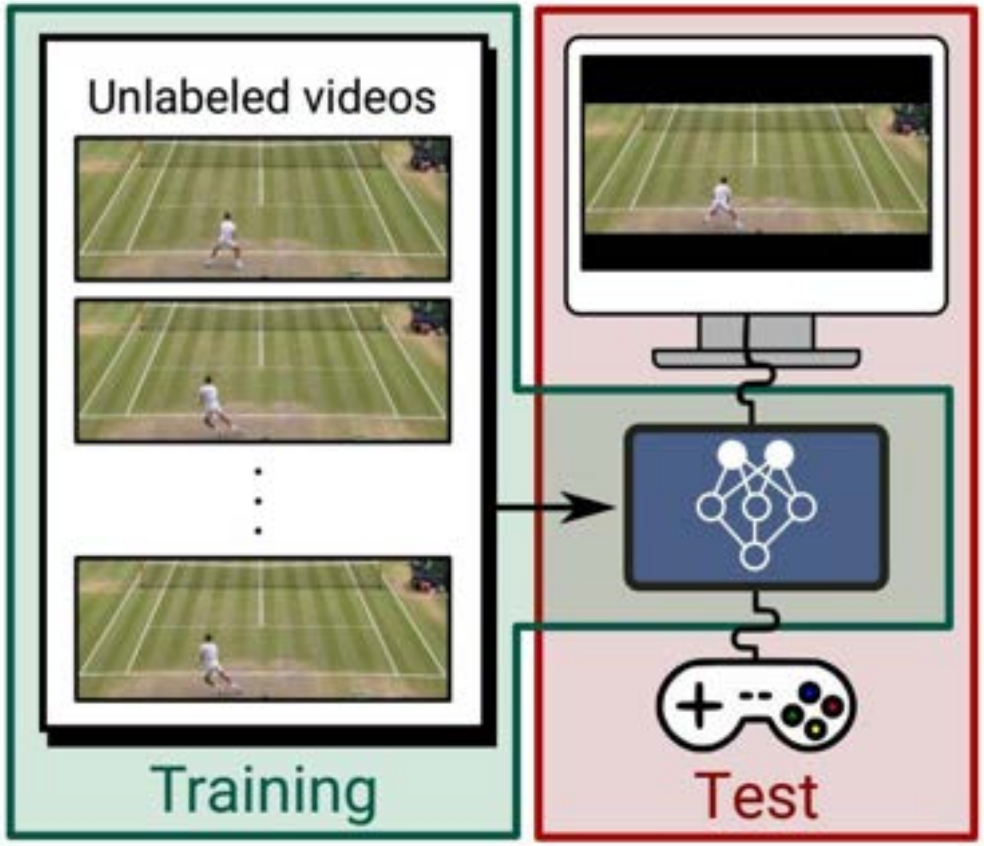
https://github.com/willi-menapace/PlayableVideoGeneration

Playable Video Generation



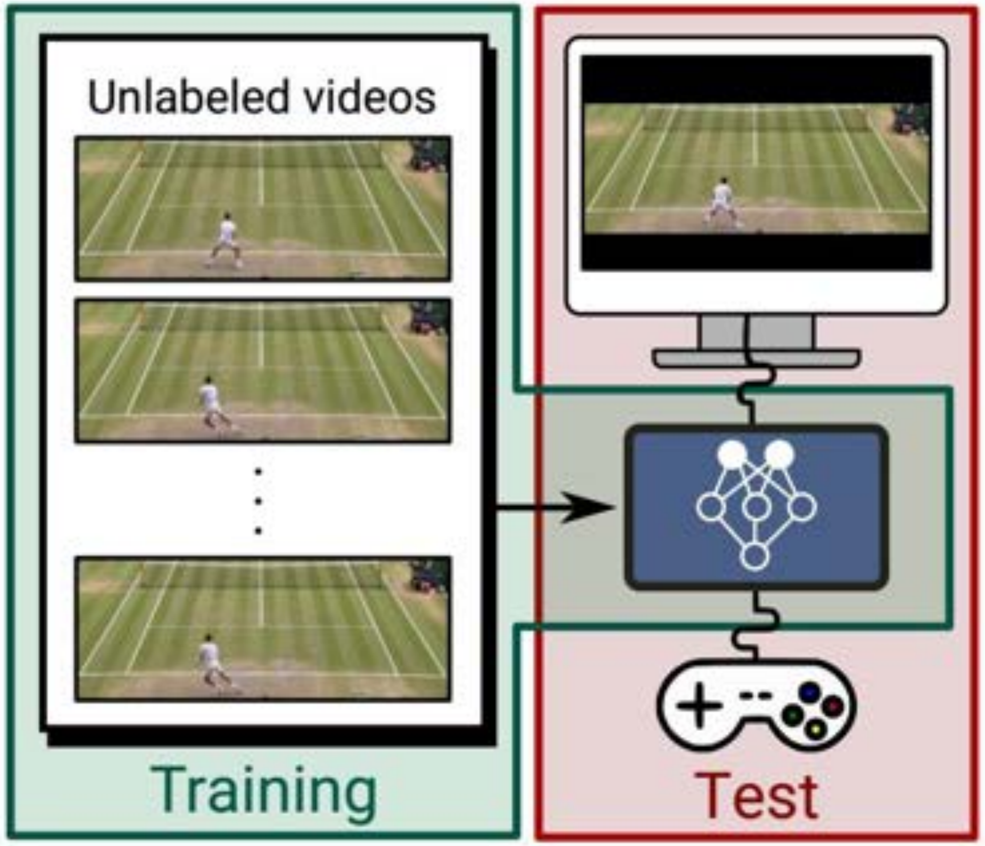
- Consider a set of videos depicting an agent acting in an environment
- Differently from other methods that use frame by frame action annotations, we assume no annotation is present

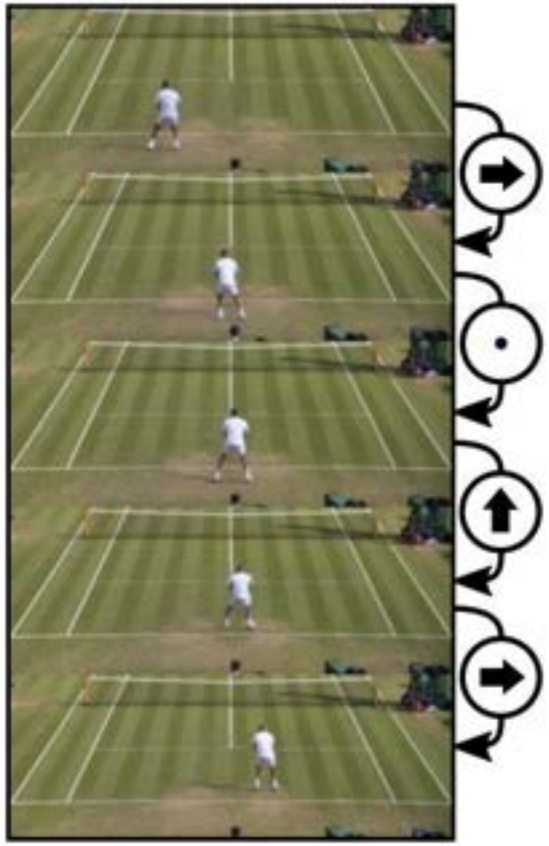
Playable Video Generation



- Learn a model that represents the observed environment.
- Allow the user to input actions to the model through a controller at test time

Playable Video Generation





- Produce a video where the agent acts according to the actions specified by the user