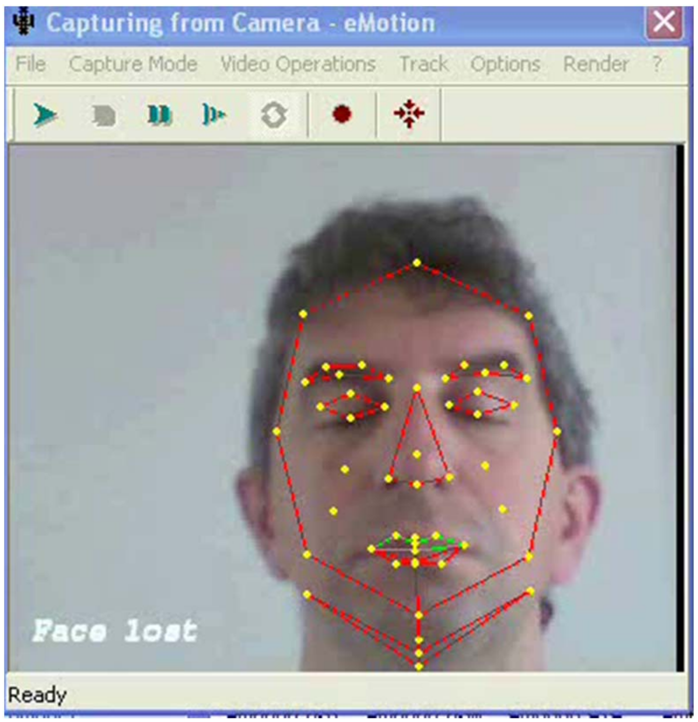
AIDA
ARTIFICIAL INTELLIGENCE DOCTORAL ACADEMY
www.i-aida.org

Image and Video Generation
A Deep Learning Approach

Nicu Sebe
University of Trento
niculae.sebe@unitn.it

Collaborators: Xavier Alameda-Pineda, Stephane Lathuilière, Willi Menapace, Elisa Ricci, Subhankar Roy, Enver Sangineto, Aliaksandr Siarohin, Hao Tang, Sergey Tulyakov, Wei Wang, Dan Xu, etc.

A Bit of History



... about 2008

... about 2018

A Bit of History



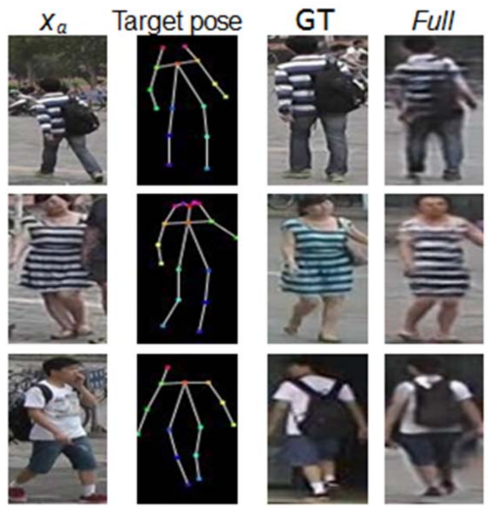
… about 2019

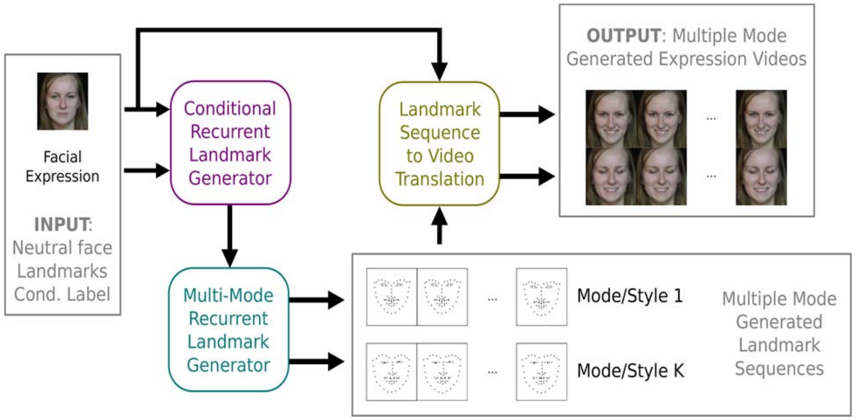
A Bit of History

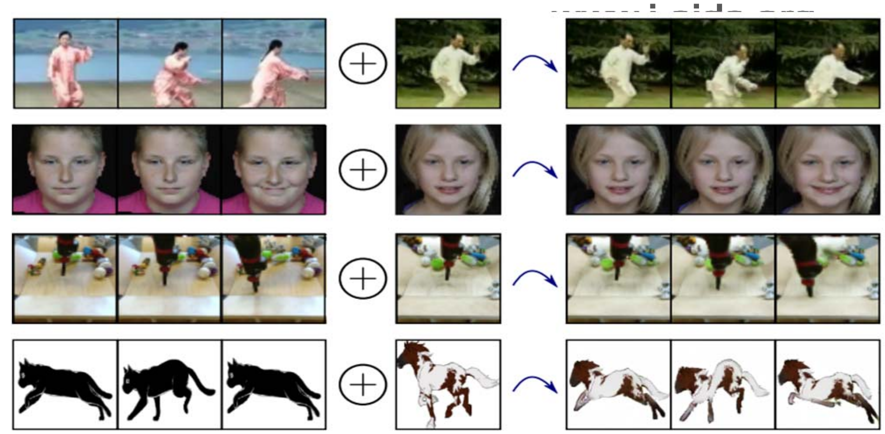
… nowadays

Image and Video Generation

Deep Generative Models for Image/Video Generation & Animation







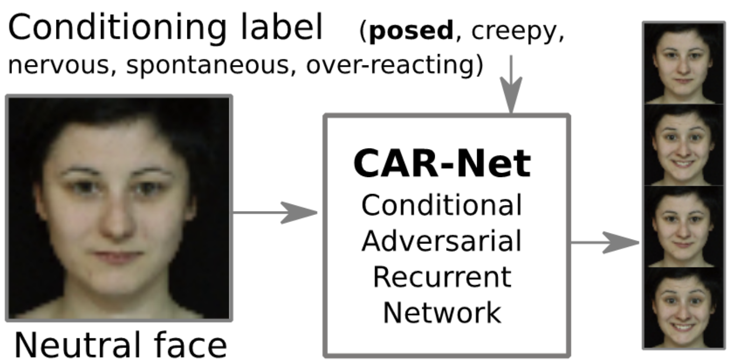
Arbitrary Object Animation without 3D modeling

Playable and Multimodal Video Generation

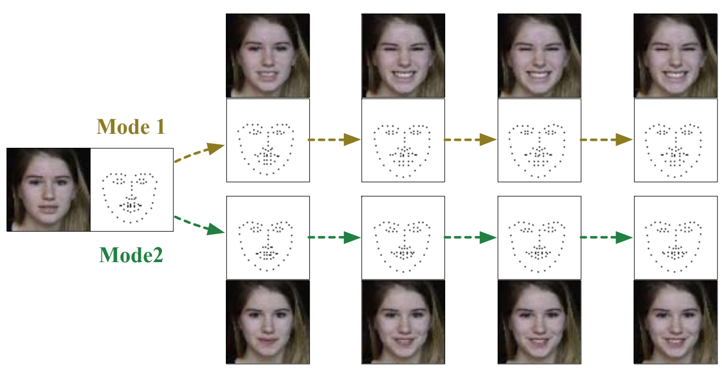
Diverse Smile Video Generation

- Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. “Every Smile is Unique: Landmark-Guided Diverse Smile Generation”, in CVPR 2018
- Wei Wang, Xavier Alameda-Pineda, Dan Xu, Elisa Ricci, and Nicu Sebe. “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

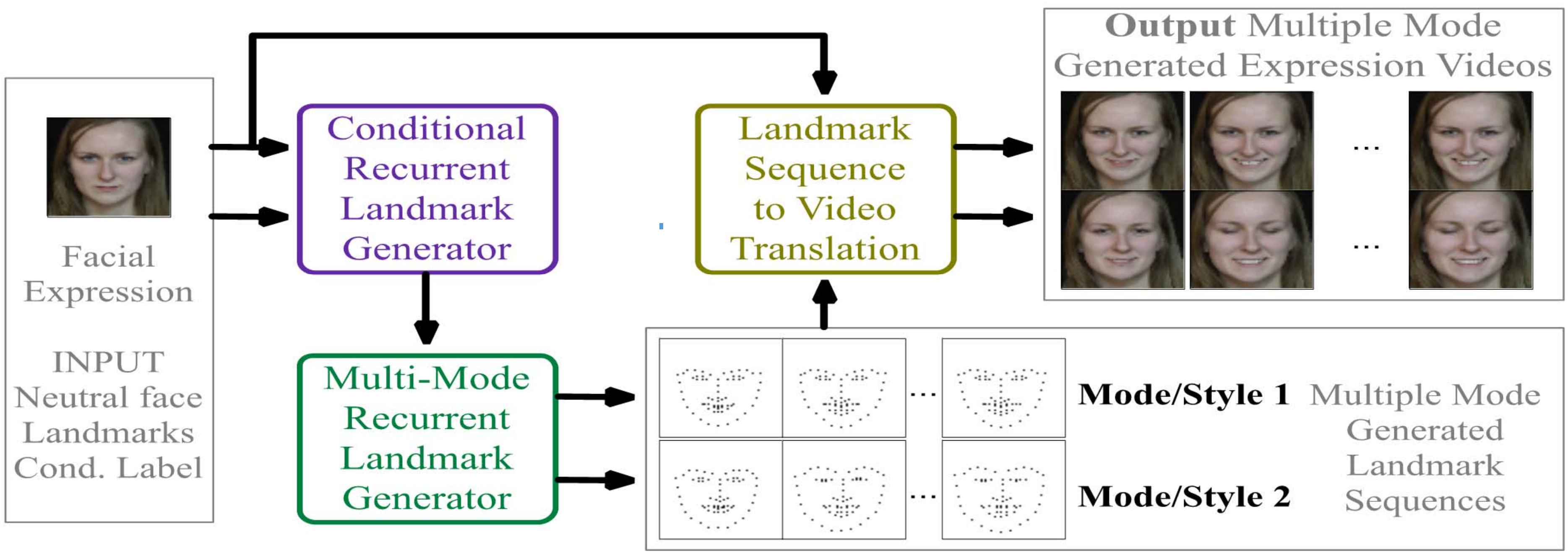


(b) Generate K different sequences of smiles

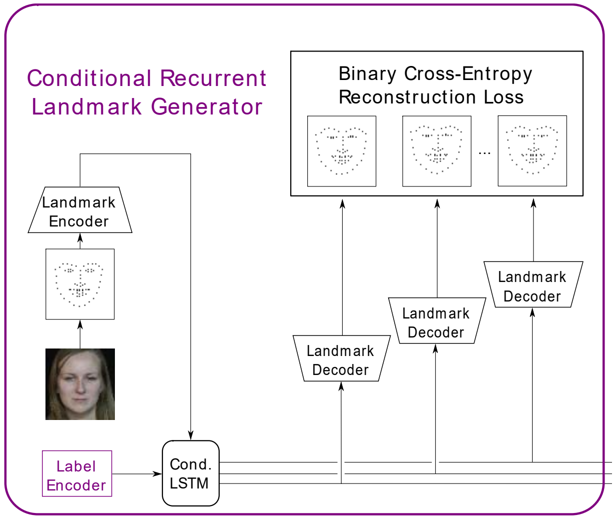
Challenges

- 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

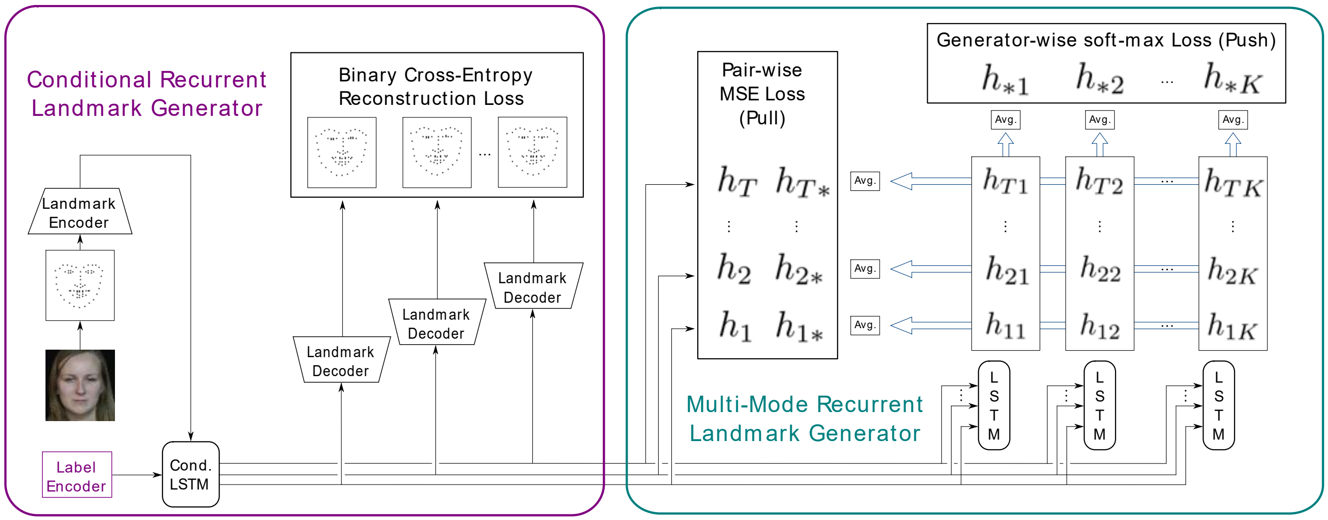


Landmark-Guided Diverse Smile Generation



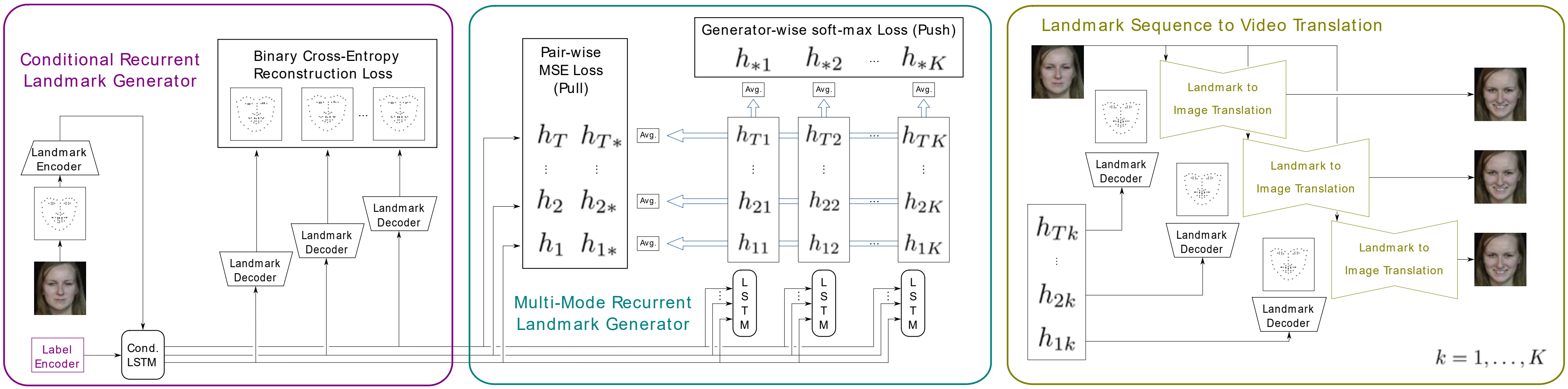
- Encode the landmark image and generate a sequence of landmark embeddings according to the conditioning label

Landmark-Guided Diverse Smile Generation



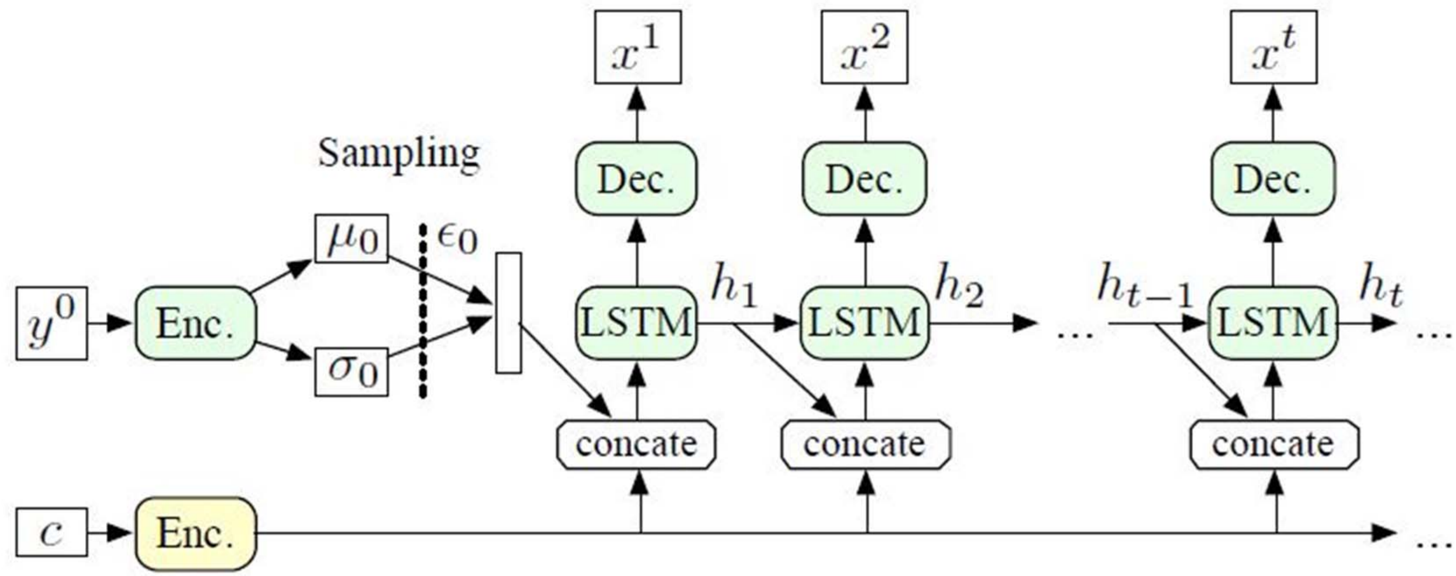
- Encode the landmark image and generate a sequence of landmark embeddings according to the conditioning label
- Generate K different landmark embedding sequences

Landmark-Guided Diverse Smile Generation



- Encode the landmark image and generate a sequence of landmark embeddings according to the conditioning label
- Generate K different landmark embedding sequences
- Translate each of the sequences into a face video

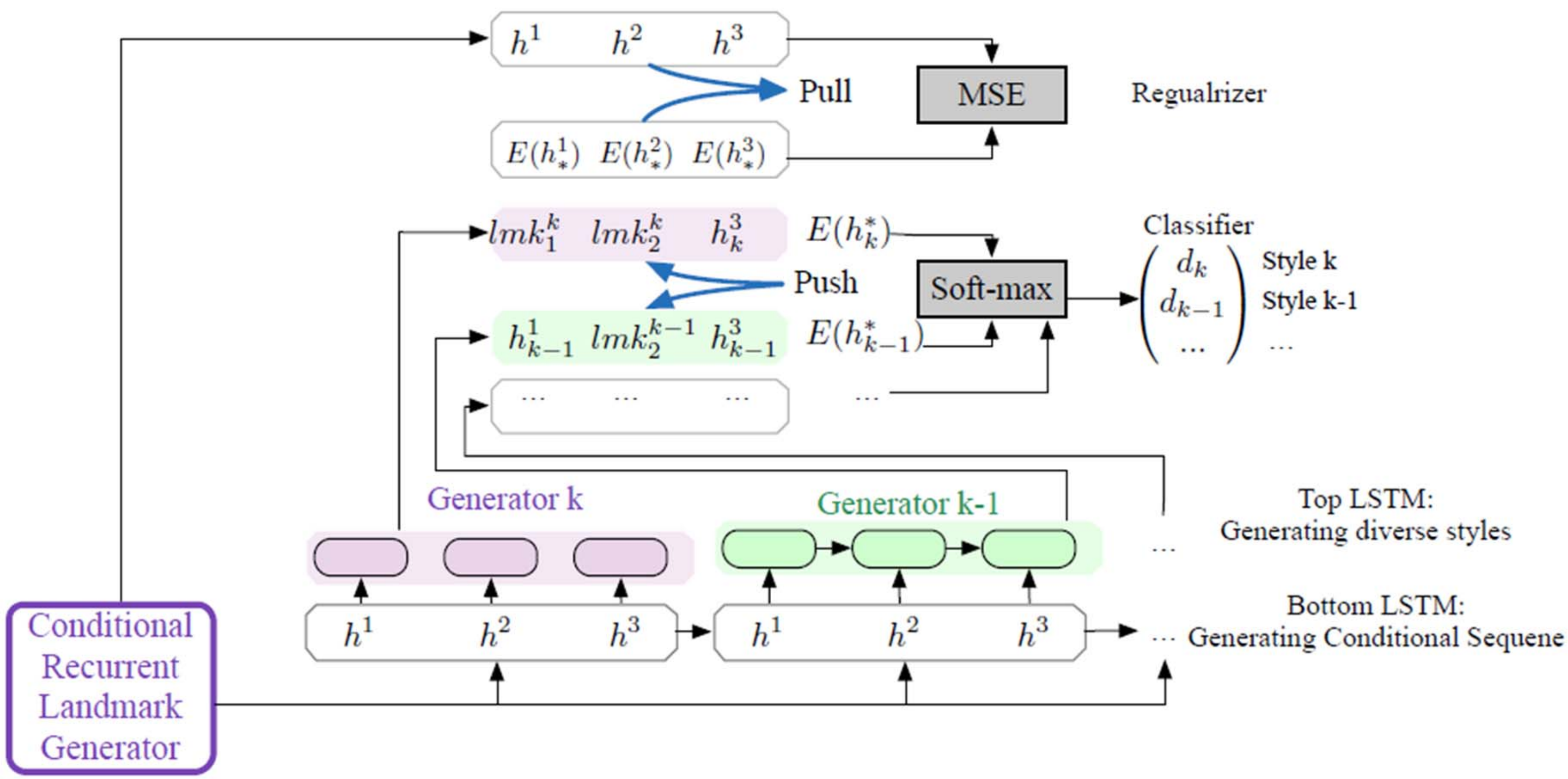
Landmark-Guided Diverse Smile Generation



(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 the conditioning label c

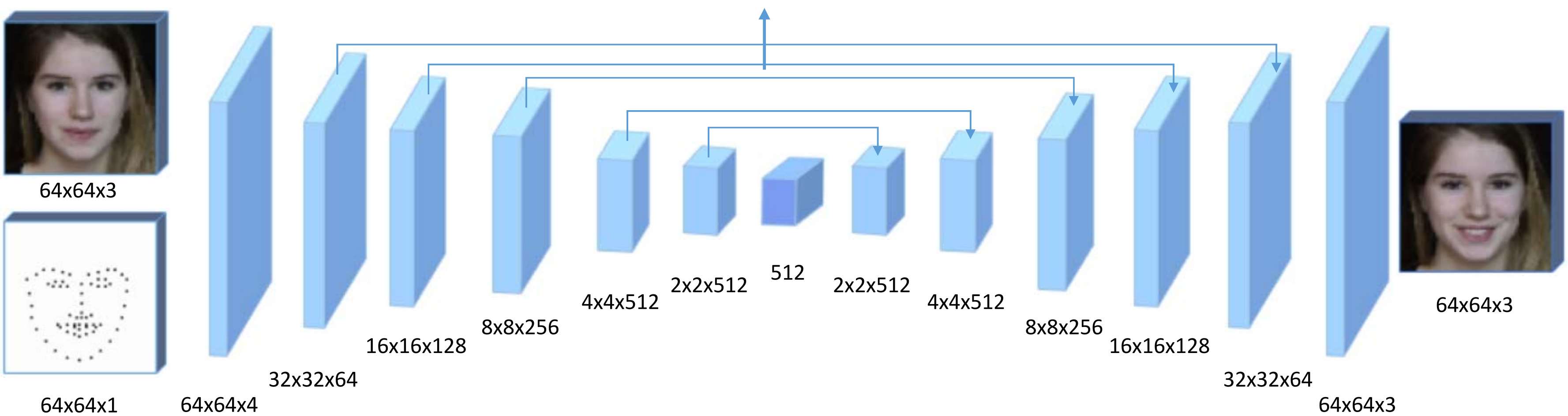
Landmark-Guided Diverse Smile Generation



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

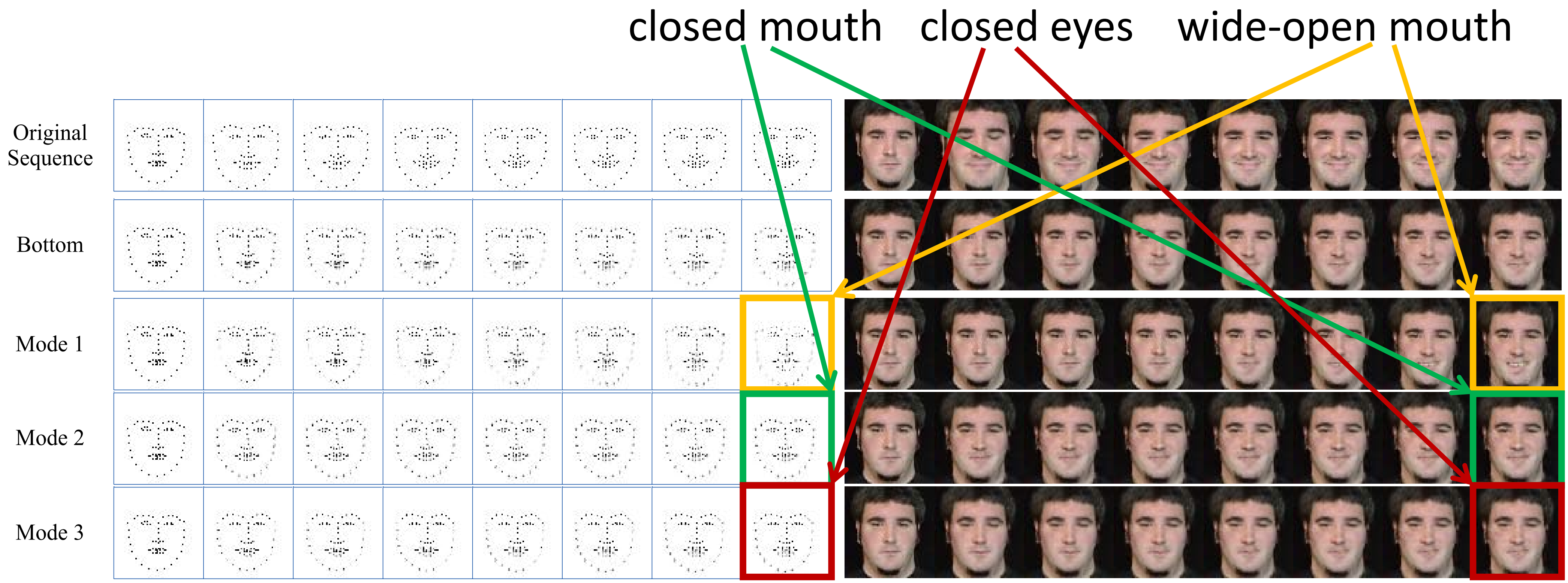
Landmark-Guided Diverse Smile Generation

Skip Connections allow texture passing from source to target to preserve the identity



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

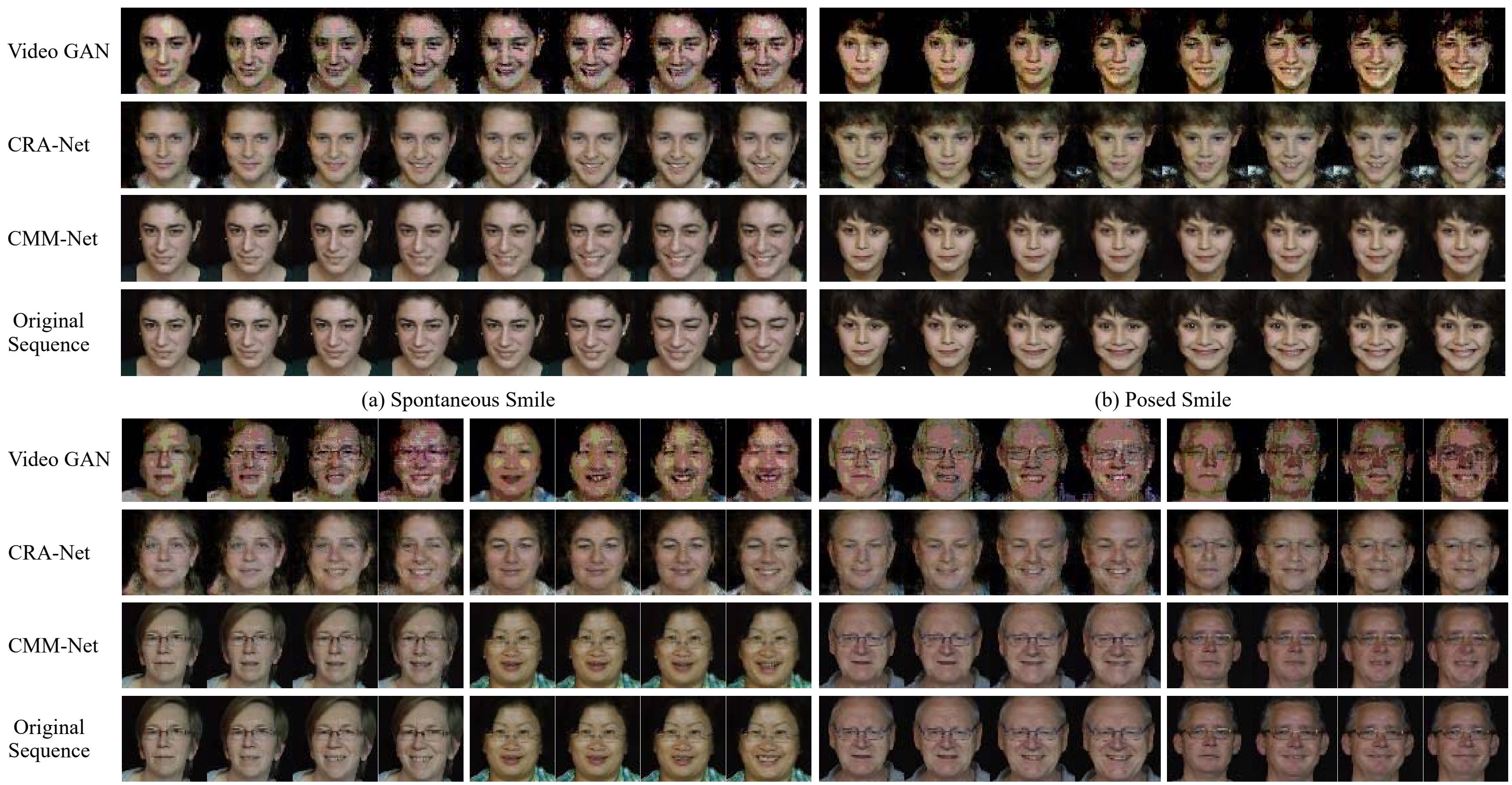
Landmark-Guided Diverse Smile Generation



Multi-Mode Generation

Landmark-Guided Diverse Smile Generation

Comparison with the state-of-the-art



(a) Spontaneous Smile

(b) Posed Smile

(c) Spontaneous Smile with Glasses

(d) Posed Smile with Glasses

Example 1: Neutral -> Smile -> Neutral

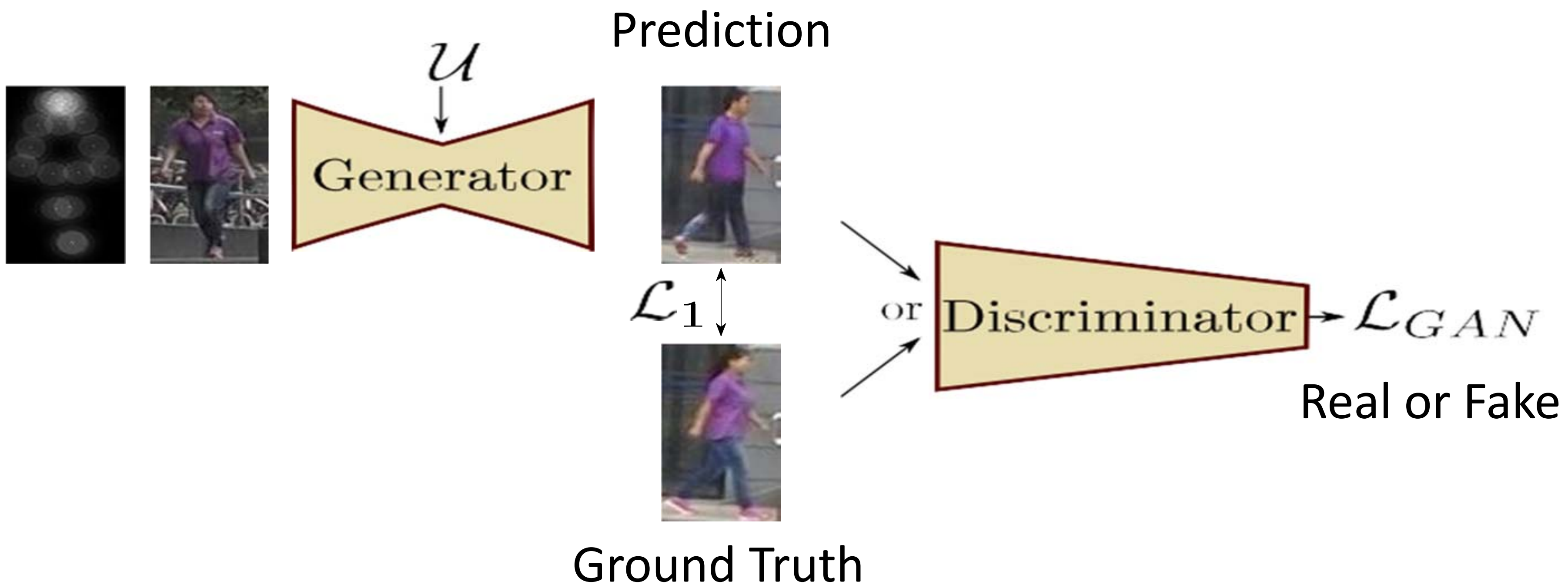
Speed: 12fps

Pose-based Human Image Generation

- Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuilière, and Nicu Sebe. "Deformable GANs for Pose-based Human Image Generation", in CVPR 2018
- Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuilière, and Nicu Sebe "Appearance and Pose-Conditioned Human Image Generation using Deformable GANs", in IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(4):1156-1171, April 2021

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

Pose-based Human Image Generation [1]

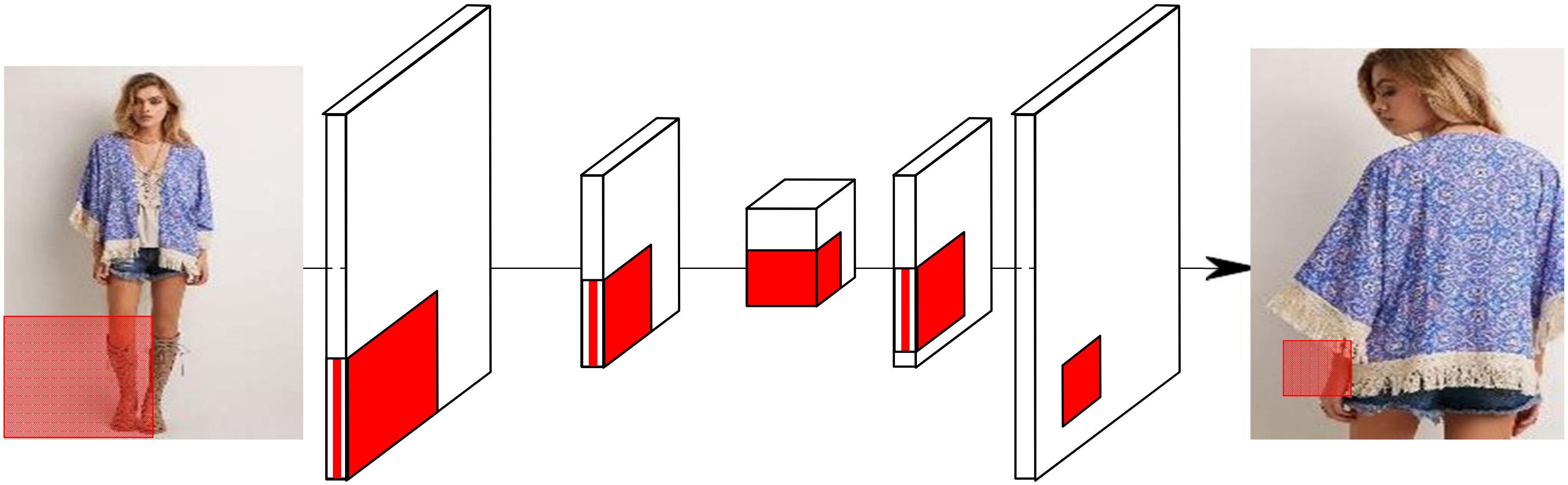


[1] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, Pose-guided person image generation, NeurIPS, 2017

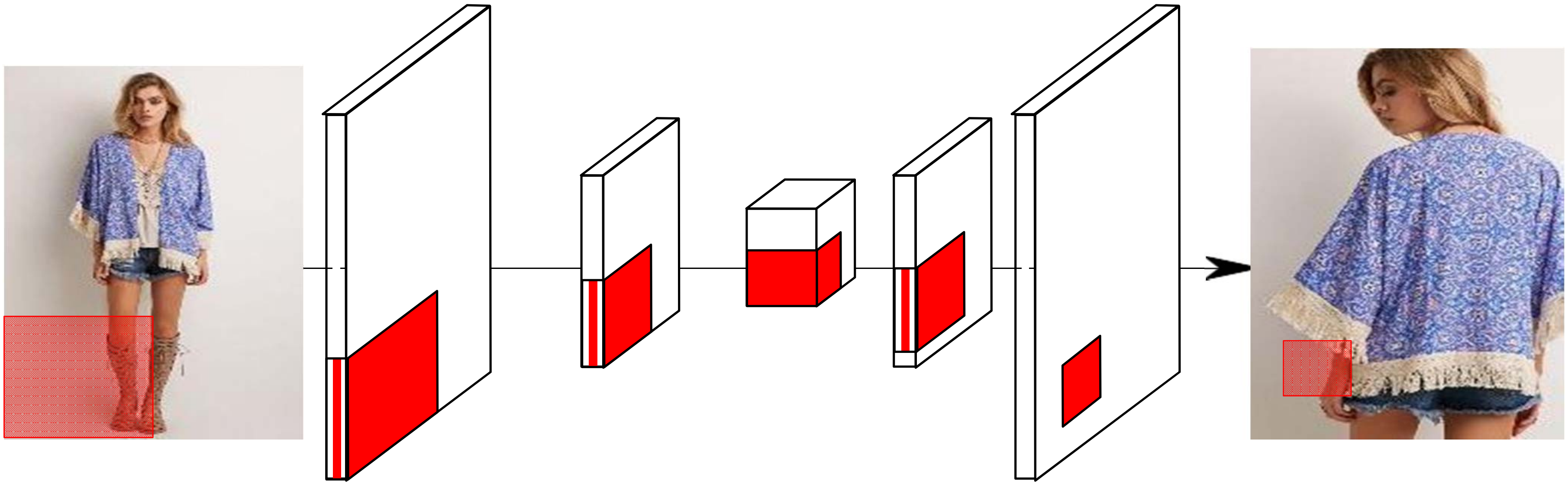
Pose-based Human Image Generation

(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

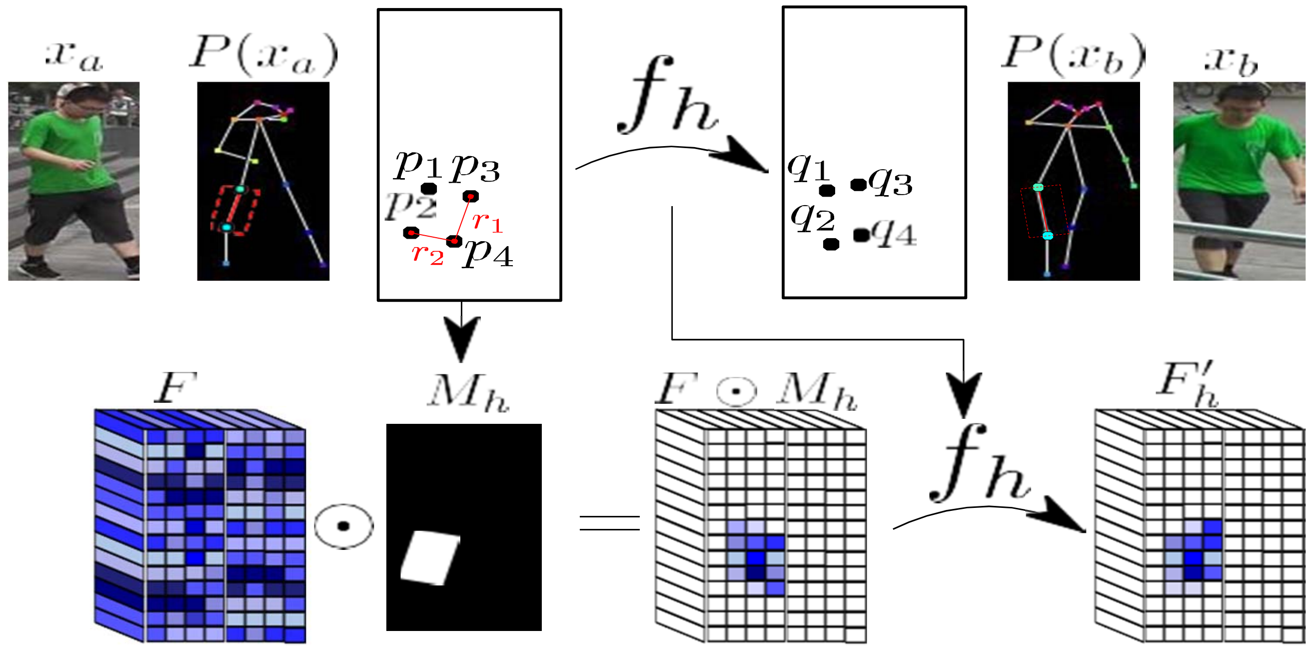


Pose-based Human Image Generation



We need a deformation model

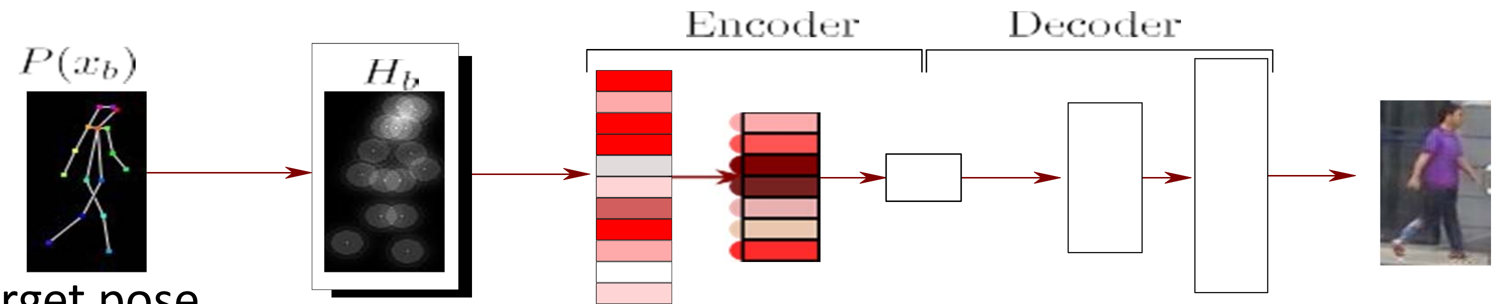
Pose-based Human Image Generation



- 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



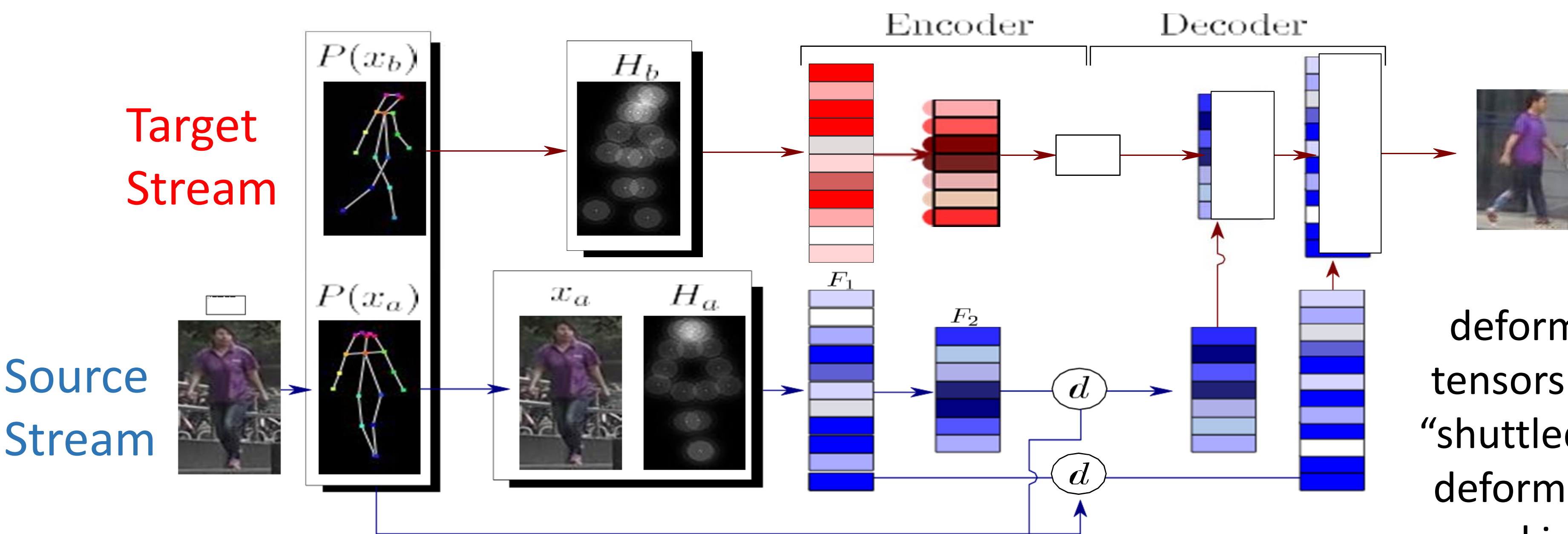
target pose

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

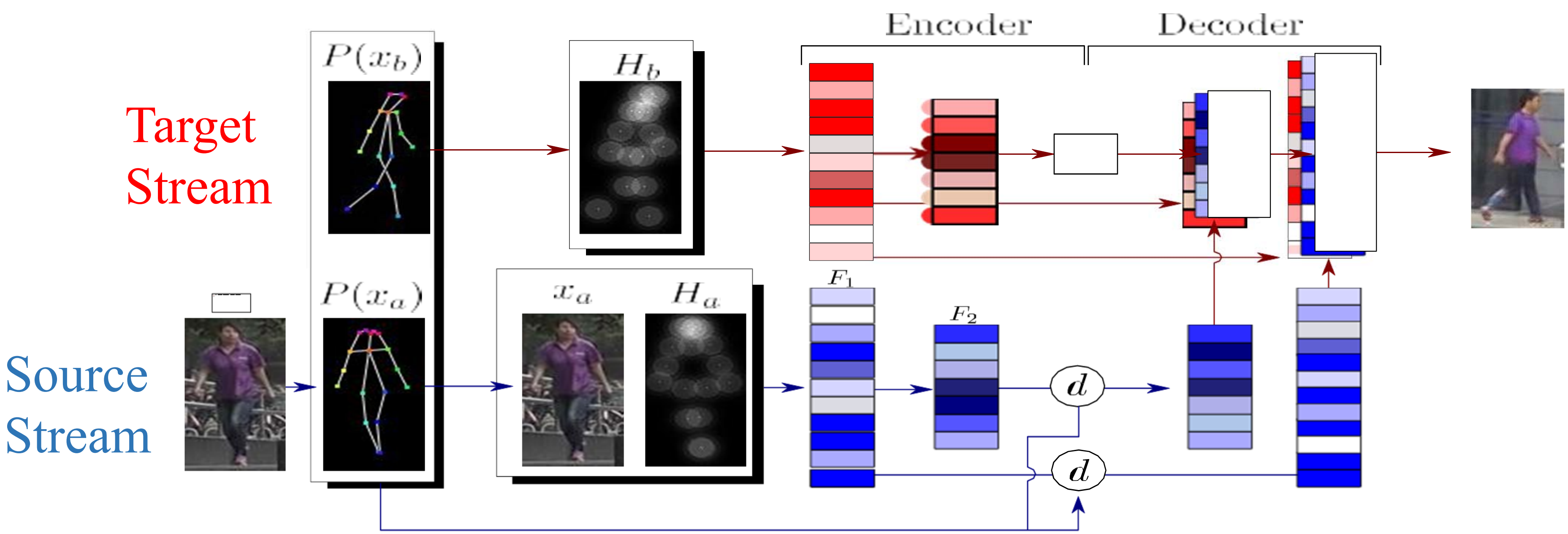
Pose-based Human Image Generation



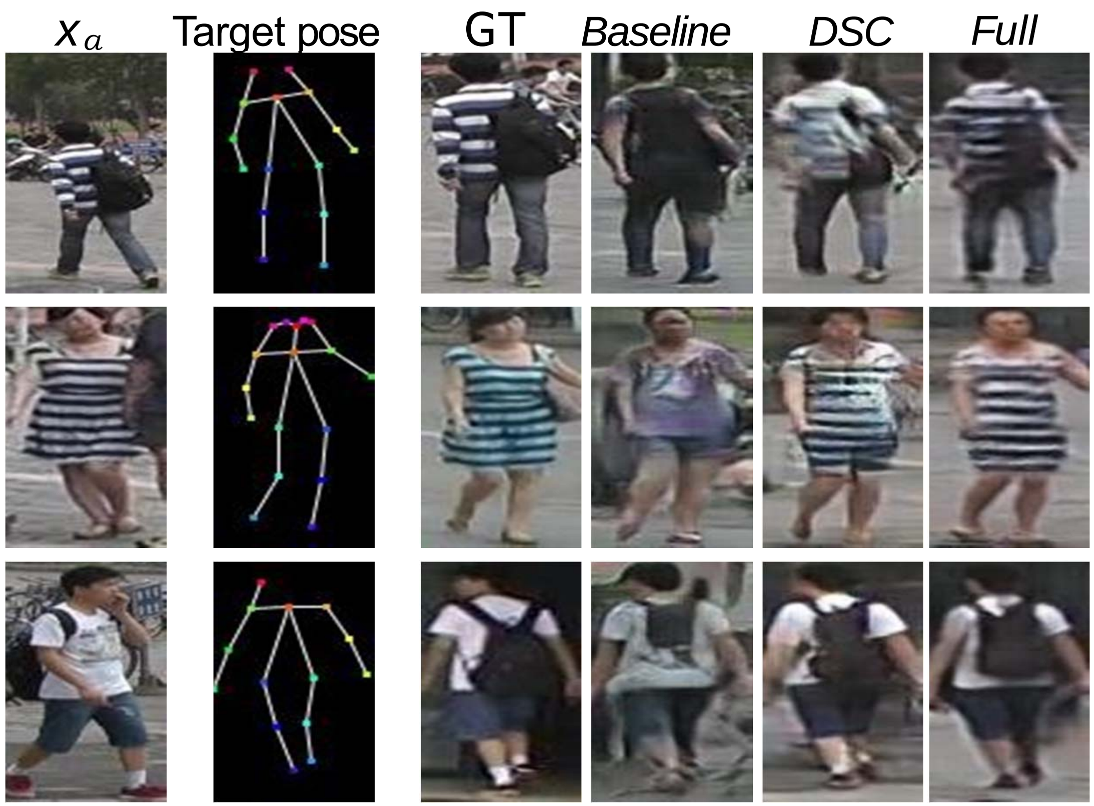
deformed tensors d(F) “shuttled” by deformable skip connections from (x_a, H_a) stream

- 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

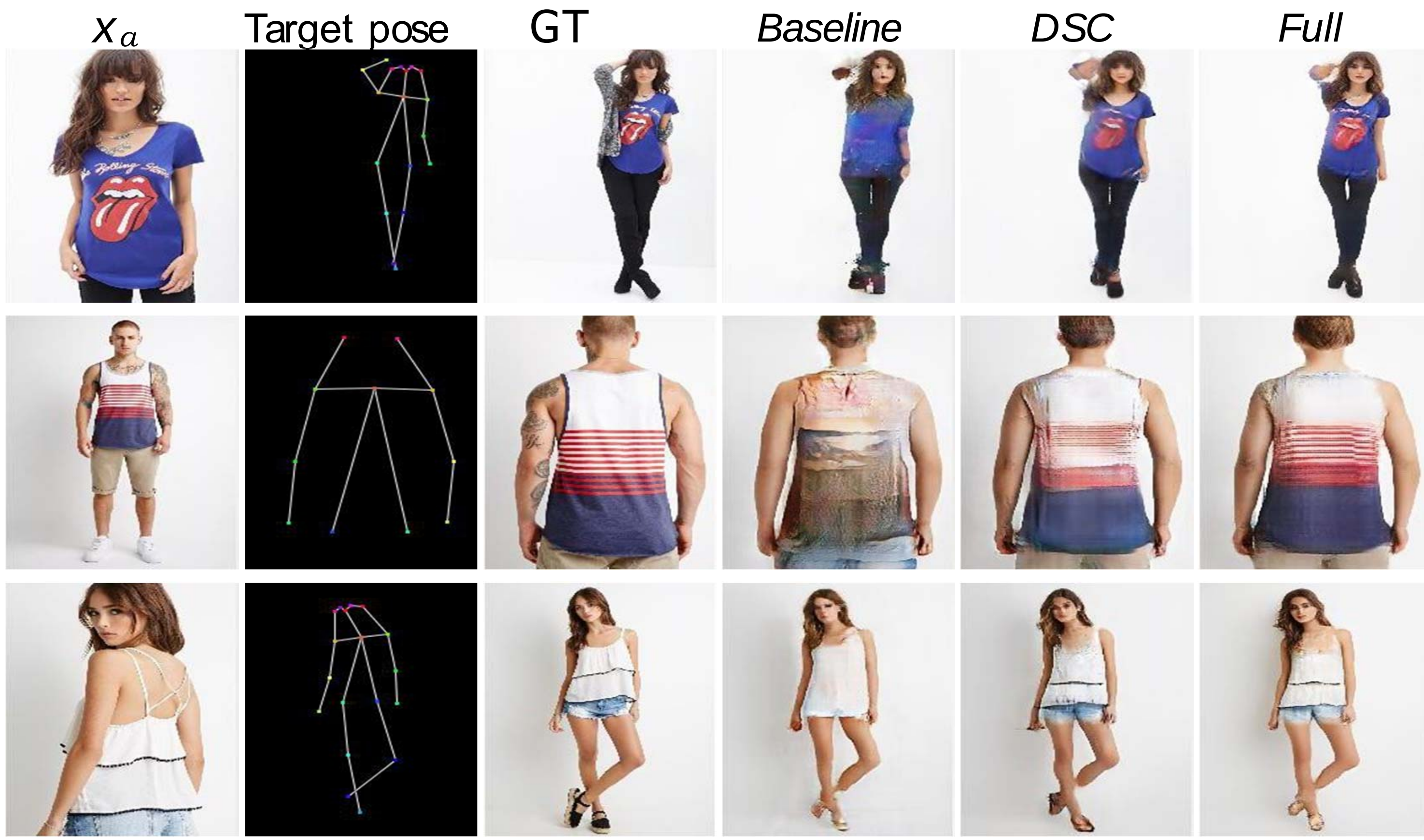
Pose-based Human Image Generation



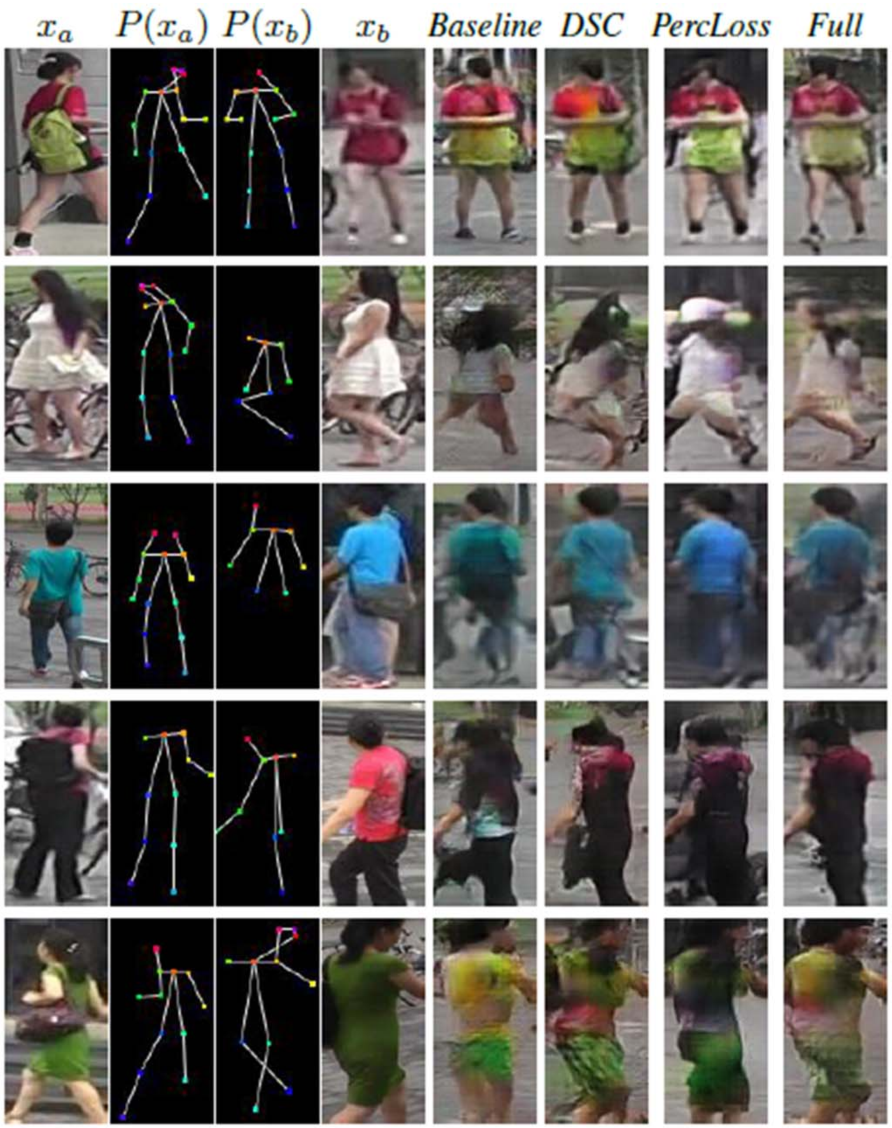
Conditional Image Generation



Qualitative results on the Market-1501 dataset



Qualitative results on the DeepFashion dataset



Badly generated images

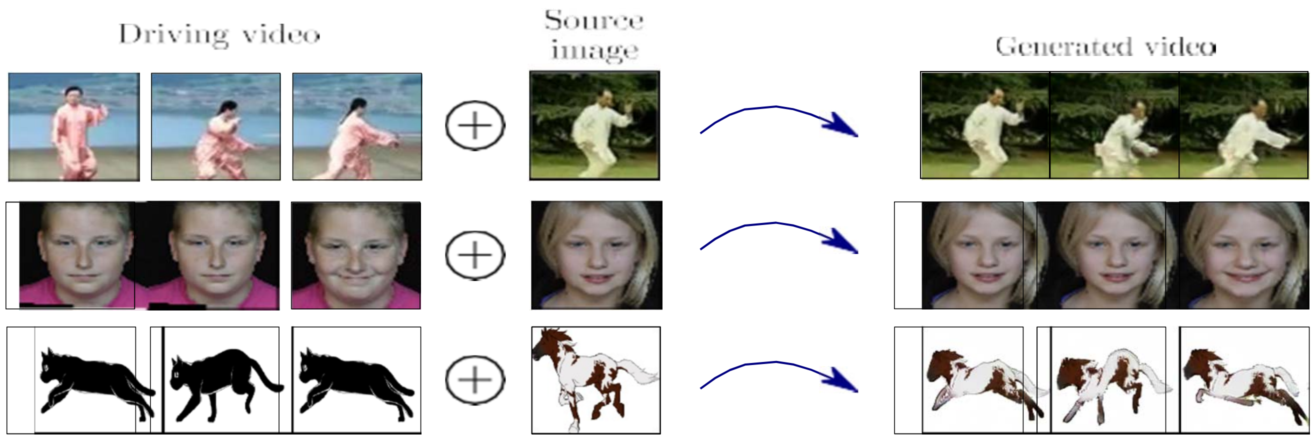
- errors of the pose estimation
- ambiguity of the pose estimation
- rare object appearance
- rare poses

Image Animation

- Aliaksandr Siarohin, Stephane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, "Animating Arbitrary Objects via Deep Motion Transfer", in CVPR,2019
- Aliaksandr Siarohin, Stephane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, "First Order Motion Model for Image Animation", in NeurIPS, 2019

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

Image Animation: Appearance or Motion Transfer?



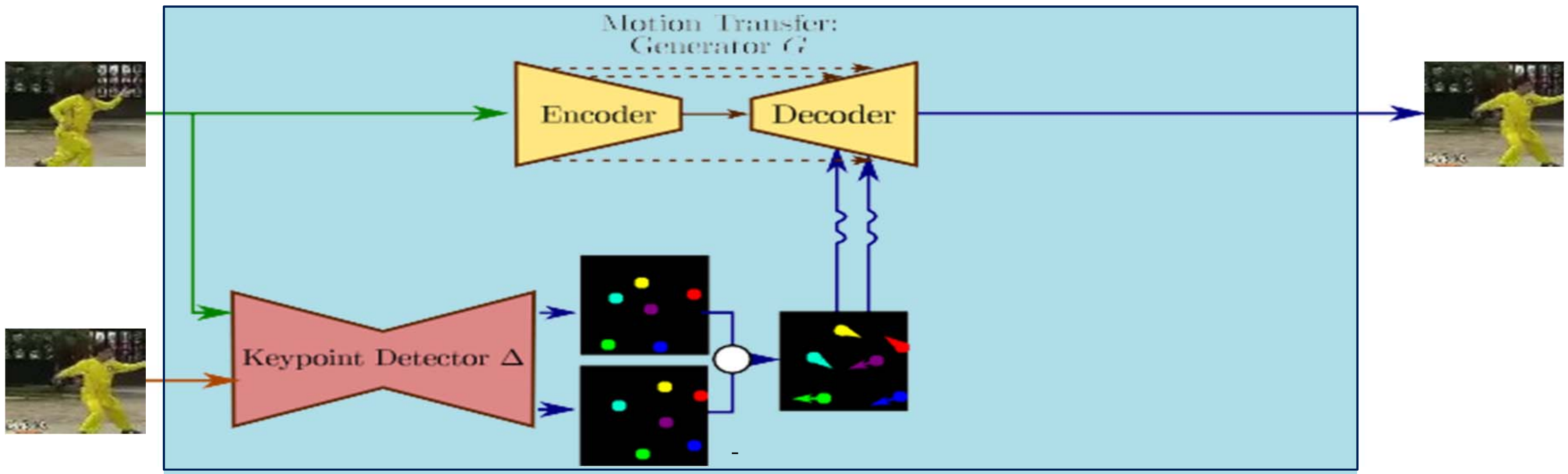
Appearance transfer

Detect pose in each frame of the driving video

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

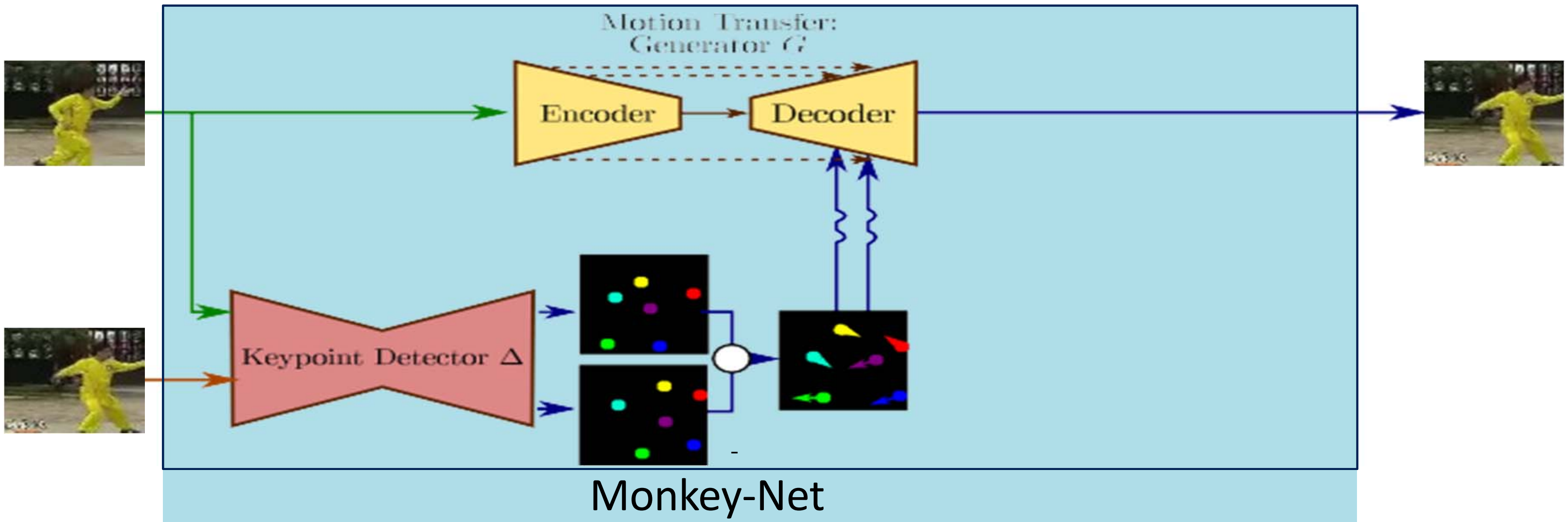
Problems: requires a detector, does not work when the shapes of the object are different (ie. short to tall persons) => **Use Unsupervised Transfer Motion**

Image Animation with MOviNg KEYpoints



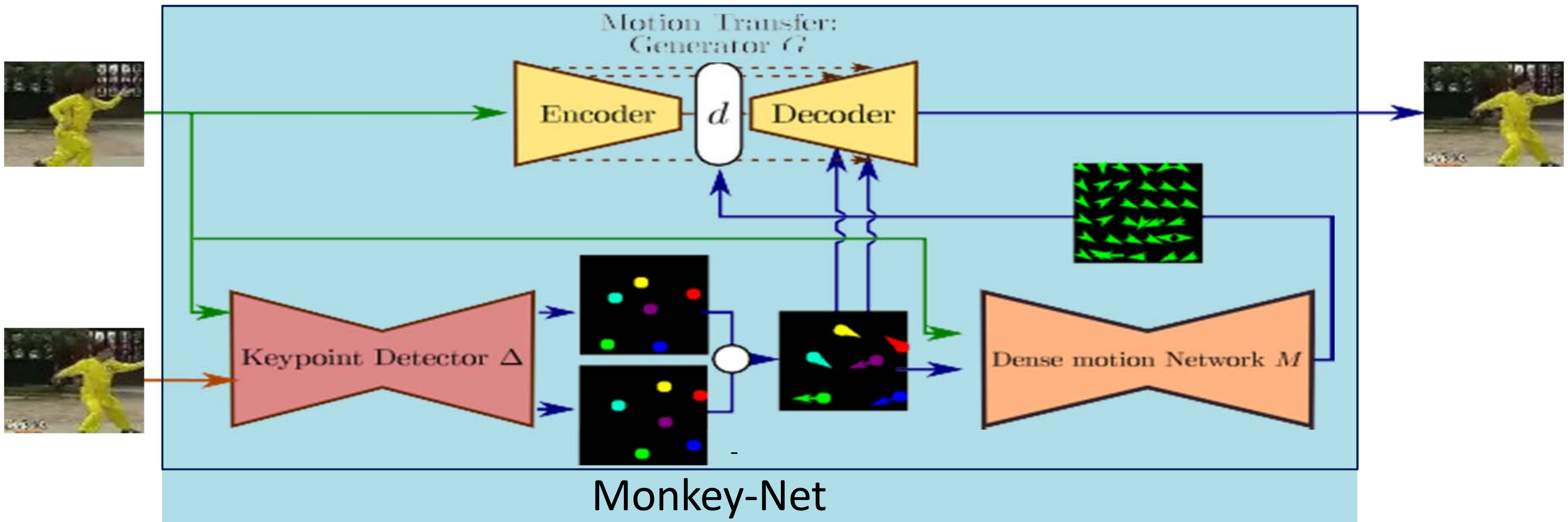
Monkey-Net

Image Animation with MOviNg KEYpoints



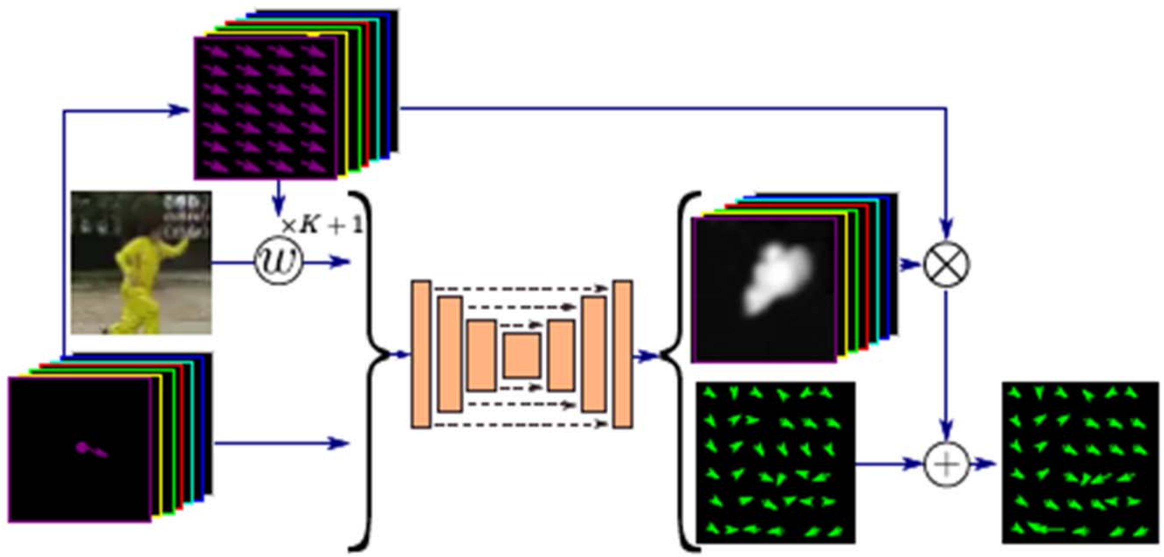
Again, we have an alignment problem

Image Animation with MOviNg KEYpoints



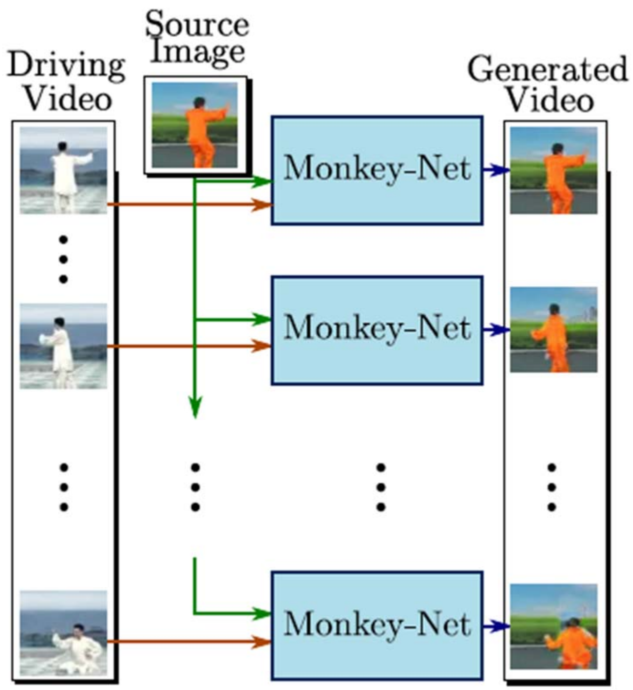
- Monkey-Net has a motion-specific keypoint detector Δ, a motion prediction network M, and an image generator G (reconstructs the image x′ from the keypoint positions Δ(x) and Δ(x′)); Optical flow computed by M is used by G to handle misalignments between x and x′.
- The model is learned with a self-supervised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

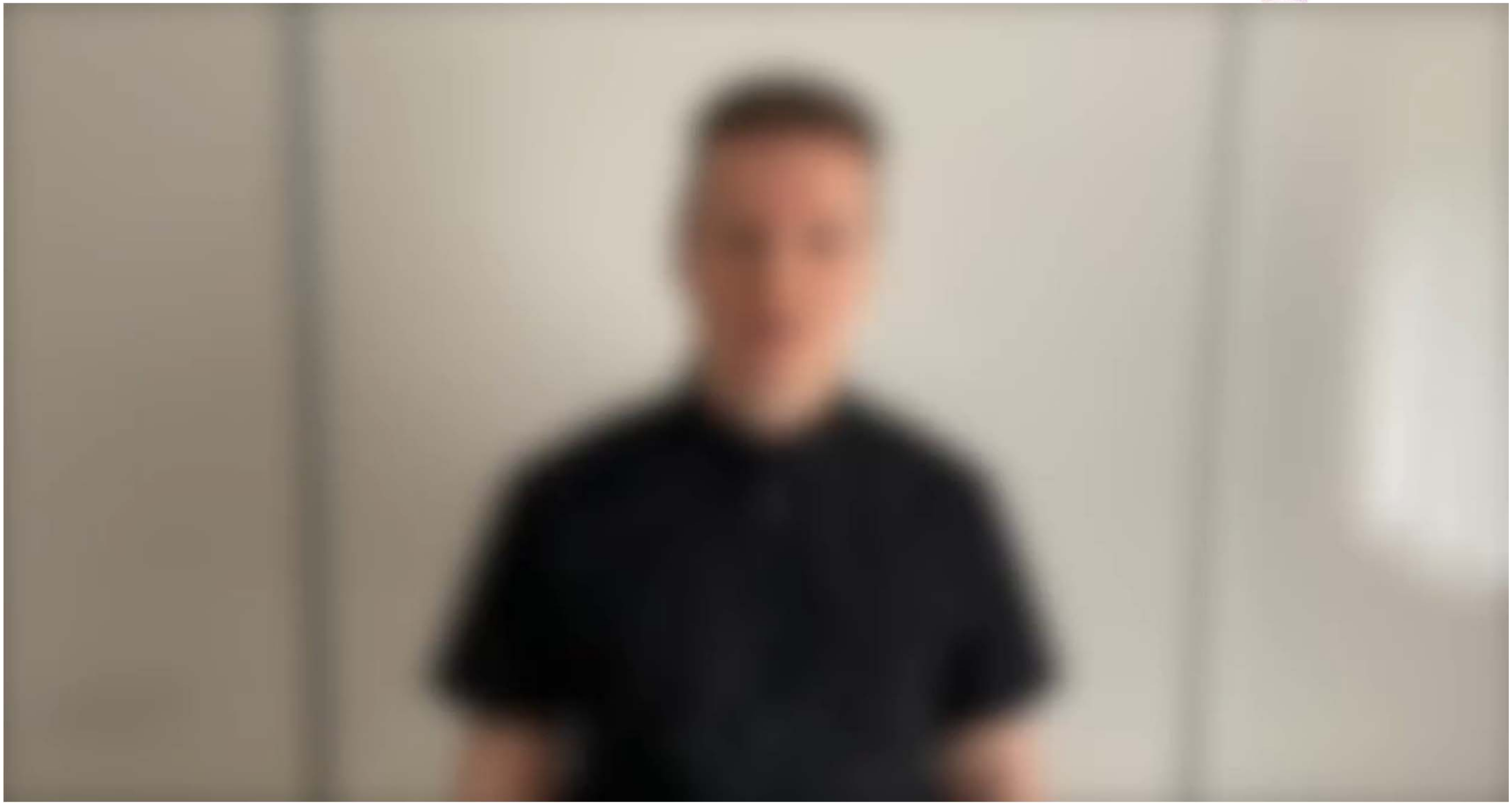
Image Animation Evaluation

	Tai-Chi			Nemo			Bair
	$\mathcal{L}_1$	AKD	AED	$\mathcal{L}_1$	AKD	AED	$\mathcal{L}_1$
X2Face [7]	0.068	4.50	0.27	0.022	0.47	0.140	0.069
Ours	**0.050**	**2.53**	**0.21**	**0.017**	**0.37**	**0.072**	**0.025**

AKD: Average Keypoint Distance; AED: Average Euclidean Distance

Tai-Chi	*Nemo*	*Bair*
85.0%	79.2%	90.8%

User study. Proportion of times our approach is preferred over X2face

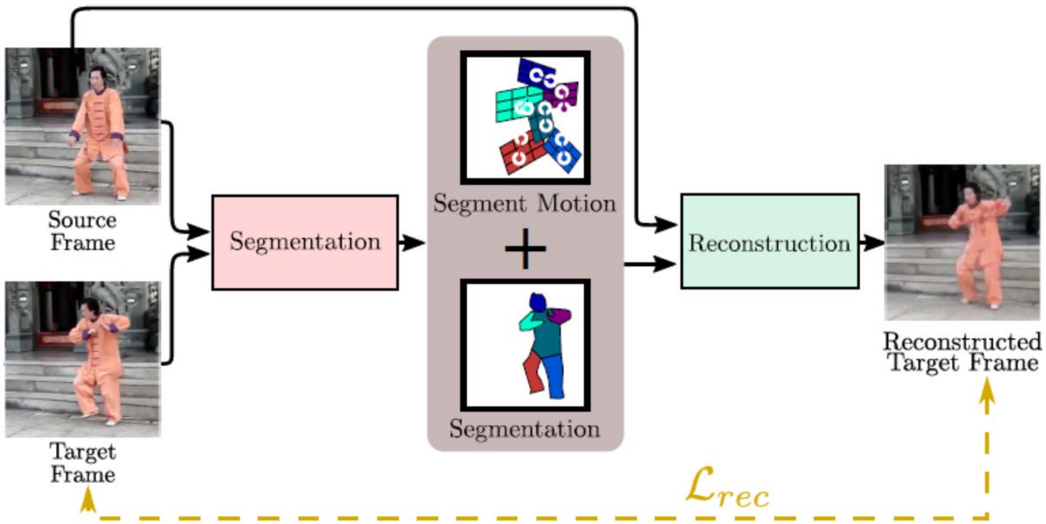


Motion-supervised Co-Part Segmentation

- Aliaksandr Siarohin, Subhankar Roy, Stephane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, “Motion Supervised Co-Part Segmentation”, in ICPR 2020

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

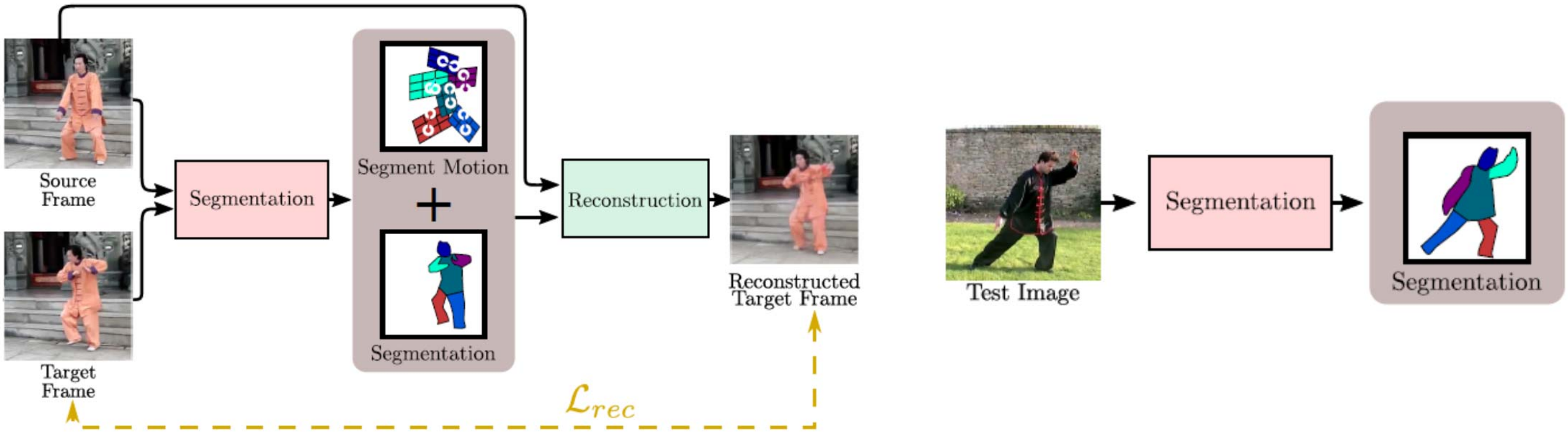
Self-supervised Co-Part Segmentation



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

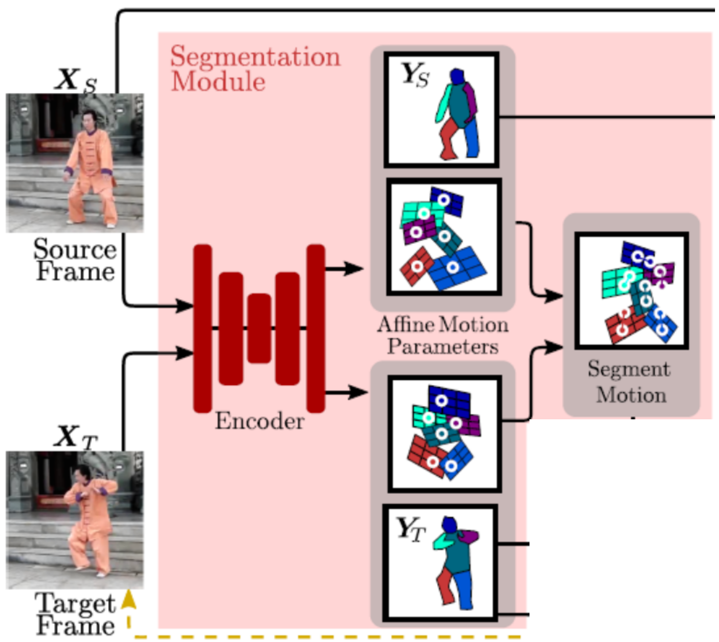
Self-supervised Co-Part Segmentation



Leverage motion info to train a segmentation network without annotation

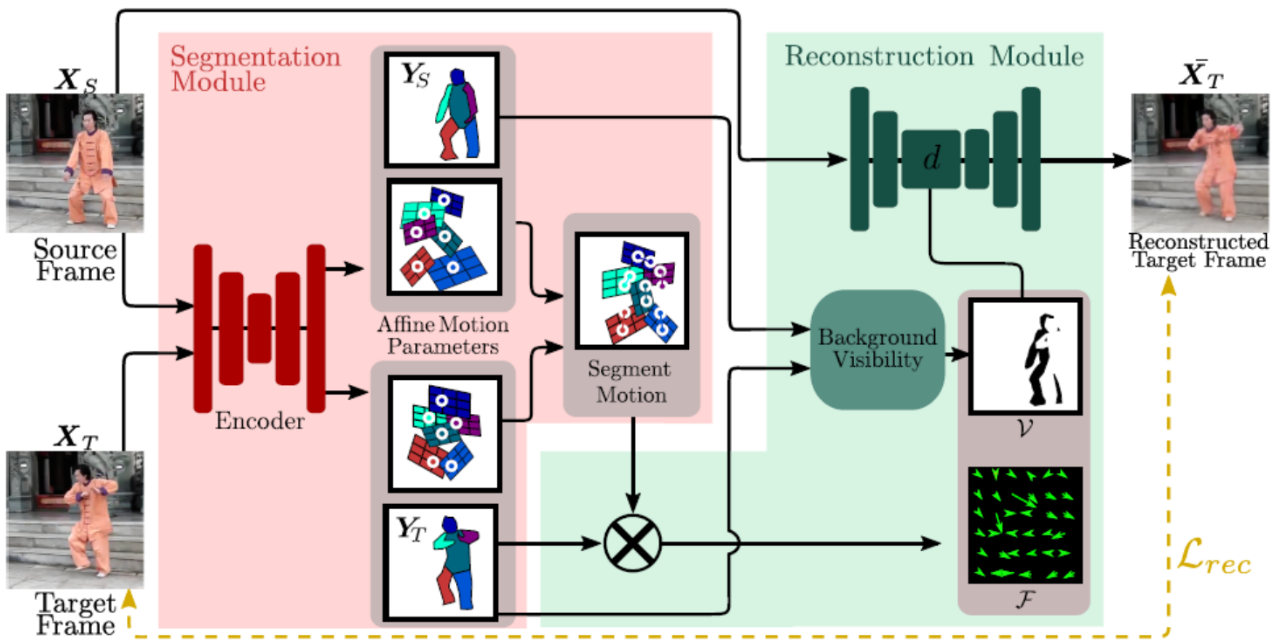
- At training, use frame pairs (source and target) extracted from the same video => predict segments in target that can be combined with a motion representation between the two frames to reconstruct the target frame
- At inference, use the trained segmentation model to predict object parts segments

Self-supervised Co-Part Segmentation



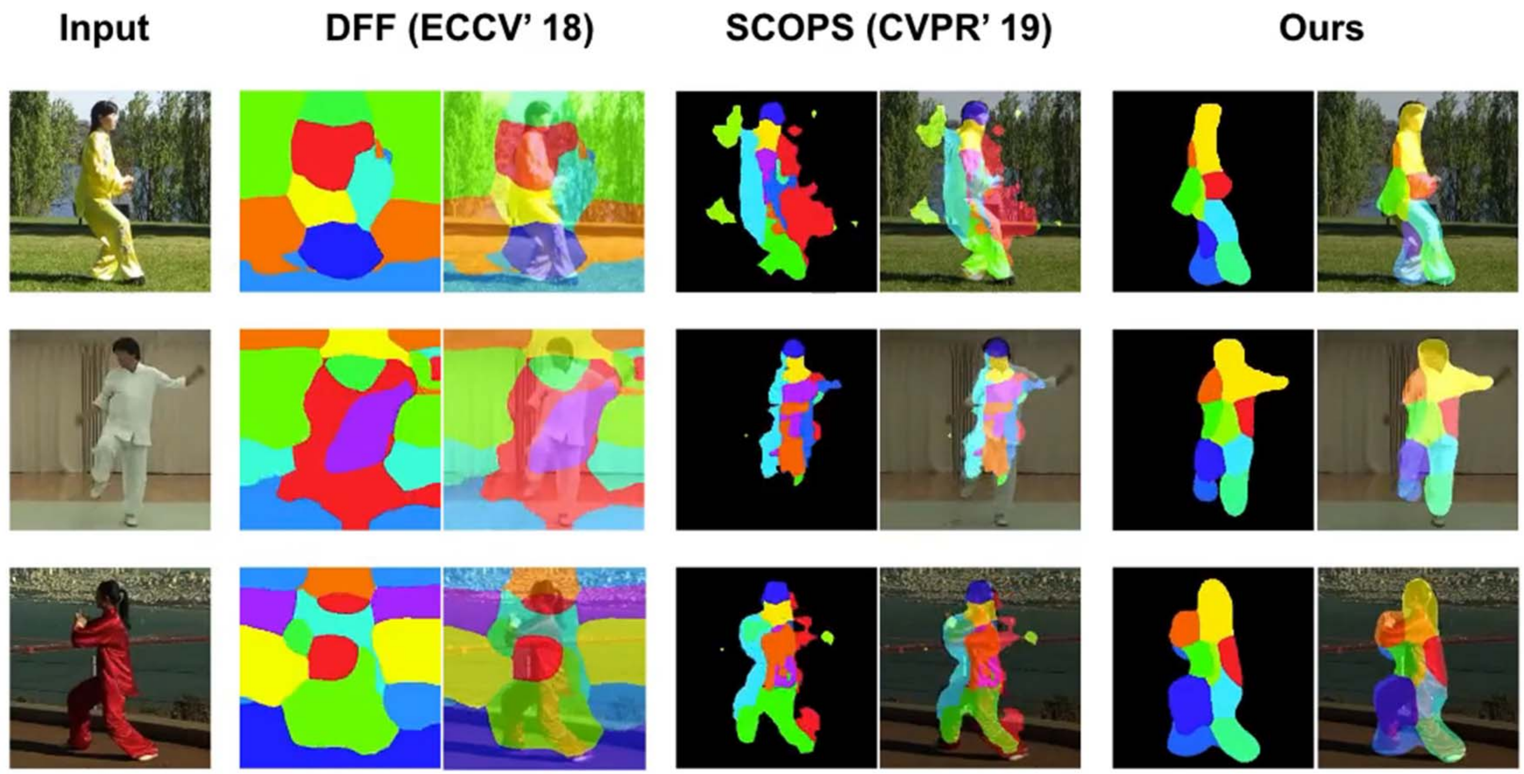
- **Segmentation Module** predicts the segmentation maps Y_S and Y_T, and the affine motion parameters

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 the optical flow F; (2) reconstructs the target frame X_T by warping the features of the source frame X_S and masking the occluded features

Tai-Chi-HD

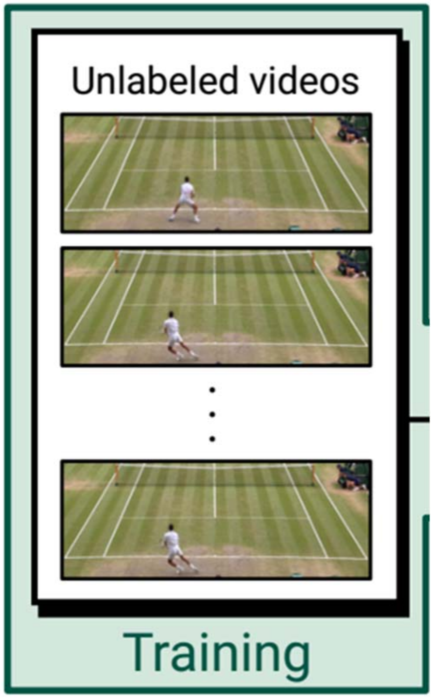


Playable Video Generation

- Willi Menapace, Stephane Lathuilière, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci , “Playable Video Generation”, in CVPR 2021

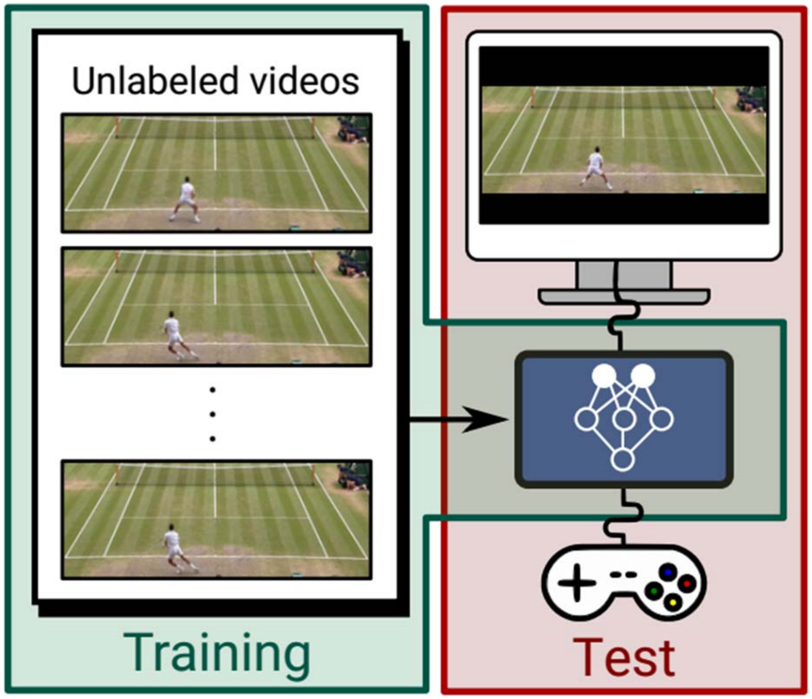
https://github.com/willi-menapace/PlayableVideoGeneration

Playable Video Generation



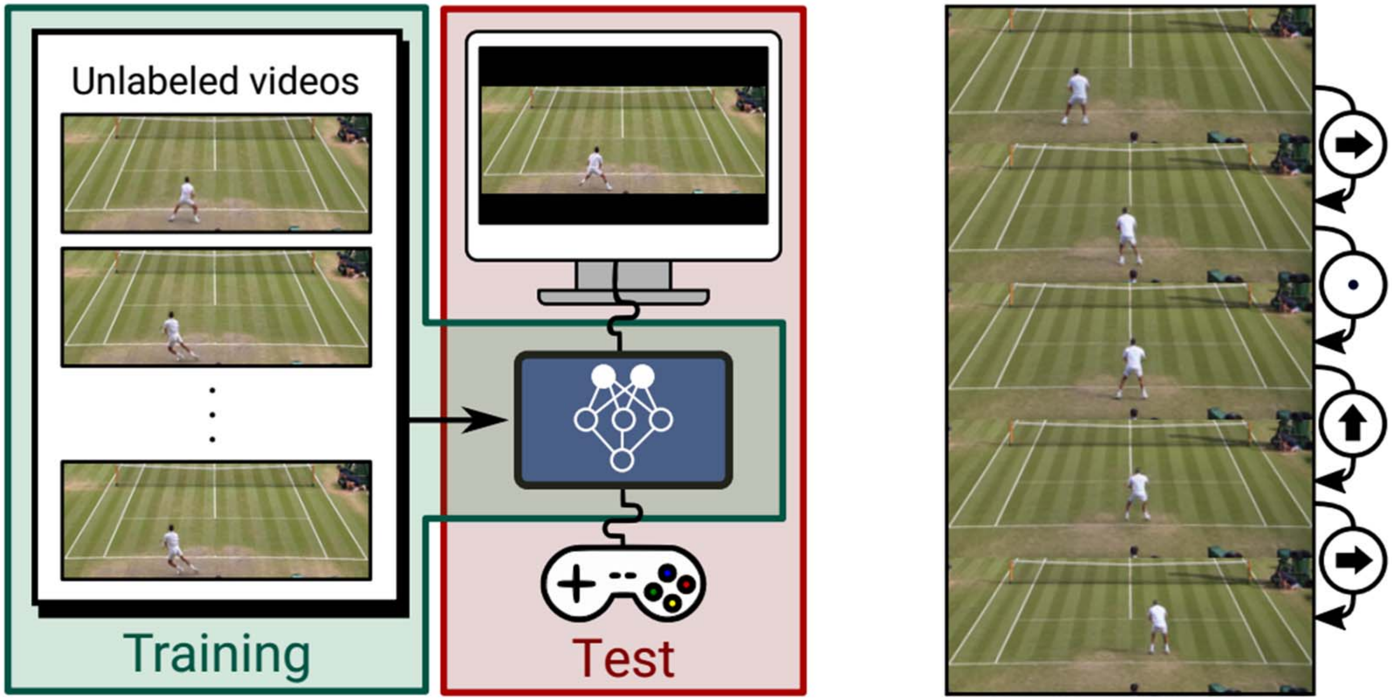
- Consider a set of videos depicting an agent acting in an environment
- Differently from other methods that use frame by frame action annotations, no annotation is present

Playable Video Generation



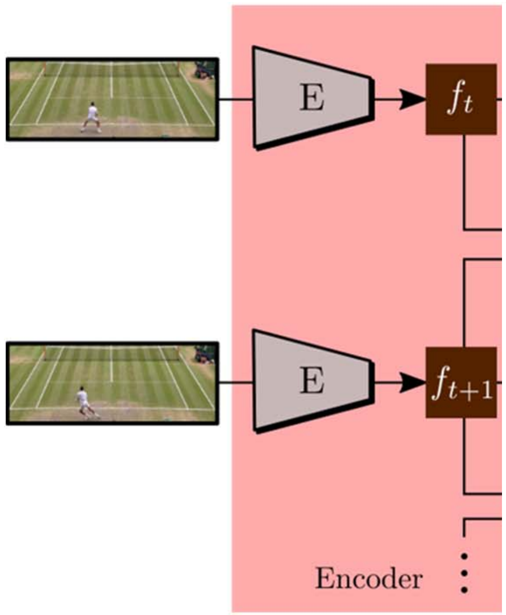
- Learn a model that represents the observed environment
- Allow the user to input actions to the model through a controller at the testing time

Playable Video Generation



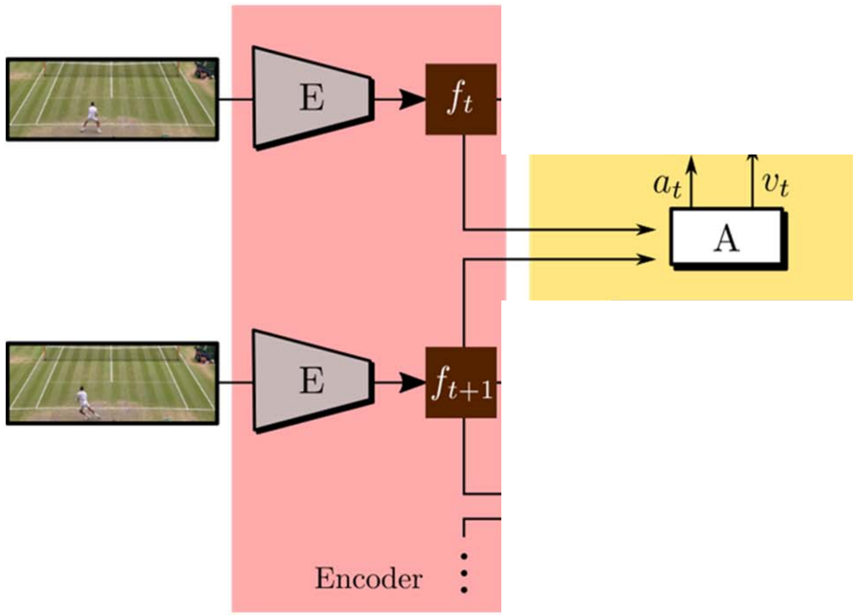
- Produce a video where the agent acts according to the actions specified by the user

Architecture



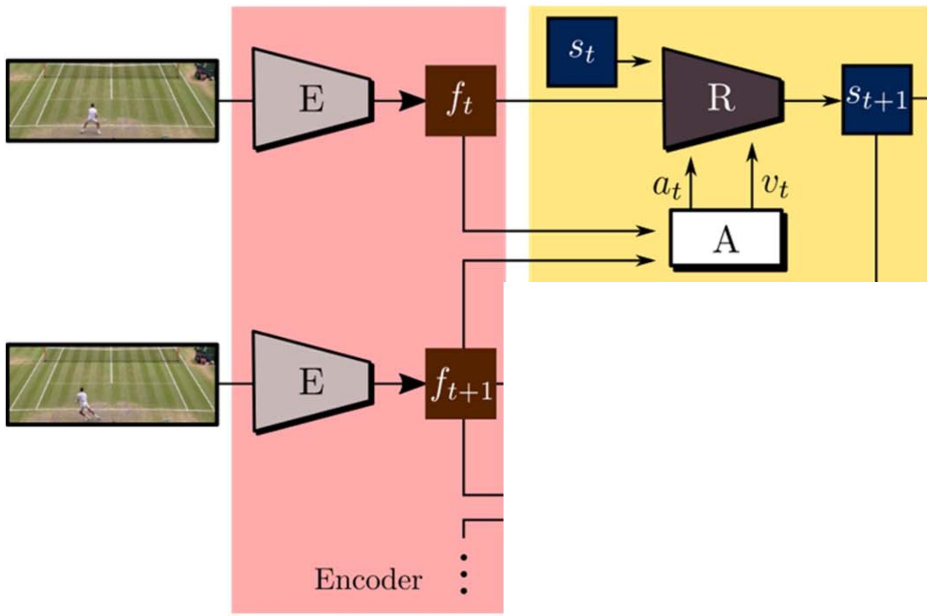
- Sample an input sequence and use an encoder network to extract frame features

Architecture



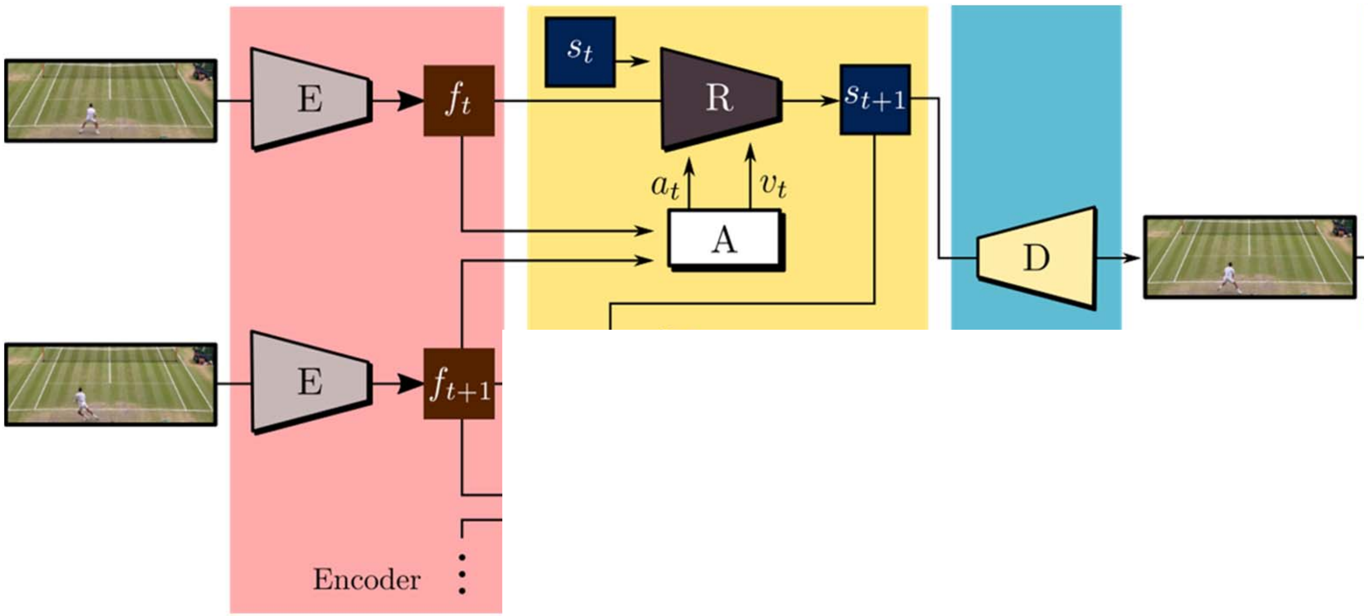
- Use then pairs of successive features to infer the action that was performed by the agent in the corresponding transition using an action network

Architecture



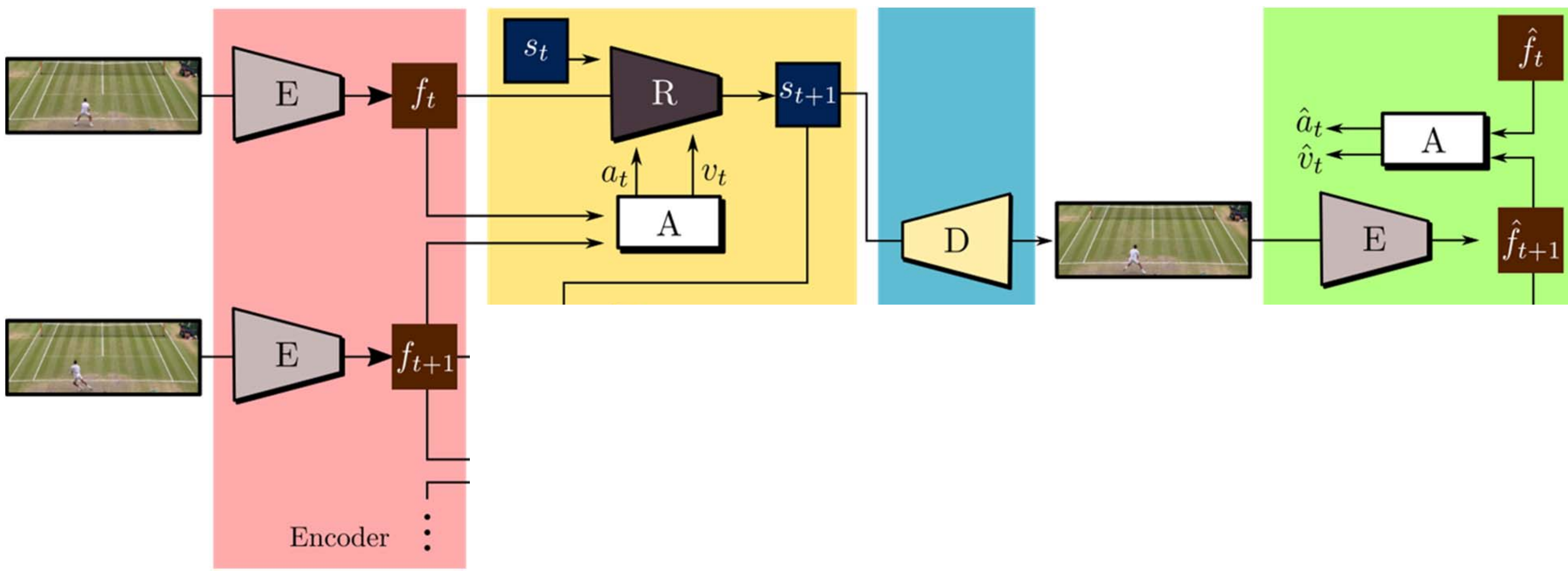
- Given the frame features and the action, a recurrent model is used to produce features representing the successive state

Architecture



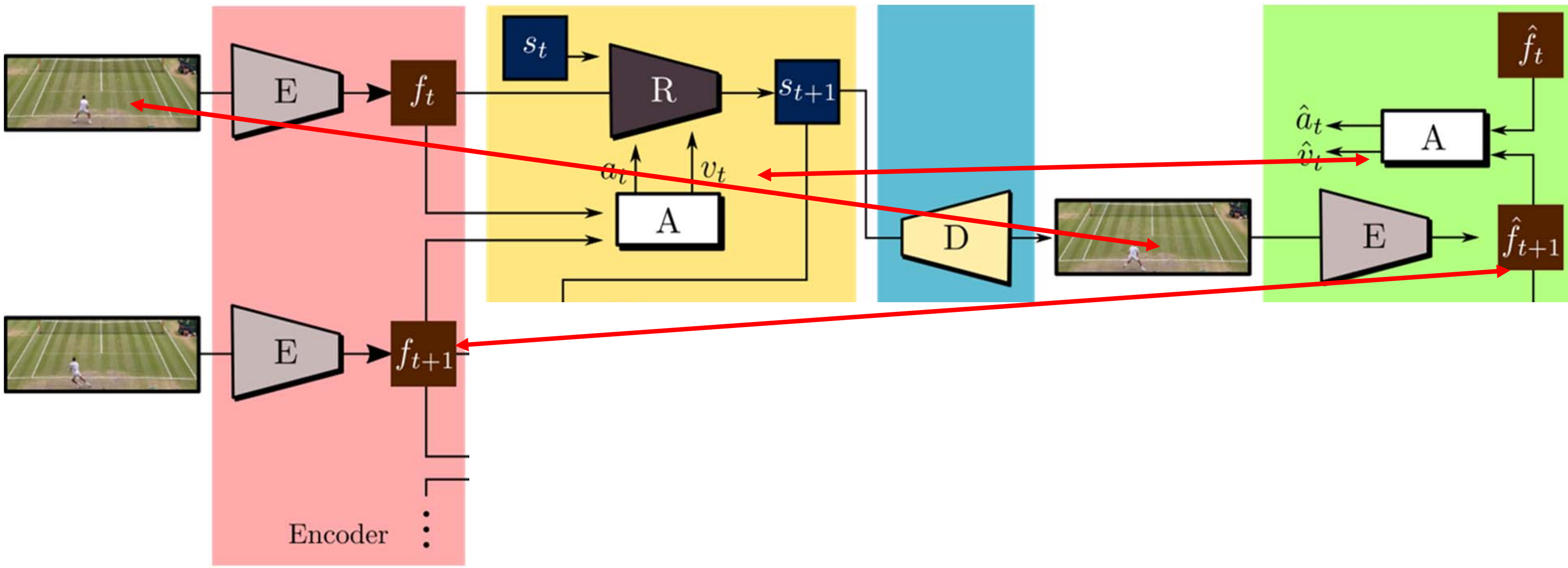
- The successive state is translated back to an image using a decoder network

Architecture



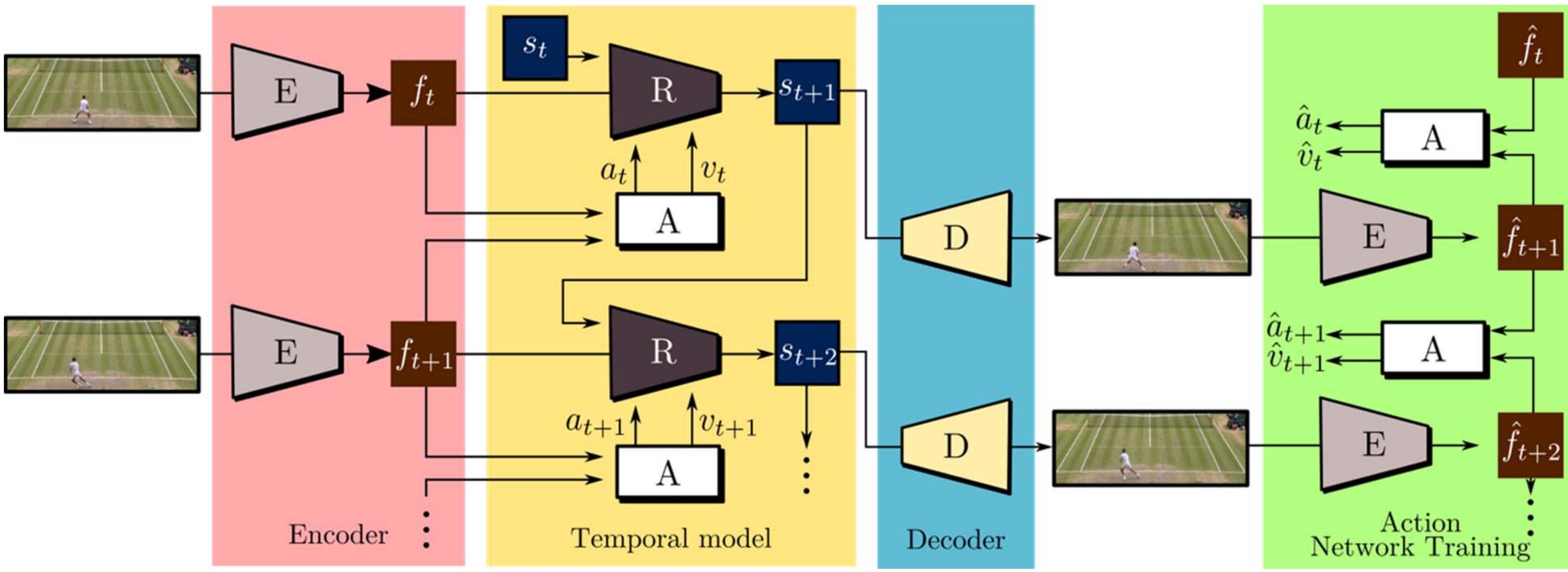
- For extra supervision, we encode back the produced frame using the encoder and the action network

Architecture



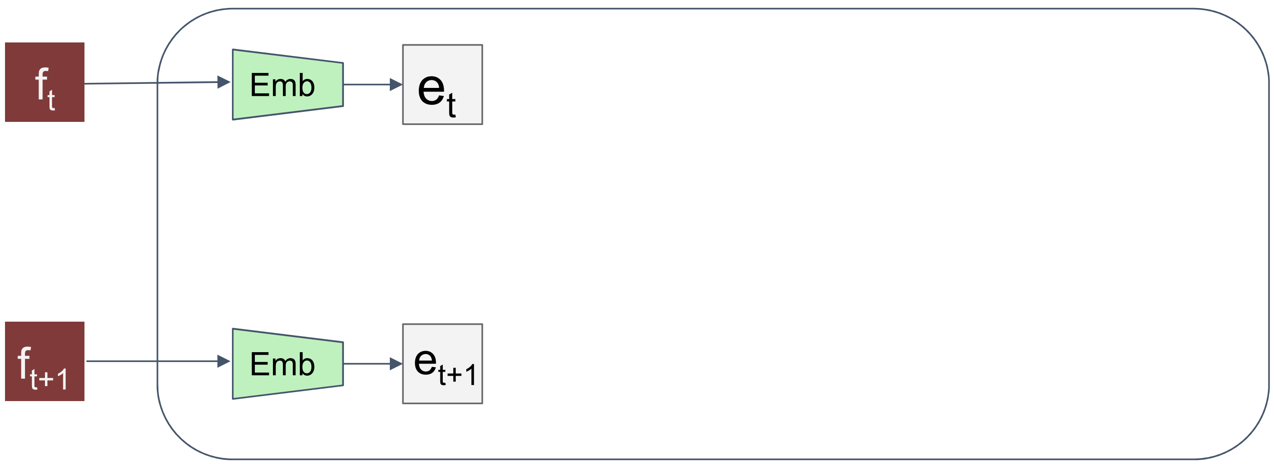
- Impose different self supervision losses on the frames, the frame features and the produced actions: use a mutual information maximization loss between actions and reconstructed actions as the main driving loss for action learning

Architecture



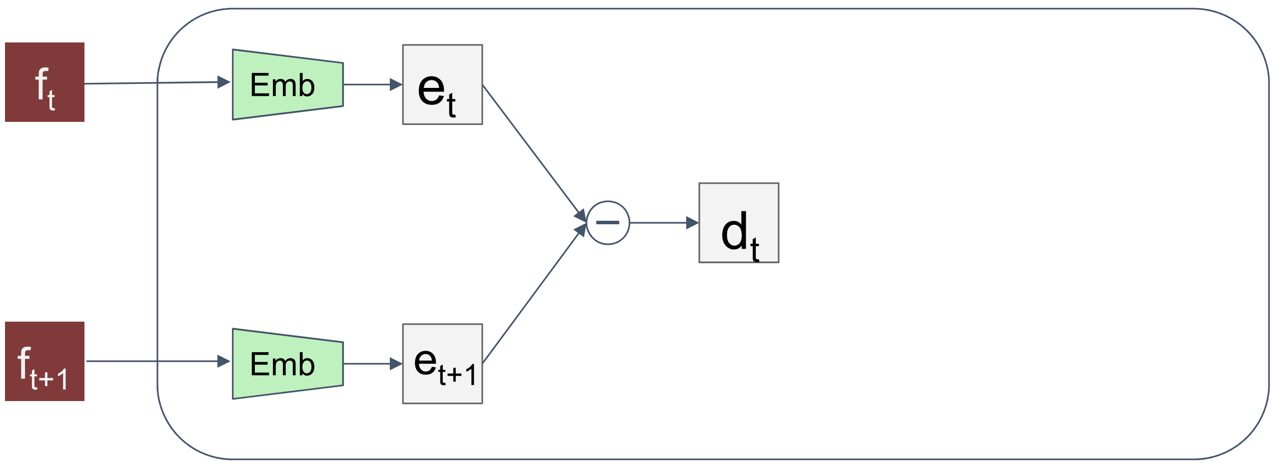
- The model is then unrolled over the whole sequence

Action Network



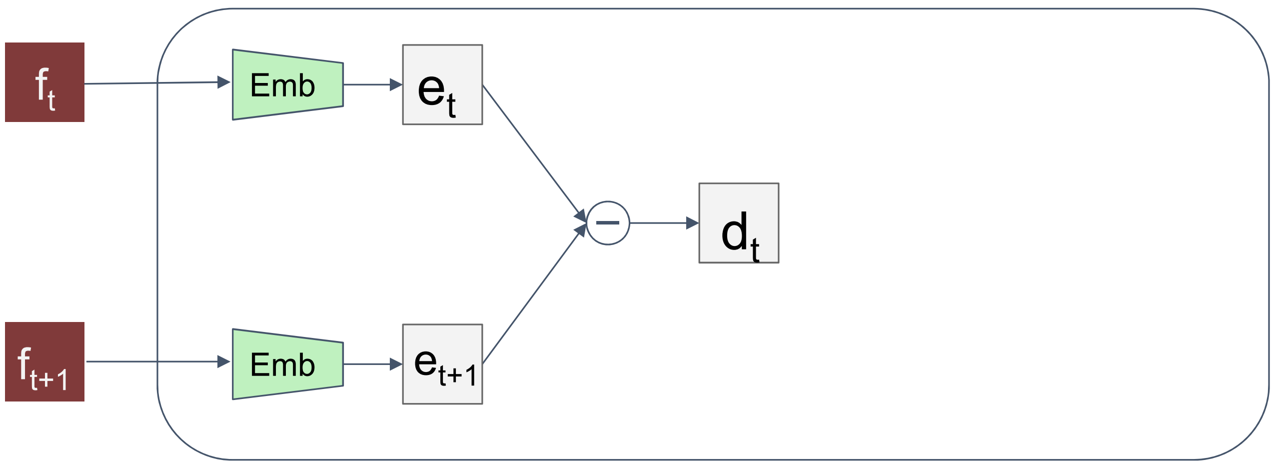
- The action network first encodes the frame features using a Multi Layer Perceptron to produce two embeddings

Action Network

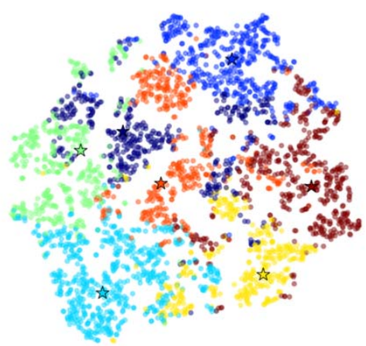


- Take the difference between these embedding as the representation of the transition between two frames: action direction d_t

Action Network

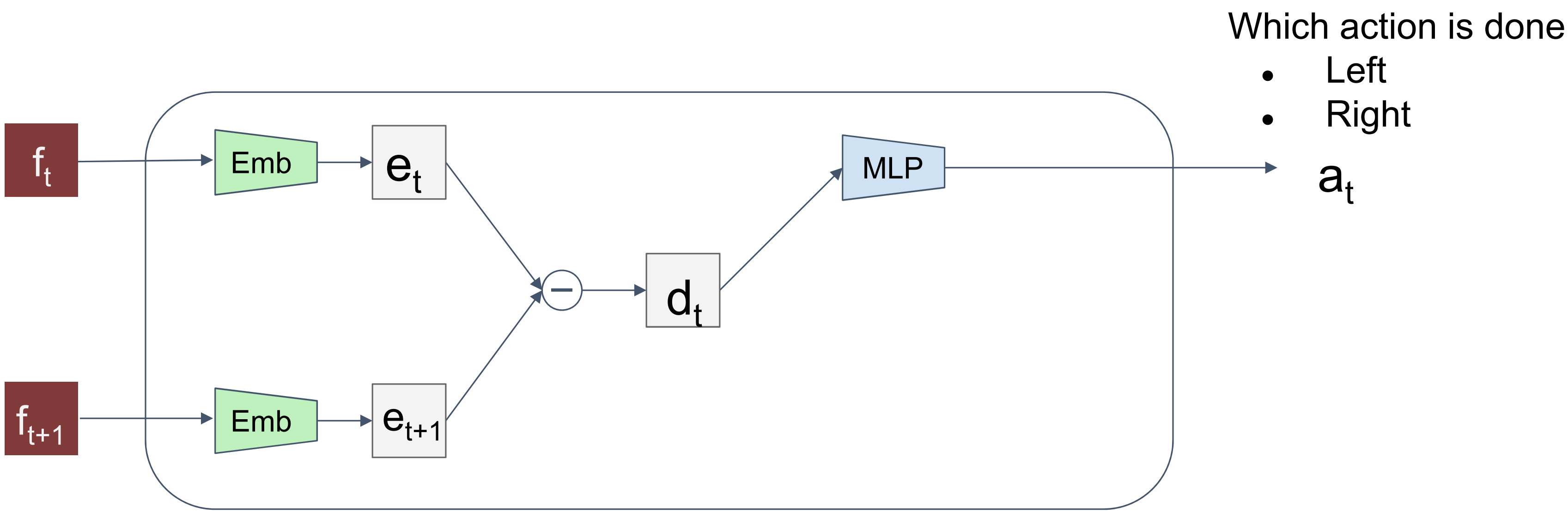


t-SNE plot of d_t

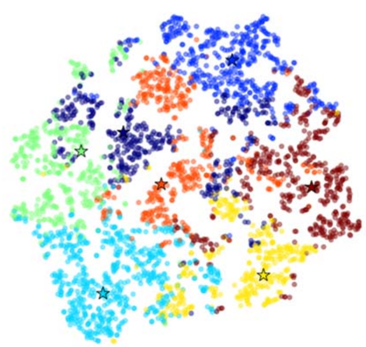


- When visualized, the learned space of action directions is a representation of the different types of transitions that are observed in the training videos

Action Network

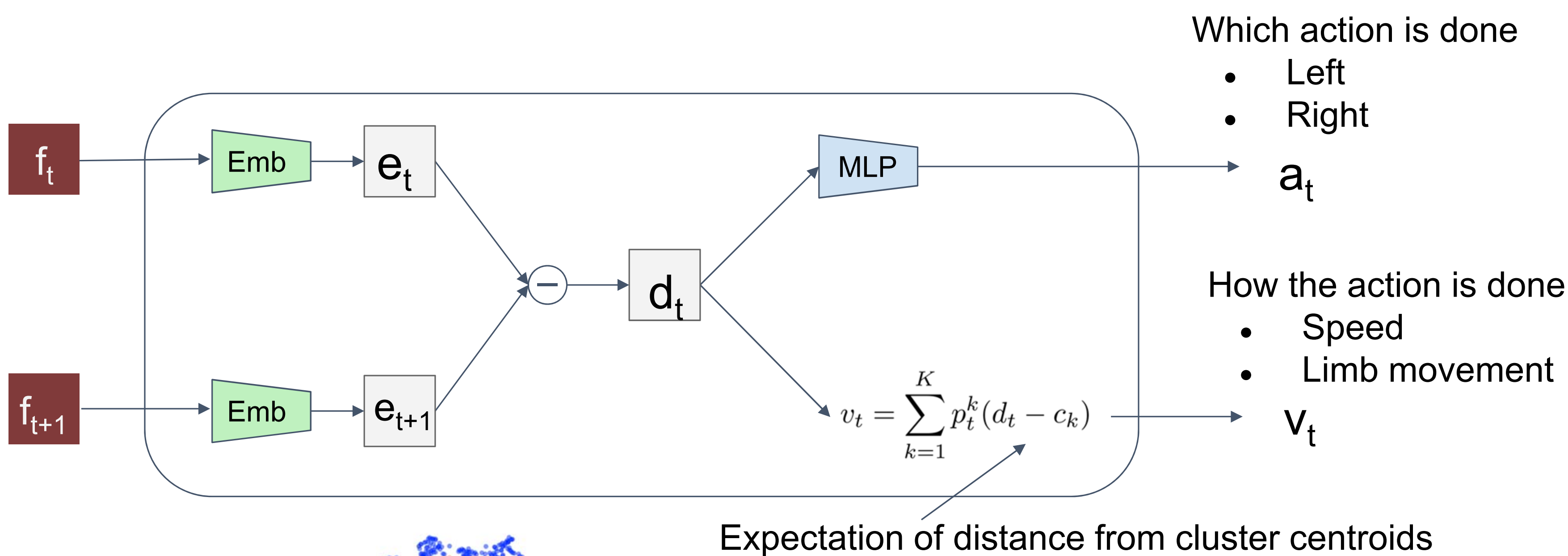


t-SNE plot of d_t

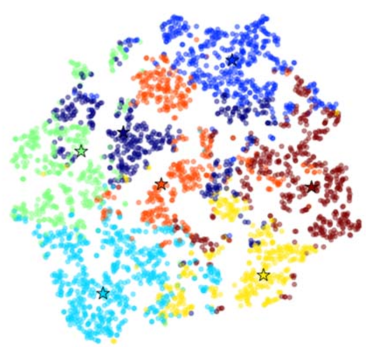


- Use an MLP to assign a label to each point d_t: the high level action associated to the current frame
- Use of action variability embeddings to ensure a well-posed reconstruction loss on the frames

Action Network

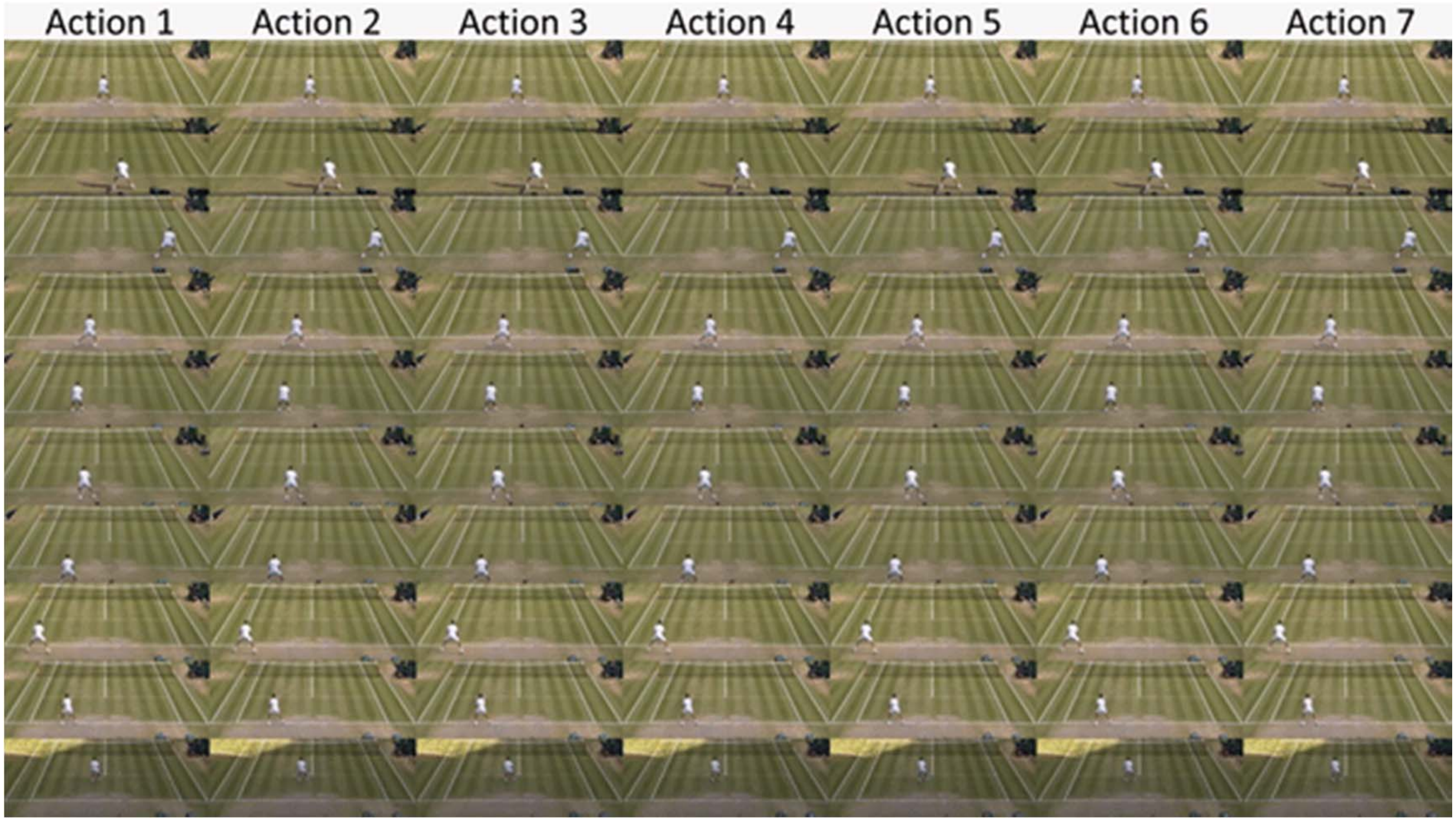


t-SNE plot of d_t



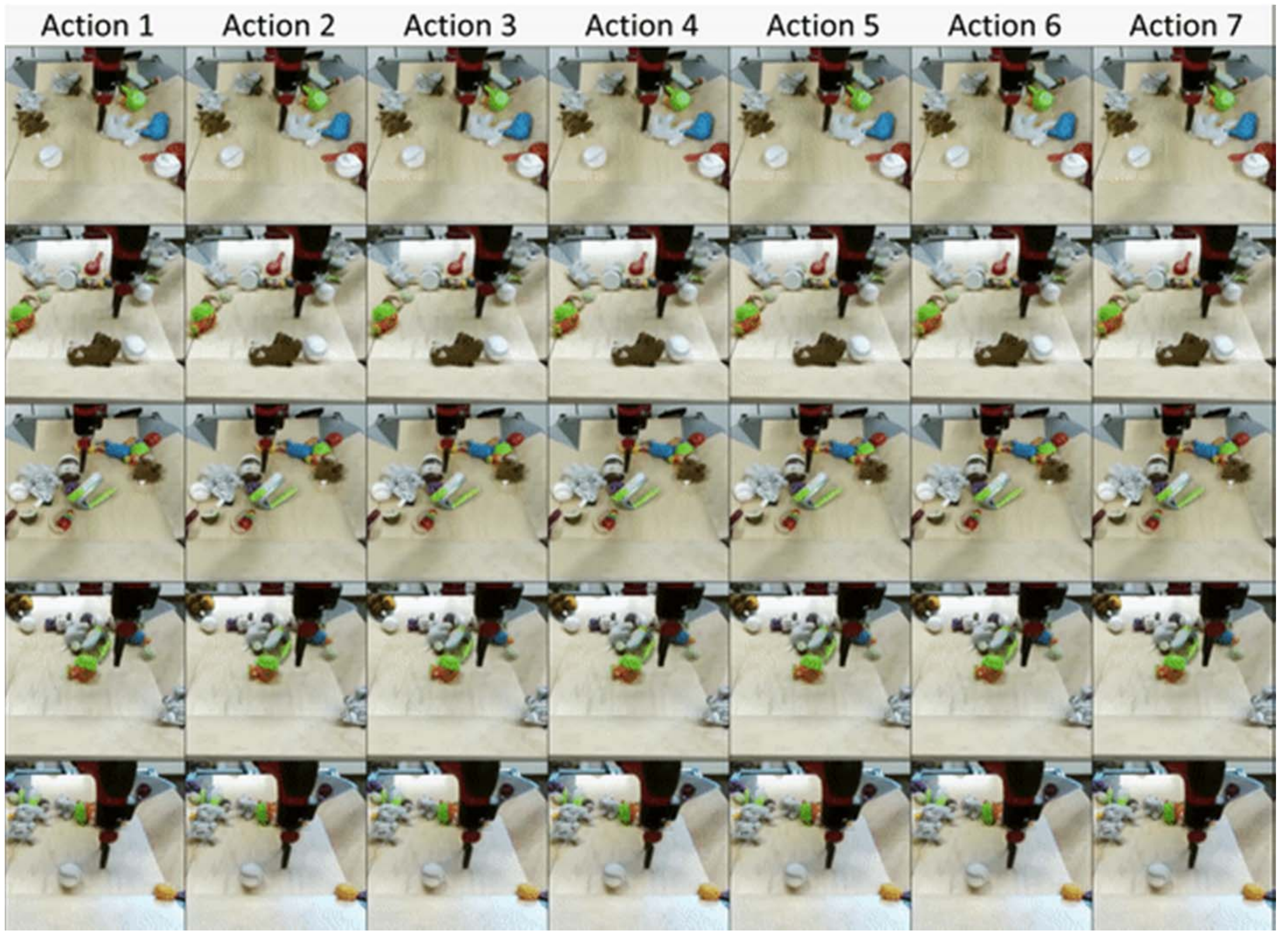
- For each d_t compute the expectation of its distance from the cluster centroids: variability embedding v_t => the particular way in which an action is performed

Results

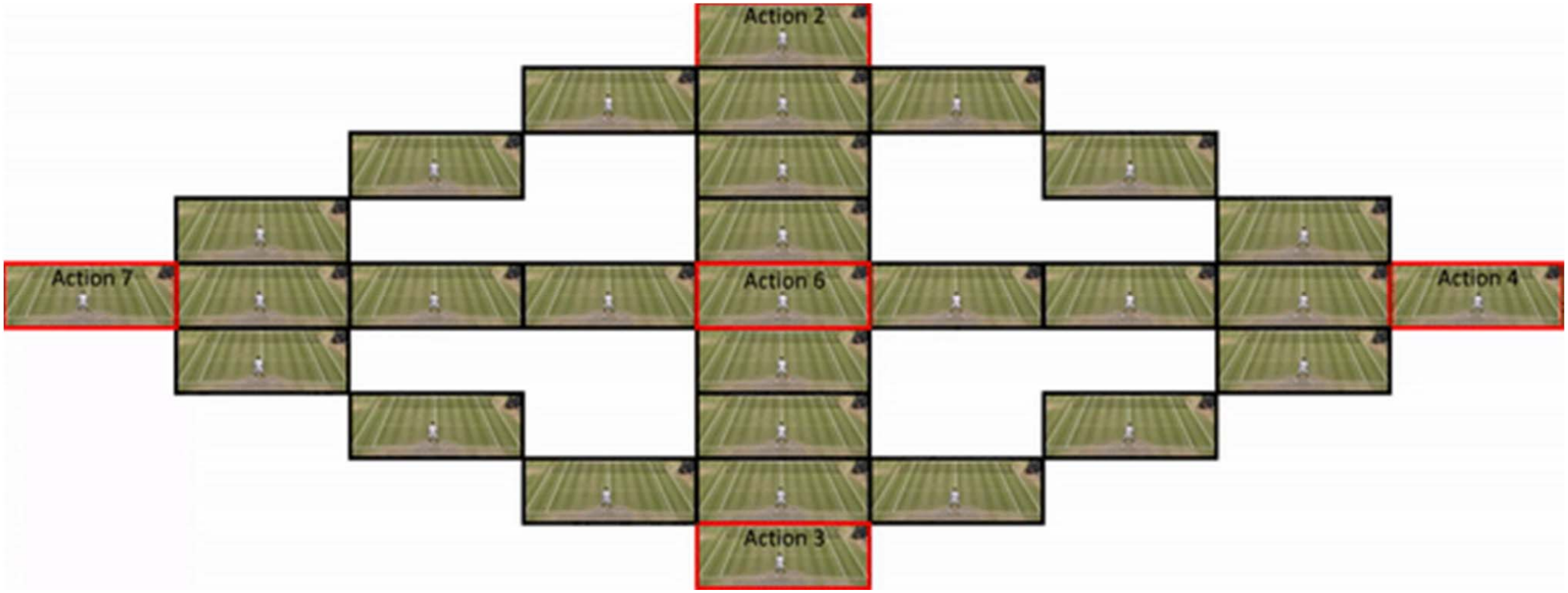


- We learn a wide range of actions. The meaning of actions is consistent, independently from the starting frame the action is applied to

Results



Action Interpolation

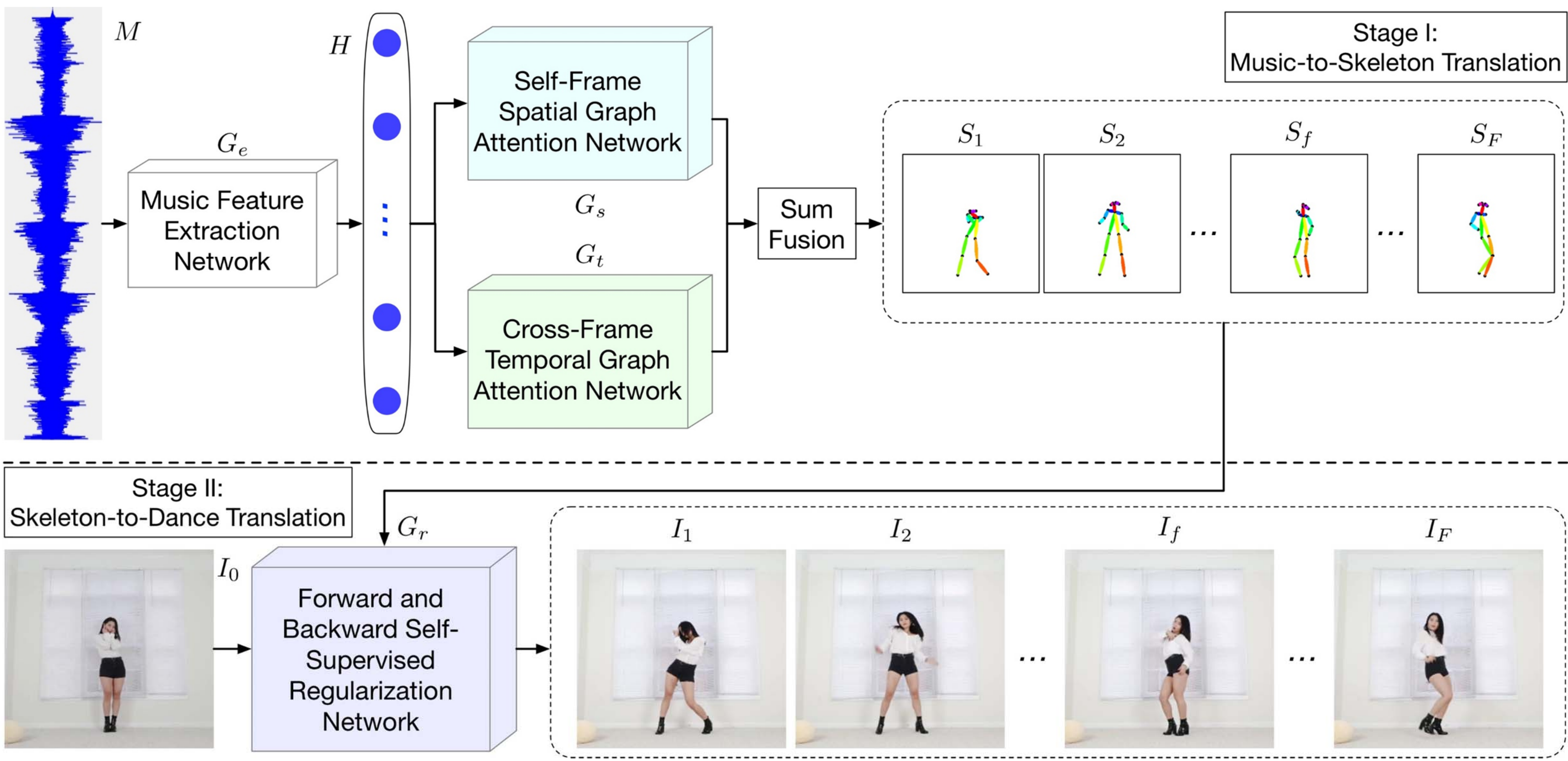


- At inference, typically v_t = 0 and user is specifying actions a_t at each time step
- v_t can also be obtained from an action direction d_t that moves between the centroids of different actions => generate a variety of different movement directions, eg. diagonal movements


Action

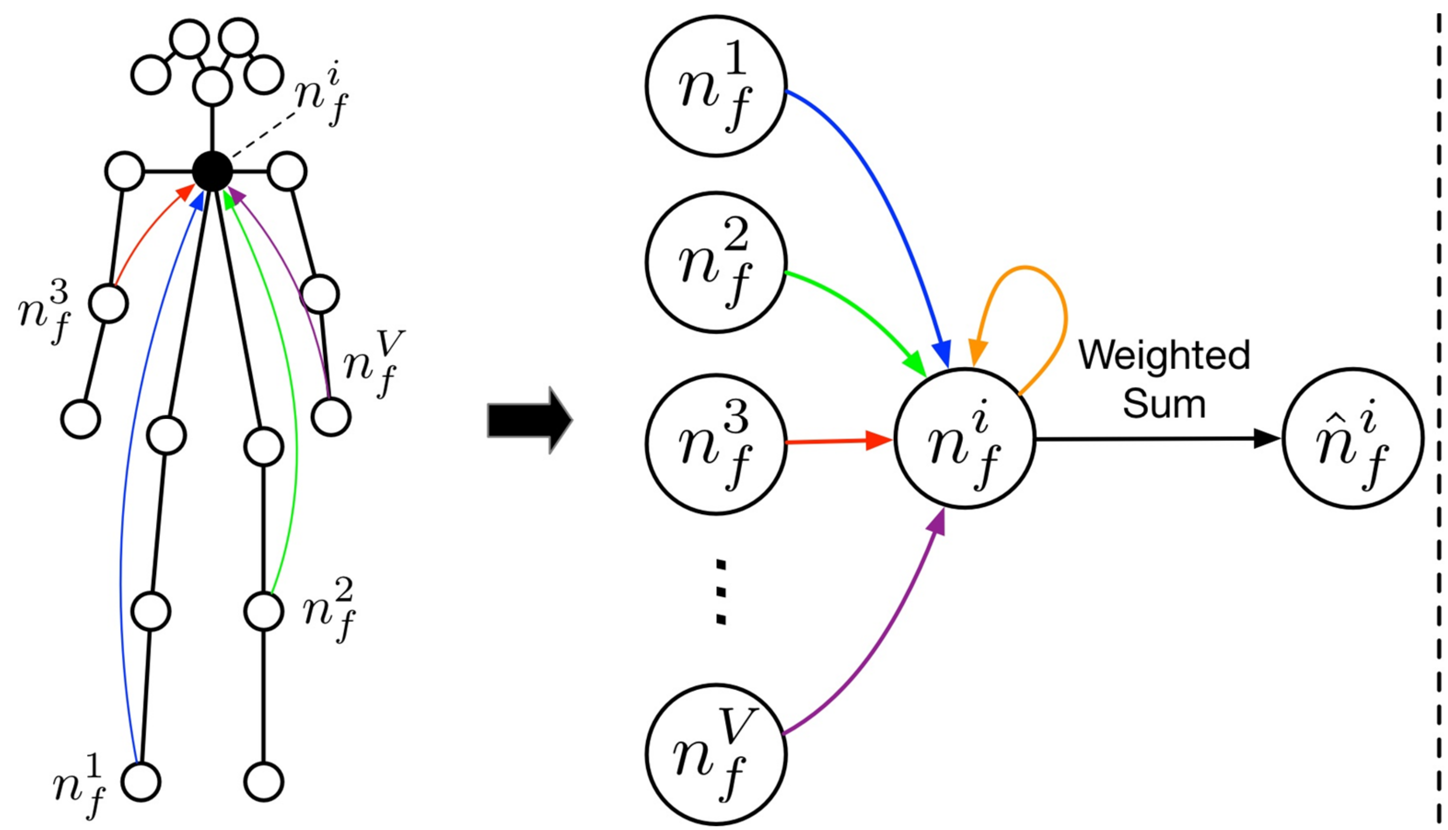
Music-Guided Dance Video Synthesis

DanceGAN

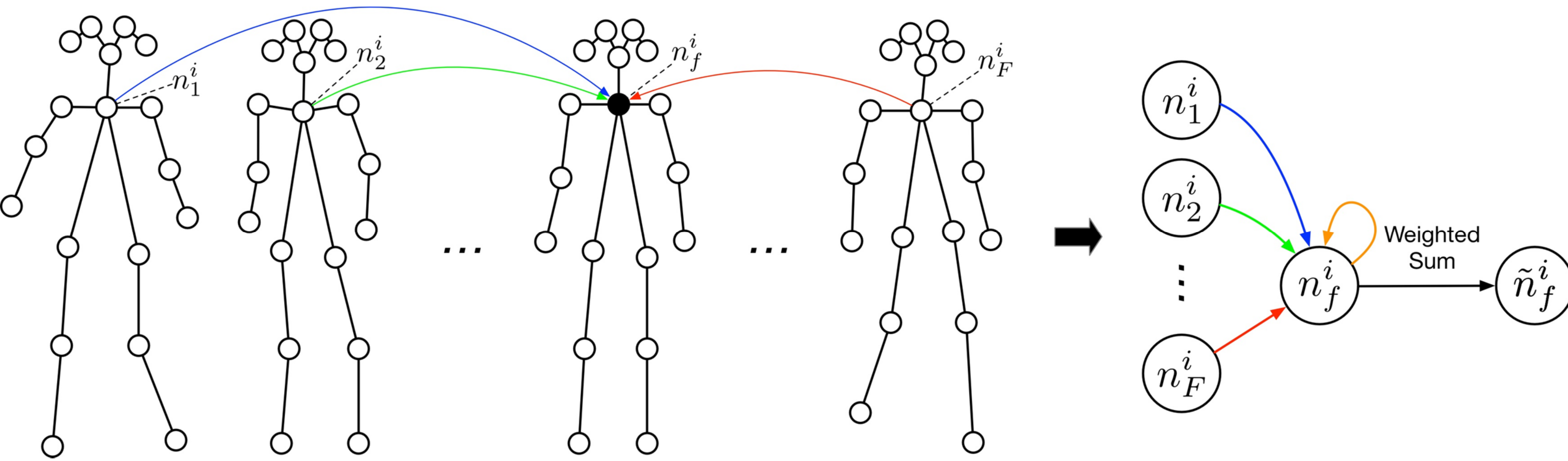


M
H
G_e
Music Feature Extraction Network
Self-Frame Spatial Graph Attention Network
G_s
G_t
Cross-Frame Temporal Graph Attention Network
Sum Fusion
Stage I: Music-to-Skeleton Translation
S_1
S_2
S_f
S_F
Stage II: Skeleton-to-Dance Translation
G_r
I_0
Forward and Backward Self-Supervised Regularization Network
I_1
I_2
I_f
I_F

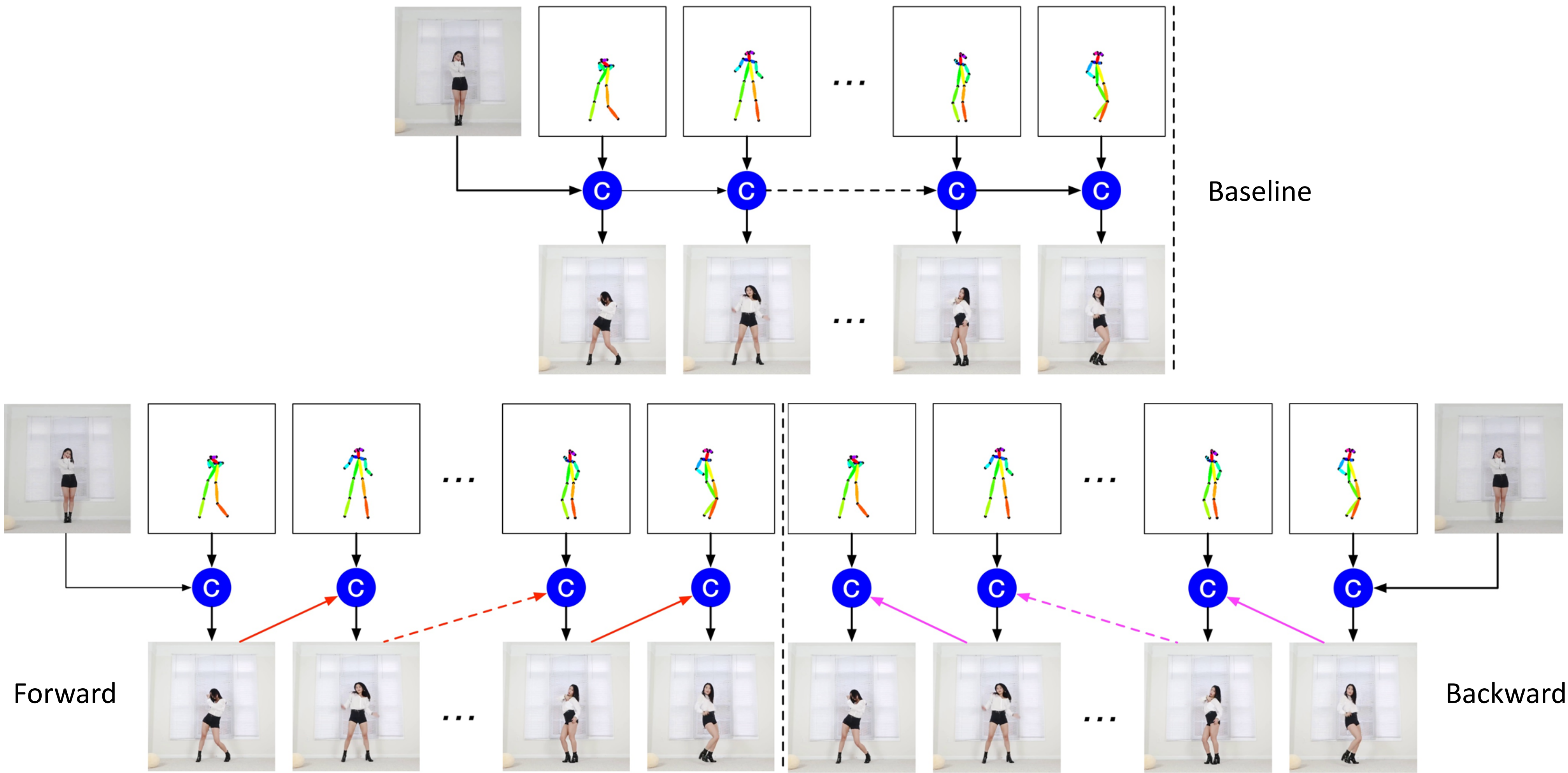
Self-Frame Spatial Graph Attention Network



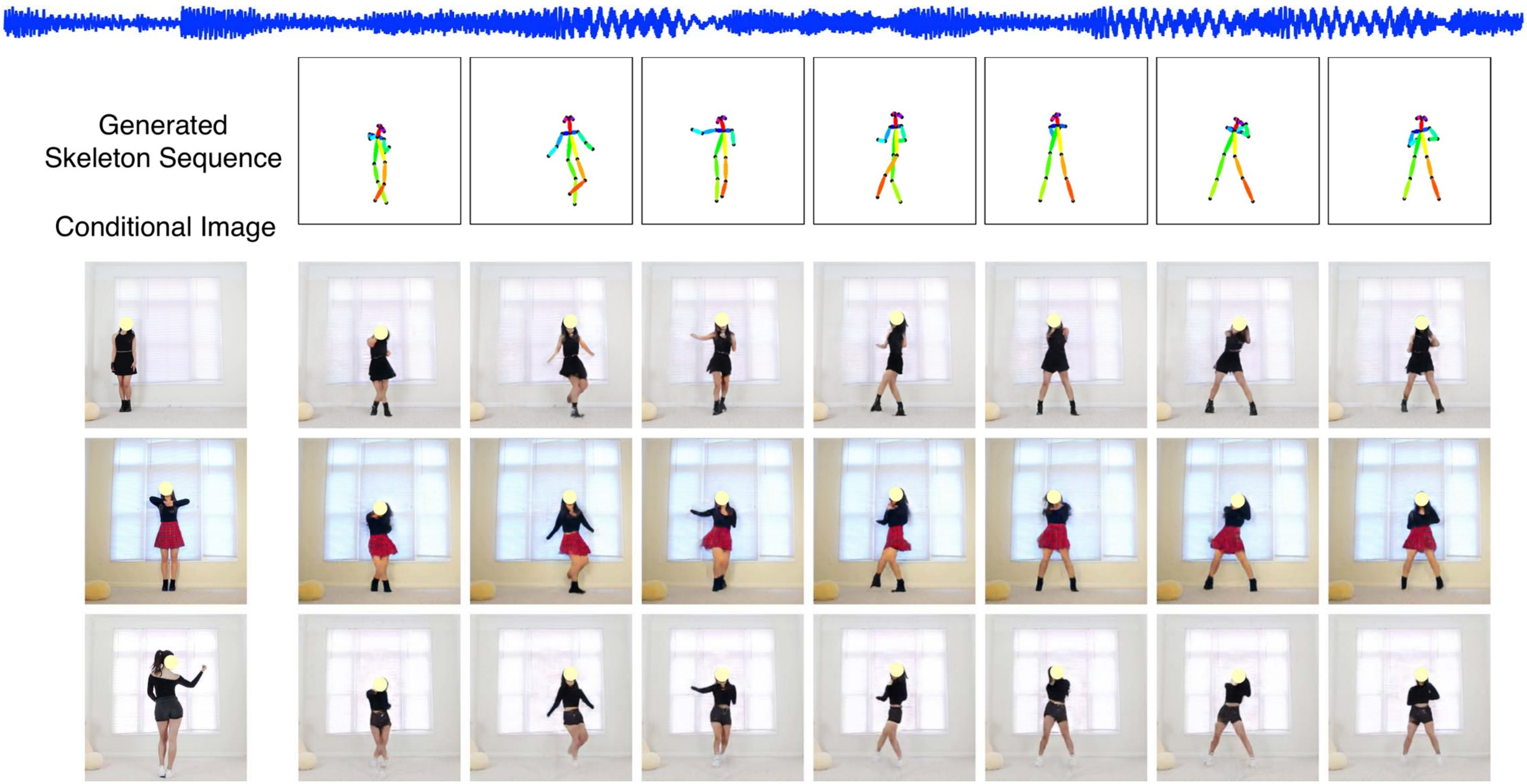
Cross-Frame Temporal Graph Attention Network



Self-Supervised Regularization Network



Music-Guided Dance Video Synthesis



Can't Stop Dancing: Music-Guided Dance Video Synthesis

Paper ID 3316

Limitations and extensions

- Issues with 3D movements => incorporate the modeling of 3D keypoints or other 3D information
- So far we are animating single objects => animate multiple objects and consider also the interactions/constraints between them, e.g, people interactions, complex surveillance scenes, etc.
- Interactive video generation
- Video2video translation => repurpose video generation to different domains, e.g., Comics2Video and Video2Comics
- Possible ethical issues => deep fake forensics

The International AI Doctoral Academy (AIDA) is a joint initiative of the European R&D Projects: AI4media, ELISE, HumanE-AI Net, TAILOR, VISION

www.ai4media.eu www.elise-ai.eu www.humane-ai.eu tailor-network.eu www.vision4eu.eu

These projects have received funding from the European Union's Horizon 2020 research and innovation programme under the following Grant agreements: No 951911 (AI4media), No. 952070 (VISION) No. 952026 (HumanE-AI Net) and No 951847 (ELISE)