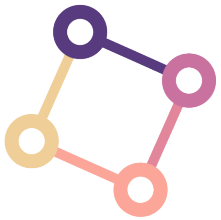
UNSUPERVISED KEYPOINT LEARNING FOR GUIDING CLASS-CONDITIONAL VIDEO PREDICTION

Yunji Kim. et al., NeuIPS, 2019

VISION STUDY 2020/03/19

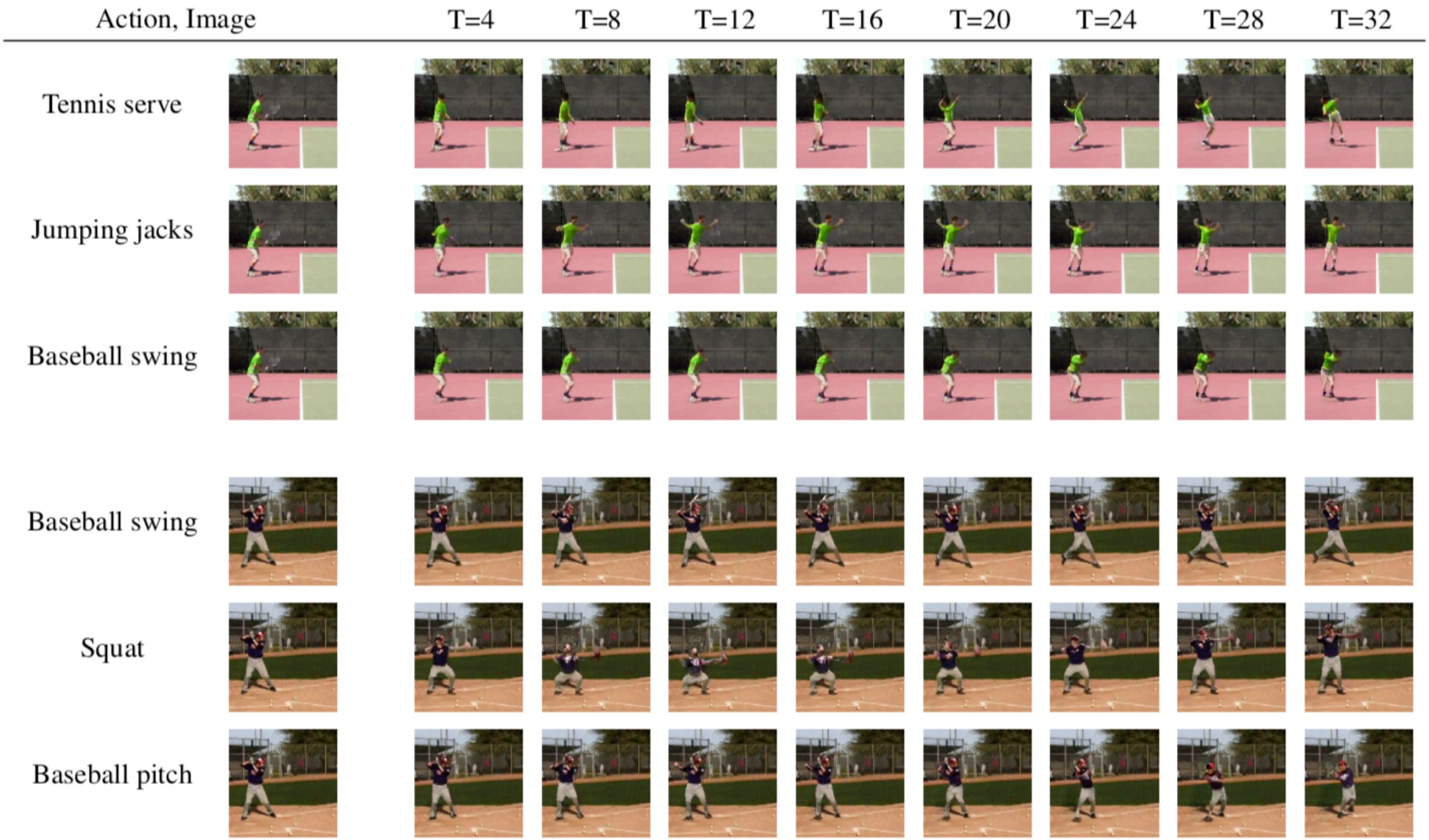


DAVIAN

Data and Visual Analytics Lab

Overview

- **The outputs of paper**
- **Tackling points of paper**
- **Methods**
- **Experiments**
- **Failure cases**


Action, Image
T=4
T=8
T=12
T=16
T=20
T=24
T=28
T=32
Tennis serve
Jumping jacks
Baseball swing
Baseball swing
Squat
Baseball pitch

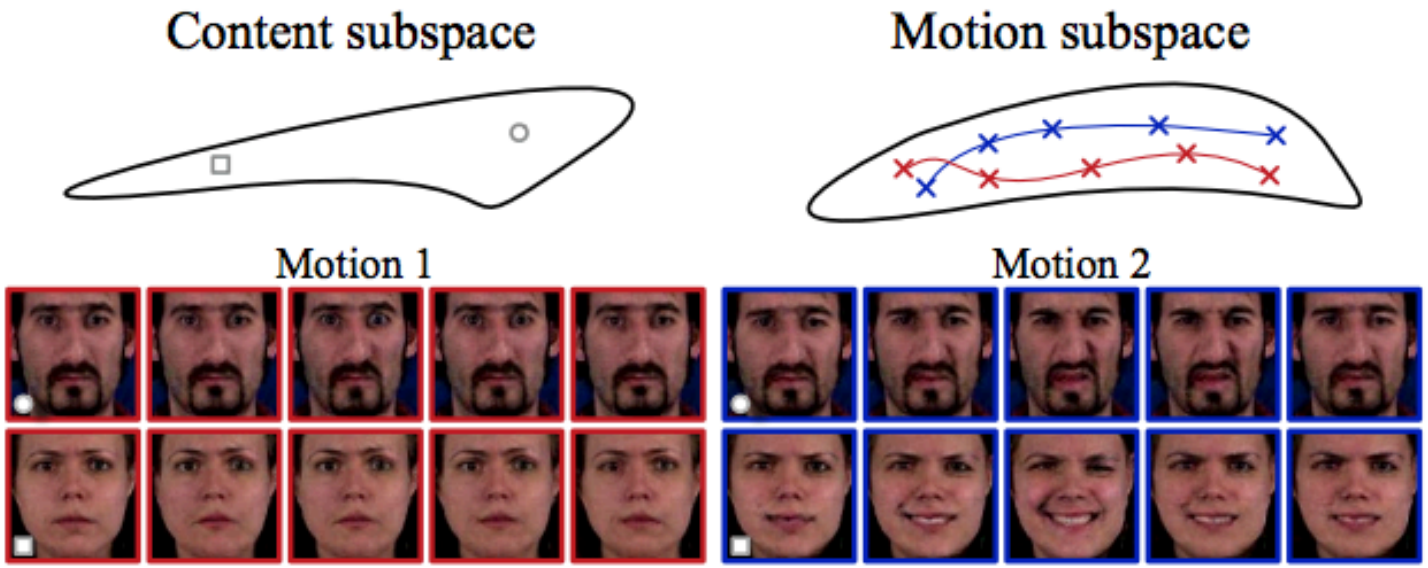
Tackling Points of Paper

GOAL: Video prediction is a task of synthesizing future video frames from a single of few images

NATURE OF VIDEO PREDICTION

Bearing uncertainty about ***future*** frame (Future is not deterministic; e.g. 1, 2, 4 (Inputs) => 7 vs 8 (Prediction))
=> Want to learn "Distribution" of future frame. => GAN and VAE are introduced.

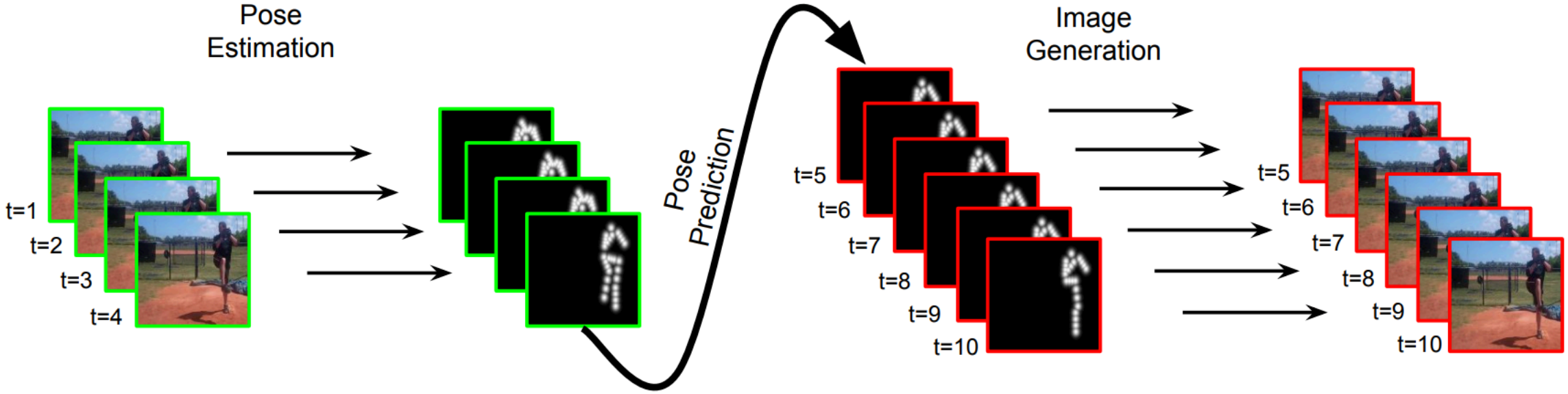
Learning spatial domain (content) and temporal domain (motion)
=> Content is still, but motion is dynamic => **Disentangling** these 2 components



S. Tulyakov, MoCoGAN: Decomposing Motion and Content for Video Generation, CVPR, 2018

Tackling Points of Paper

Since landmarks (i.e. keypoints) are **constrained representation of object's motion**, **They are often leveraged to represent the dynamics of video and utilized to generate video frames *for long-term prediction (?)*.**



R Villegas et al., Learning to Generate Long-term Future via Hierarchical Prediction, ICML, 2018

WEAKNESS : It requires frame-by-frame keypoints labeling <= Major tackling point

Tackling Points of Paper

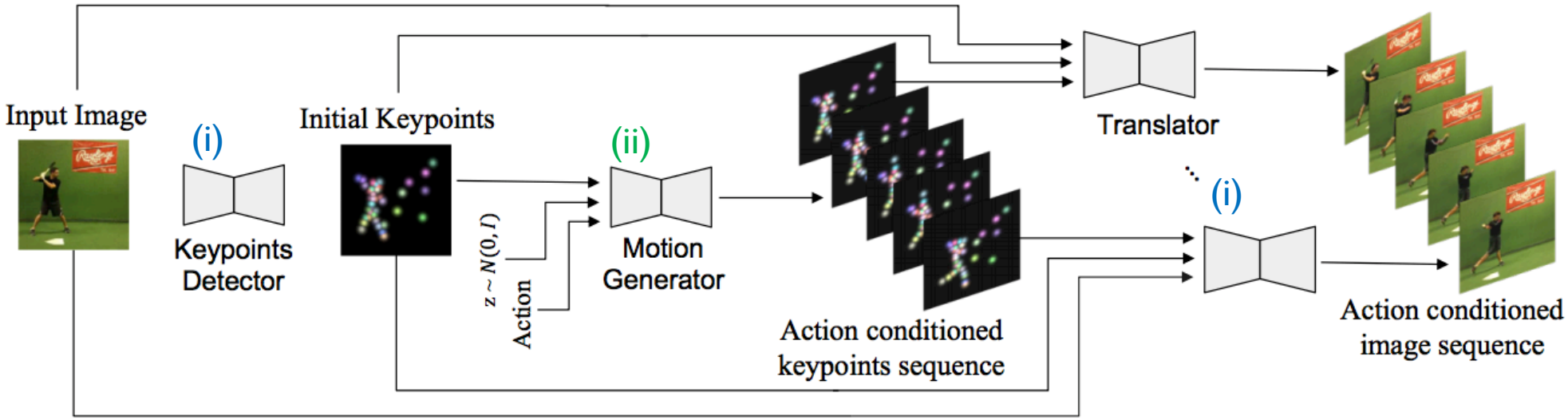
Contributions

- Deep generative method for **class-conditional video prediction from a single image**. Proposed method **internally generates keypoints** of the foreground object to guide the synthesis of future motion

- Proposed method **learns to generate a variety of keypoints sequences from data without labels**, which enables our method to model the motion of arbitrary objects including human, animal, and etc

- Proposed method is robust to the noise of data such as distracting backgrounds, allowing our method to work robustly on challenging datasets

Methods: Overview



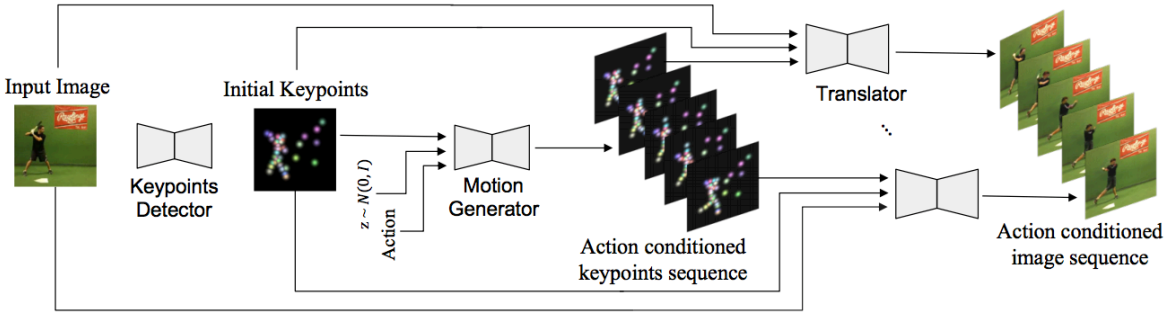
[INPUT] A source image $v_0 \in R^{H*W*3}$ and a target action vector $a \in R^C$
[OUTPUT] Future frames $\hat{v}_{1:T} \in R^{T*H*W*3}$

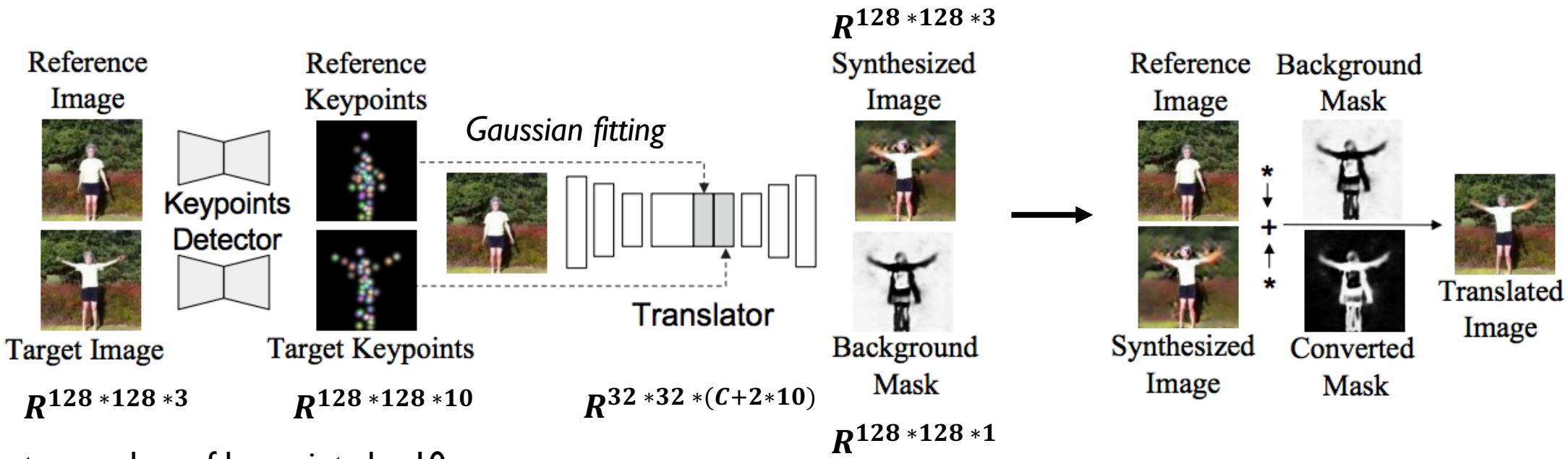
Training Procedure

(i) Learning the keypoints detector with the image translator

(ii) Learning the motion generator with pseudo-labeled data

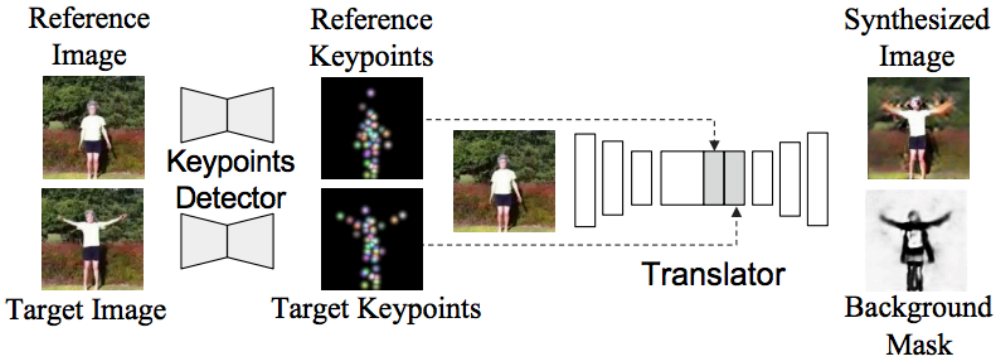
Methods: Learning Detector/Translator





- Let a number of keypoints be 10
- The paper insists that through mask, their method is robust to the distracting background.
- Training Losses
  - $L_{D_{im}} = -logD_{im}(v') - \log(1 - D_{im}(\hat{v}))$, GAN Loss
  - $L_{D_{im}} = -\log\big(D_{im}(\hat{v})\big) + \lambda_1 E_l || \ \Phi_l(\hat{v}) - \ \Phi_l(v') \ ||$ } GAN + Perceptual Loss

Methods: Learning Detector/Translator



(Differentiable) Gaussian fitting 1D Example

Coordinate 1 2 3 4 5

Softmax 0.01 0.08 0.1 0.8 0.01

weighted avg ⇒ 3.72 (treat it as a mean)

extracted coordinate

Expected dist. with argmax

⇒ 7.4 2.9 0.52 0.08 1.6

(Coordinate − mean)²

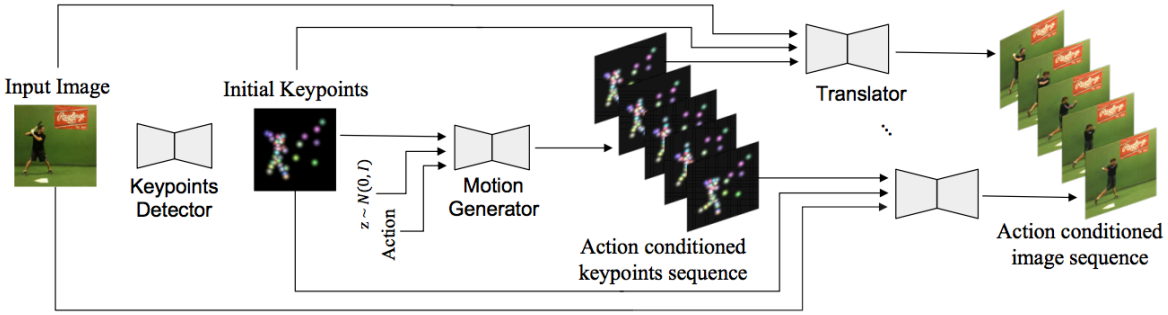
⇒ 0.0006 0.05 0.59 0.92 0.2

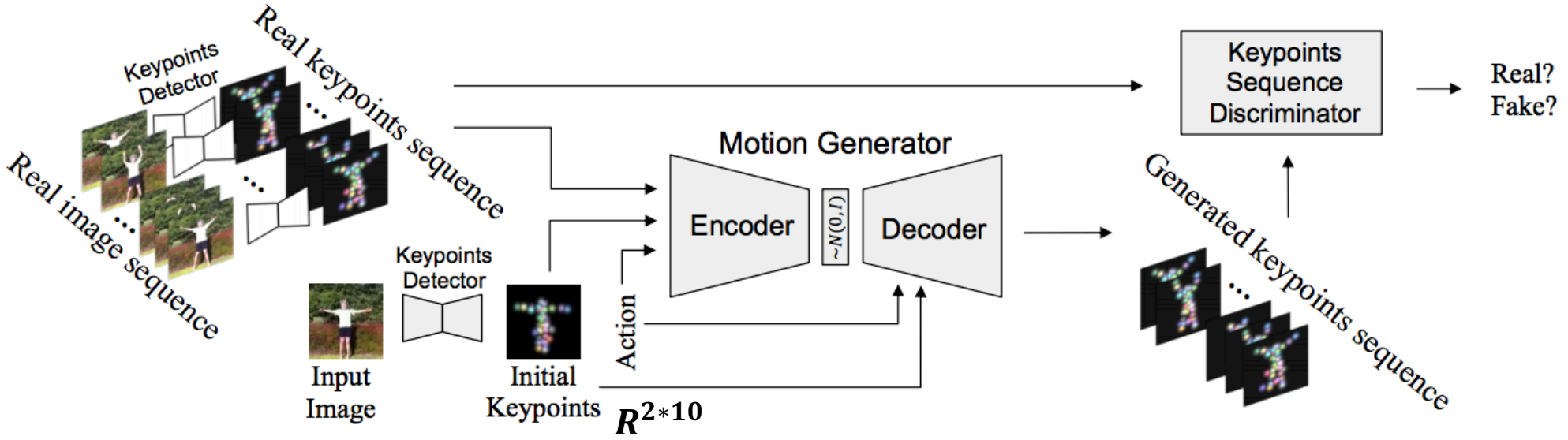
gaussian kernel with fixed sigma = 1/√2

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$

output

Methods: Learning Motion Generator





- Training with pseudo-label generated from pre-trained keypoints detector
- cVAE architecture / LSTM encoder, decoder / Sequence discriminator
- Training Losses
  - $L_{D_{seq}} = -logD_{seq}(\hat{k}_{1:T}) - \log(1 - D_{seq}(\tilde{k}_{1:T}))$, GAN Loss
  - $L_{D_{im}} = D_{KL}(q_\phi(z|\hat{k}_{1:T}; \hat{k}_0; a)) - \lambda_2||\tilde{k}_{1:T} - \hat{k}_{1:T}|| - \lambda_3 \log\left(D_{seq}(\tilde{k}_{1:T})\right)$ KL + Recon + GAN Loss

Experiment: Dataset

Penn Action

Baseball pitch

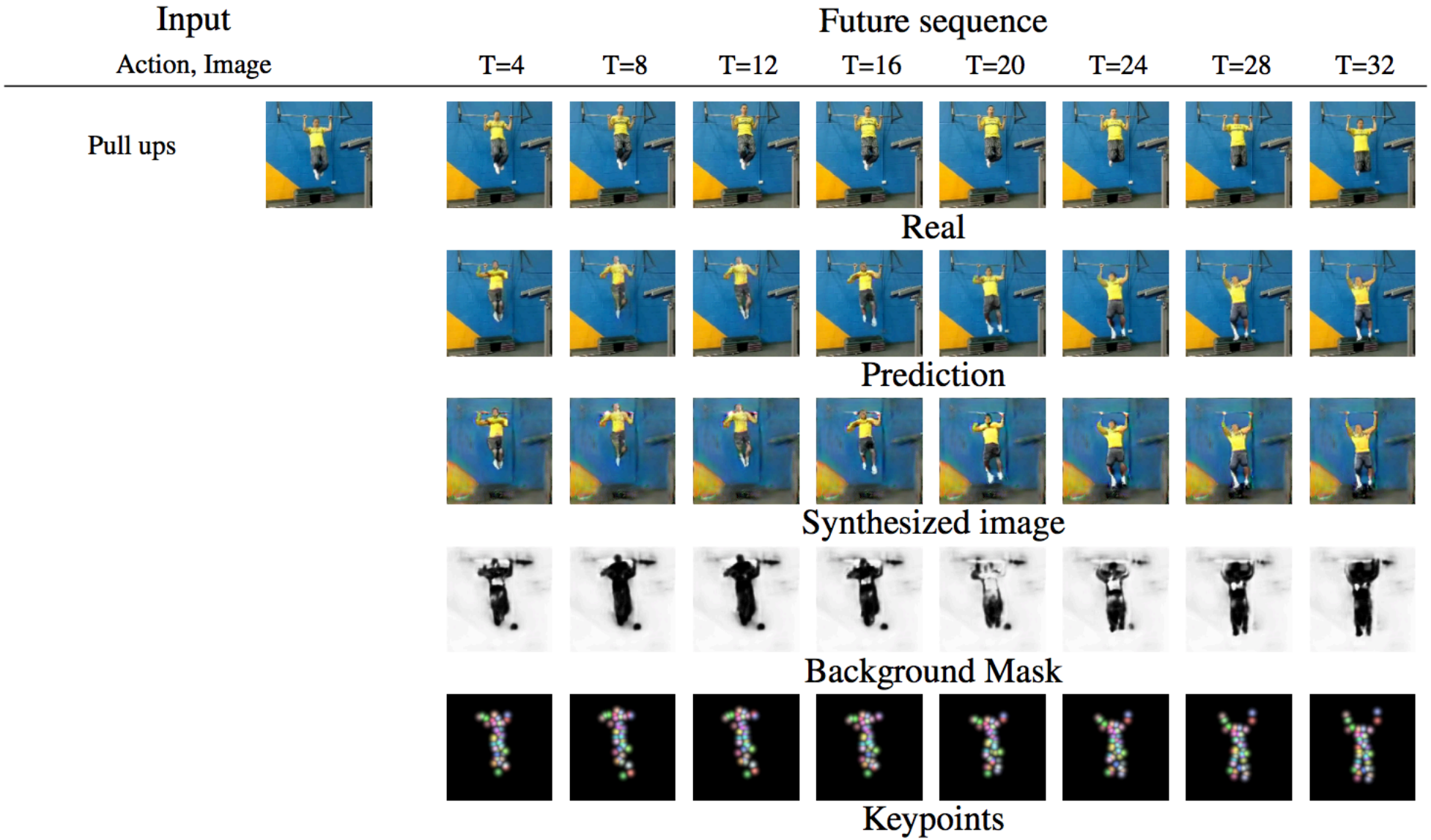
UvA-NEMO

Real

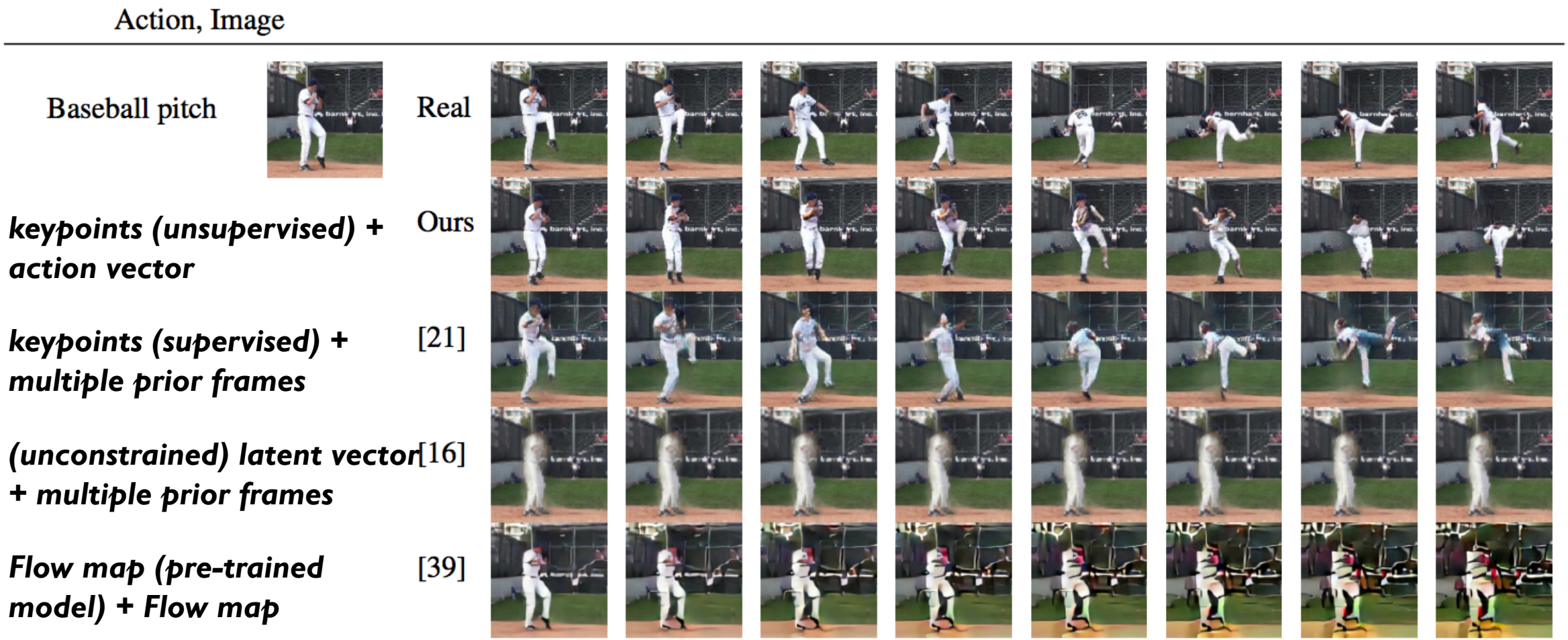
MGIF

Real

Experiment: Inference result

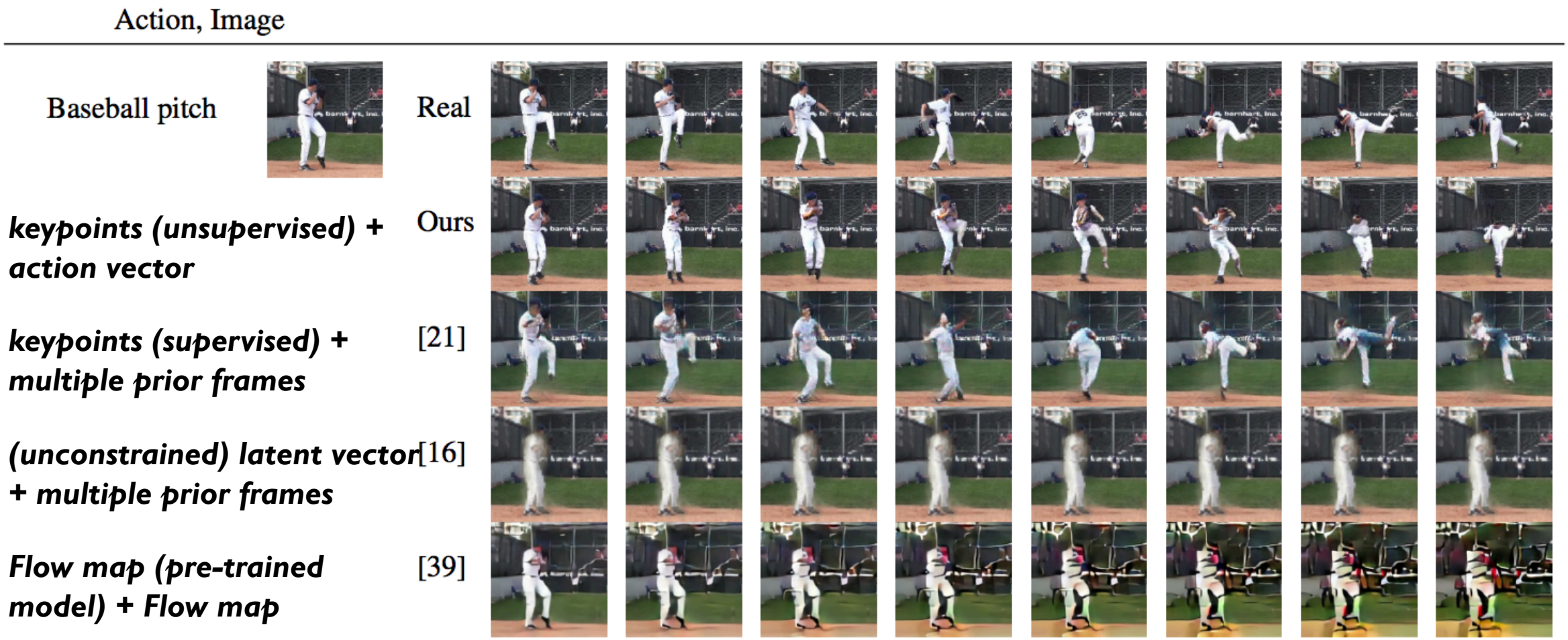


Experiment: Baselines comparison



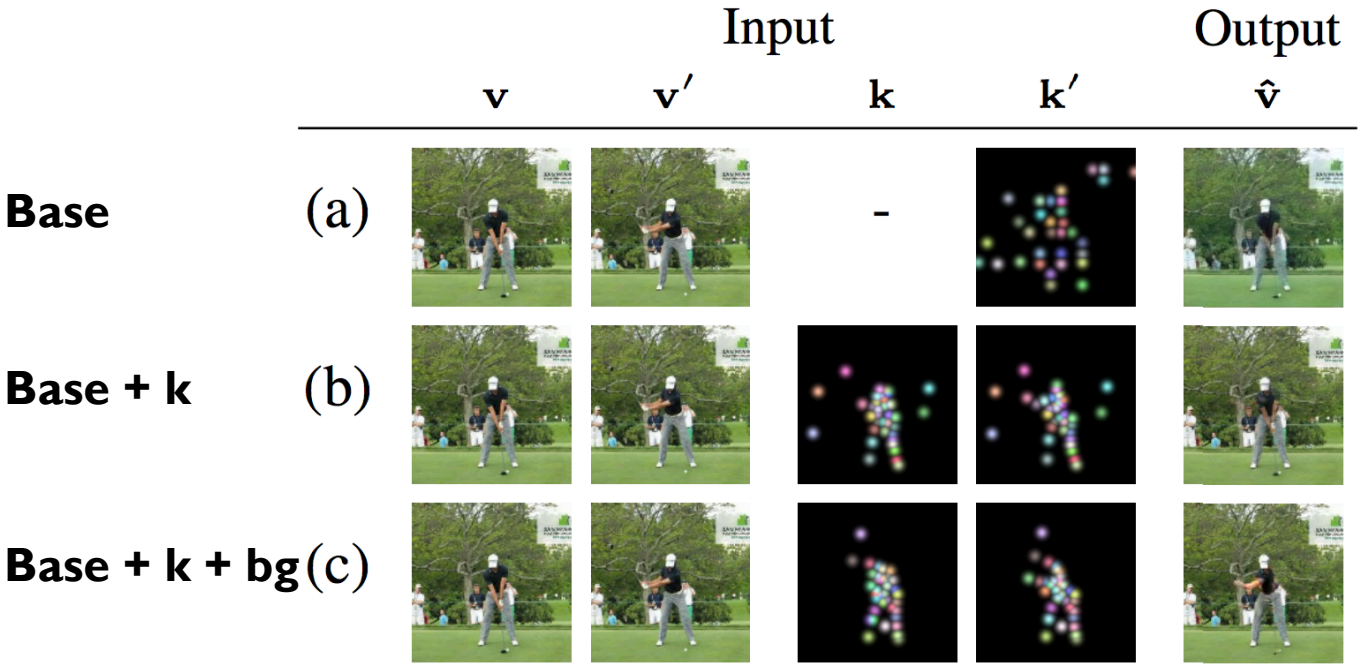
* motion feature form + cues for motion

Experiment: Baselines comparison



* motion feature form + cues for motion

Experiment: Components of Keypoints Detector



(+) Learning with keypoints information of original image

- With keypoints in both images, the translator can synthesize the foreground object in the target pose by **inferring the analogical relationship** like *"A is to B as C is to what? "; v + (k – k') = v_hat*
- If not, the translator would have to find the region to translate independently which is redundant and inefficient set-up
- *Question? Just simple concatenation is presented here.*

(+) Generating a mask and utilizing it to synthesize image

- The mask generation is effective when **only a specific part of the image needs to be translated**
- **Synthesizing only the foreground object** is beneficial for the network to fool the image discriminator by **reducing the complexity of the modeling compared to synthesizing the entire scene**

Experiment: Qualitative results

- **User study**
- **FVD measurement**

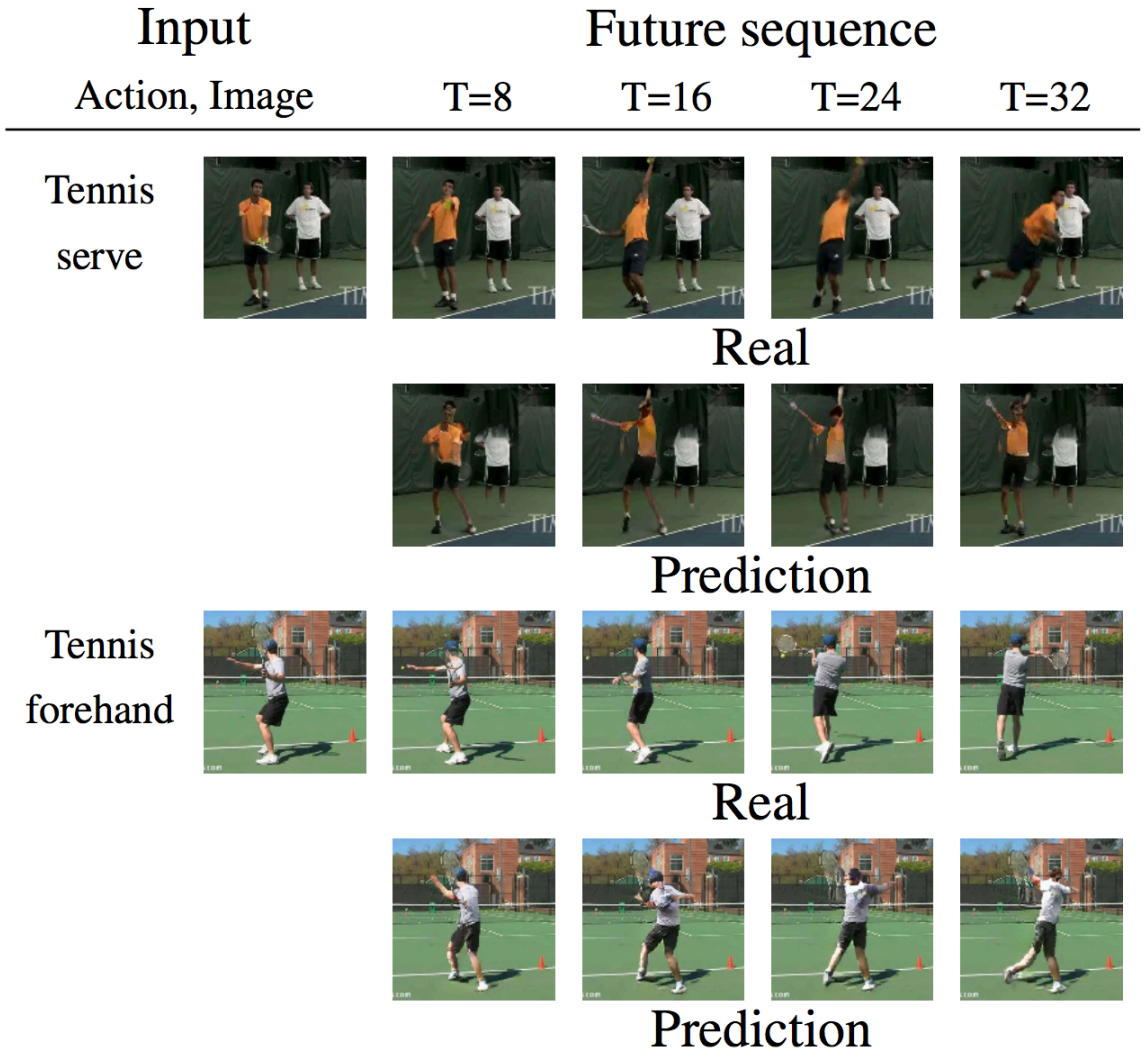
Method	Ranking
Ours	**1.81** $\pm$ 1.02
[21]	2.44 $\pm$ 0.98
[16]	3.14 $\pm$ 1.09
[39]	2.61 $\pm$ 0.96

Table 3: Quantitative result of the user study. The values refer to average rankings.

Dataset	[39]	[16]	[21]	Ours
Penn Action [25]	4083.3	3324.9	2187.5	**1509.0**
UvA-NEMO [38]	666.9	265.2	-	**162.4**
MGIF [31]	683.1	1079.6	-	**409.1**

Table 1: Fréchet Video Distance (FVD) [40] of generated videos. On every datasets, our method achieved the best score. (The lower is better.)

Experiment: Failure Cases



- Multiple object exist.
- Object moves in the opposite direction