Unsupervised Keypoint Learning for Guiding Class-Conditional Video Prediction

Yunji Kim[1] Seonghyeon Nam[1] In Cho[1] Seon Joo Kim[1, 2]
[1]Yonsei University [2]Facebook

Our project page

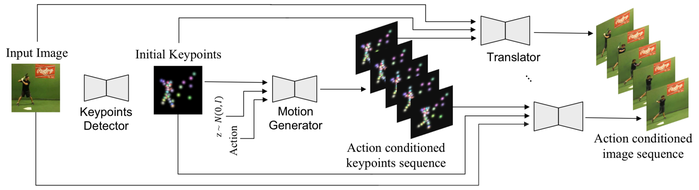
I. GOALS

- Generate future video given a single image and target action class.
- Learn motion with keypoints sequences.
- Train model without demanding for any human-annotated labels.

II. METHODS

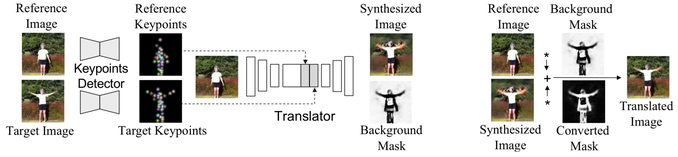
Generation process

(i) Detect keypoints of the input image.
(ii) Generate keypoints motion following the target action class.
(iii) Generate image sequence following the keypoints motion.



Training stage 1

Keypoints Detector $(Q) * $ Translator $(T) * $ Image Discriminator (D_{im})



- Randomly sample two images from video.
- Learn to detect keypoints that guides translating of one image to another.

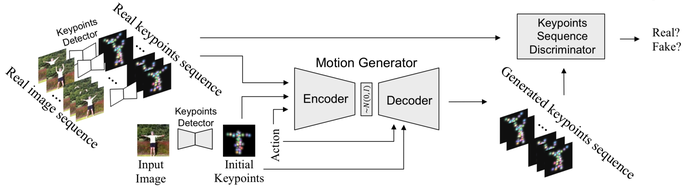
OPTIMIZE adversarial loss & perceptual loss

$$L_{D_{im}} = -\log D_{im}(v') - \log(1 - D_{im}(\hat{v}))$$
$$L_{Q,T} = -\log D_{im}(\hat{v}) + \lambda \mathbb{E}_l \|\Phi_l(v') - \Phi_l(\hat{v})\|$$

v': target image, $\hat{v}$: translated image, Φ_l: l^{th} layer of the pretrained VGGNet

Training stage 2

Motion Generator $(M) * $ Keypoints sequence Discriminator (D_{seq})



- Make pseudo keypoints labels using trained keypoints detector.
- With cVAE as a framework, learn to generate many possible keypoints motions.

OPTIMIZE adversarial loss & variational lower bound of the cVAE

$$L_{D_{seq}} = -\log D_{seq}(\hat{k}_{1:T}) - \log(1 - D_{seq}(\tilde{k}_{1:T}))$$
$$L_M = D_{KL}(q_\phi(z|\hat{k}_{1:T}, \hat{k}_0, a) \parallel p_z(z)) + \lambda_1 \|\hat{k}_{1:T} - \tilde{k}_{1:T}\|_1 - \lambda_2 \log D_{seq}(\tilde{k}_{1:T})$$

$\hat{k}_{0:T}$: pseudo keypoints label, $\tilde{k}_{1:T}$: generated keypoints sequence, z: latent variables, a: target action class

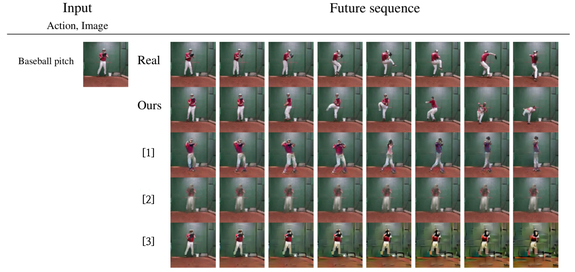
III. RESULTS

- We experimented with Penn Action, UvA-Nemo and MGIF datasets.
- All baselines generate video in two stages.

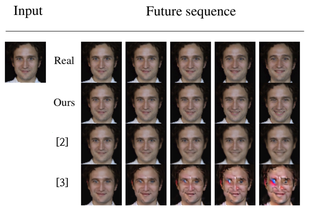
[1] "Learning to generate long-term future via hierarchical prediction", Villegas et. al., *ICML*, 2017.
[2] "Hierarchical long-term video prediction without supervision", Wichers et. al., *ICML*, 2018.
[3] "Flow-grounded spatial-temporal video prediction from still images", Li et. al., *ECCV*, 2018.

Qualitative and quantitative comparison of the results

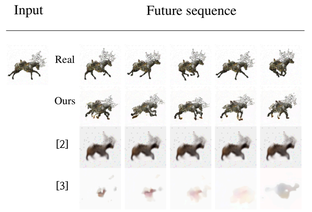
Penn Action



UvA-NEMO



MGIF



Quantitative Results

		ours	[1]	[2]	[3]
User study (Averaged ranking)		**1.81**	2.44	3.14	2.61
Action Recognition (Accuracy)		**68.89**	47.14	40.00	15.55
Fréchet Video Distance (The lower, the better.)	Penn Action	**1509.0**	2187.5	3324.9	4083.3
	UvA-NEMO	**162.4**	-	265.2	666.9
	MGIF	**409.1**	-	1079.6	683.1

Various results generated from identical initial image