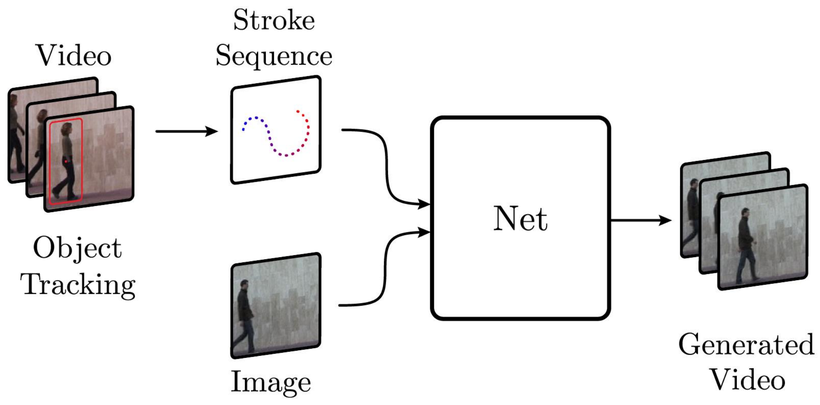
Learning to Take Directions One Step at a Time

Qiyang Hu[1], Adrian Wälchli[1], Tiziano Portenier[1], Matthias Zwicker[2] and Paolo Favaro[1]
[1]University of Bern, Switzerland, [2]University of Maryland, USA

[1]{qiyang.hu, adrian.waelchli, tiziano.portenier, paolo.favaro}@inf.unibe.ch
[2]zwicker@cs.umd.edu

Video From a Motion Stroke Sequence



Goal

- Generate video conditioned on single image
- Give user control over generated trajectory of object

Challenges

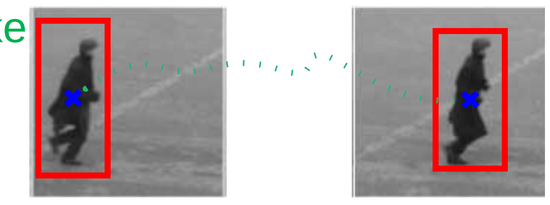
- Generating future without a “memory” of the past
- A single image contains much ambiguity for motion
- Temporally coherent sequence

Contribution

We address these challenges with a novel system that recurrently generates a video of arbitrary length from a single image and a sequence of strokes that guides the motion.

Motion Stroke Sequence Creation

Training: The motion stroke is a sequence of 2D points extracted from the bounding box center.



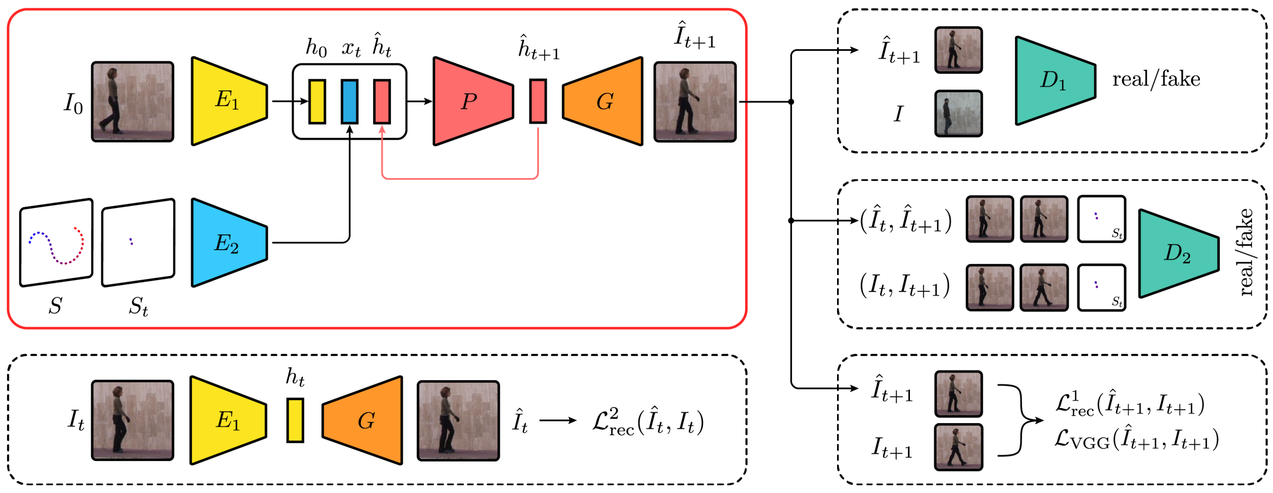
Test: Any stroke, but must start at center of object

Smoothness of Generated Motion

	walking		*jogging*		*running*		det.
	mean	std	mean	std	mean	std	
Denton *et al.* [1]	7.5	10.0	9.9	11.9	10.7	11.5	54.2
Li *et al.* [2]	7.4	9.1	10.1	11.3	**8.7**	**9.9**	54.9
Ours	**7.2**	**7.7**	**8.2**	**9.1**	9.2	10.5	**87.1**
Ground truth	4.3	5.7	5.3	5.8	7.4	6.8	100.0

We measure the smoothness of the generated motion as the rel. mean Euclidean distance (%) and std. of pose joints in consecutive frames for comparison.

Proposed Framework



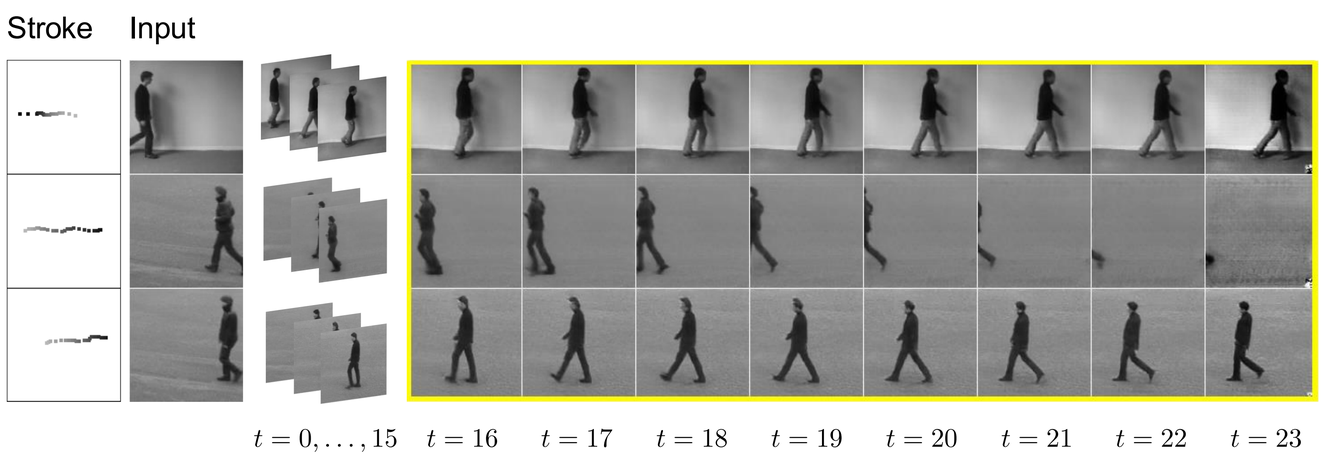
- Encoder E_1 extracts texture and initial conditions
- Encoder E_2 provides motion encoding from stroke
- Predictor P is applied recursively on features
- G generates temporally consistent image sequence
- D_1 and D_2 force distribution of pairs and single frames

Training

$$\min_{\theta_1} \max_{\theta_2} \mathcal{L}^1_{\text{GAN}} + \lambda_0 \mathcal{L}^2_{\text{GAN}} + \lambda_1 \mathcal{L}_{\text{VGG}} + \lambda_2 \mathcal{L}^1_{\text{rec}} + \lambda_3 \mathcal{L}^2_{\text{rec}}$$

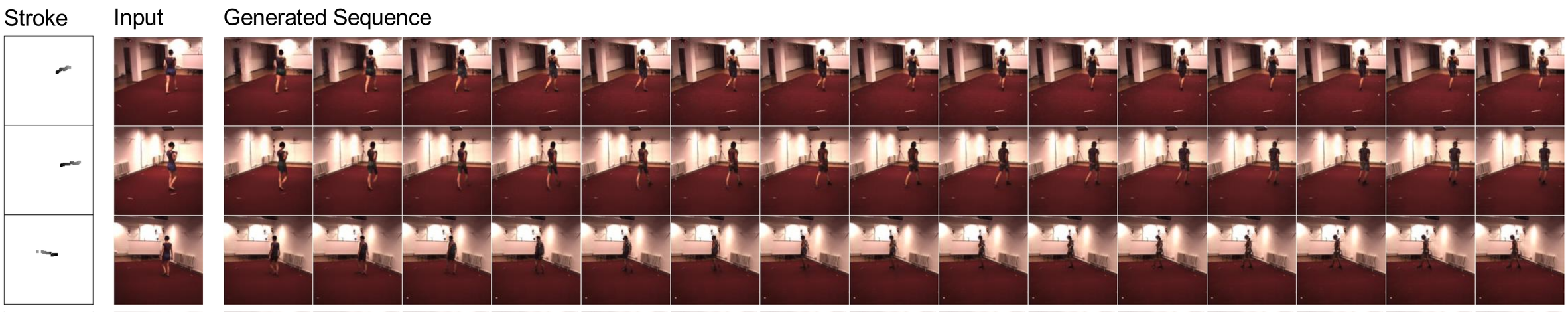
- Adversarial training of generator $\theta_1 = \{\theta_{E_1}, \theta_{E_2}, \theta_P, \theta_G\}$ with discriminator $\theta_2 = \{\theta_{D_1}, \theta_{D_2}\}$
- Reconstruction loss on full frame and center-crop
- Perceptual loss using VGG features

Long Sequences



Our method is able to generate realistic sequences (yellow) beyond the training regime with 16 frames (middle).

Qualitative Results on Human3.6M



References
[1] E. Denton and R. Fergus. Stochastic video generation with a learned prior. arXiv preprint arXiv:1802.07687, 2018.
[2] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Flow-grounded spatial-temporal video prediction from still images. In ECCV, 2018.

PAPER/CODE