Motion Transformer for Unsupervised Image Animation

Jiale Tao[1] Biao Wang[2] Tiezheng Ge[2] Yuning Jiang[2] Wen Li[1] Lixin Duan[1]
[1]University of Electronic Science and Technology of China [2]Alibaba Group

ECCV TEL AVIV 2022

Introduction

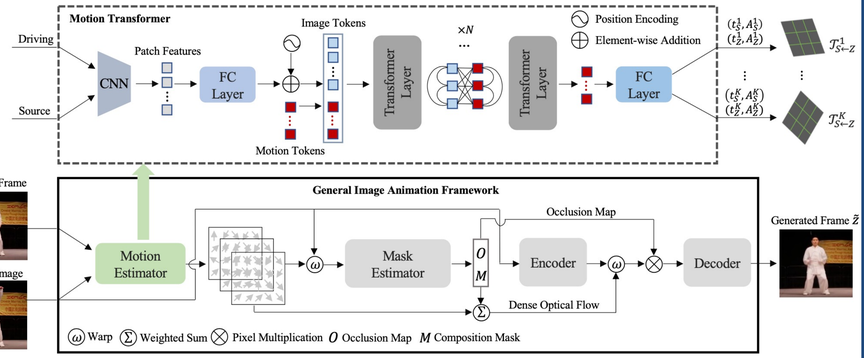
1. Image Animation

- **Input:** a source image and a driving video; **Output:** a video with the motion from the driving video and the appearance from the source image.
- **General animation:** learn part motion representations (keypoints and affine transformations) via self-supervised image reconstruction.

2. Motivation

- **Current motion estimators:** built by CNNs, do not explicitly model the interactions/relationship between motions, which can potentially lead to noticeable artifacts being produced in the generated animation video.

Methodology



$$P_l^i = \sum_j \mathrm{MSA}(Q_{P_{l-1}^i}, K_{P_{l-1}^j}, V_{P_{l-1}^j}) + \sum_j \mathrm{MSA}(Q_{P_{l-1}^i}, K_{I_{l-1}^j}, V_{I_{l-1}^j})$$

- Motions are represented as learnable tokens (**Motion Tokens**).
- Image is divided into patches with position encoding (**Image Tokens**).
- Motion tokens are decoded to final motion information (affine matrix) via **motion2image cross attention.**
- The relationship between motions are modeled by **motion2motion self attention.**
- The two types of attention are unified in a single transformer architecture.

Experiments

1. Quantitative and Qualitative Comparison with Existing Methods

Quantitative comparisons on the video self-reconstruction task.

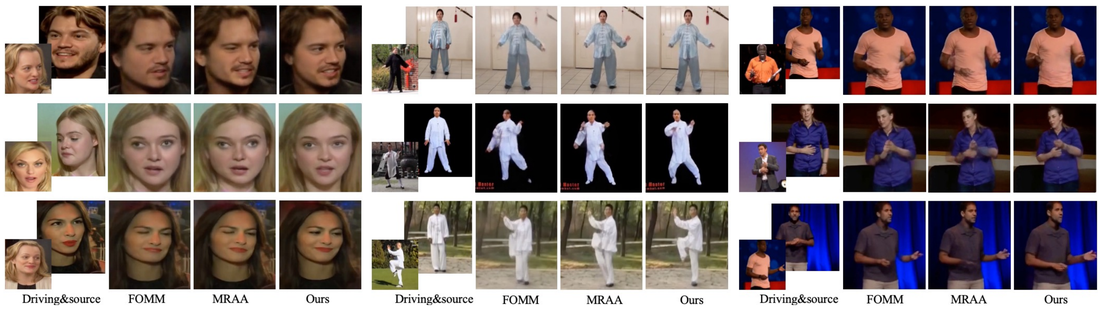
	TaiChiHD			TEDTalks			VoxCeleb			MGIF
	L1	(AKD, MKR)	AED	L1	(AKD, MKR)	AED	L1	AKD	AED	L1
FOMM	0.057	(6.649, 0.036)	0.172	0.029	(4.382, 0.008)	0.127	0.041	1.29	0.133	0.0224
MRAA	0.048	(5.246, 0.024)	0.150	0.027	(3.955, **0.007**)	0.118	0.040	1.28	0.133	0.0274
Ours	**0.045**	(**4.670**, **0.021**)	**0.148**	**0.026**	(**3.456**, **0.007**)	**0.113**	**0.038**	**1.18**	**0.116**	**0.0200**

Model capacity.

	Parameters	FLOPs
ImageGenerator	45.57M	53.64G
MotionEstimator-FOMM	14.21M	1.28G
MotionEstimator-MRAA	14.20M	1.26G
MotionEstimator-Ours	12.23M	7.54G

User preferences.

	TaiChiHD	TEDTalks	VoxCeleb
FOMM	96.5%	66.4%	60.8%
MRAA	68.5%	57.1%	69.8%



Qualitative comparison on cross-identity image animation

2. Ablation study

On position encoding.

	L1	(AKD, MKR)	AED
w/o PE	0.047	(5.482, 0.028)	0.158
w PE	**0.045**	(**4.670**, **0.021**)	**0.148**

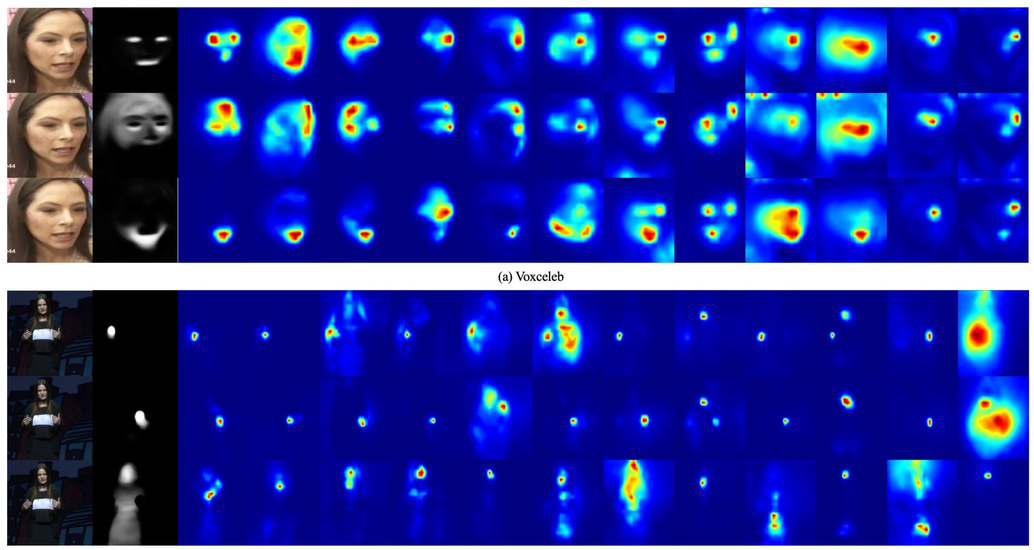
On CNN backbone.

CNN	Param.	L1	(AKD, MKR)	AED
Stem	5.56M	0.048	(6.056, 0.030)	0.161
HR-W32	12.23M	**0.045**	(**4.670**, 0.021)	**0.148**
HR-W48	21.30M	**0.045**	(4.829, **0.020**)	0.149

On transformer layers.

Layers	L1	(AKD, MKR)	AED
4	0.046	(5.320, 0.027)	0.155
8	0.046	(5.226, 0.025)	0.154
12	**0.045**	(**4.670**, **0.021**)	**0.148**

3. Visualizations



(a) Voxceleb

(b) TEDTalks

Visual attention visualization.