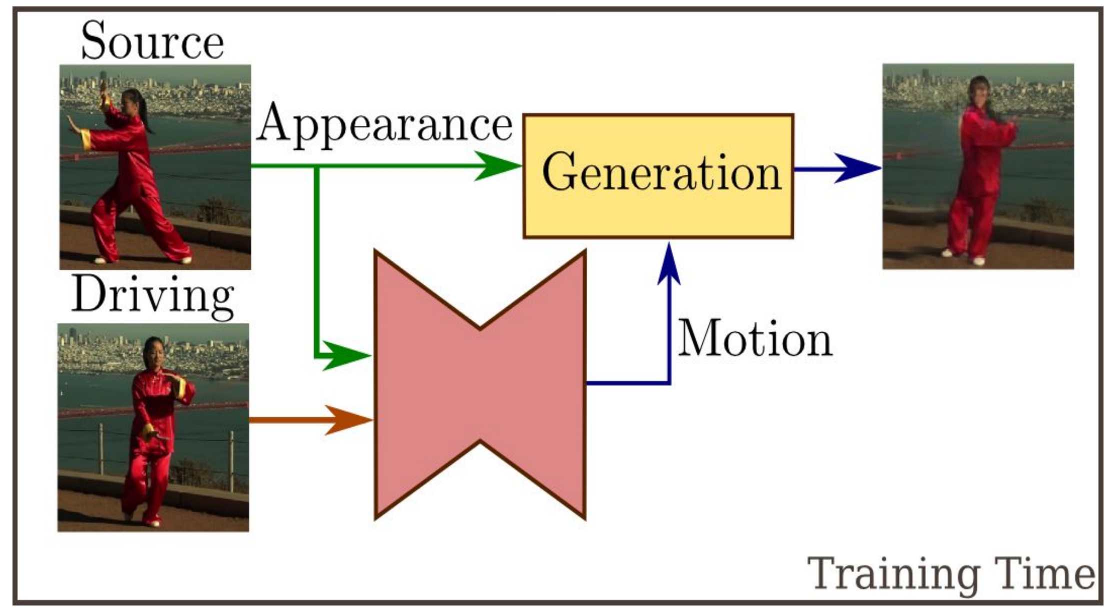
First Order Motion Model for Image Animation

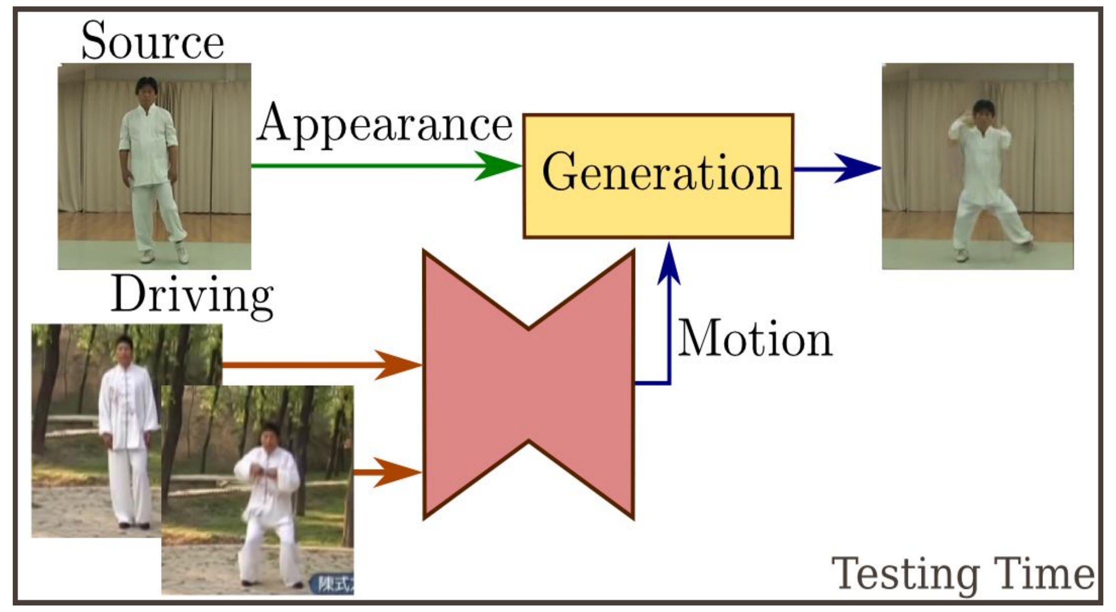
Aliaksandr Siarohin[1], Stephane Lathuiliere[1,4], Sergey Tulyakov[2], Elisa Ricci[1,3] and Nicu Sebe[1]

[1]DISI, University of Trento; [2]Snap Inc, [3]Fondazione Bruno Kessler, [4]LTCI, Institut polytechnique de Paris

Self-Supervised Image Animation



- Training time: we learn a self-supervised motion representation, using image reconstruction objective
- Testing time: we extract motion from driving video and appearance from source



Proposed Method

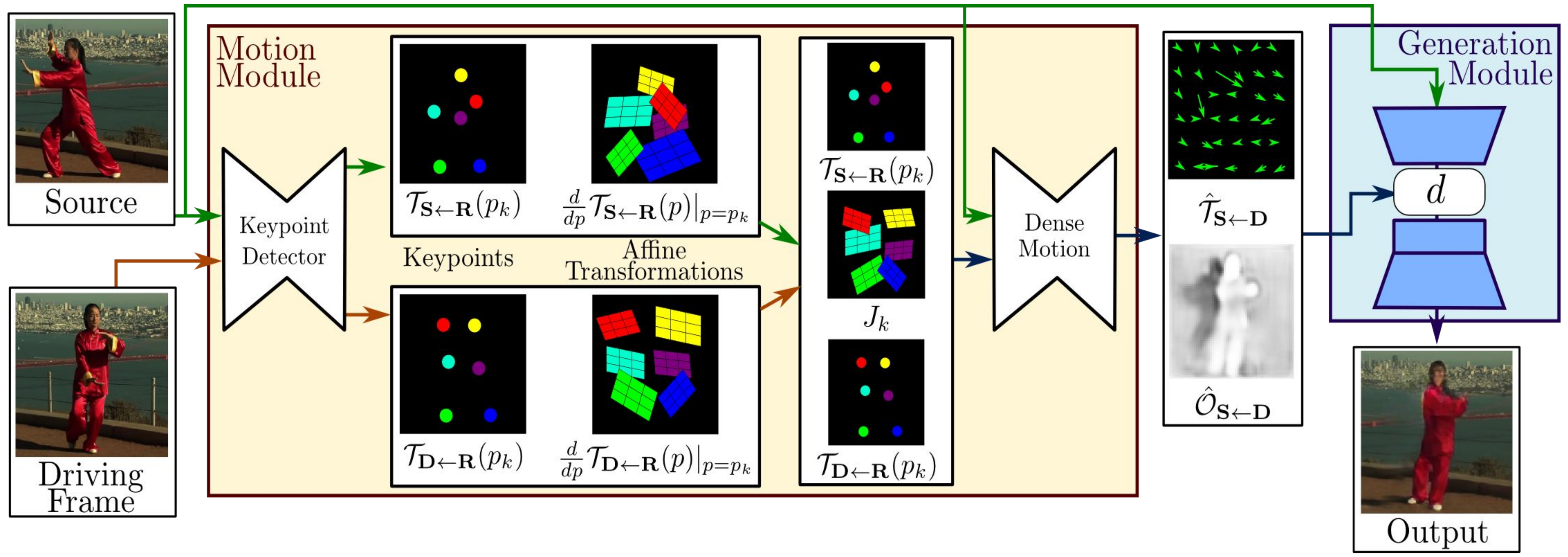
- We assume existence of abstract reference frame. We estimate reference to source $\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p)$ and reference to driving $\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p)$ motion representation using first order approximation:

$$\mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p) = \mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p_k) + \left(\frac{d}{dp}\mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p)\bigg|_{p=p_k}\right)(p-p_k) + o(\|p-p_k\|)$$

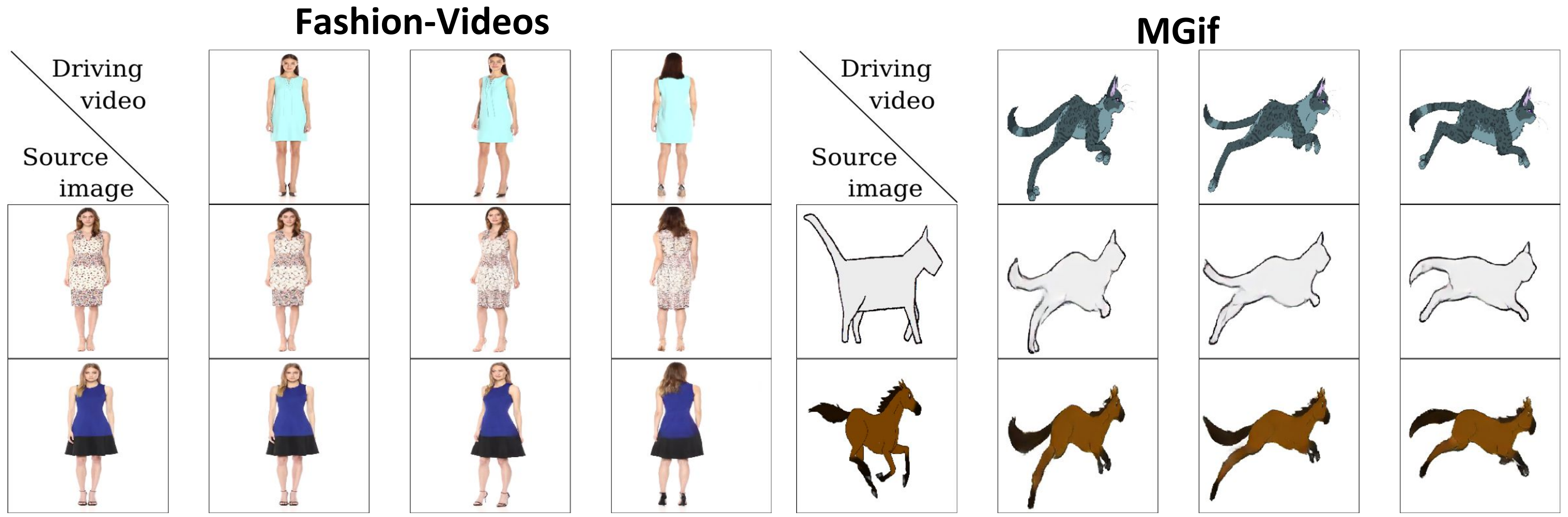
- Source $\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p)$ and driving $\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p)$ motion representations are combined:

$$\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{D}}(z) \approx \mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p_k) + J_k(z - \mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p_k)); J_k = \left(\frac{d}{dp}\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p)\bigg|_{p=p_k}\right)\left(\frac{d}{dp}\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p)\bigg|_{p=p_k}\right)^{-1}$$

- From $\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{D}}(z)$ optical flow and occlusion mask is predicted
- Representation of the source image is warped and missing parts are inpainted



Results on different datasets



User Study	Tai-Chi-HD	Nemo	Bair	VoxCeleb
X2Face vs First Order Model	92.0%	79.8%	95.0%	95.8%
MonkeyNet vs First Order Model	80.6%	60.6%	67.0%	68.4%

Our code is publicly available:
https://github.com/AliaksandrSiarohin/first-order-model