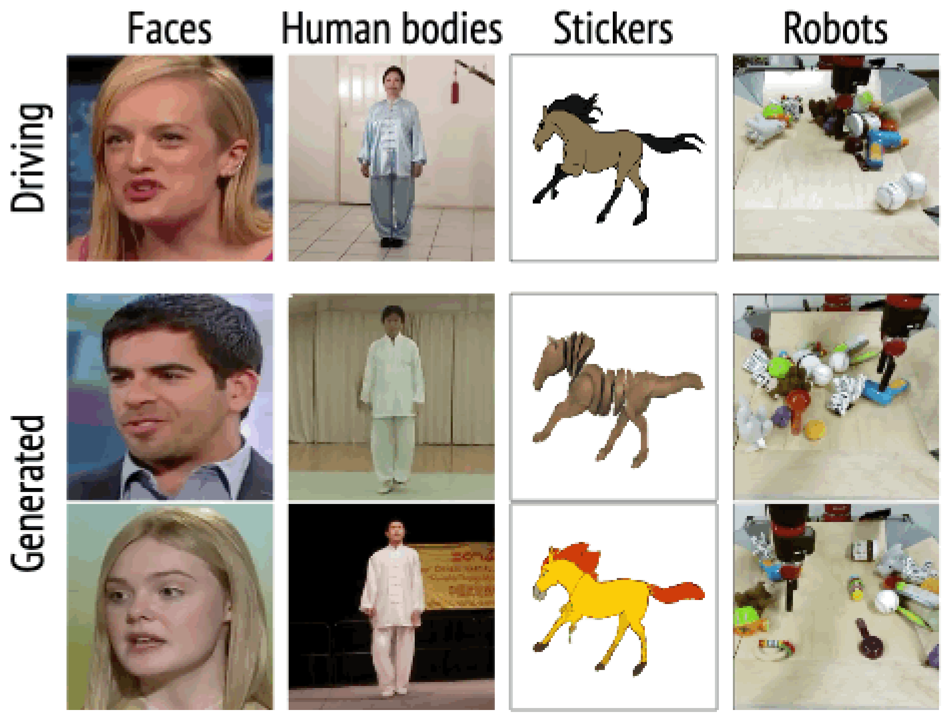
First Order Motion Model for Image Animation

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, Nicu Sebe

Kihong Kim

Introduction

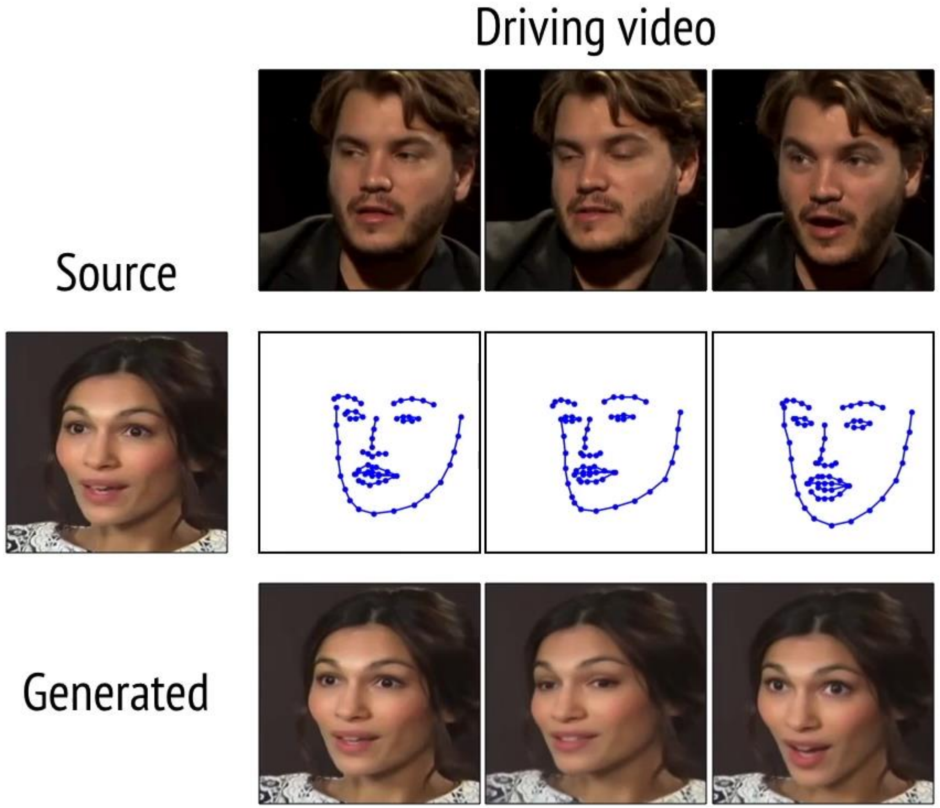


- Image animation
  - Animating object in a source image according to the motion of a driving video

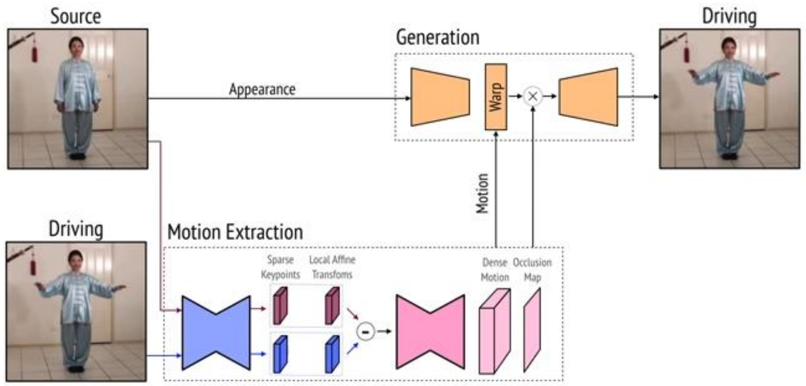
Background

- Prior works
  - Are object specific
  - Require landmark detectors
  - Impose too strong motion prior

- Our Work
  - Object-agnostic model
  - Does not require object-specific prior
  - Animates multiple objects categories



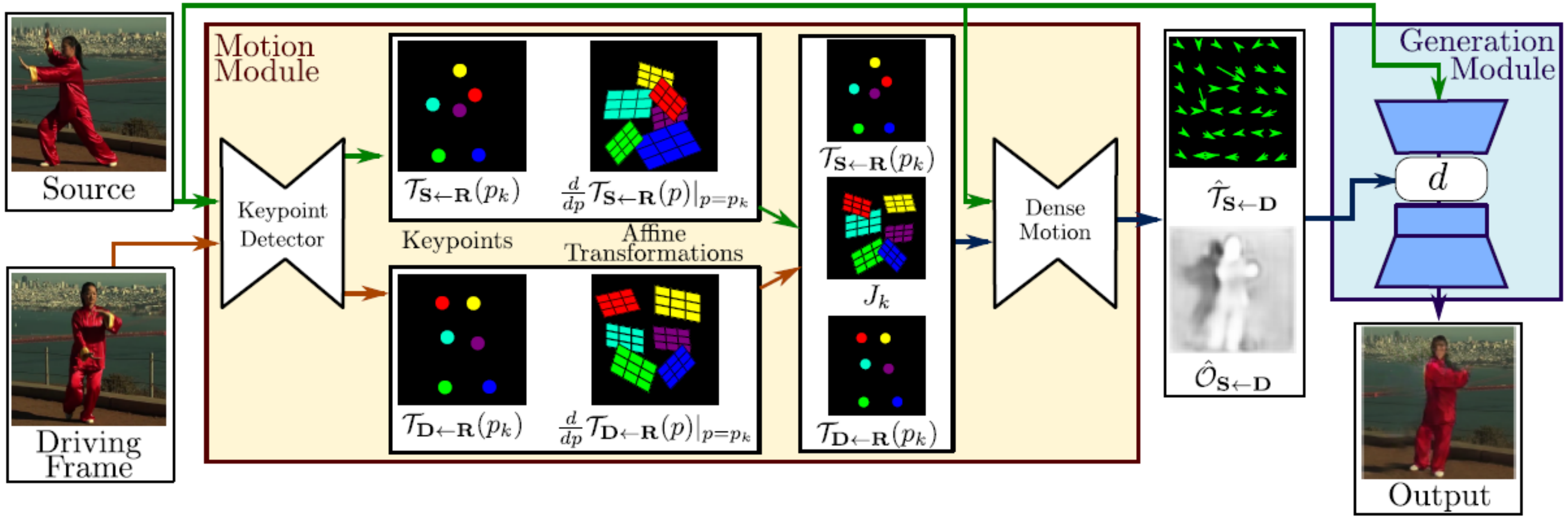
Self-supervised Training



- Direct supervision is not available
  - Train Keypoint Detector in a self-supervised manner

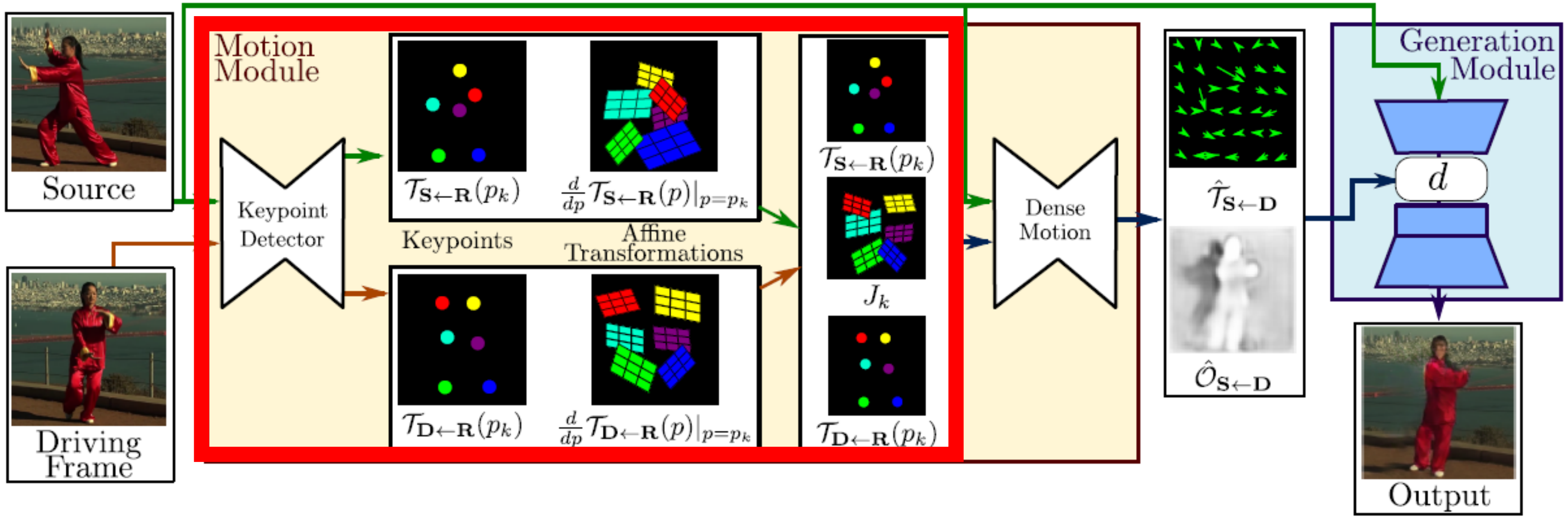
- Extracting each source & target frame from the same video
  - Reconstruct the training videos by combining a single frame and a learned latent motion

Overview



- Motion Module $\hat{T}_{S\leftarrow D}$
  - Predict a dense motion field from a source and driving frame ($T_{S\leftarrow D}$)
- Image Generation Module
  - Renders an image of the source object moving as provided in the driving video

Motion Module – Keypoint Detector



- Keypoint Detector predict **keypoint displacement** and **local affine transformation**
- Approximate $T_{S\leftarrow D}$ by its first order Taylor expansion in a neighborhood of the keypoint locations
- $T_{S\leftarrow D(p)} = T_{S\leftarrow D(p_k)} + \left(\frac{d}{dp} T_{S\leftarrow D(p)}|_{p=p_k}\right)$

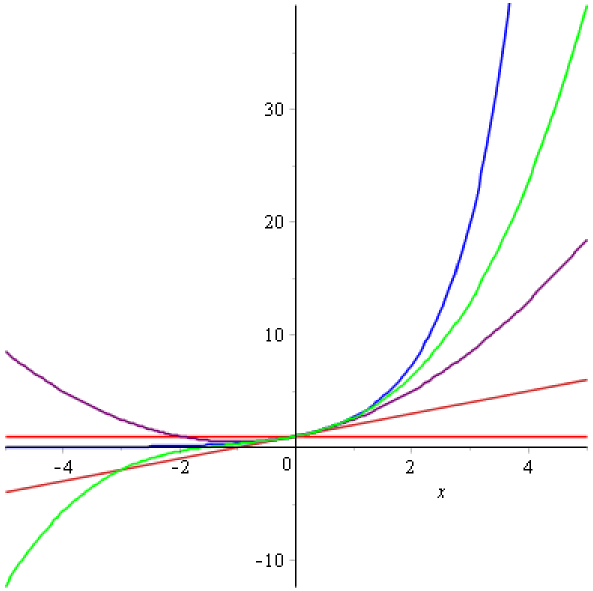
Taylor expansion

$$f(x) \longrightarrow \boxed{\text{Taylor Expansion}} \longrightarrow \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k$$

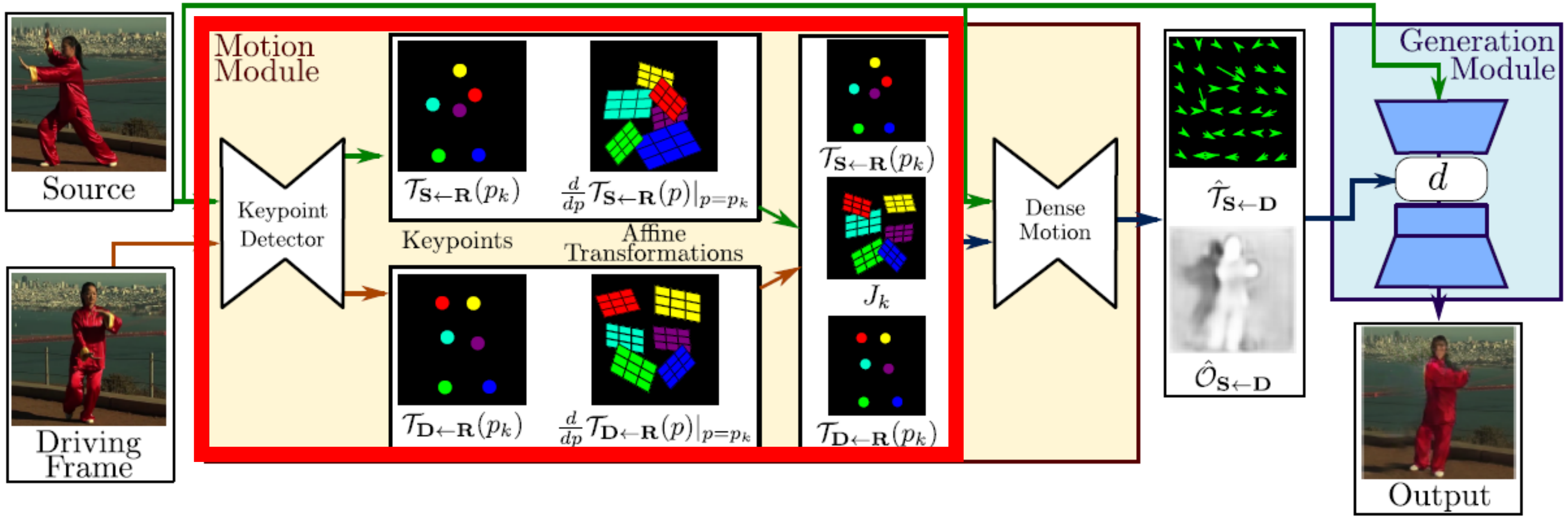
complex or unknown

simple to deal with

$$\begin{aligned} e^x &= f(0)\frac{x^0}{0!} + f'(0)\frac{x^1}{1!} + f''(0)\frac{x^2}{2!} + f'''(0)\frac{x^3}{3!} + f^{(4)}(0)\frac{x^4}{4!} + f^{(5)}(0)\frac{x^5}{5!} + \cdots \\ &= \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots \\ &= \sum_{n=0}^{\infty} \frac{x^n}{n!} \end{aligned}$$

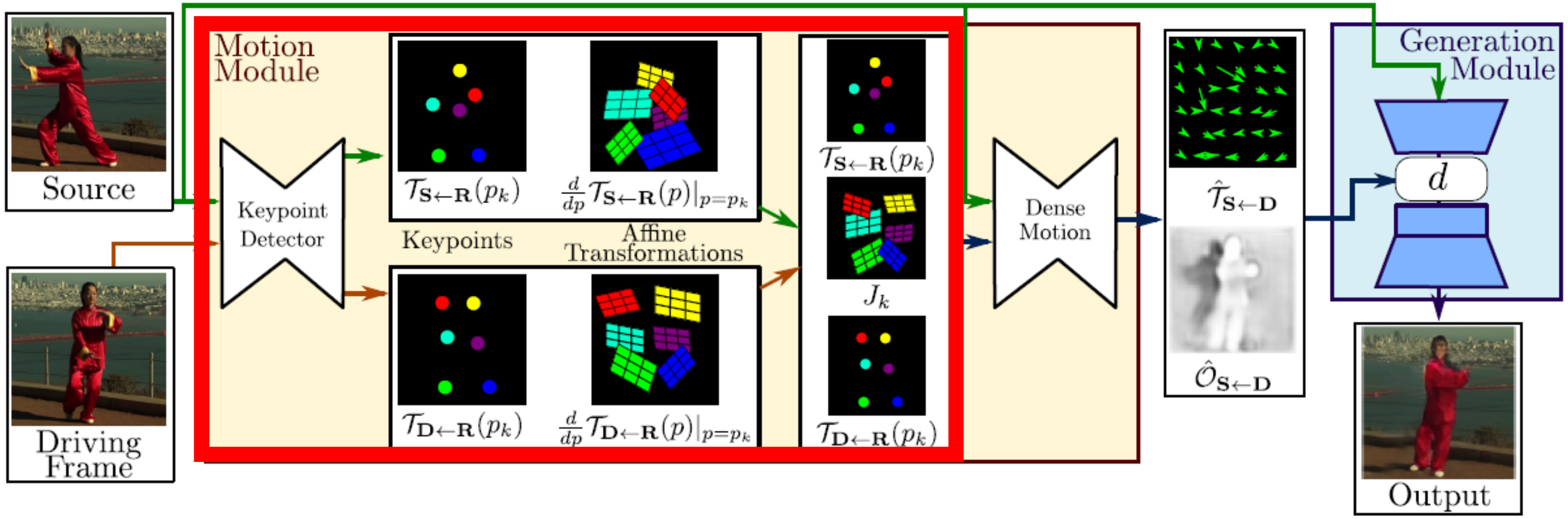


Motion Module – Keypoint Detector



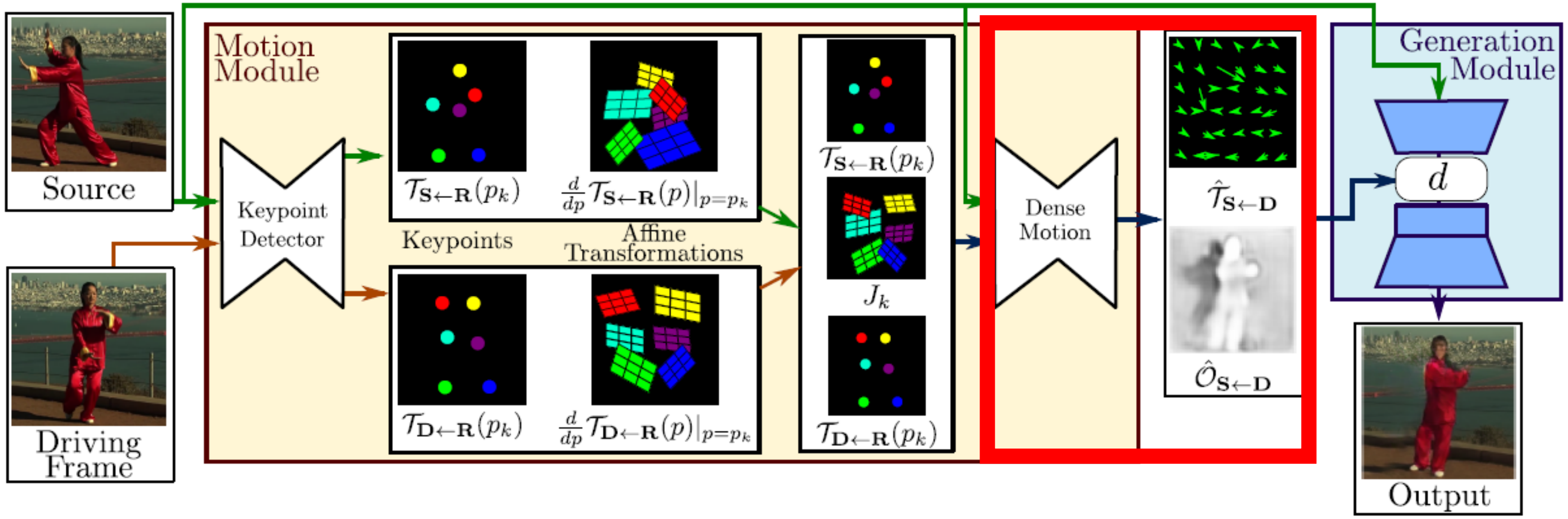
- $T_{X\leftarrow R(p)} = T_{X\leftarrow R(p_k)} + \left(\frac{d}{dp} T_{X\leftarrow R(p)}|_{p=p_k}\right)$
- Assume an abstract reference frame **R**
  - $T_{S\leftarrow D} = T_{S\leftarrow R} \circ T_{R\leftarrow D} = T_{S\leftarrow R} \circ T_{D\leftarrow R}^{-1}$
  - **R** allow us to independently process **S** and **D**

Motion Module – Keypoint Detector



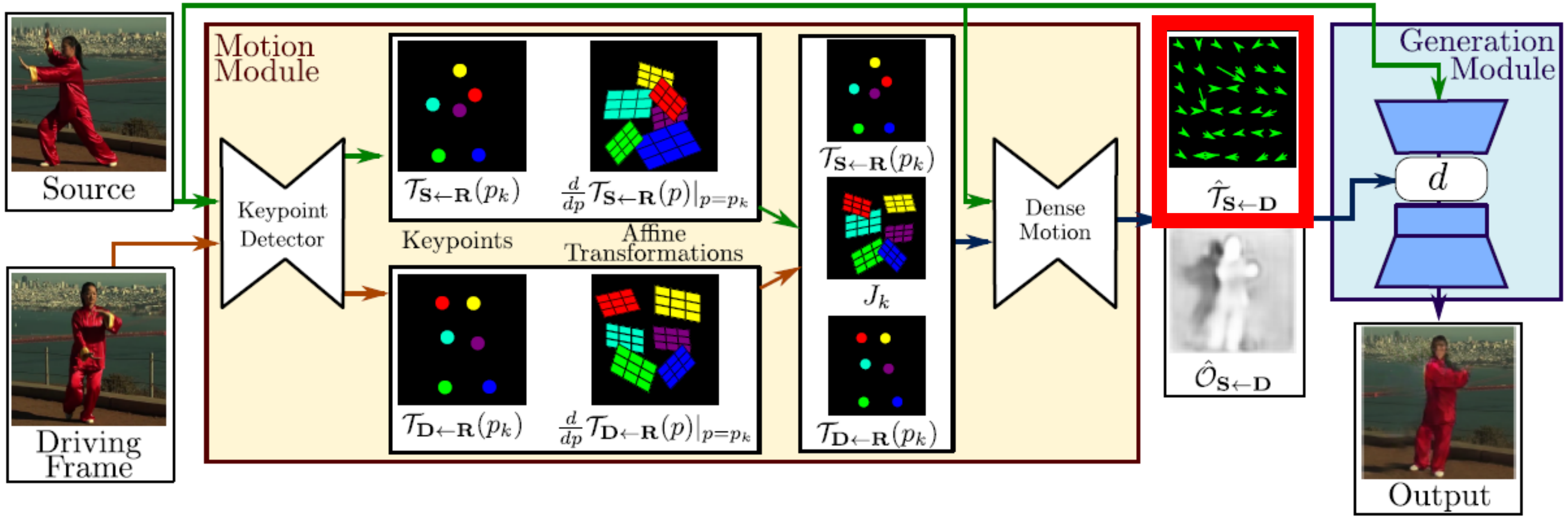
- $T_{S\leftarrow D} = T_{S\leftarrow R} \circ T_{R\leftarrow D} = \boldsymbol{T_{S\leftarrow R}} \circ \boldsymbol{T_{D\leftarrow R}^{-1}}$
- $T_{X\leftarrow R(p)} \cong \left\{T_{X\leftarrow R(p_k)}, \frac{d}{dp} T_{X\leftarrow R(p)}|_{p=p_1}\right\} + \ldots + \left\{T_{X\leftarrow R(p_k)}, \frac{d}{dp} T_{X\leftarrow R(p)}|_{p=p_k}\right\}$
- $T_{S\leftarrow D(z)} \approx T_{S\leftarrow R(p_k)} + J_k\left(z - T_{D\leftarrow R(p_k)}\right)$
  - $J_k = \left(\frac{d}{dp} T_{S\leftarrow R(p)}|_{p=p_k}\right)\left(\frac{d}{dp} T_{D\leftarrow R(p)}|_{p=p_k}\right)^{-1}$

Motion Module - Dense Motion



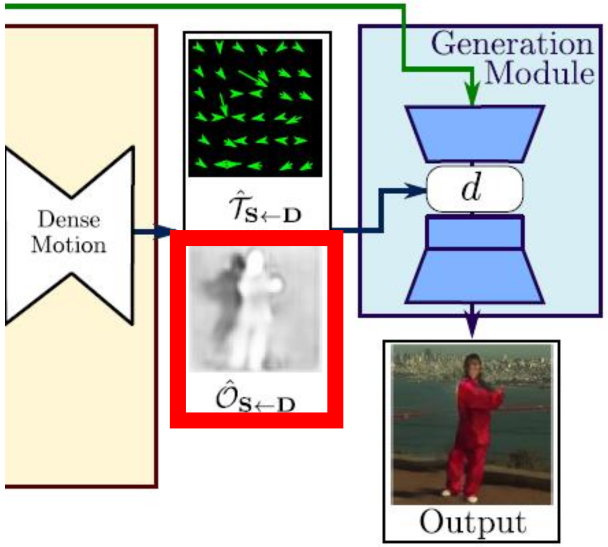
- $T_{S\leftarrow D(z)} \approx T_{S\leftarrow R(p_k)} + J_k\left(z - T_{D\leftarrow R(p_k)}\right)$
- Dense Motion network combines the local approximations to obtain dense motion field $\hat{T}_{S\leftarrow D(z)}$
- Dense Motion network outputs an occlusion mask $\hat{O}_{S\leftarrow D}$

Motion Module - Dense Motion



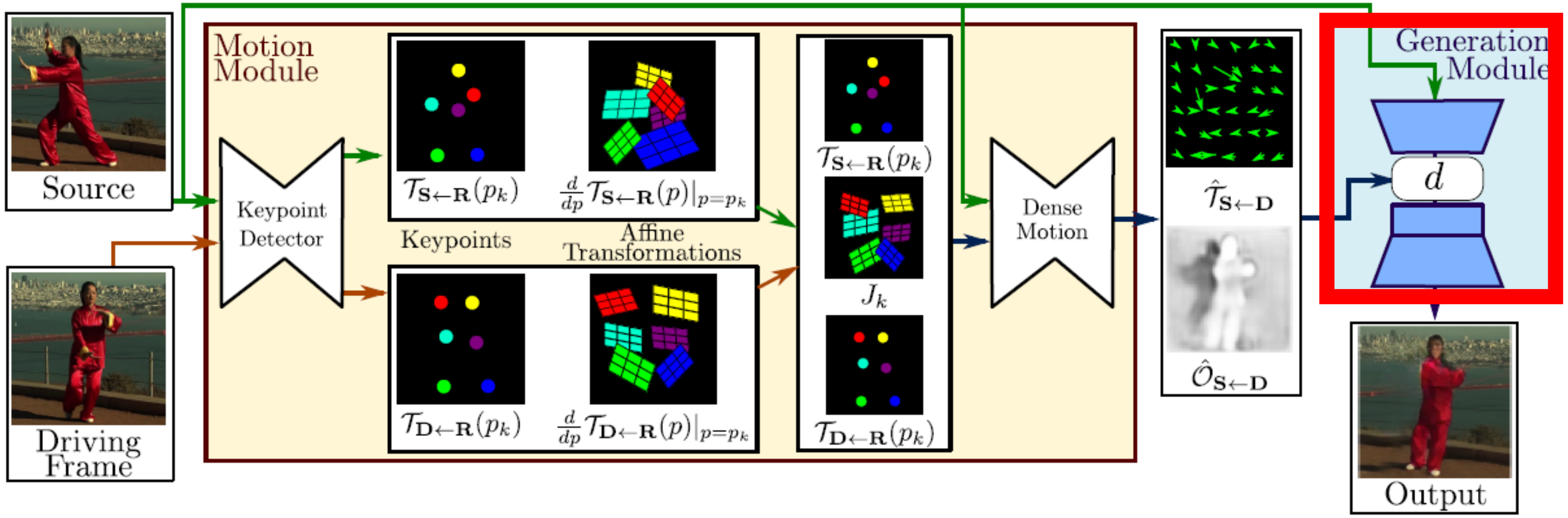
- $\hat{T}_{S\leftarrow D(z)} = M_0 z + \sum_{k=1}^{K} M_k (T_{S\leftarrow R(p_k)} + J_k(z - T_{D\leftarrow R(p_k)}))$
- $M_k(z) = \exp\left(\frac{\left(T_{D\leftarrow R(p_k)} - z\right)^2}{\sigma}\right) - \exp\left(\frac{\left(T_{S\leftarrow R(p_k)} - z\right)^2}{\sigma}\right)$
- Heatmap indicate to the dense motion network where each transformation happens

Occlusion-Aware Image Generation



- $\xi' = \hat{O}_{S\leftarrow D} \odot f_w(\xi, \hat{T}_{S\leftarrow D})$
- Occluded parts in S cannot be recovered by image-warping and thus should be inpainted
- Mask out feature map regions that should be inpainted

Generation Module



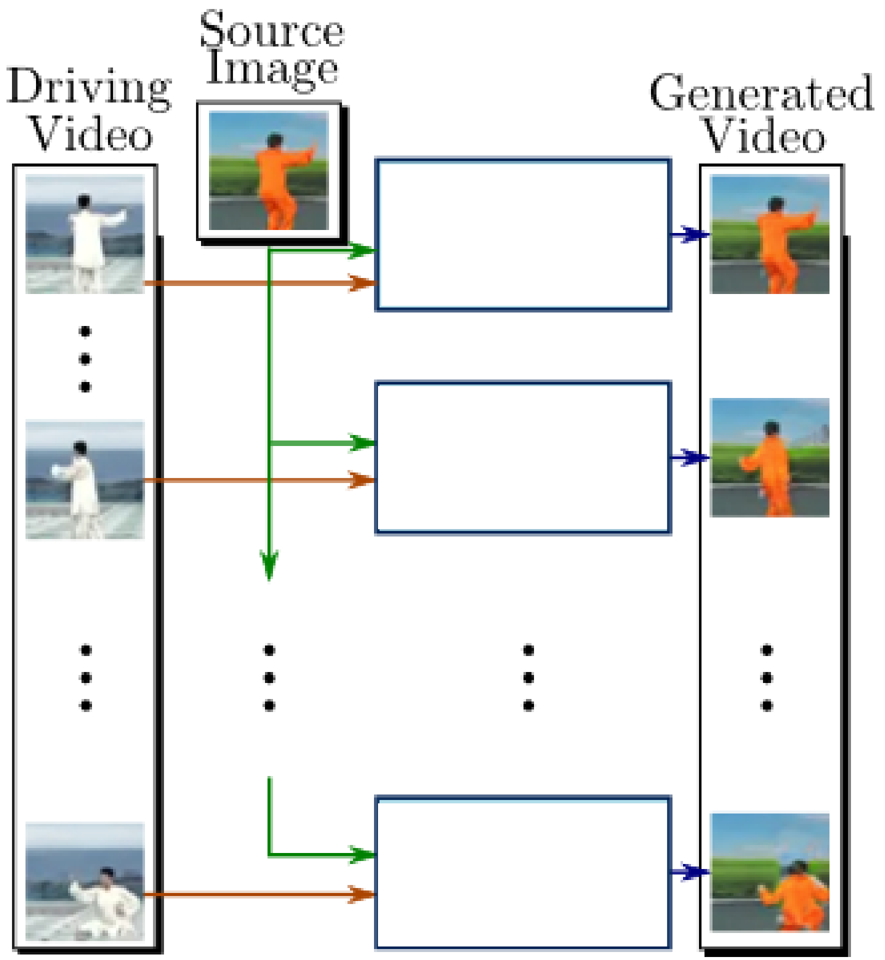
- Generation Module renders an image of the source object moving as provided in the driving video
- warps the source image according to $\hat{T}_{S\leftarrow D}$ and inpaints the image parts that are occluded in the source image

Loss function

- **Reconstruction Loss**
  - $L_{rec}\left(\widehat{D}, D\right) = \sum_{i=1}^{I} |N_i\left(\widehat{D}\right) - N_i(D)|$
  - Perceptual loss using the pre-trained VGG-19
  - Multiple resolution
    - 256x256, 128x128, 64x64, 32x32

- **Equivariance Constraint**
  - Our keypoint predictor doesn't require any keypoint annotations during training.
  - This may lead to unstable performance
  - $T_{X \leftarrow R}\ (p_k) \equiv T_{X \leftarrow Y} \circ T_{Y \leftarrow R}\ (p_k)$

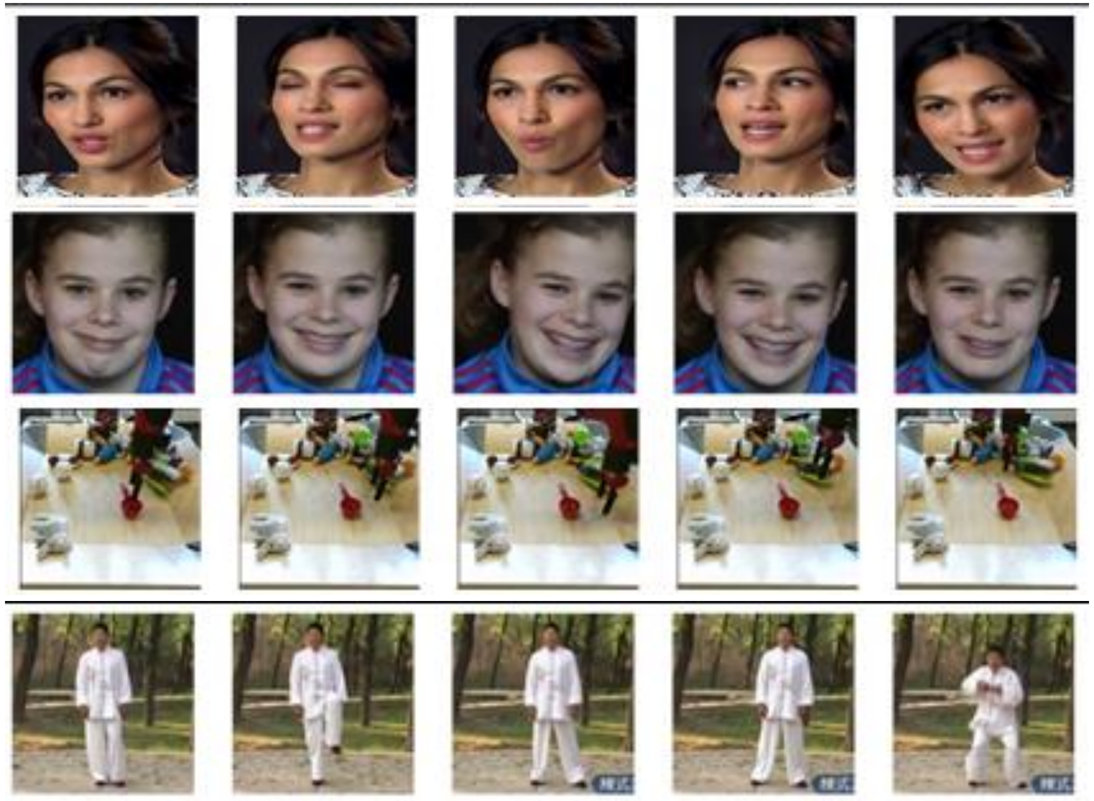
Testing Stage – Relative Motion Transfer



- **Animation using absolute coordinates**

- **Animation using relative coordinates**

Experiment



- **VoxCeleb**
  cropping face from image using bounding box
  19,522 training videos and 525 test videos
- **UvA-Nemo**
  Facial analysis dataset
  1116 training videos and 124 test videos
- **BAIR robot pushing**
  42,880 training and 128 test videos
  30 frame long, 256x256 resolution
- **Tai-Chi-HD**
  3,049 training, 285 testing
  128 to 1024 frames

Ablation Study

Table 1: Quantitative ablation study for video reconstruction on *Tai-Chi-HD*.

	$\mathcal{L}_1$	*Tai-Chi-HD* (AKD, MKR)	AED
Baseline	0.073	(8.945, 0.099)	0.235
Pyr.	0.069	(9.407, 0.065)	0.213
Pyr.+$\mathcal{O}_{\mathbf{S}\leftarrow\mathbf{D}}$	0.069	(8.773, 0.050)	0.205
Jac. w/o Eq. (12)	0.073	(9.887, 0.052)	0.220
Full	**0.063**	**(6.862, 0.036)**	**0.179**

- L1 Distance
- AKD (Average Keypoint Distance)
  - Evaluate whether the motion of the input video is preserved
- MKR (Missing Keypoint Rate)
  - Evaluate the appearance quality of each generated frame
- AED (Average Euclidean Distance)
  - Evaluate Euclidean distance between G.T and generated frame representation

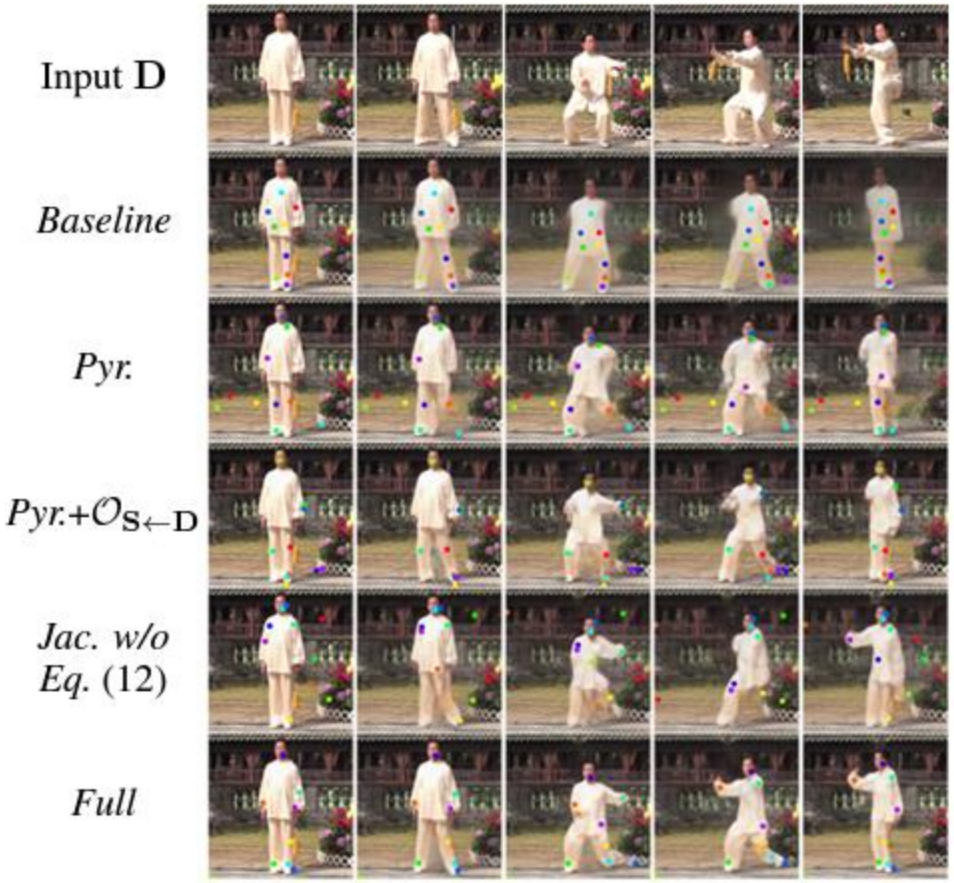


Figure 3: Qualitative ablation on *Tai-Chi-HD*.

Ablation Study

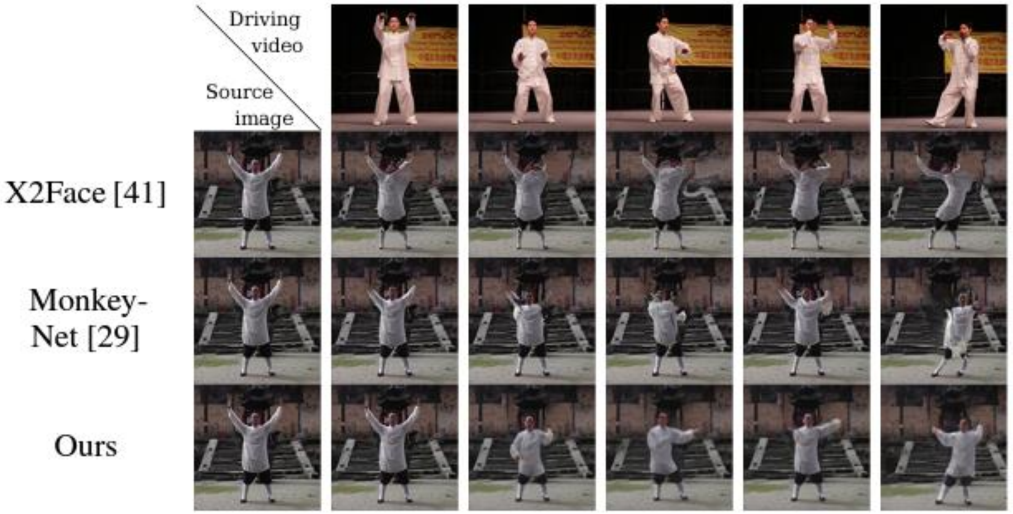


Table 3: Video reconstruction: comparison with the state of the art on four different datasets.

	Tai-Chi-HD			VoxCeleb			Nemo			Bair
	$\mathcal{L}_1$	(AKD, MKR)	AED	$\mathcal{L}_1$	AKD	AED	$\mathcal{L}_1$	AKD	AED	$\mathcal{L}_1$
X2Face [41]	0.080	(17.654, 0.109)	0.272	0.078	7.687	0.405	0.031	3.539	0.221	0.065
Monkey-Net [29]	0.077	(10.798, 0.059)	0.228	0.049	1.878	0.199	0.018	1.285	0.077	0.034
Ours	**0.063**	**(6.862, 0.036)**	**0.179**	**0.043**	**1.294**	**0.140**	**0.016**	**1.119**	**0.048**	**0.027**

- Our approach is able to generate significantly better looking videos in which each body part is independently animated

Conclusion

- Mathematical formulation describes the motion field
  - a set of keypoints displacement and local affine transformations
- Dense Motion Network produce Occlusion mask to inpaint occluded region
  - Occluded parts in S cannot be recovered by image-warping
- Test our method on four different datasets containing various objects
  - Our approach outperforms existing method

Thank You !