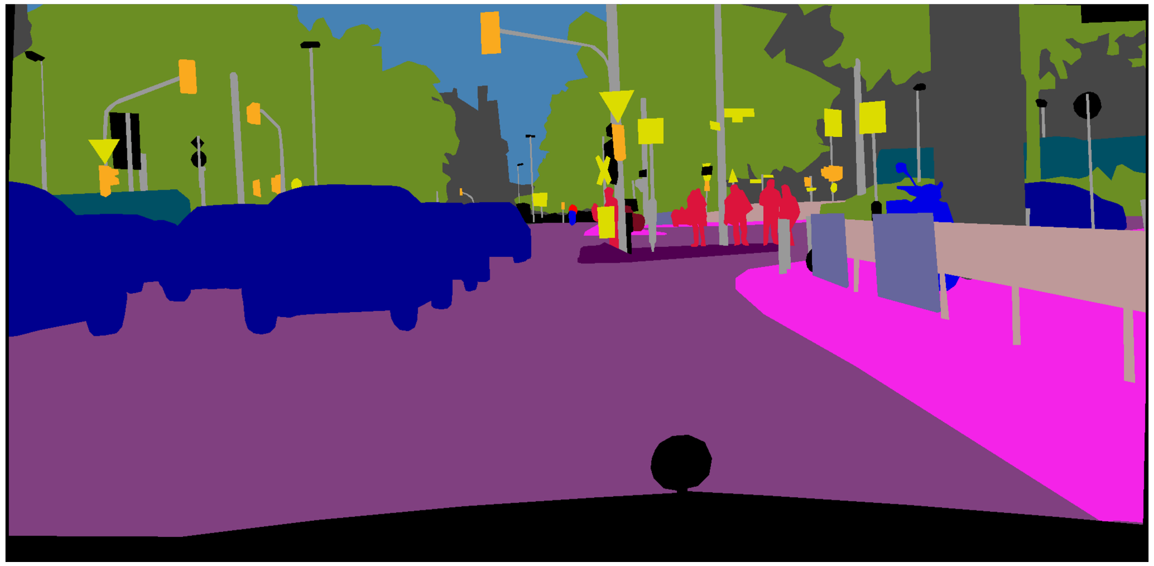
Video Synthesis

Ron Mokady

Image Synthesis

Image to Image Translation

Input labels



Synthesized image

Style Transfer

Content Transfer

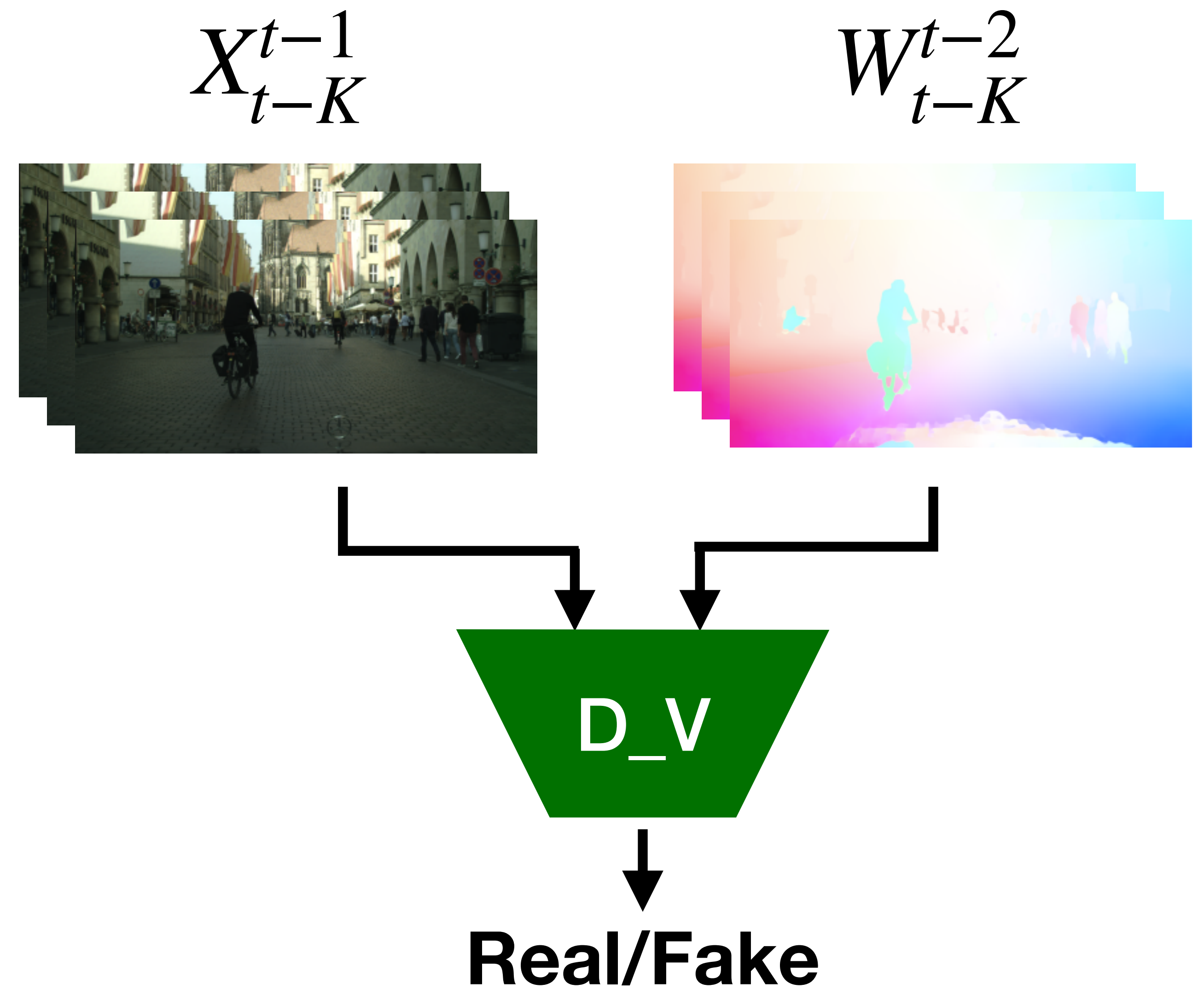
Source Target Result

How to Synthesize Video?

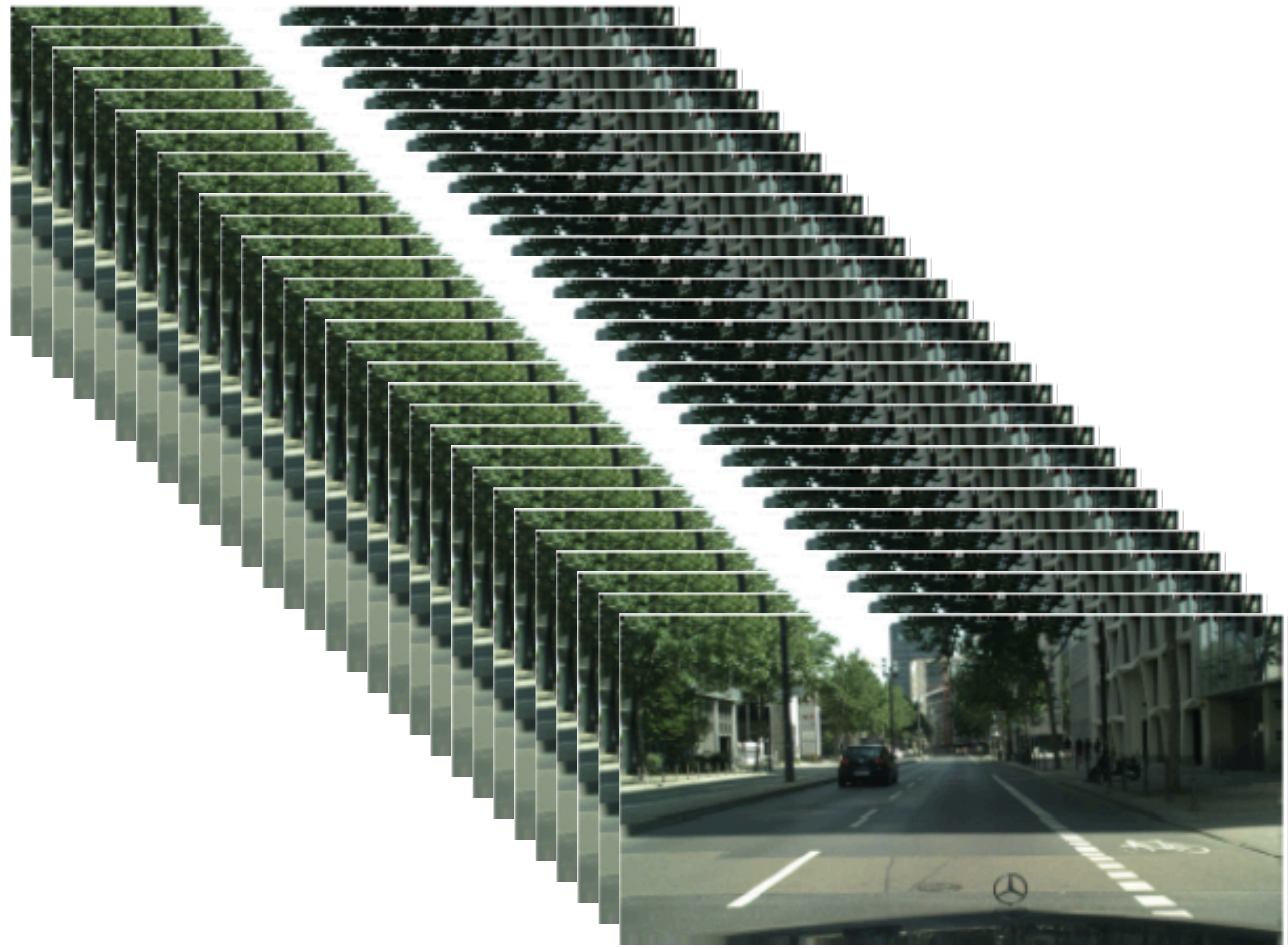
Temporal Coherence

Optical Flow

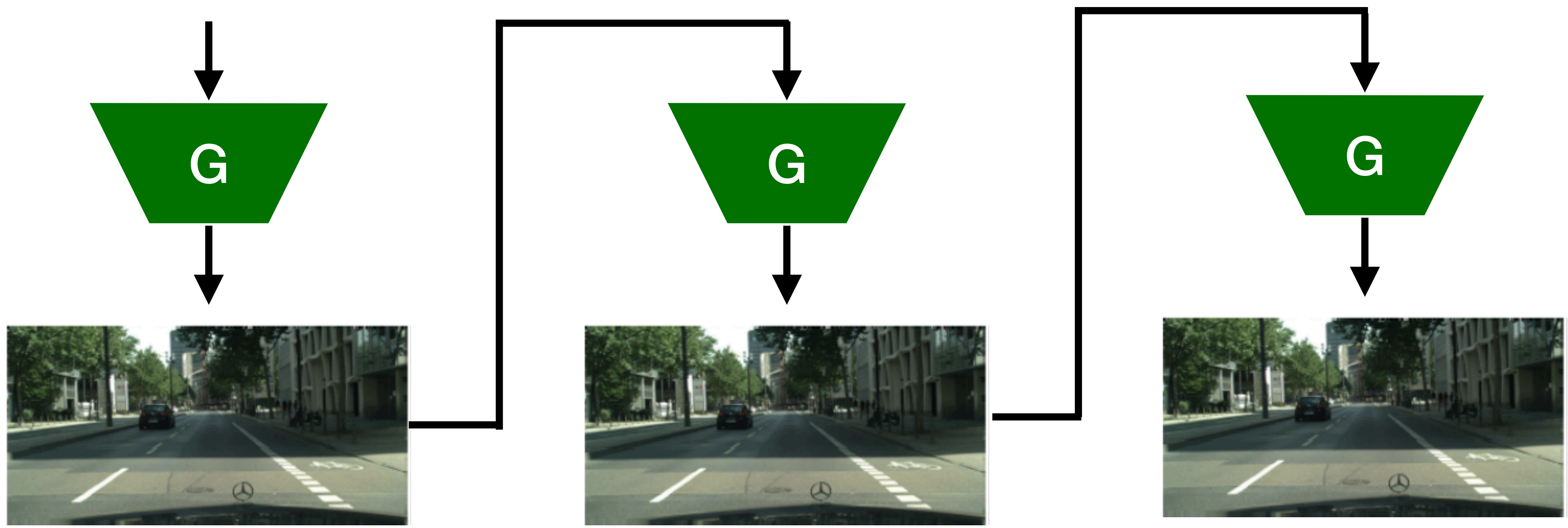
GAN Loss



Size



Sequential Generation

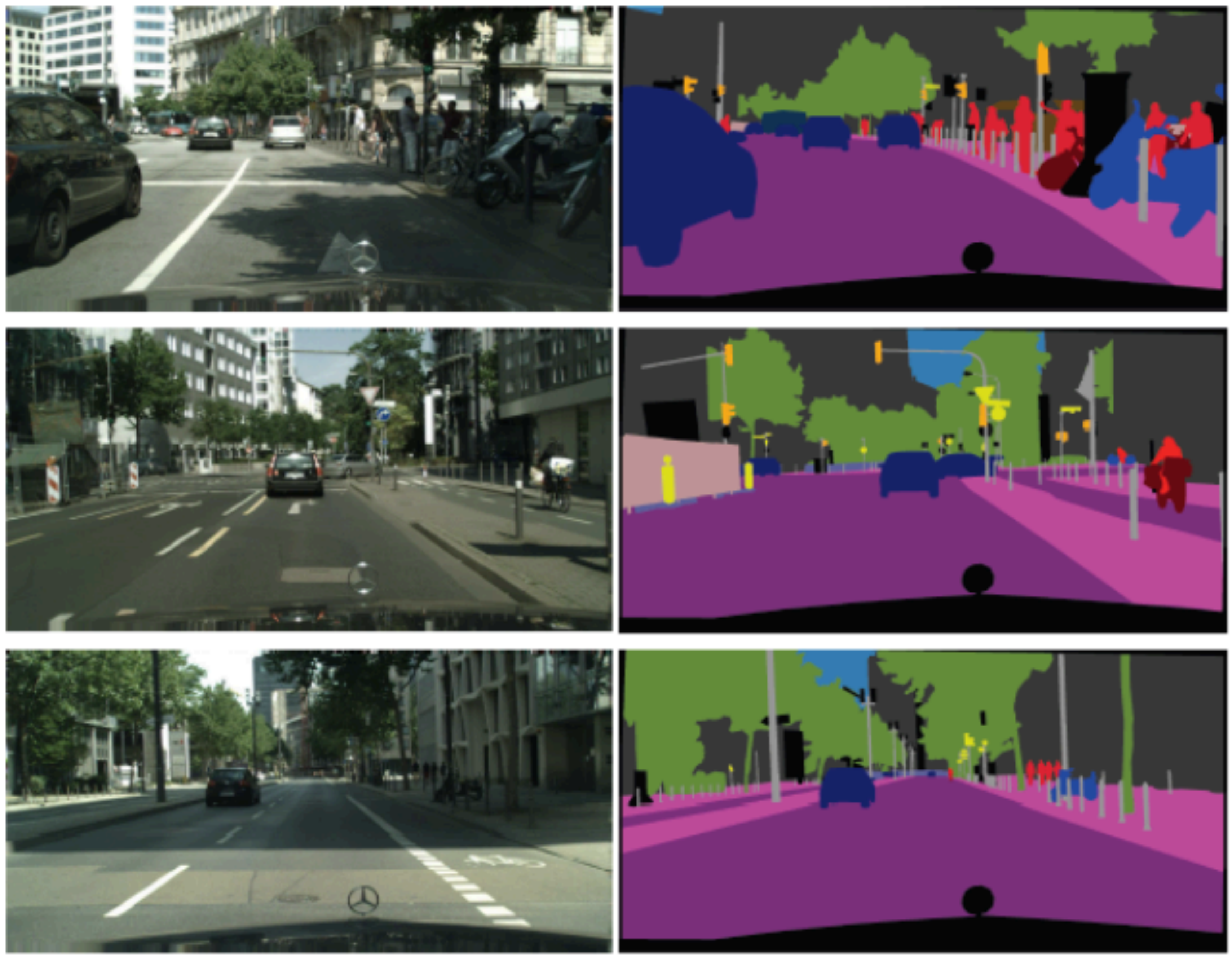


Motion Representation

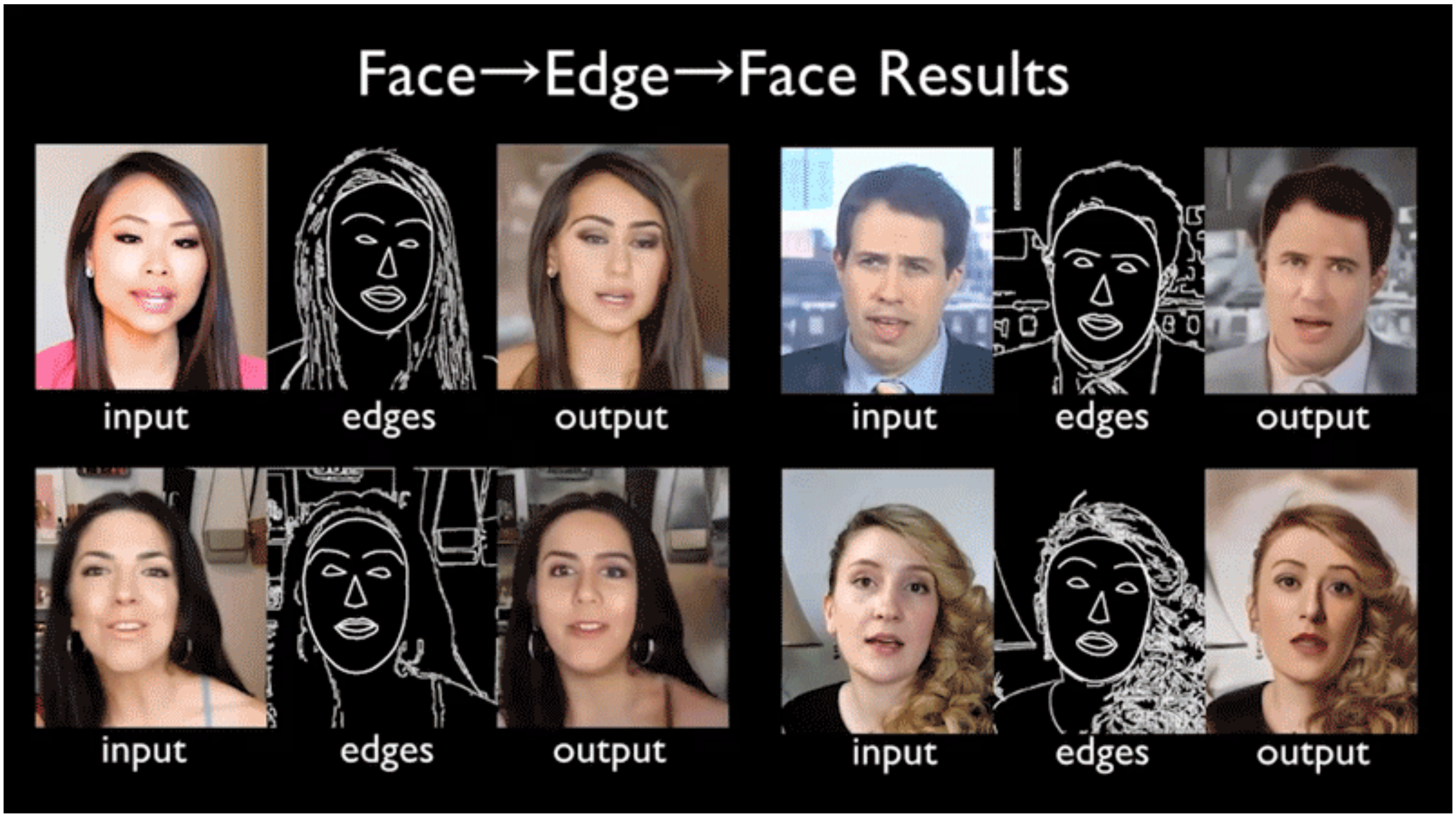
Video-to-Video Synthesis

Wang et al. NeurIPS 2018

Paired Dataset

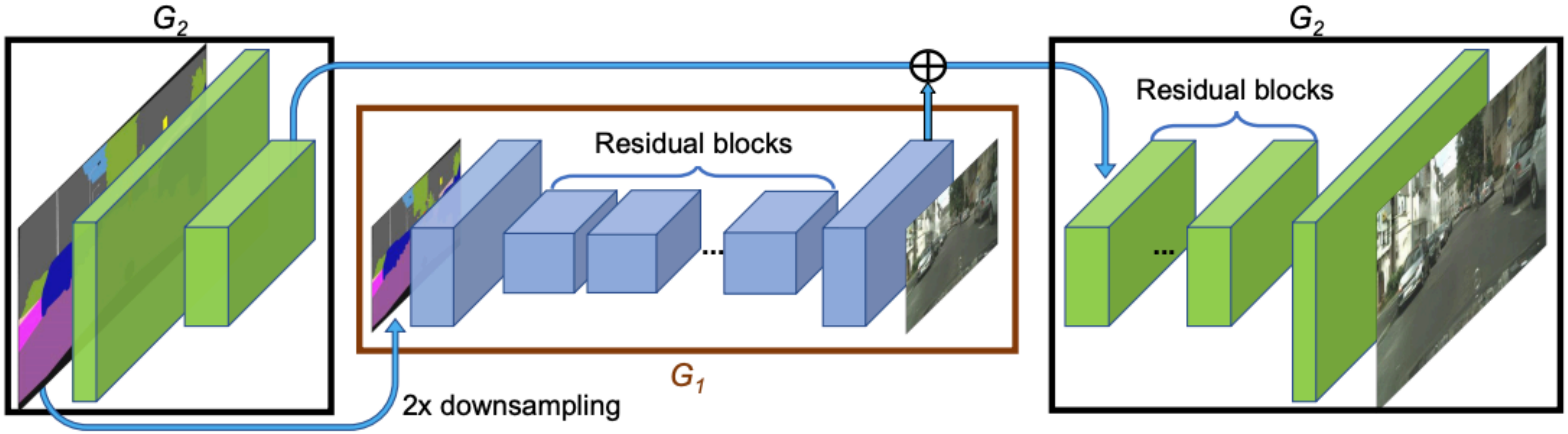


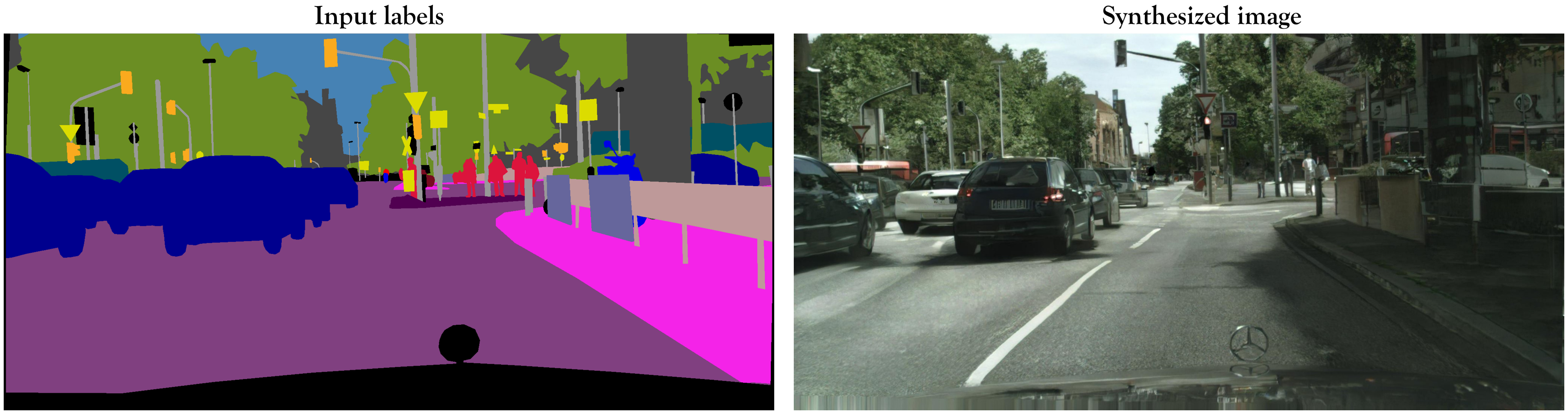
One Domain is Synthetic



Pix2PixHD

Wang et al. CVPR 2018





Sequential Generation

We generate current frame using the current source frame, the last two source frame and the last two generated frames.

$$X_t = G(S_{t-2}^{t}, X_{t-2}^{t-1}) = G(S_t, S_{t-1}, S_{t-2}, X_{t-1}, X_{t-2})$$

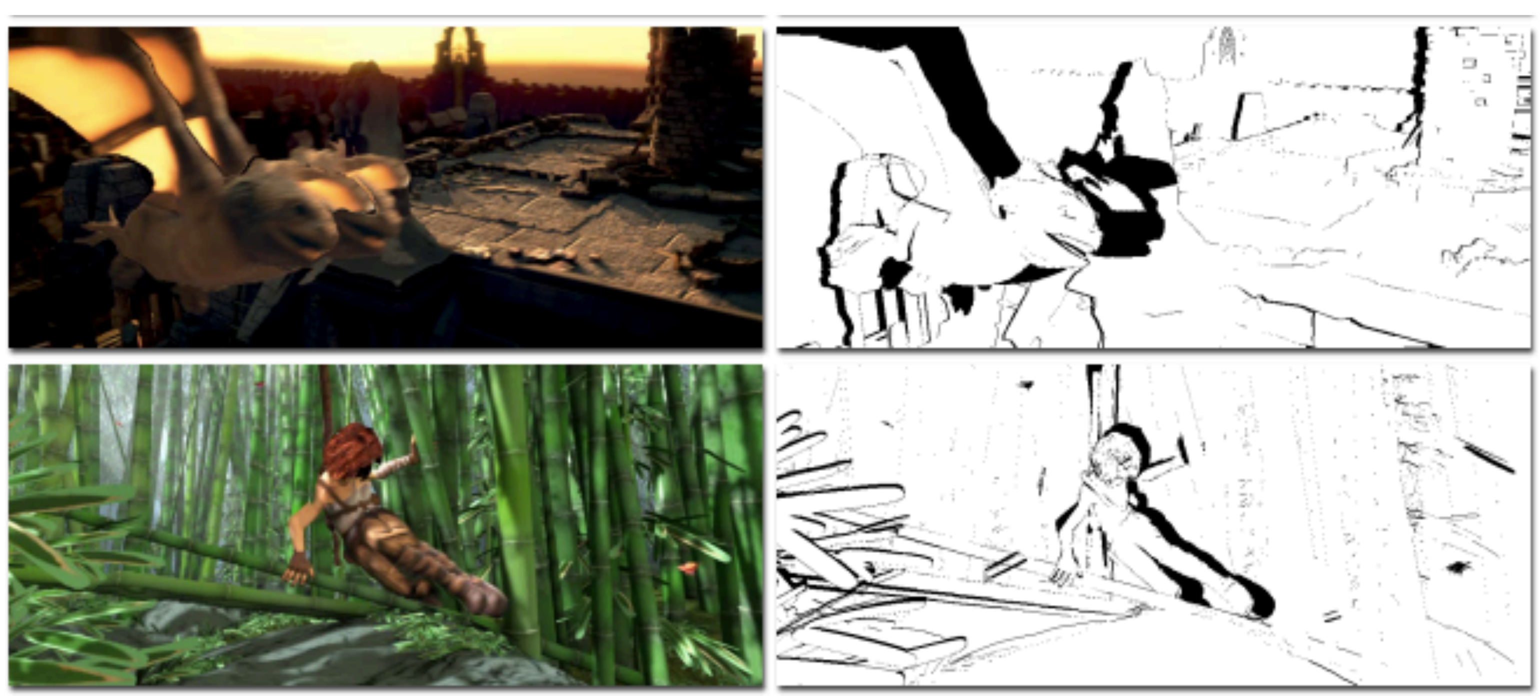
Using Optical Flow

$$X_t = (1 - M_t) \cdot W_{t-1}(X_{t-1}) + M_t \cdot H_t$$

Estimated Soft Occlusion Mask

Estimated Optical Flow

Hallucinated Image

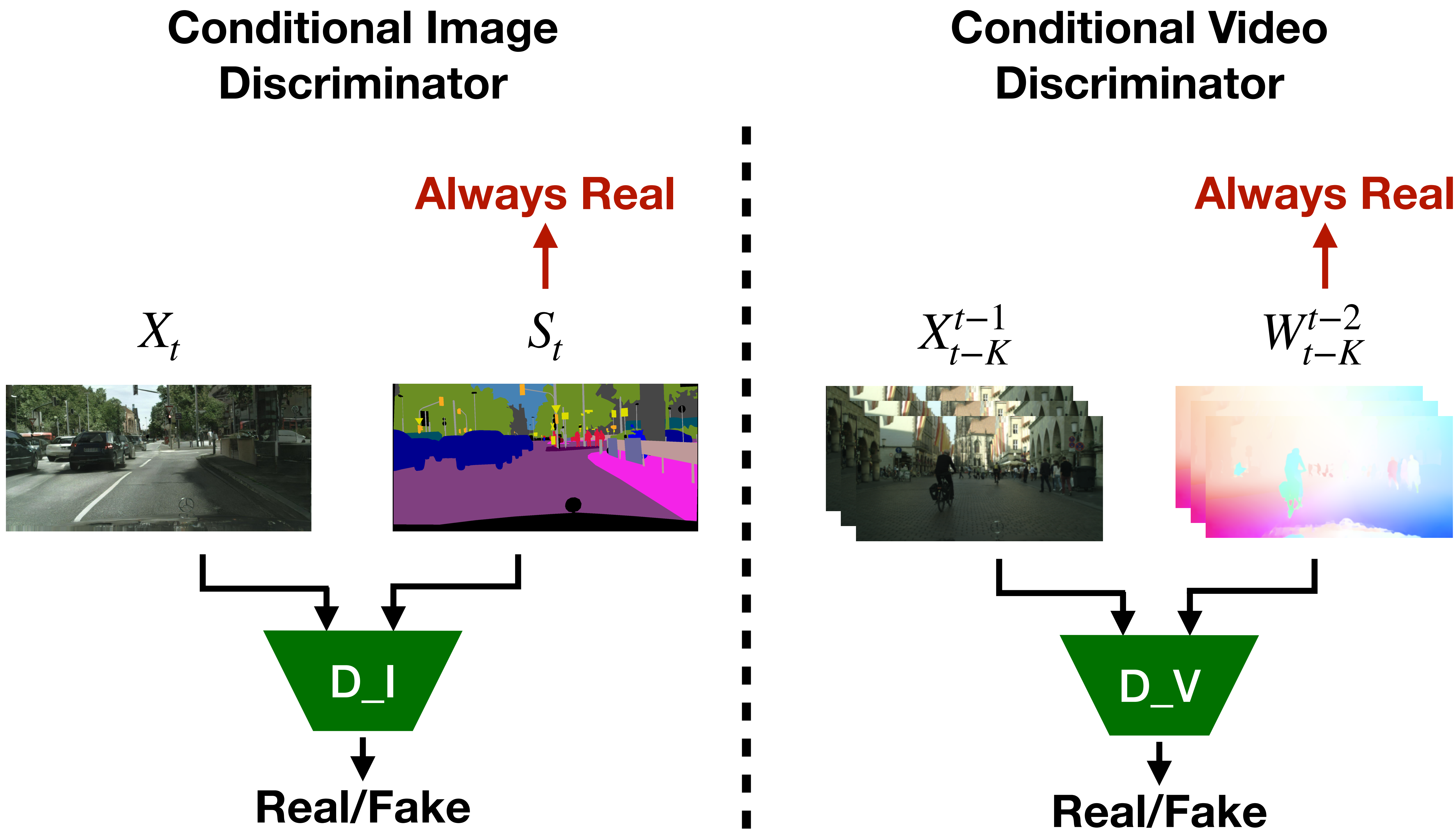


Warped Image | Inverse Ground-Truth Occlusion Map

Background Foreground Decomposition

$$H_t = (1 - \underbrace{M_{B,t}}_{\text{From the Segmentation}}) \cdot H_{F,t} + M_{B,t} \cdot H_{B,t}$$

GAN Loss


Conditional Image Discriminator
Always Real
X_t
S_t
D_I
Real/Fake
Conditional Video Discriminator
Always Real
X_{t-K}^{t-1}
W_{t-K}^{t-2}
D_V
Real/Fake

Additional Loss Terms

Optical Flow Loss

$$\mathcal{L}_W = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(||\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_1 + \|\tilde{\mathbf{w}}_t(\mathbf{x}_t) - \mathbf{x}_{t+1}\|_1\right)$$

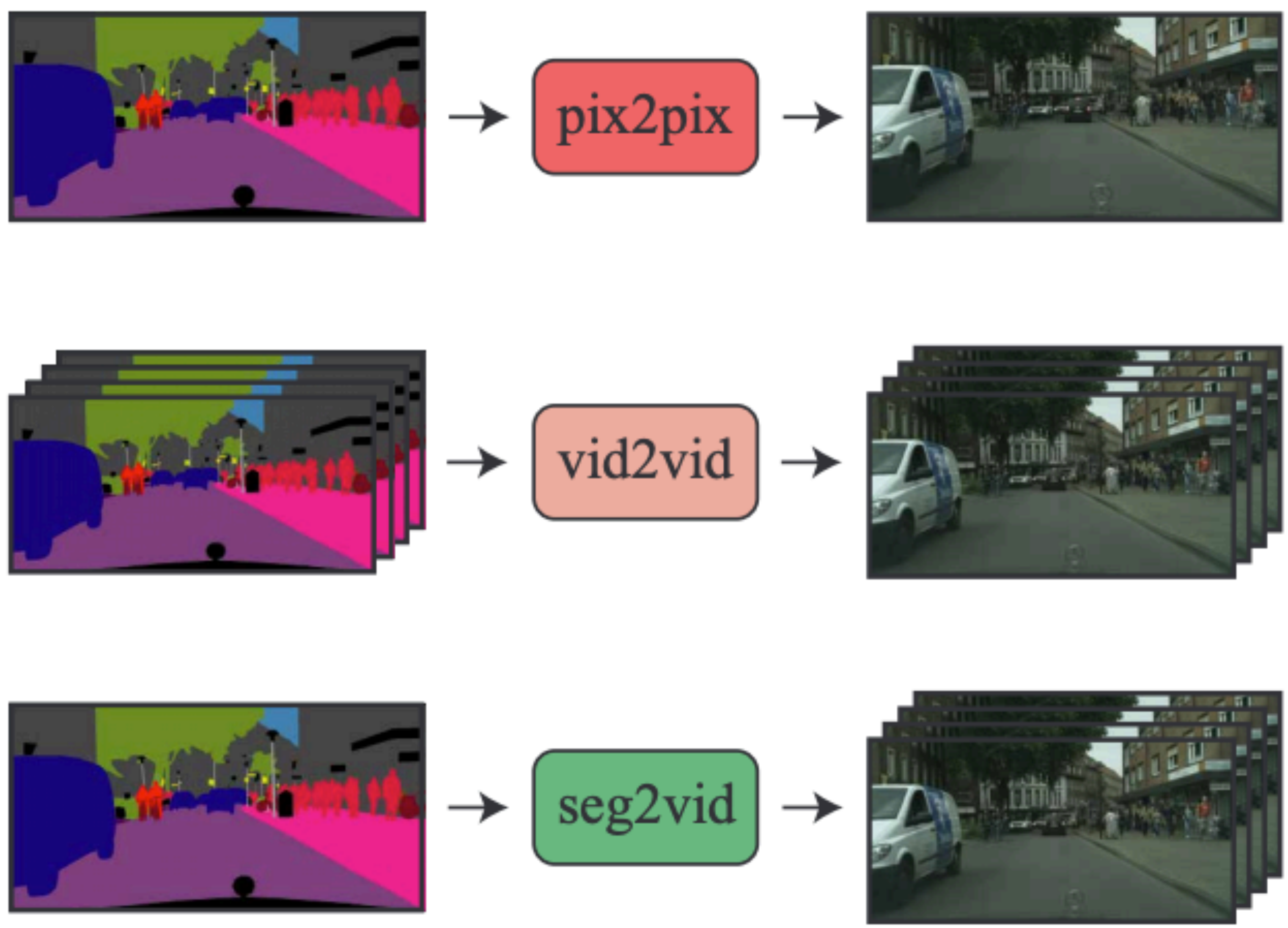
Reconstruction (VGG and discriminator features)

$$\sum_i \frac{1}{P_i}[||\psi^{(i)}(\mathbf{x}) - \psi^{(i)}(G(\mathbf{s}))||_1]$$

Questions?

Video Generation from Single Semantic Label Map

Pan et al. CVPR 2019



Generation

Single Segmentation Map

Single Segmentation Map

Prediction

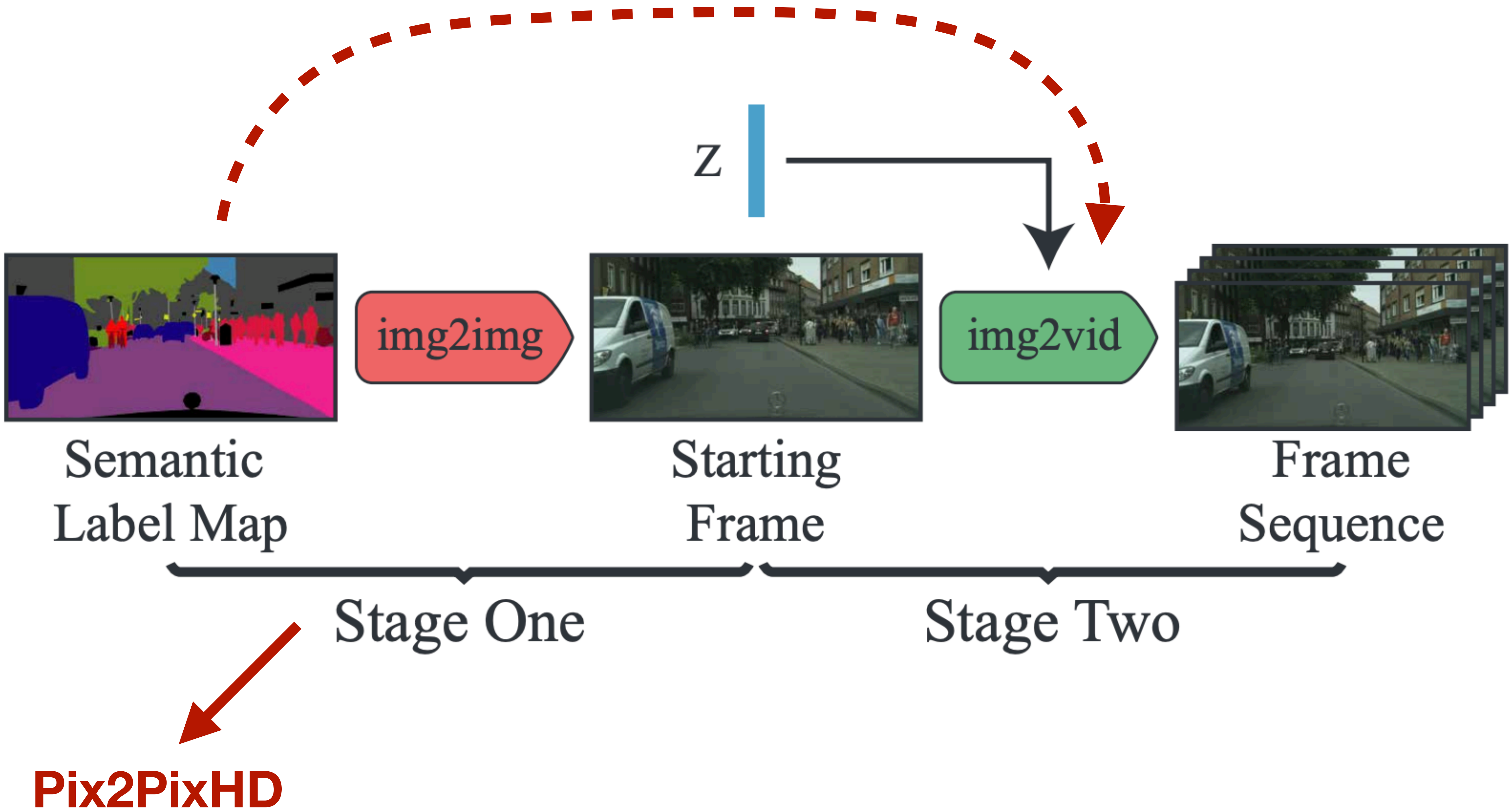
1st frame and single segmentation map

1st frame and single segmentation map

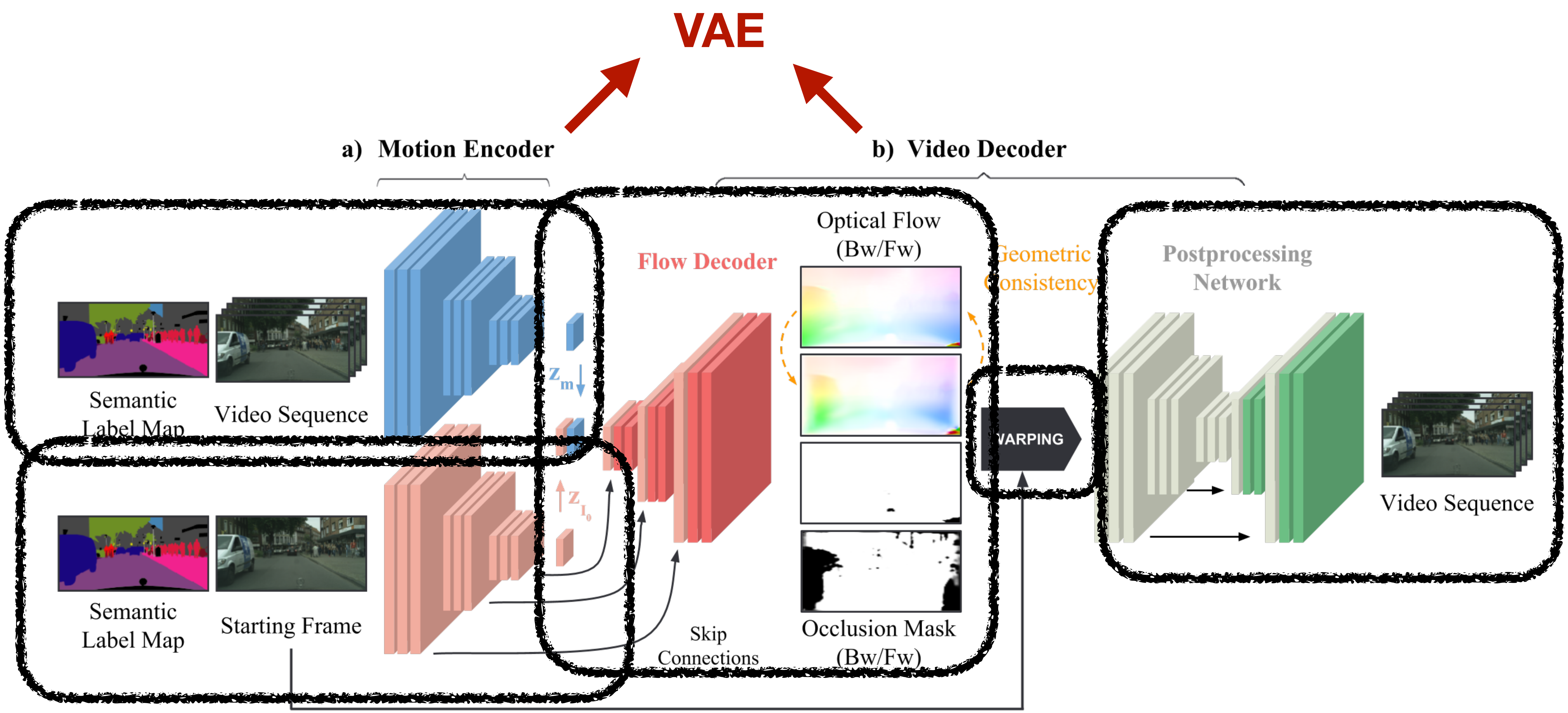
1st frame and single segmentation map

Paired Dataset

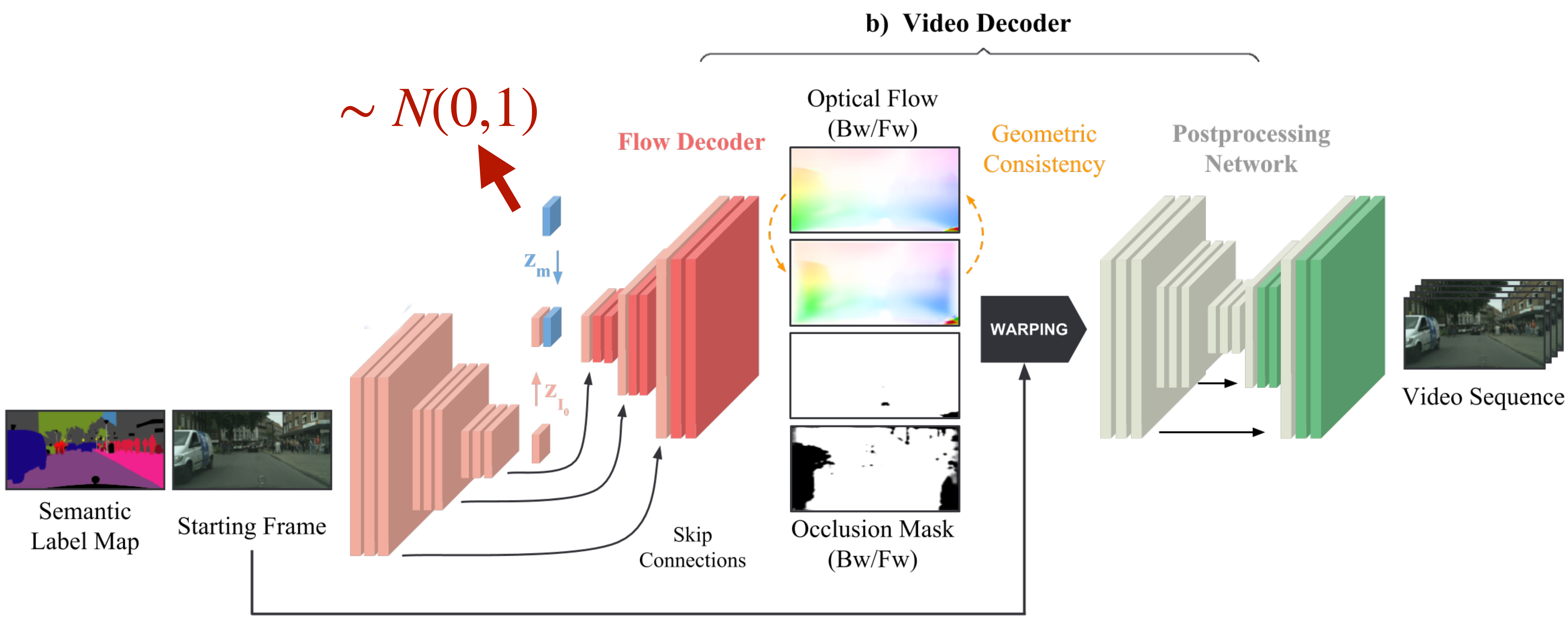
General Approach



Img2Vid Training



Img2Vid Inference



Loss Terms

Reconstruction (VGG and L1)

Optical Flow Reconstruction

$$\mathcal{L}_r(W^f, W^b, V) = \sum_t^T \sum_{\mathbf{x}} o_t^f(\mathbf{x}) |I_0(\mathbf{x}) - I_t(\mathbf{x} + \mathbf{w}_t^f(\mathbf{x}))|_1 + o_t^b(\mathbf{x}) |I_t(\mathbf{x}) - I_0(\mathbf{x} + \mathbf{w}_t^b(\mathbf{x}))|_1,$$

Optical Flow Consistency

$$\mathcal{L}_{fc}(W^f, W^b) = \sum_t^T \sum_{\mathbf{x}} o_t^f(\mathbf{x}) |\mathbf{w}_t^f(\mathbf{x}) - \mathbf{w}_t^b(\mathbf{x} + \mathbf{w}_t^f(\mathbf{x}))|_1 + o_t^b(\mathbf{x}) |\mathbf{w}_t^b(\mathbf{x}) - \mathbf{w}_t^f(\mathbf{x} + \mathbf{w}_t^b(\mathbf{x}))|_1,$$

Optical Flow Smoothness

$$\mathcal{L}_{fs}(W^f, W^b) = |\nabla W^f|_1 + |\nabla W^b|_1$$

Occlusion Mask Regularization

$$\lambda_p |1 - O^b|_1 + \lambda_p |1 - O^f|_1$$

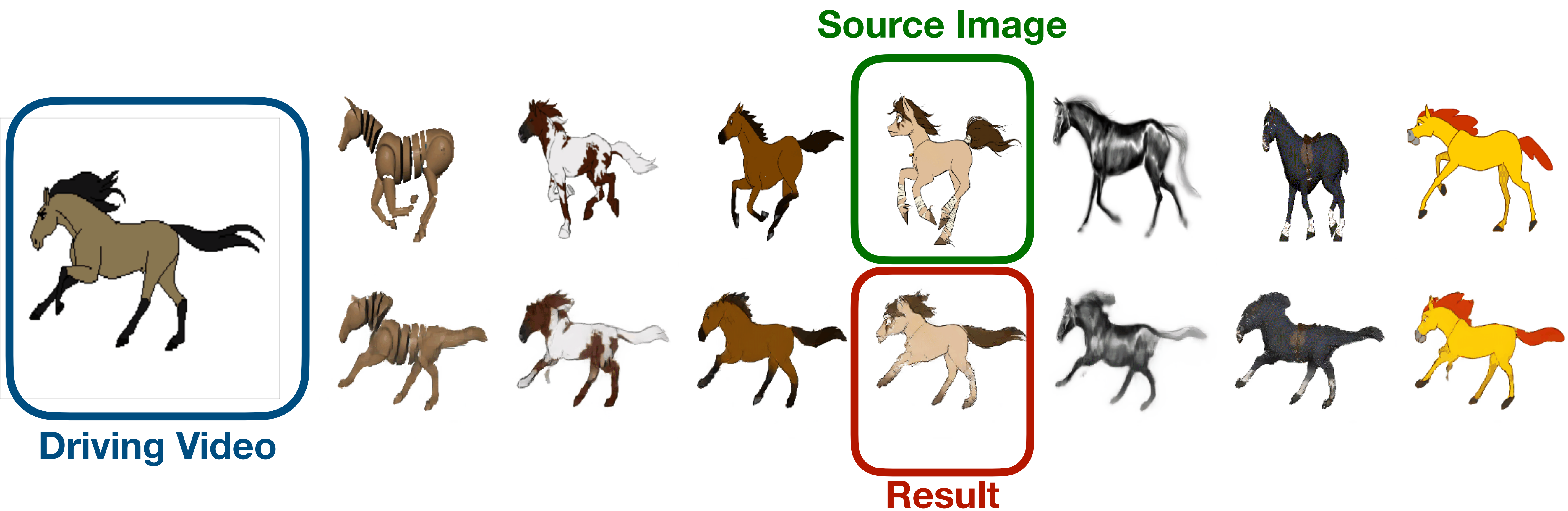
KL-divergence

Questions?

First Order Motion Model for Image Animation

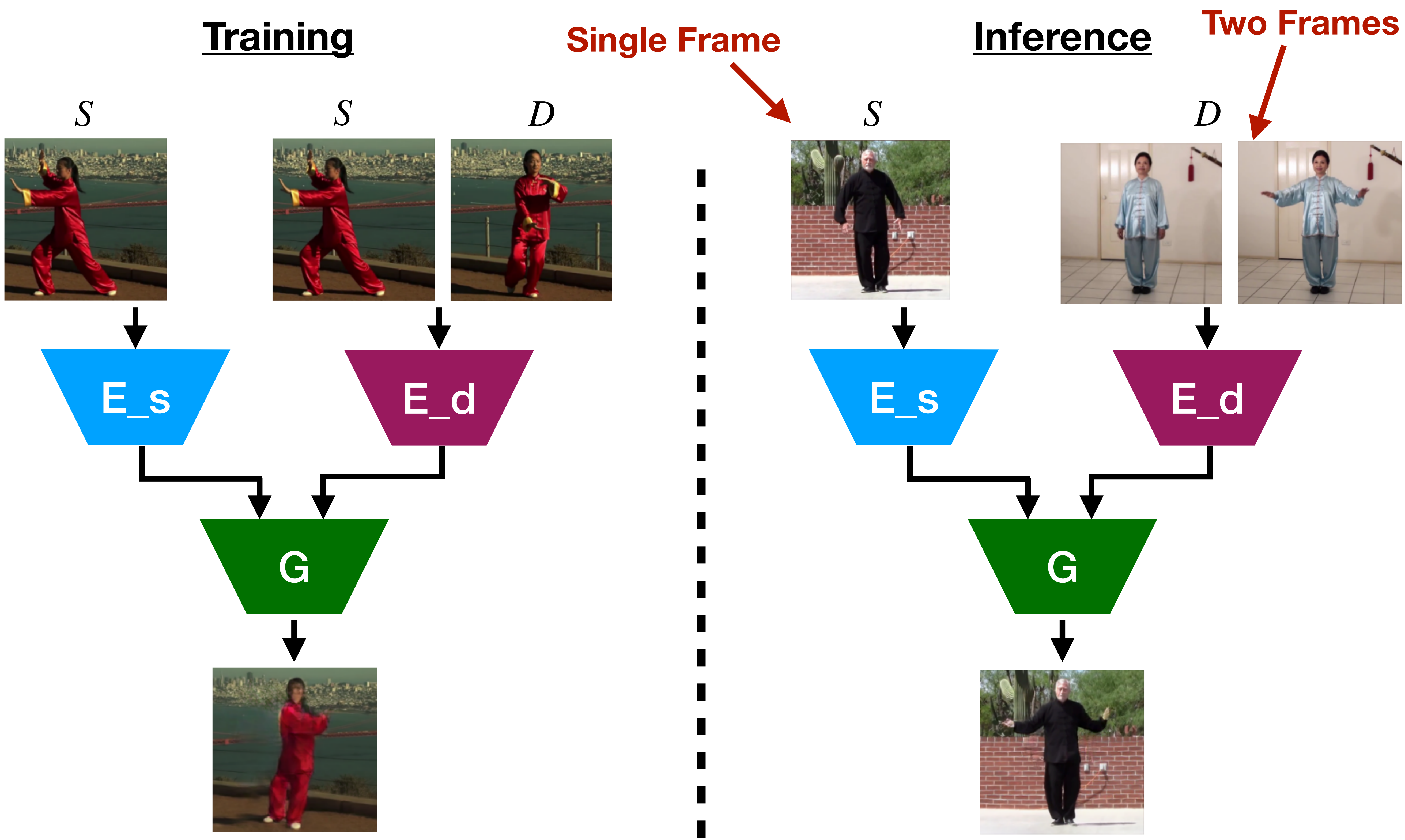
Siarohin et al. NeurIPS 2019

Motion transfer



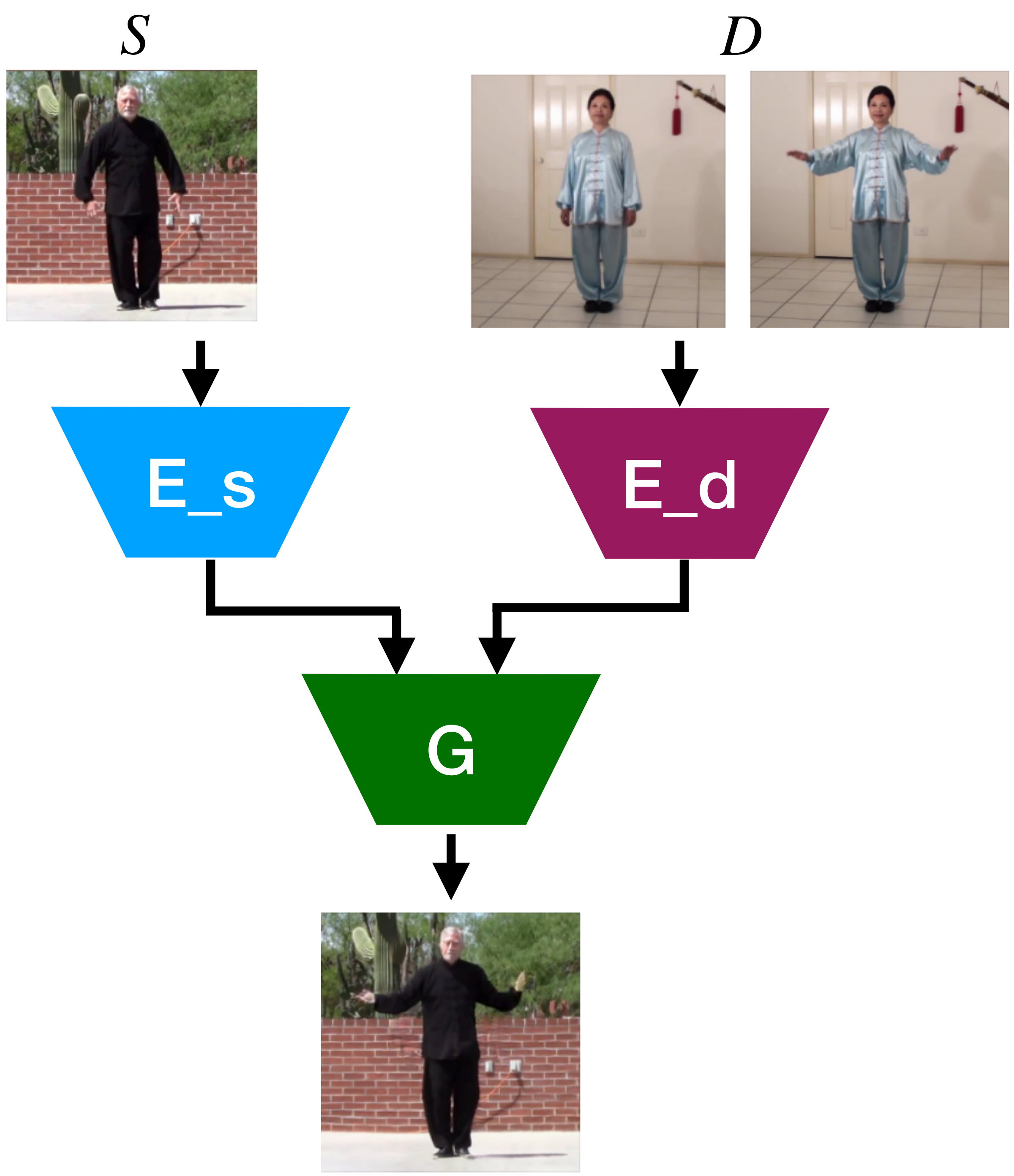
Dataset

General Approach

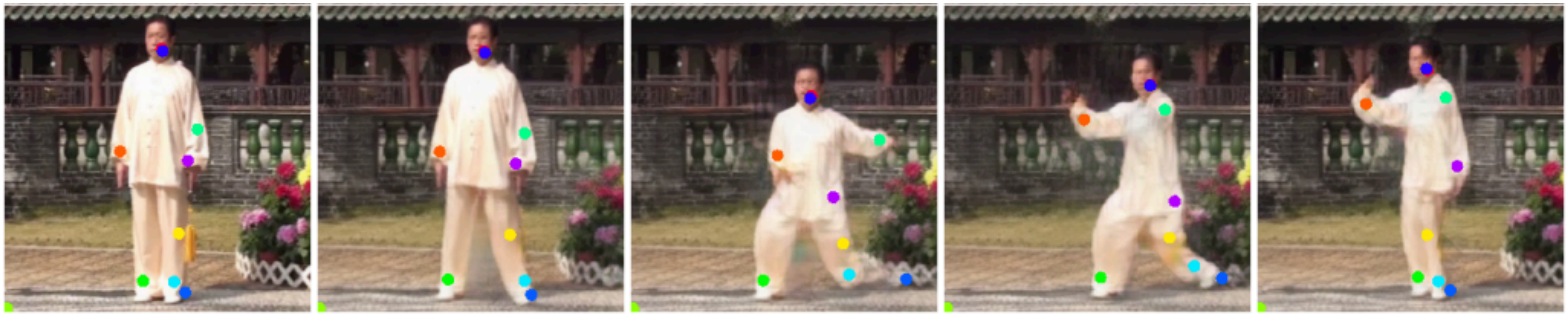


The Challenge

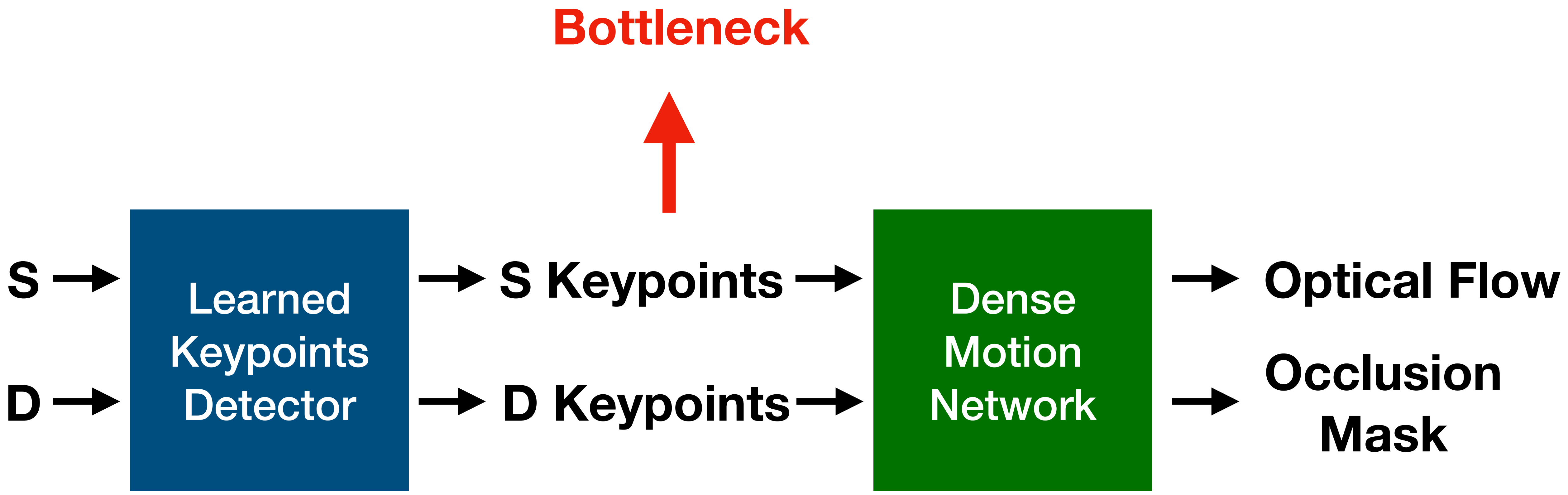
Adjust the motion representation to a different scene



Keypoints representation



Keypoints representation

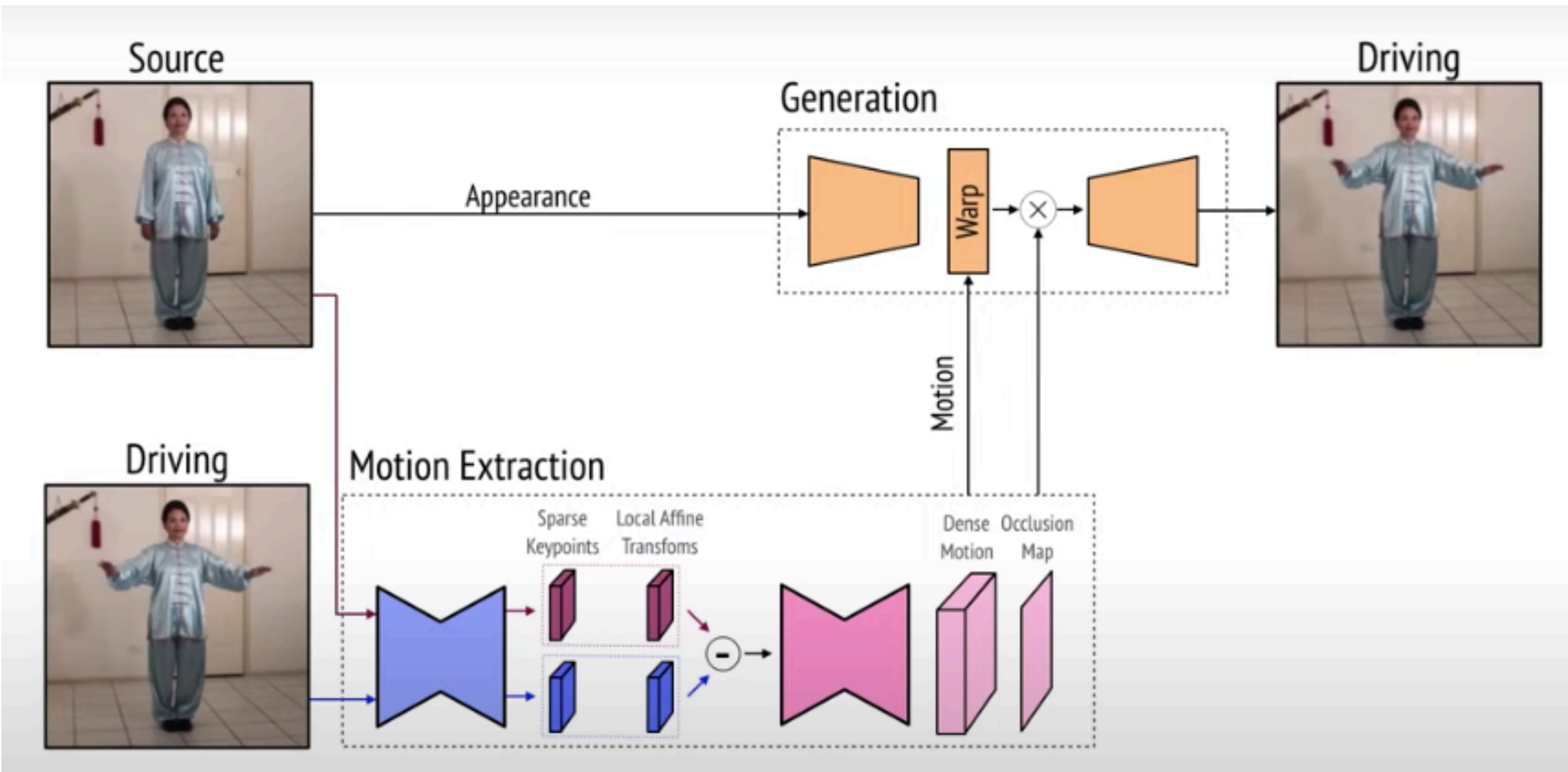


First Order Estimation

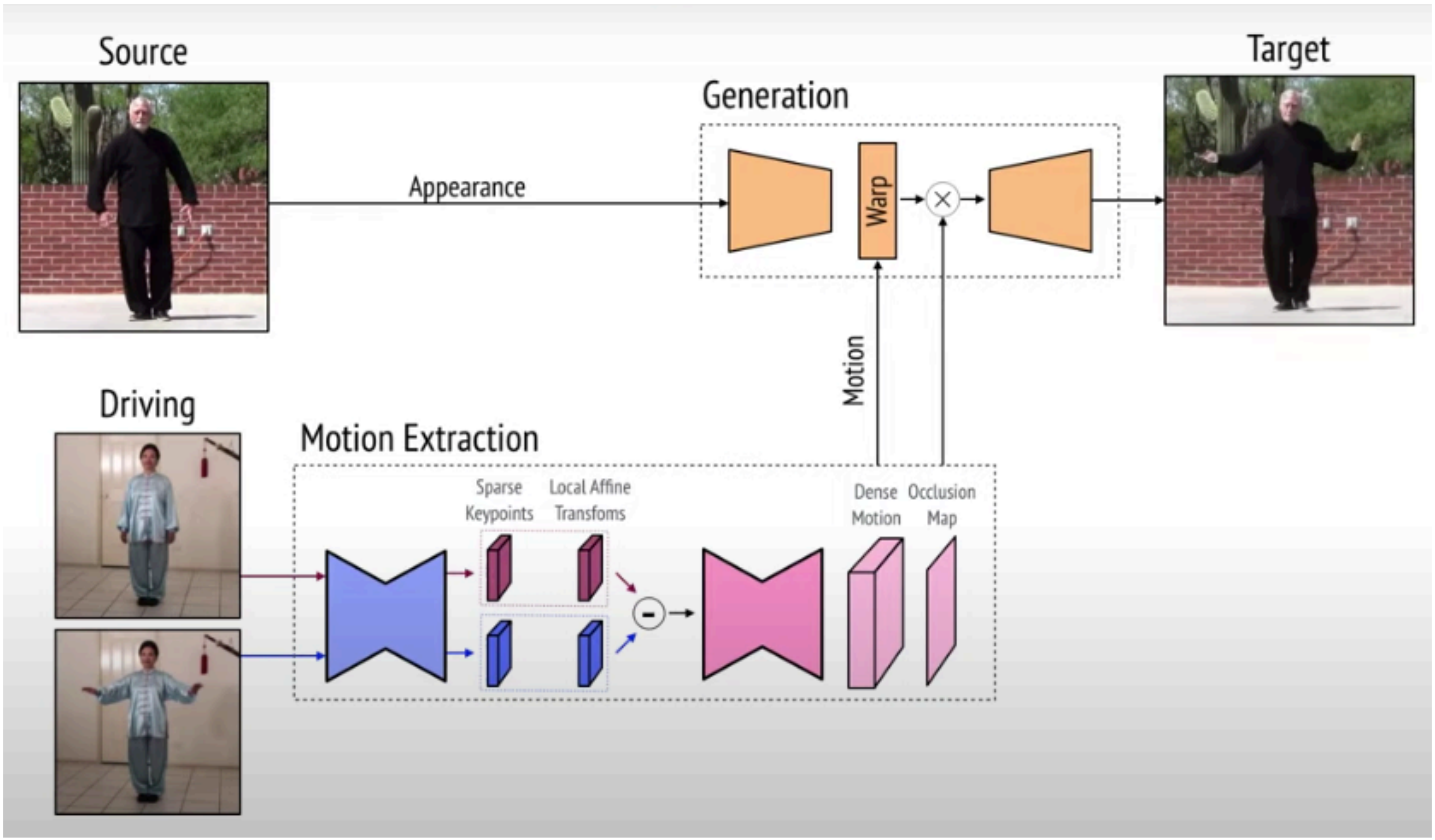
$$T(z) \approx T(p_k) + J_k(z - p_k)$$

- **Compute a transformation for each keypoint**
- **In practice, J is computed by the keypoint detector**
- **We feed the dense motion network with the transformations and source image warped according to these transformations**

Training Overview



Inference Overview



Loss Terms

- **Reconstruction**
- **Keypoints Equivariance**
(Perform transformation over the image and expect to get the keypoints after the same transformation)
- **Jacobian Equivariance**

Failure cases

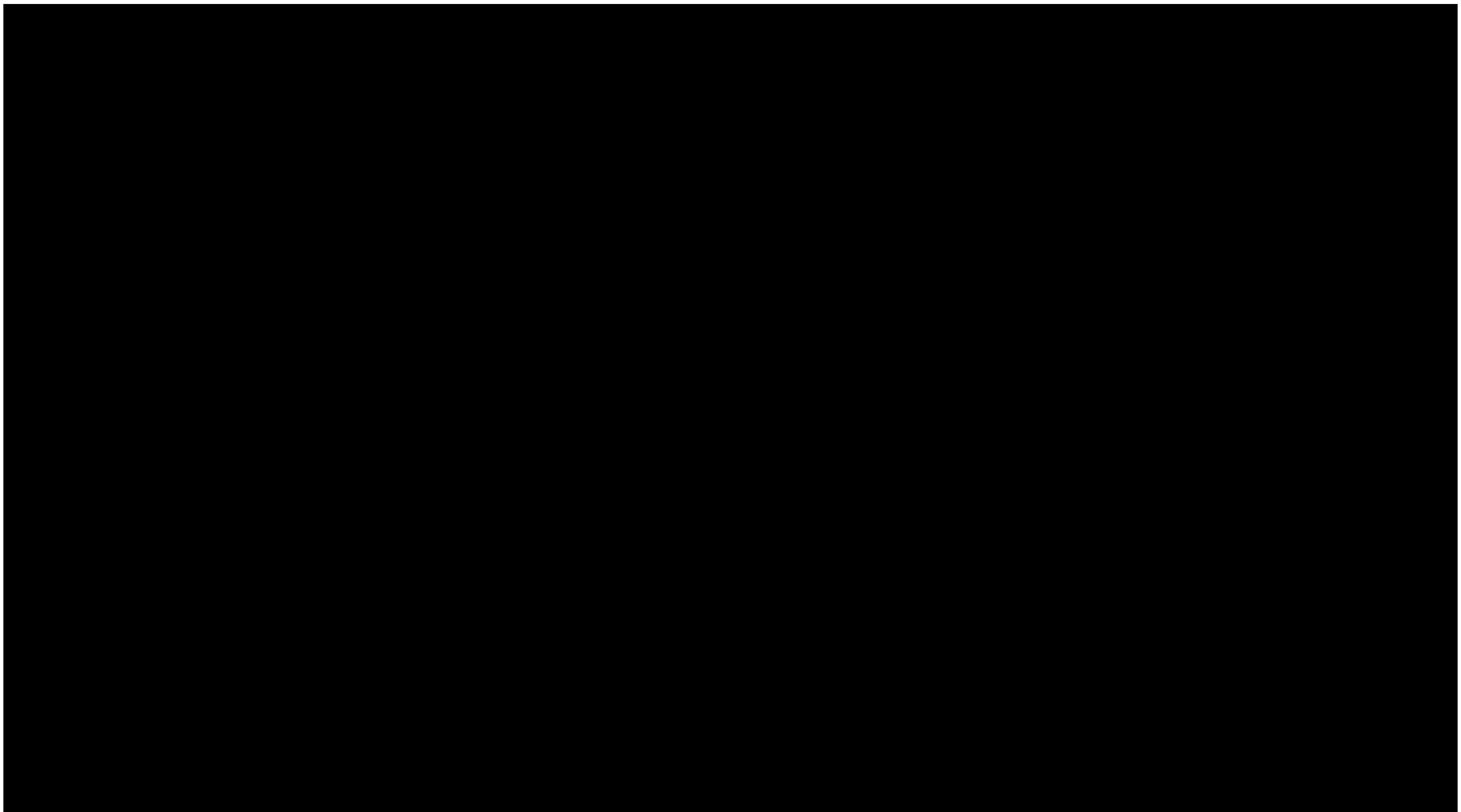
This approach assumes that the object in the first frame of the driving video and the object in the source image should be in similar poses.

Image taken from Monkey-Net paper

Questions?

Everybody Dance Now

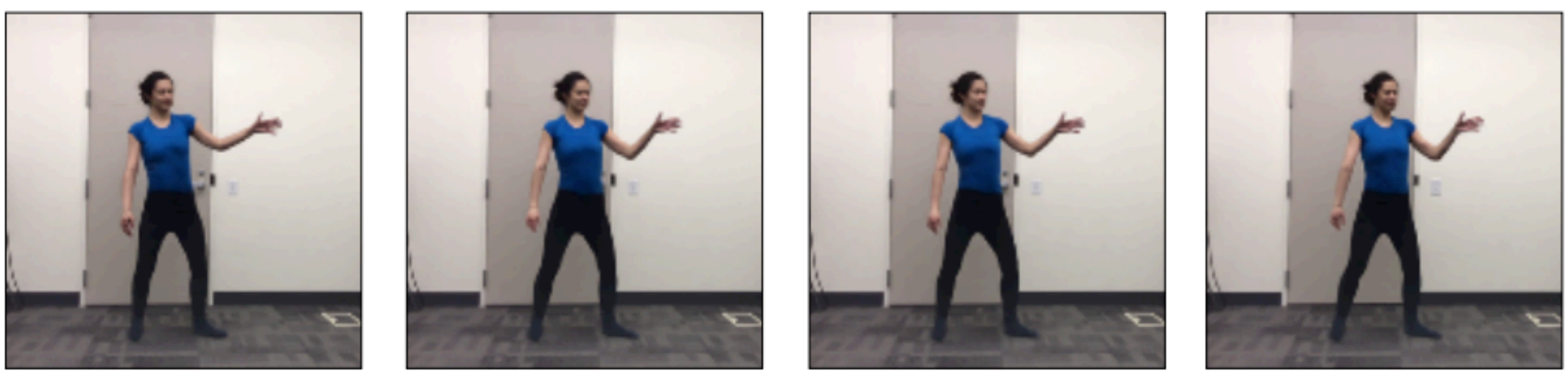
Chan et al. ICCV 2019



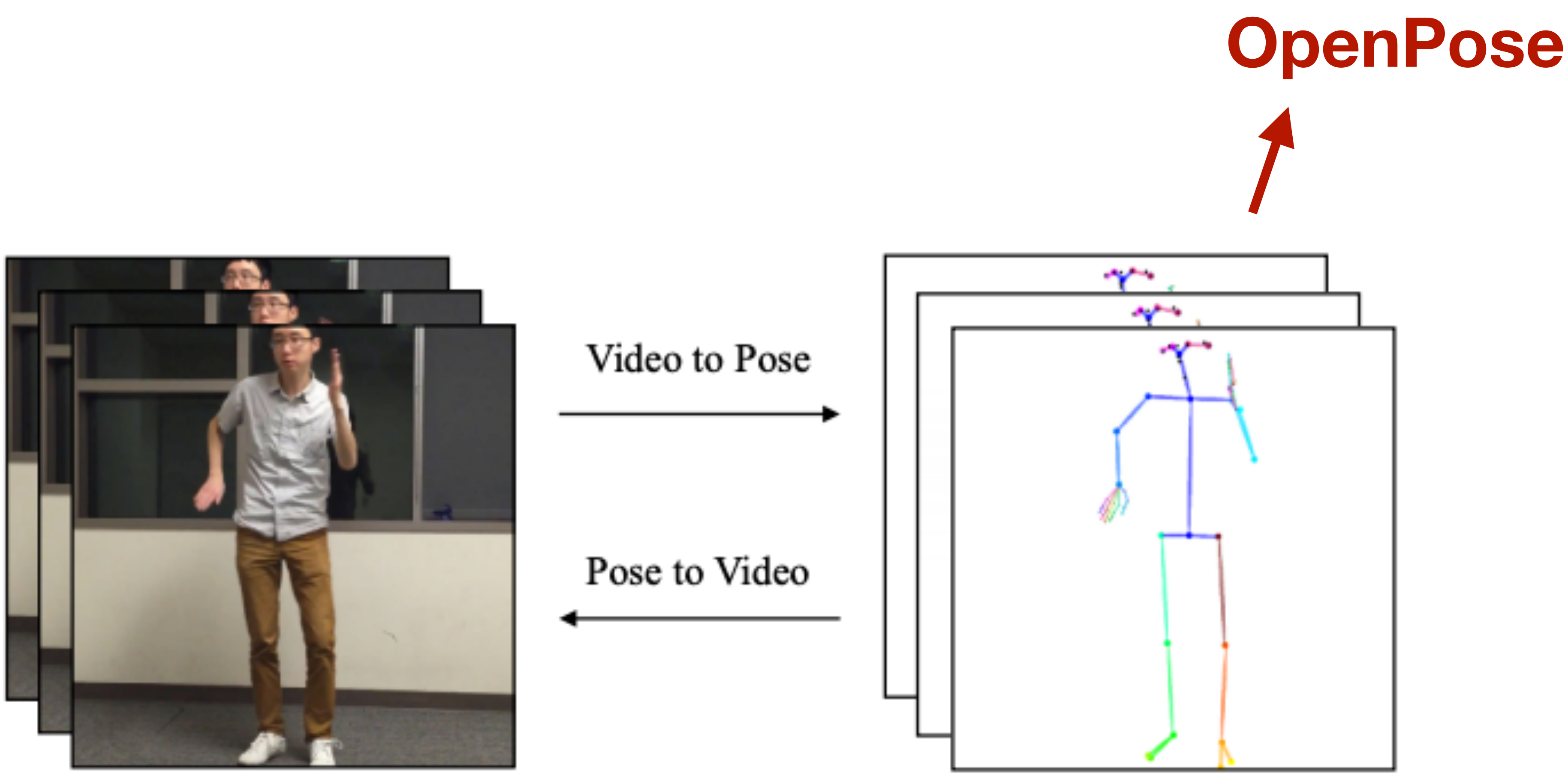
Unpaired Data

Source Video

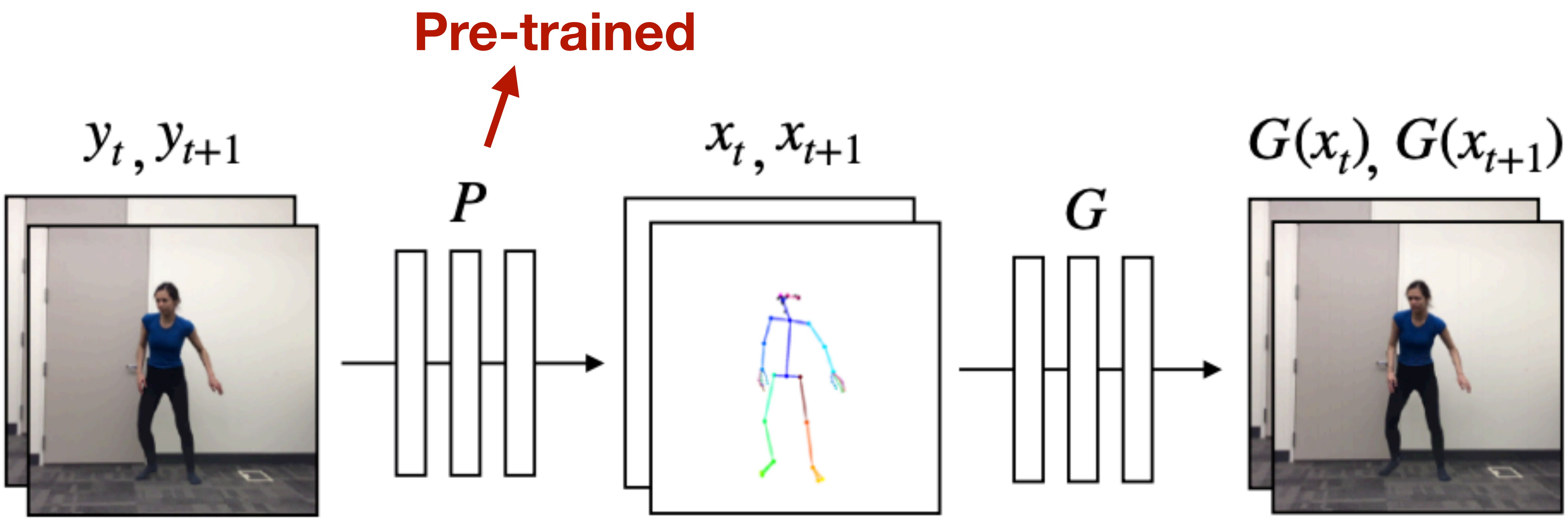
Target Video



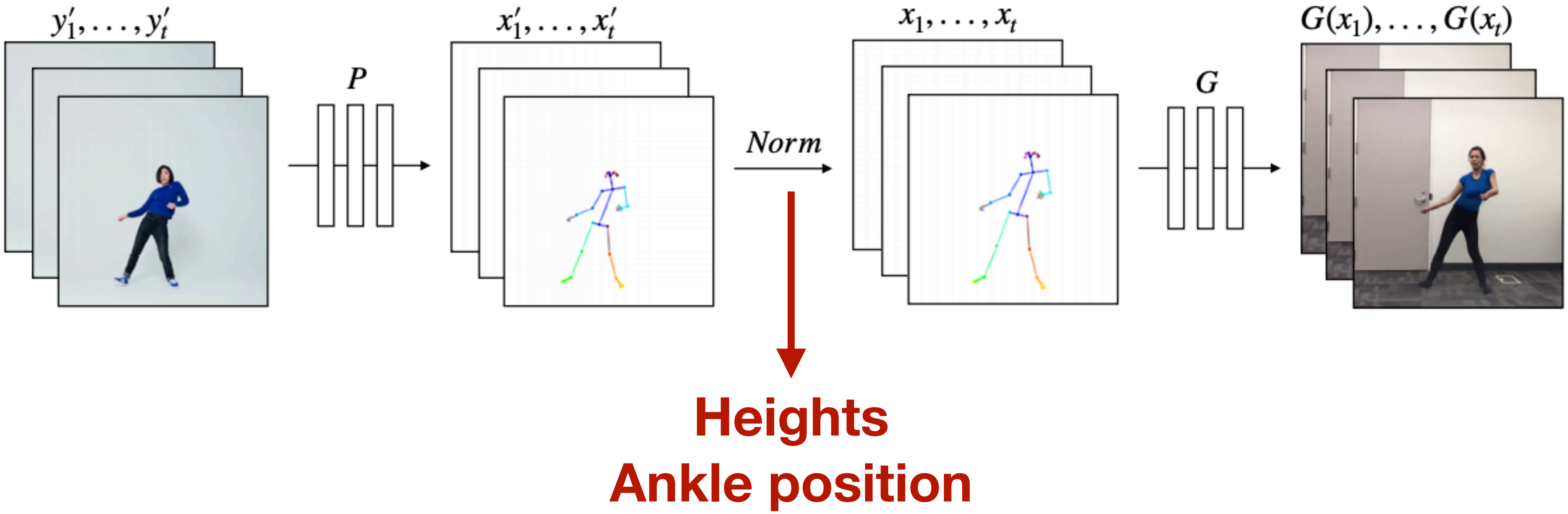
Task



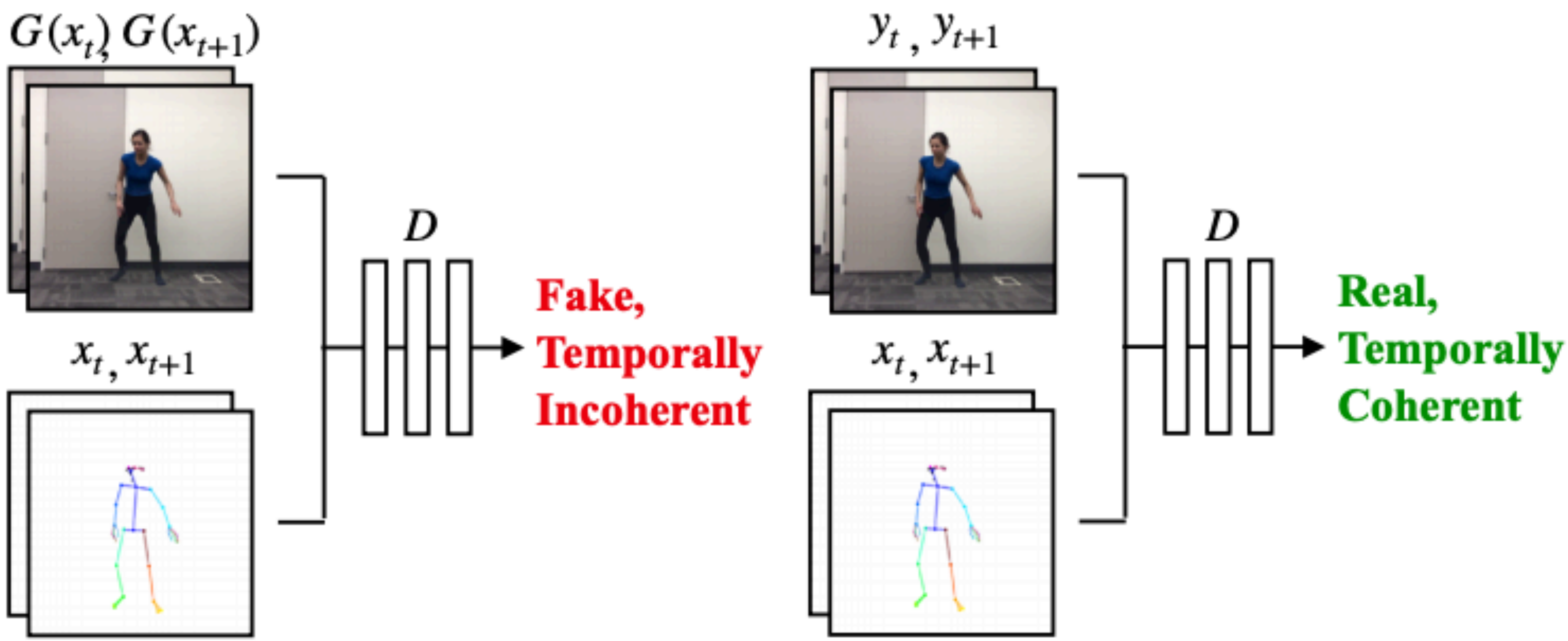
Training



Transfer


$y'_1, \ldots, y'_t$
P
$x'_1, \ldots, x'_t$
$Norm$
$x_1, \ldots, x_t$
G
$G(x_1), \ldots, G(x_t)$
Heights
Ankle position

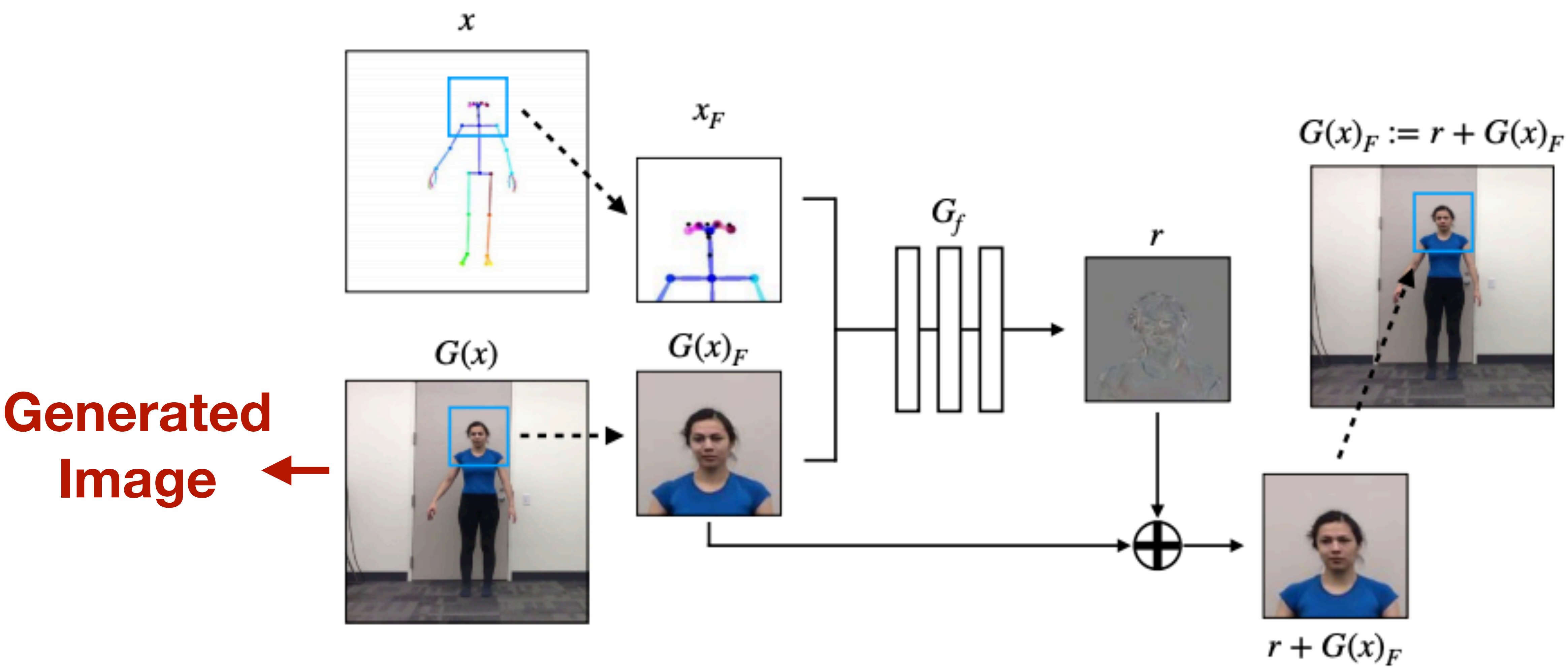
Temporal Coherence



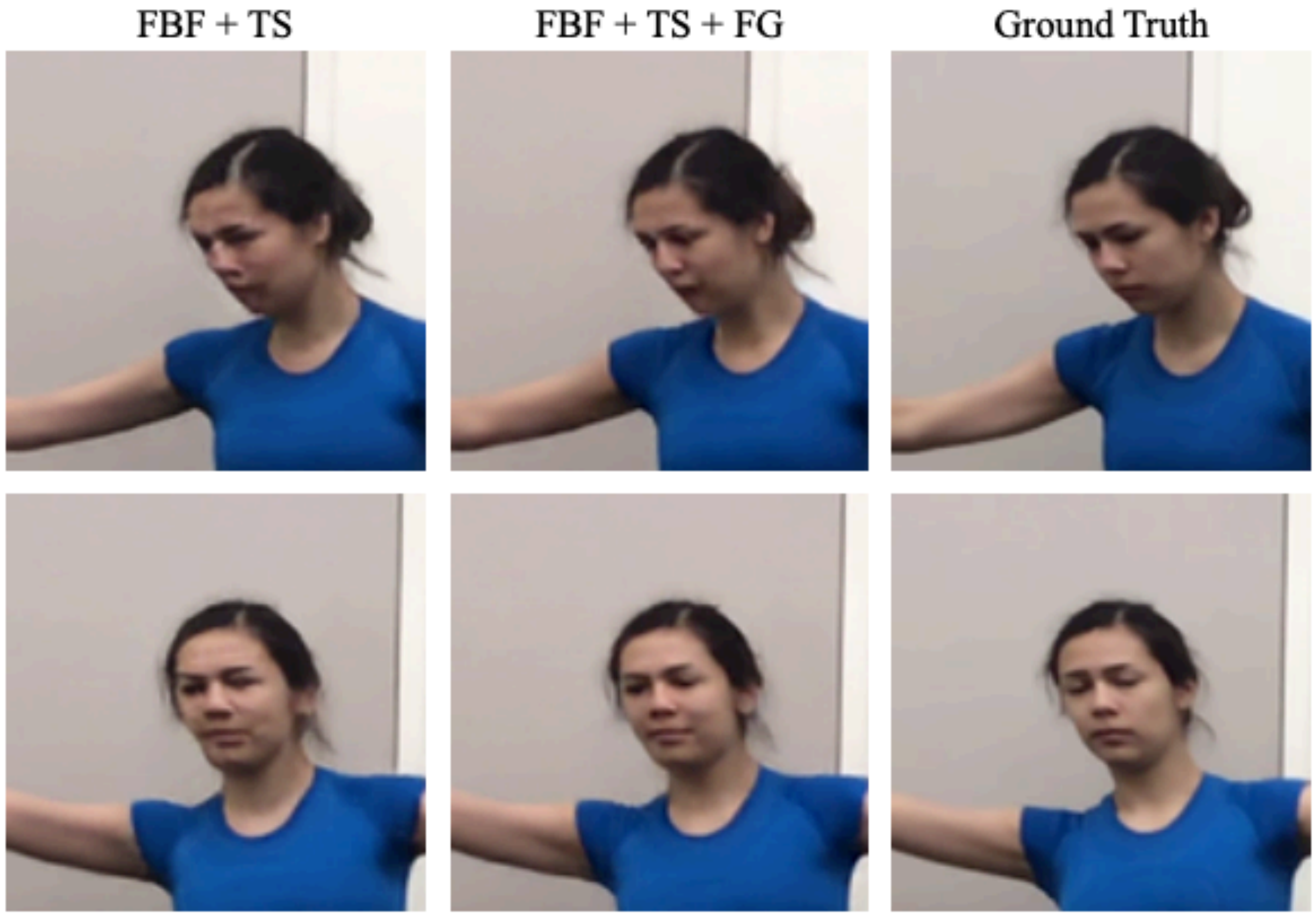
Other Loss Terms

- **Reconstruction for both frames (VGG loss)**
- **GAN loss (without temporal smoothing)**

Face GAN



- Trained using GAN
- Without temporal smoothness
- Optimize separately


FBF + TS
FBF + TS + FG
Ground Truth

Questions?