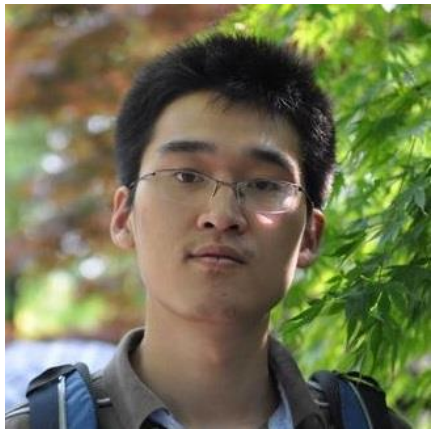
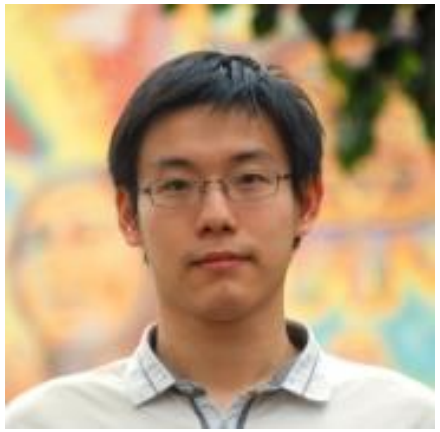
Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks



Tianfan Xue*



Jiajun Wu*

Katie Bouman

Bill Freeman

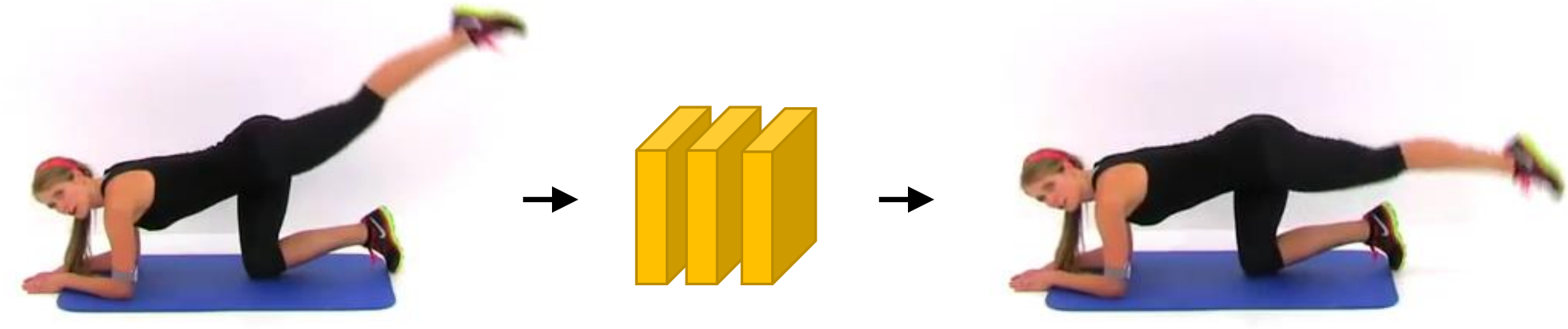
* Indicates equal contribution

Task: future frame prediction

Frame 1

Frame 2

Deterministic predictions fail to model uncertainty



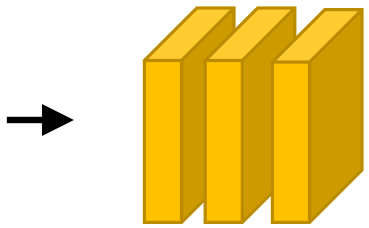
Frame 1

Deterministic neural network

Frame 2

Deterministic predictions fail to model uncertainty

Frame 1

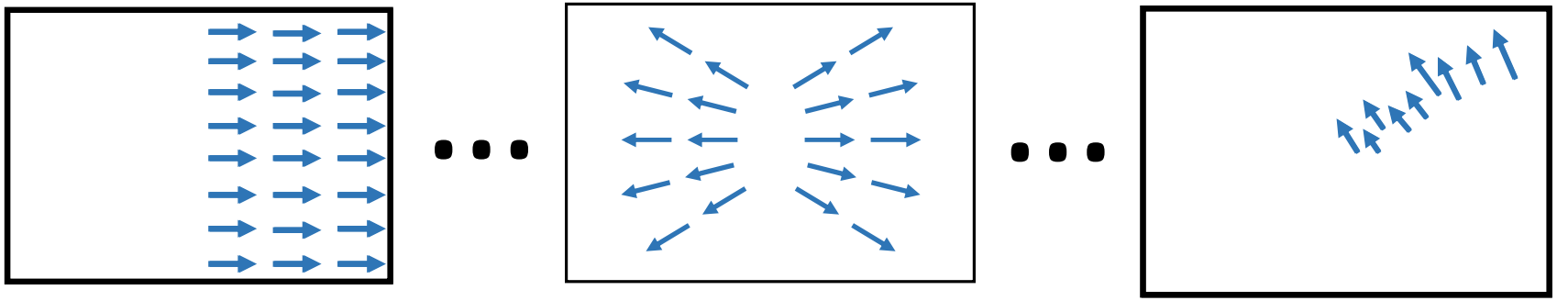


Deterministic neural network

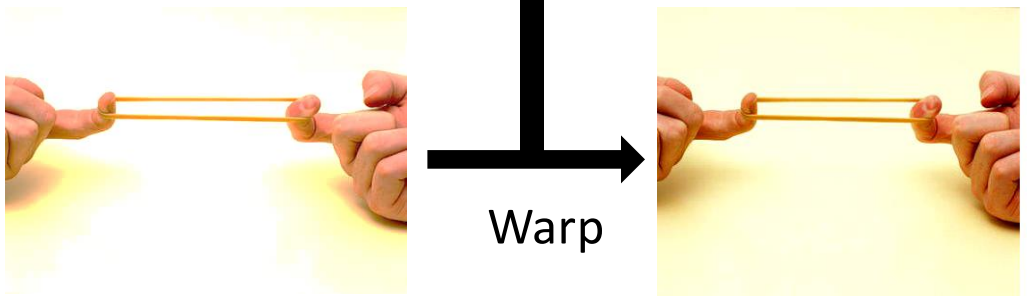
Prediction

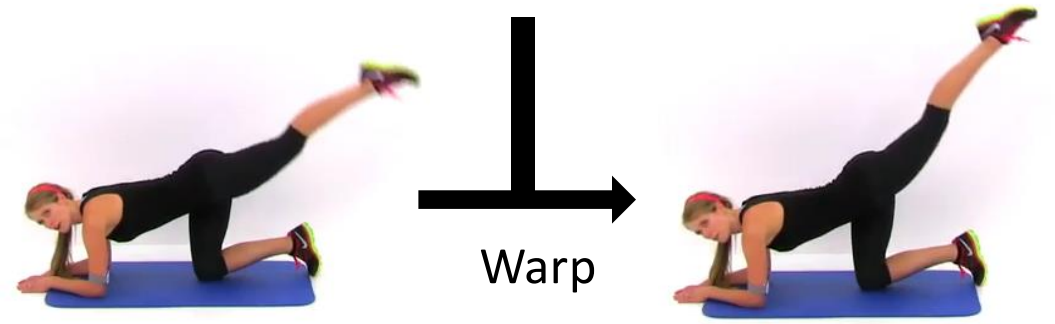
Reality

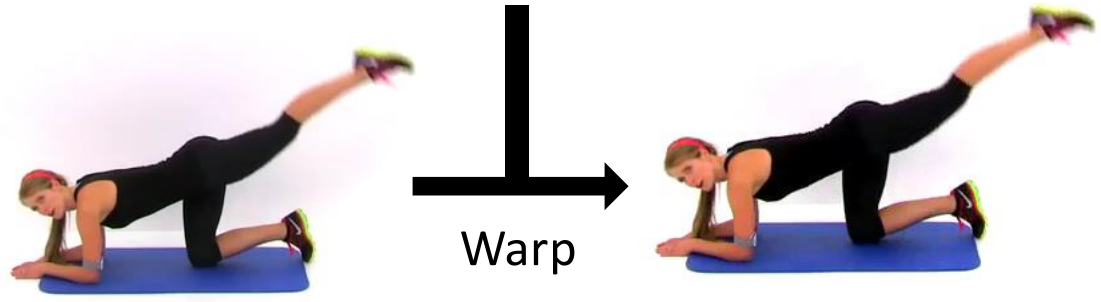
Sampling a motion field from a prior distribution



$\sim P_{prior}(\text{motion})$

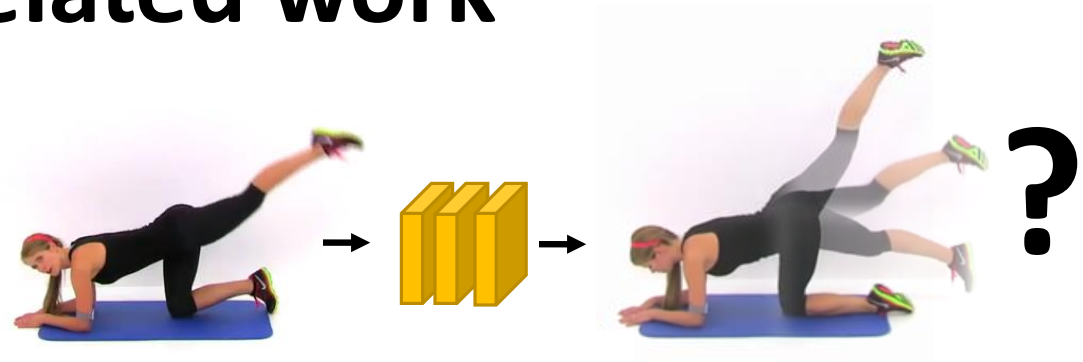




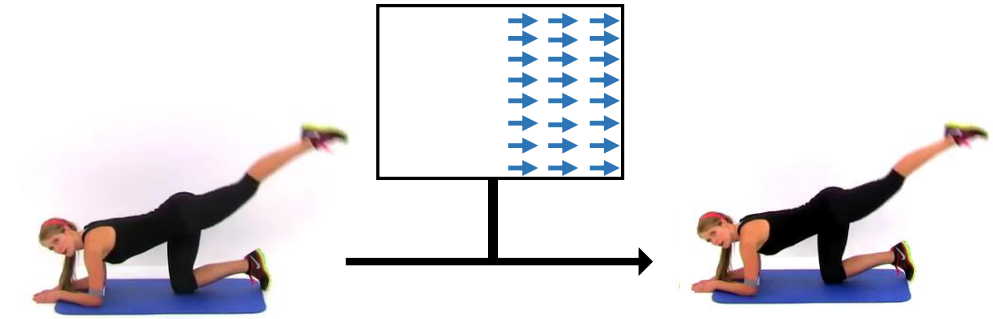


Only a few motion fields are consistent with the input image

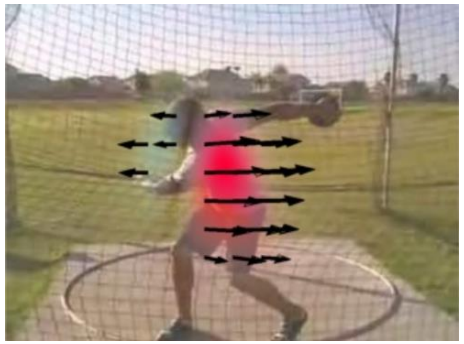
Related work

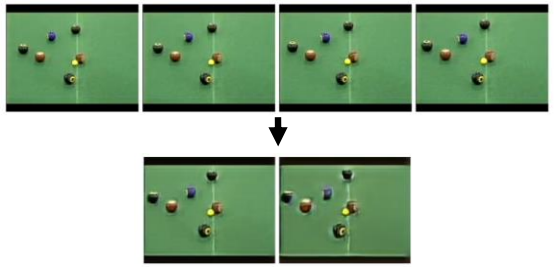


Deterministic prediction

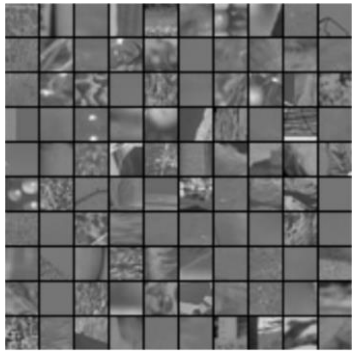


Sample from prior distribution





Motion prediction: *[Pintea et al., 2014], [Walker et al. 2015]*
Visual feature prediction: *[*Vondrick *et al., 2014]*
Future frame synthesis: *[*Mathieu *et al., 2014]*



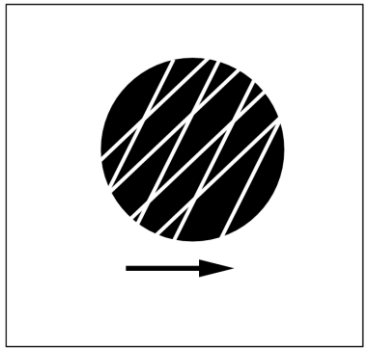
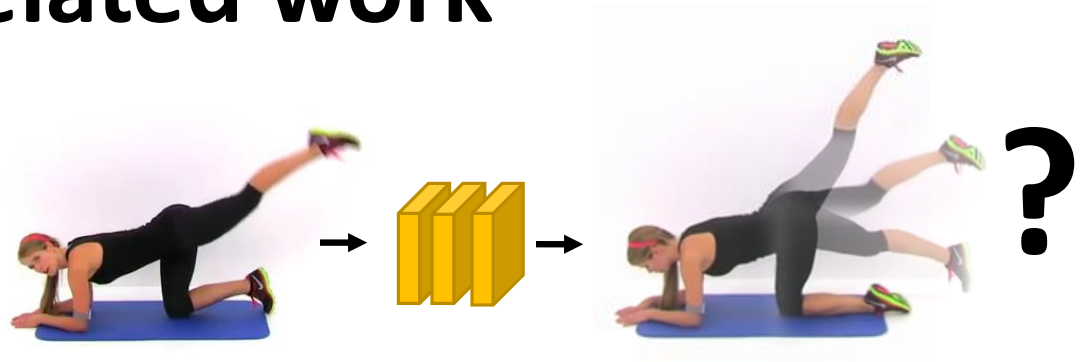


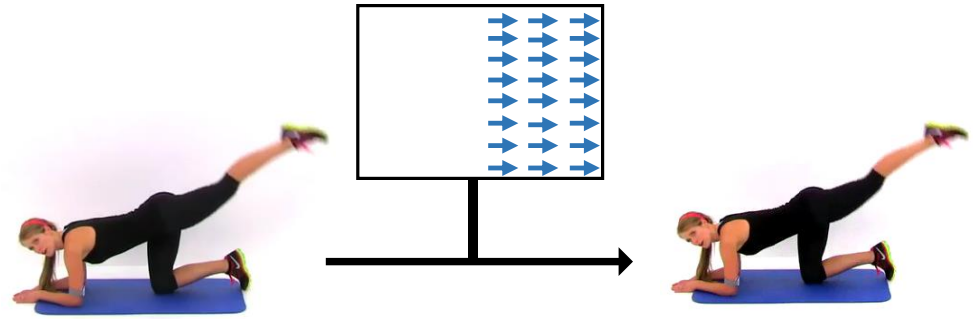
Image prior: *[Simoncelli 2001], [Zoran 2012]*
Motion prior: *[Weiss & Adelson, 1998], [Fleet 2000]*
Image synthesis: *[Portilla and Simoncelli, 2000], [Kingma and Welling, 2014], [Radford 2015], [Oord 2016]*

Probabilistic prediction: *[*Walker *et al., 2016]*

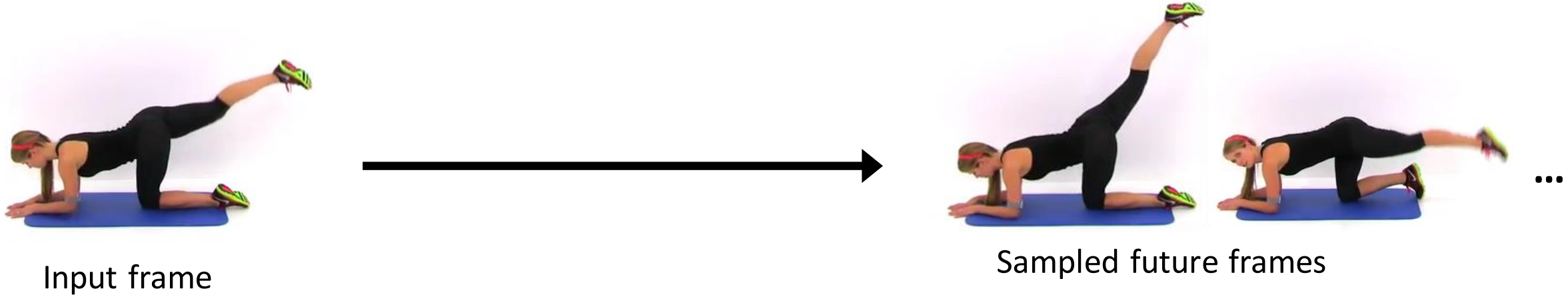
Related work



Deterministic prediction



Sample from prior distribution



Input frame

Sampled future frames

Our approach

Task: sample future frames consistent with the input

Input frame

Sampled future frames

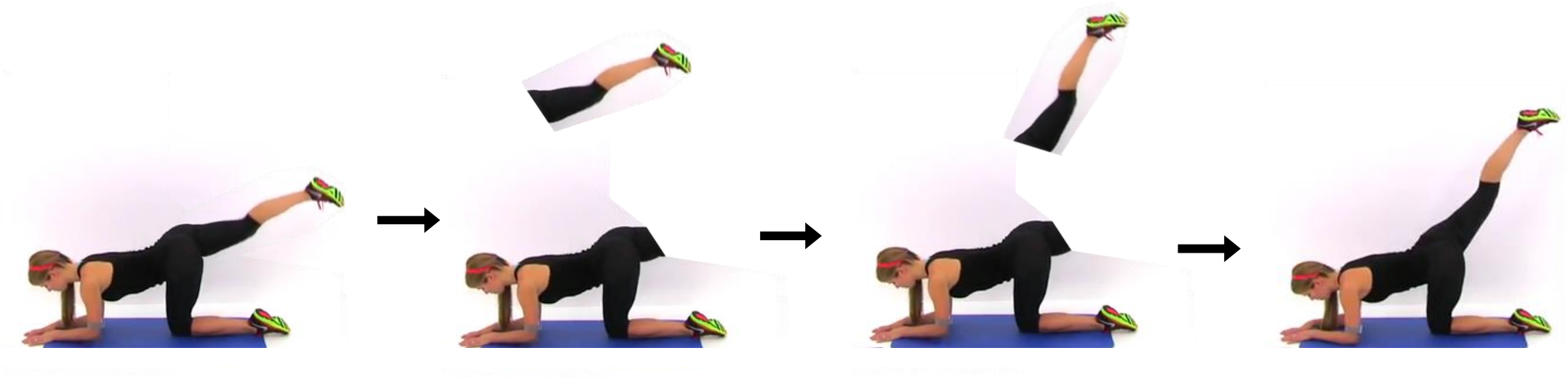
...

Segment-based synthesis

Input frame

Sampled
future frame

Segment-based synthesis



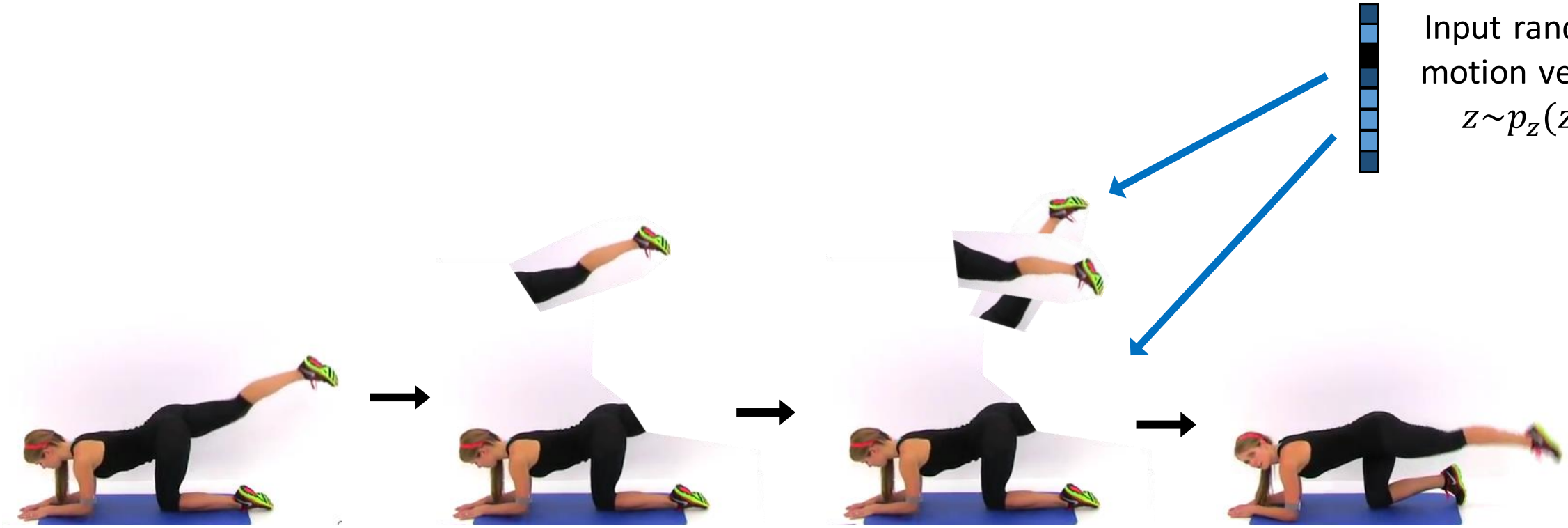
Input frame

Segments

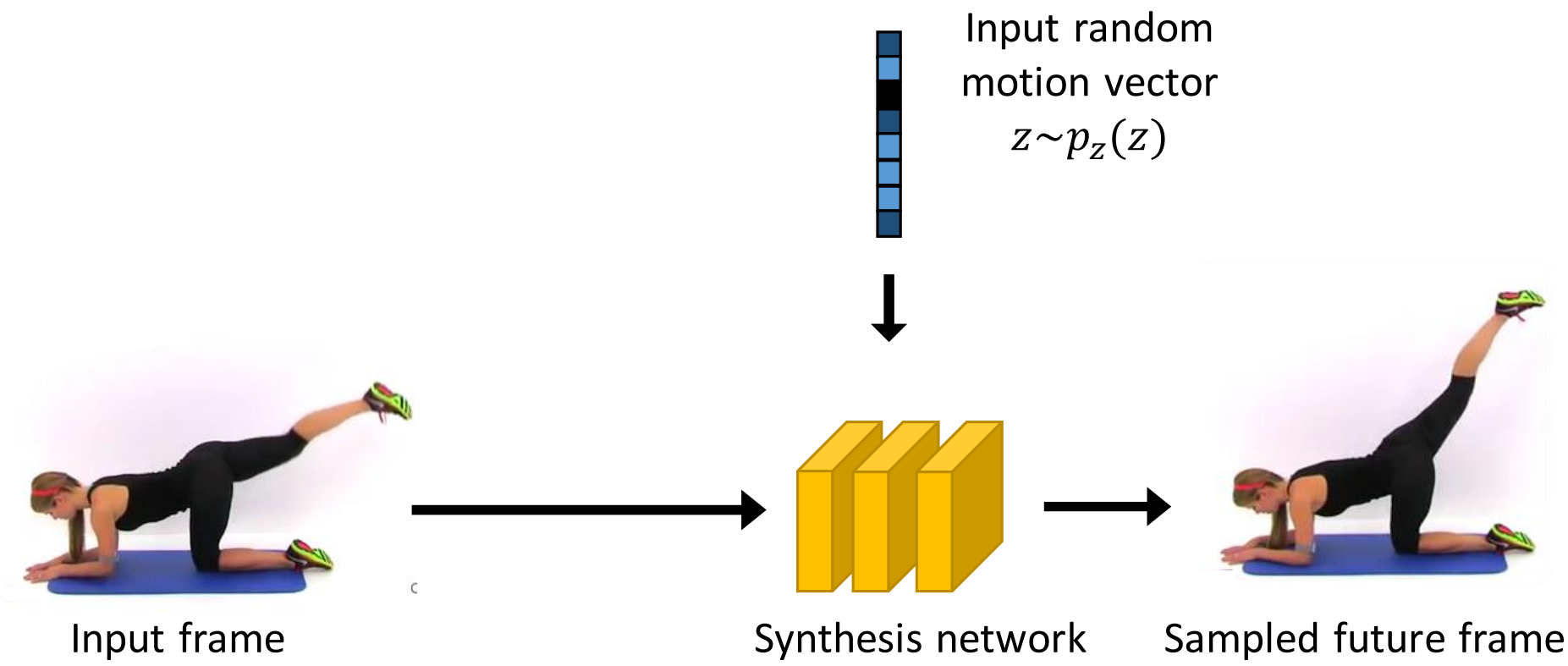
Transformed segments

Sampled future frame

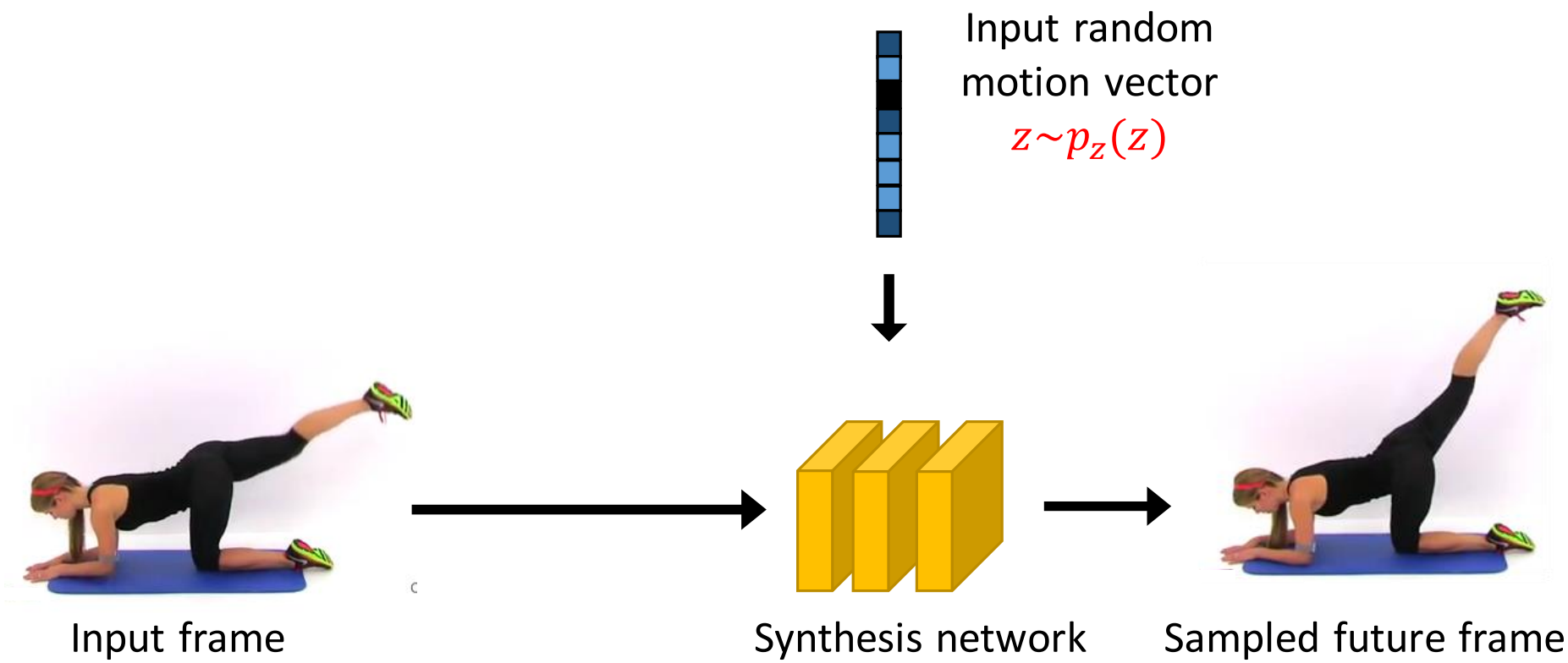
Synthesize using different transformations



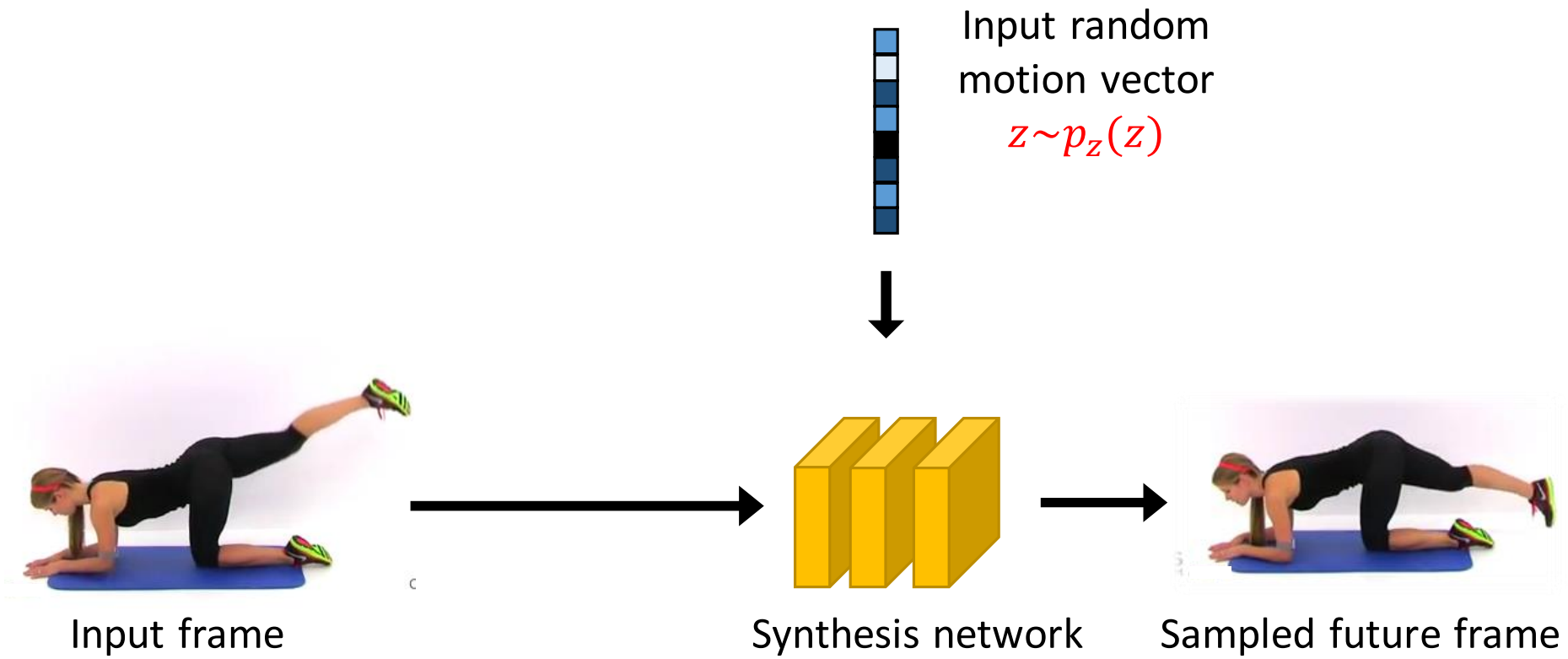
Synthesis network



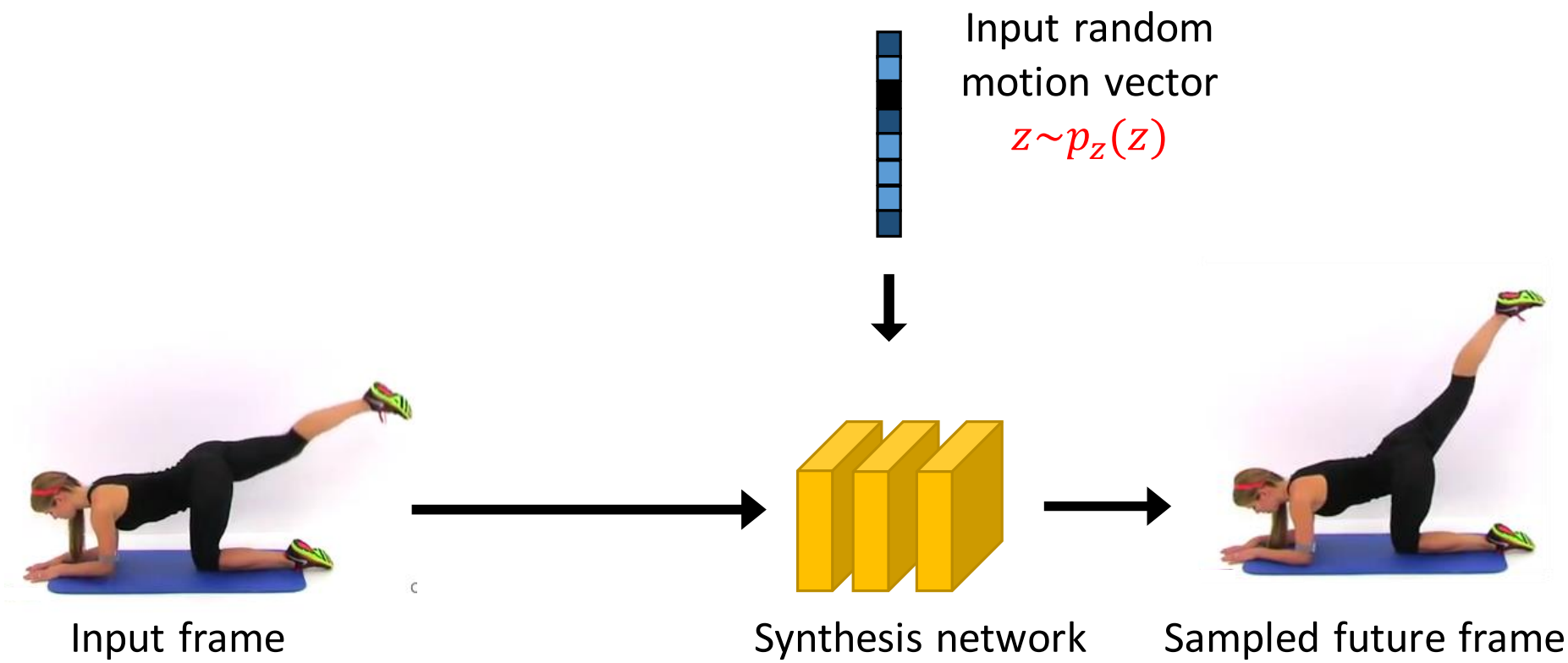
Sample different future frames



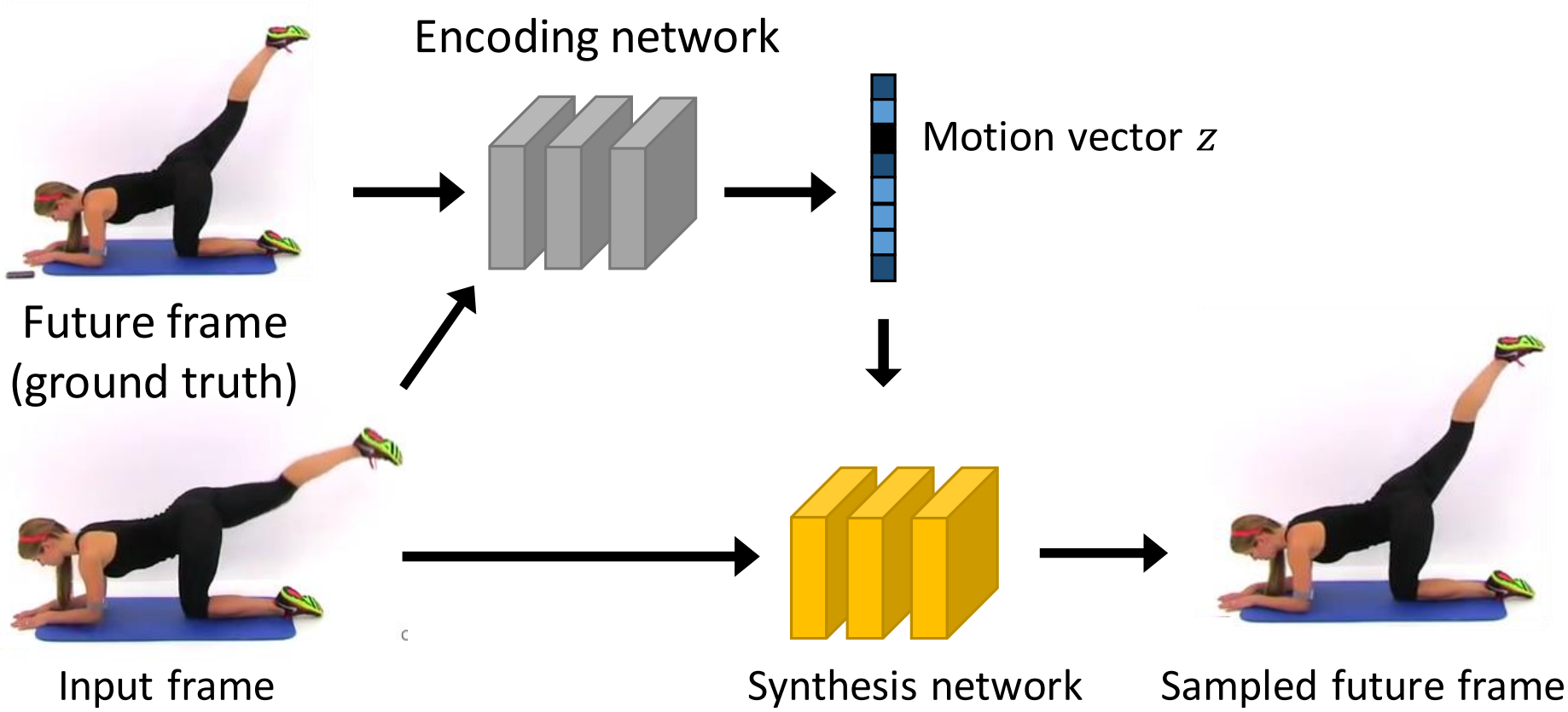
Sample different future frames



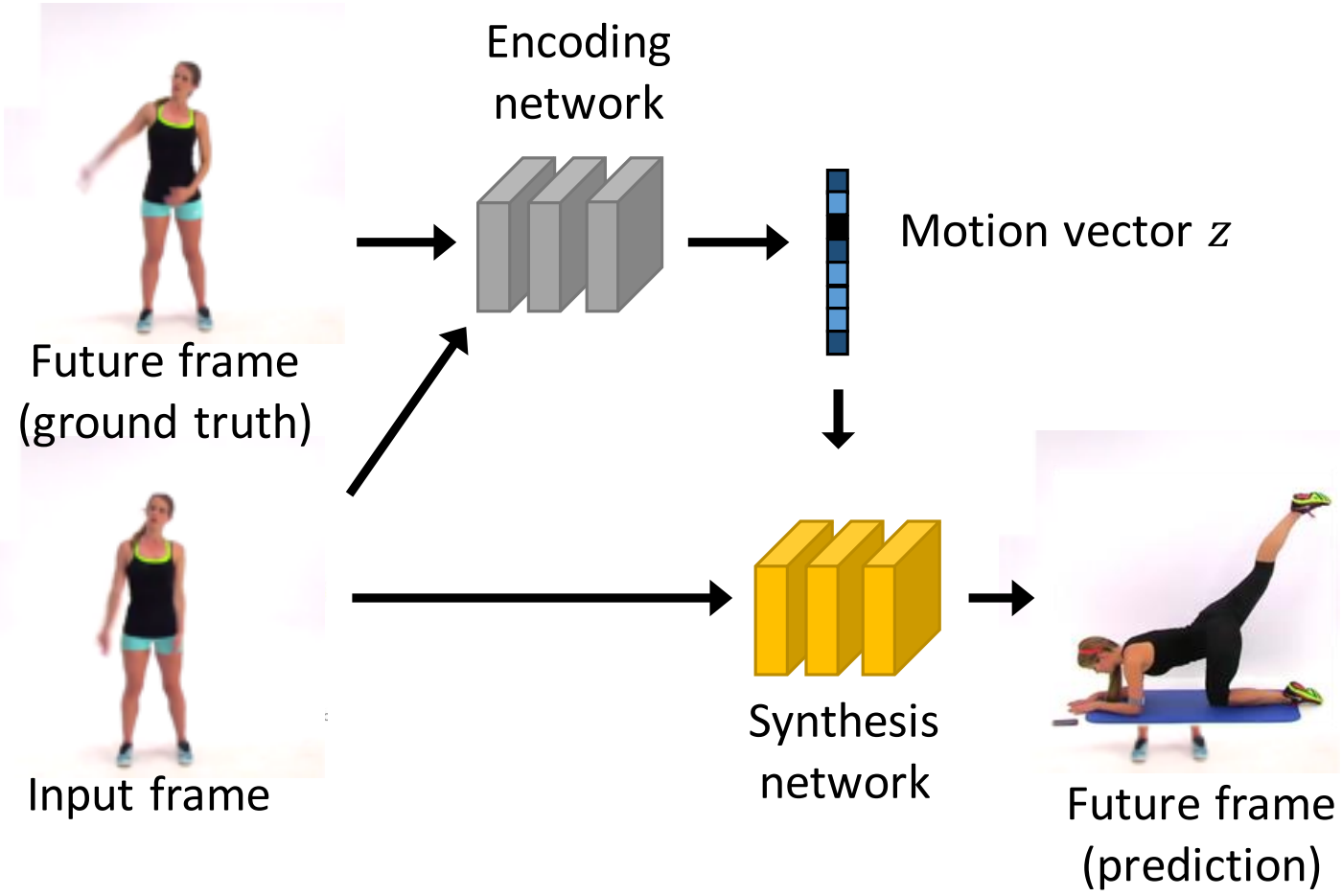
Sample different future frames



Training


Encoding network
Motion vector z
Future frame
(ground truth)
Input frame
Synthesis network
Sampled future frame

Training



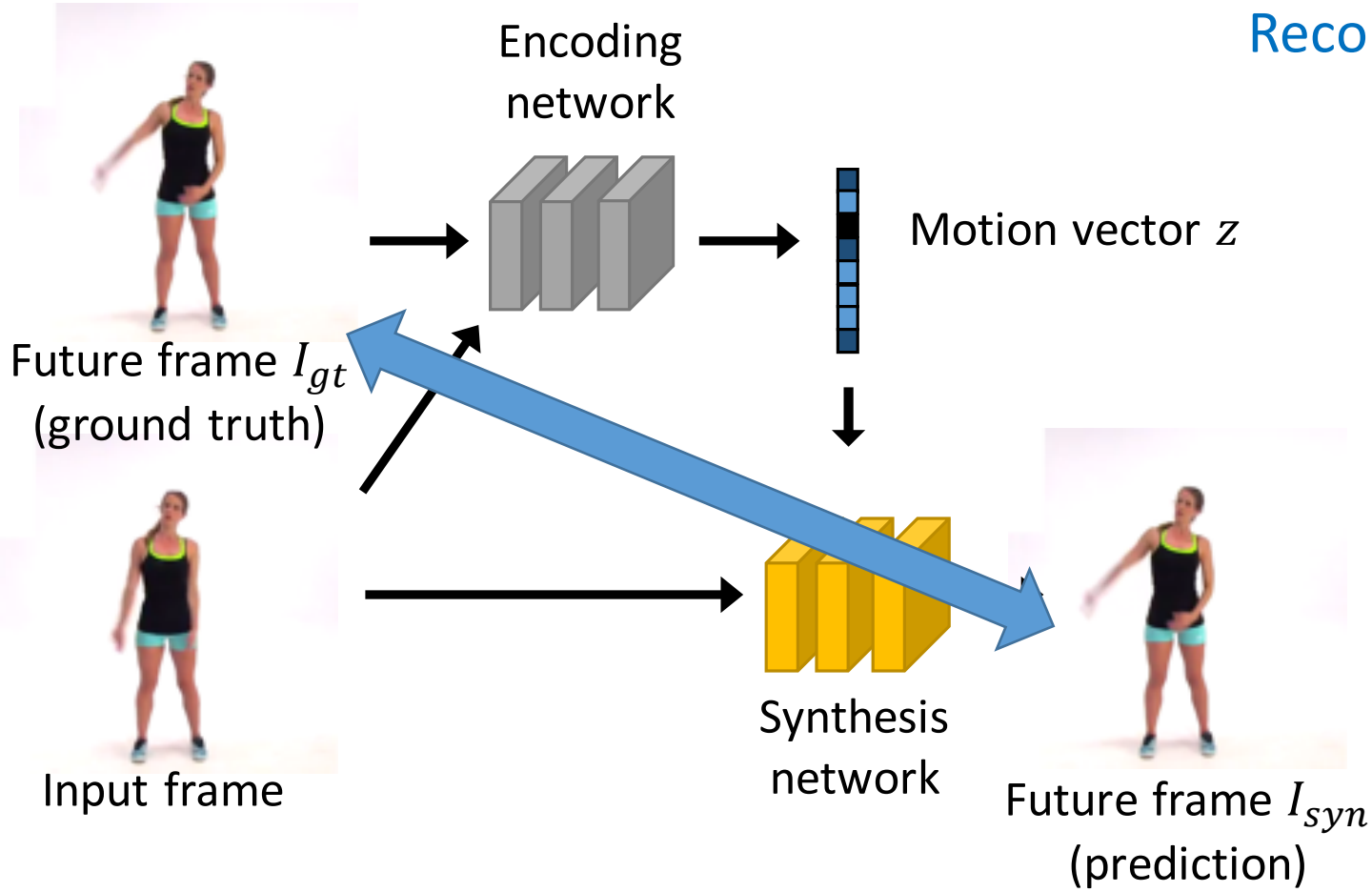
Training samples
(Label-free)

Training

Objective function:

$$\|I_{syn} - I_{gt}\| + D_{KL}(\mathbf{z}||N(\mathbf{0}, \mathbf{I}))$$

Reconstruction loss

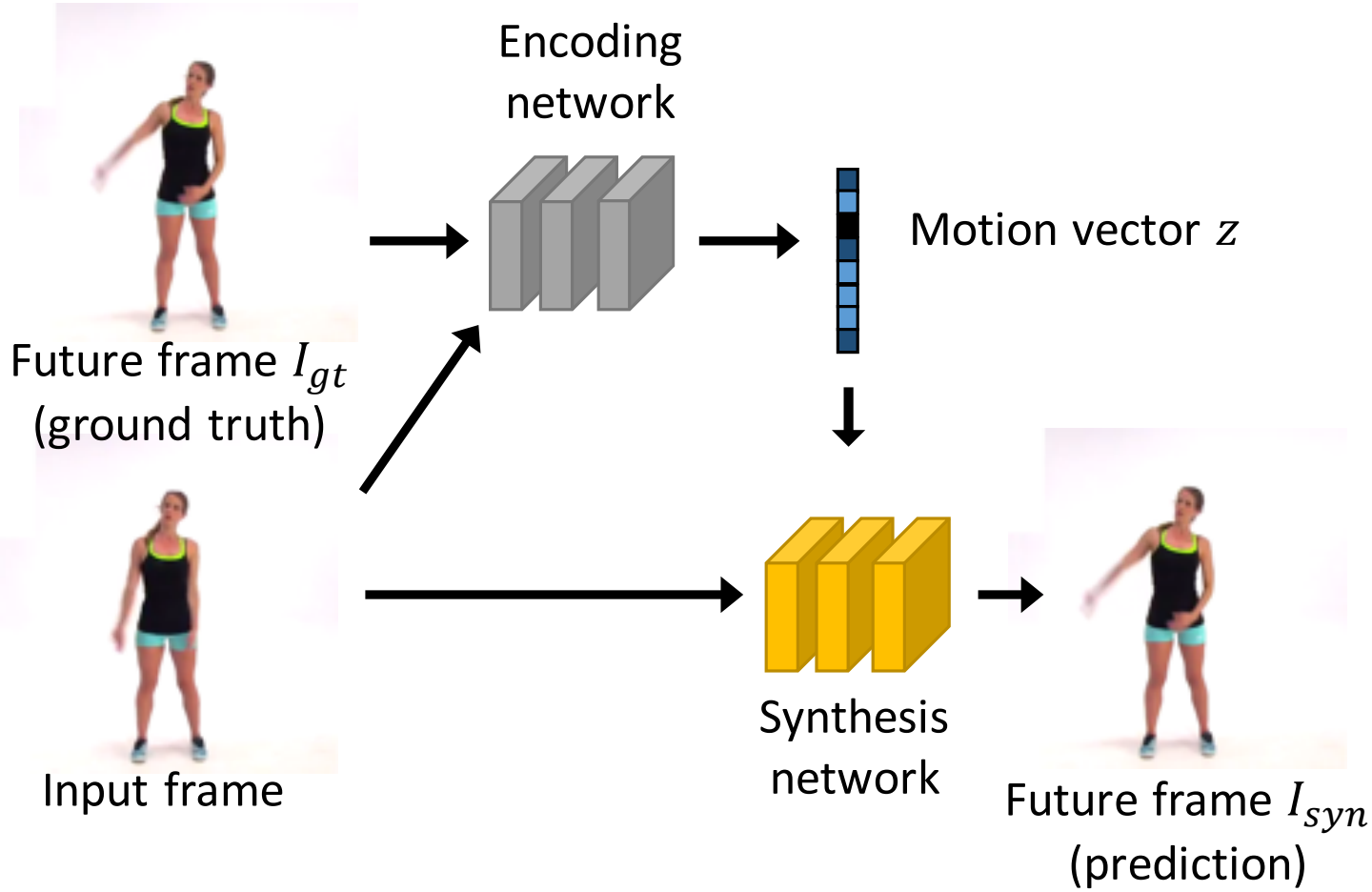


Training

Objective function:

$$\|I_{syn} - I_{gt}\| + D_{KL}(\mathbf{z}||N(\mathbf{0}, \mathbf{I}))$$

KL-divergence loss



Variational Autoencoder [Kingma and Welling, 2014]

Testing

Future frame I_{gt}
(ground truth)

Encoding network

Input random motion vector
$z \sim p_z(z)$

Input frame

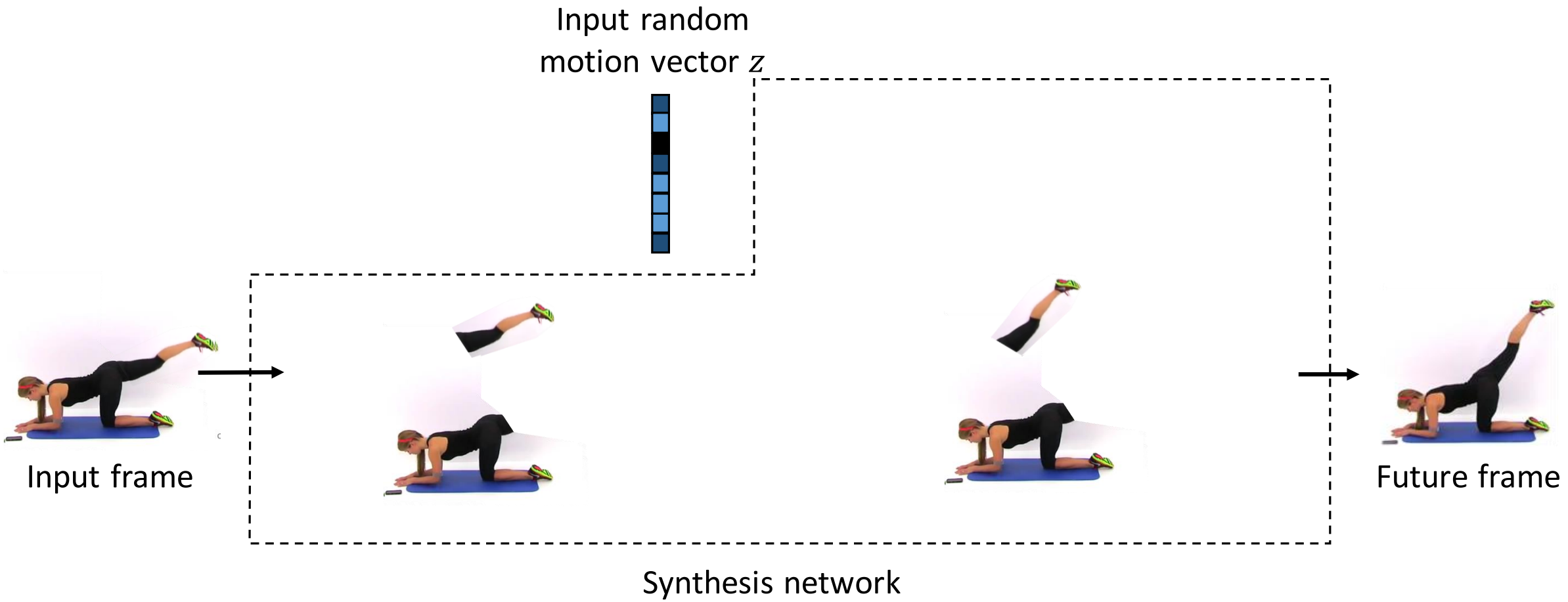
Synthesis network

Future frame I_{syn}
(prediction)

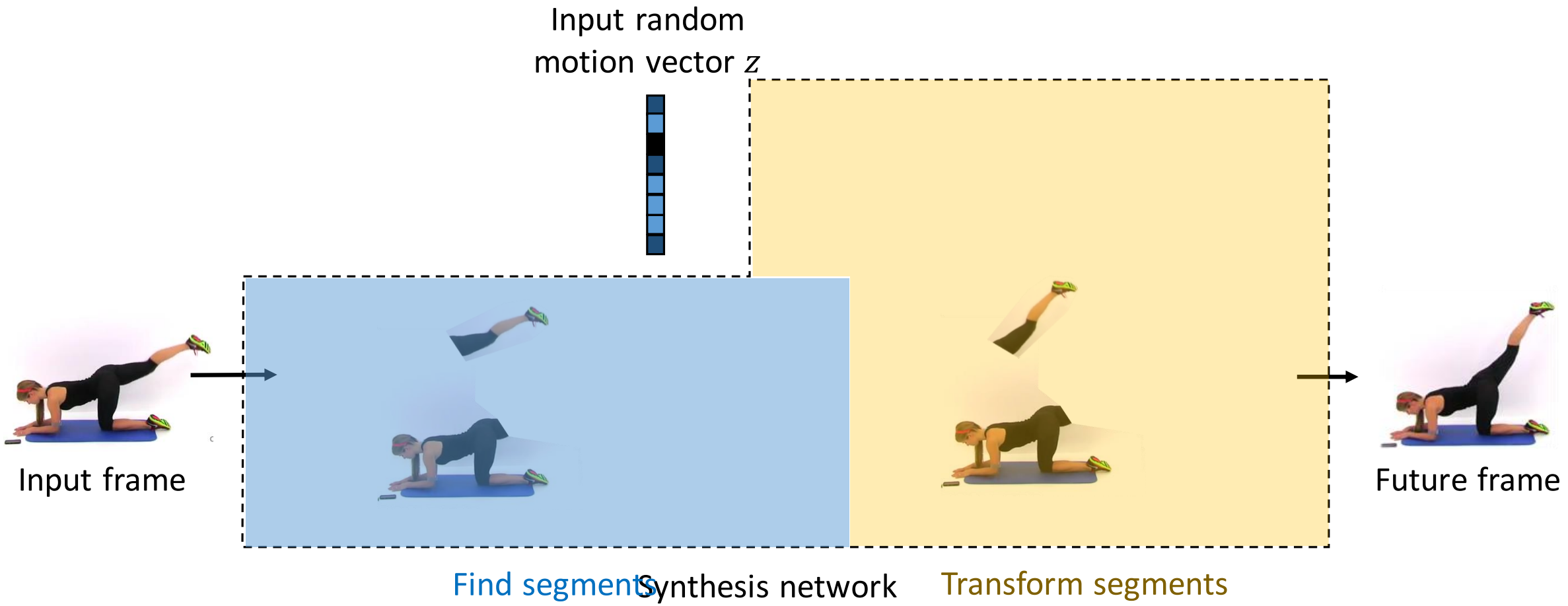
Input frame

Real output from our network

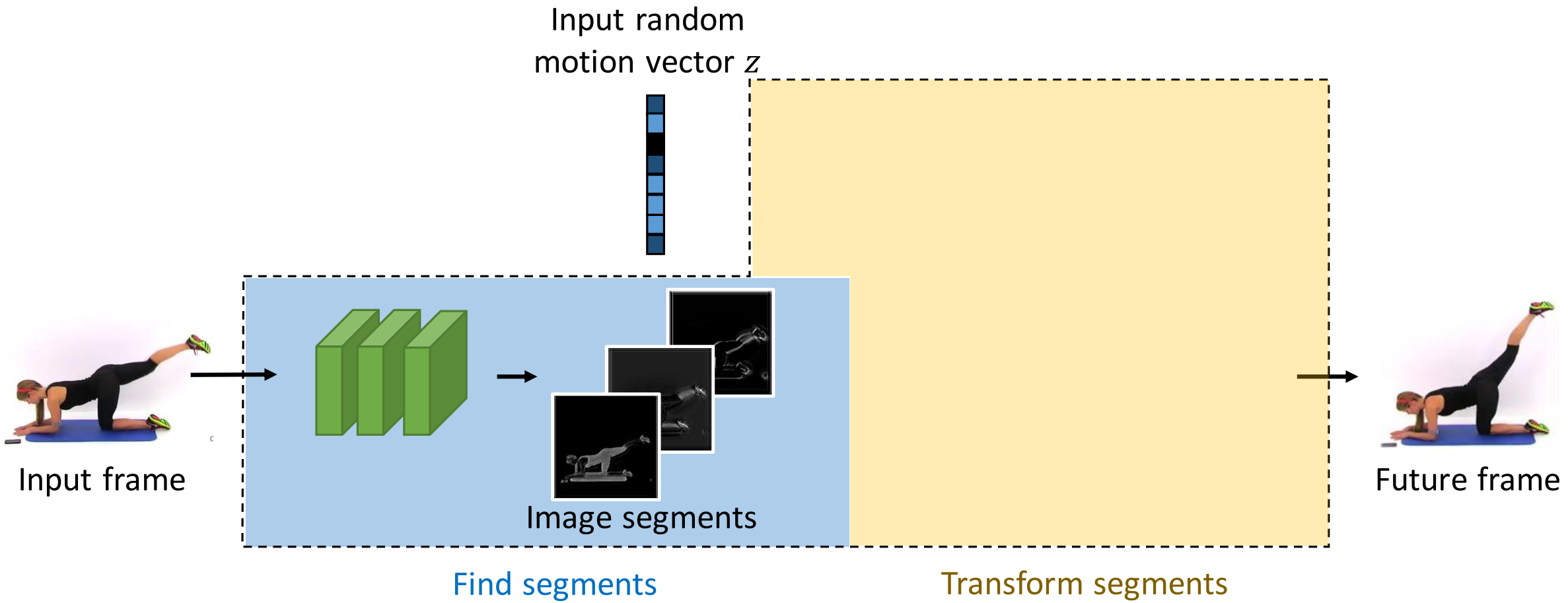
How do we design the synthesis network?



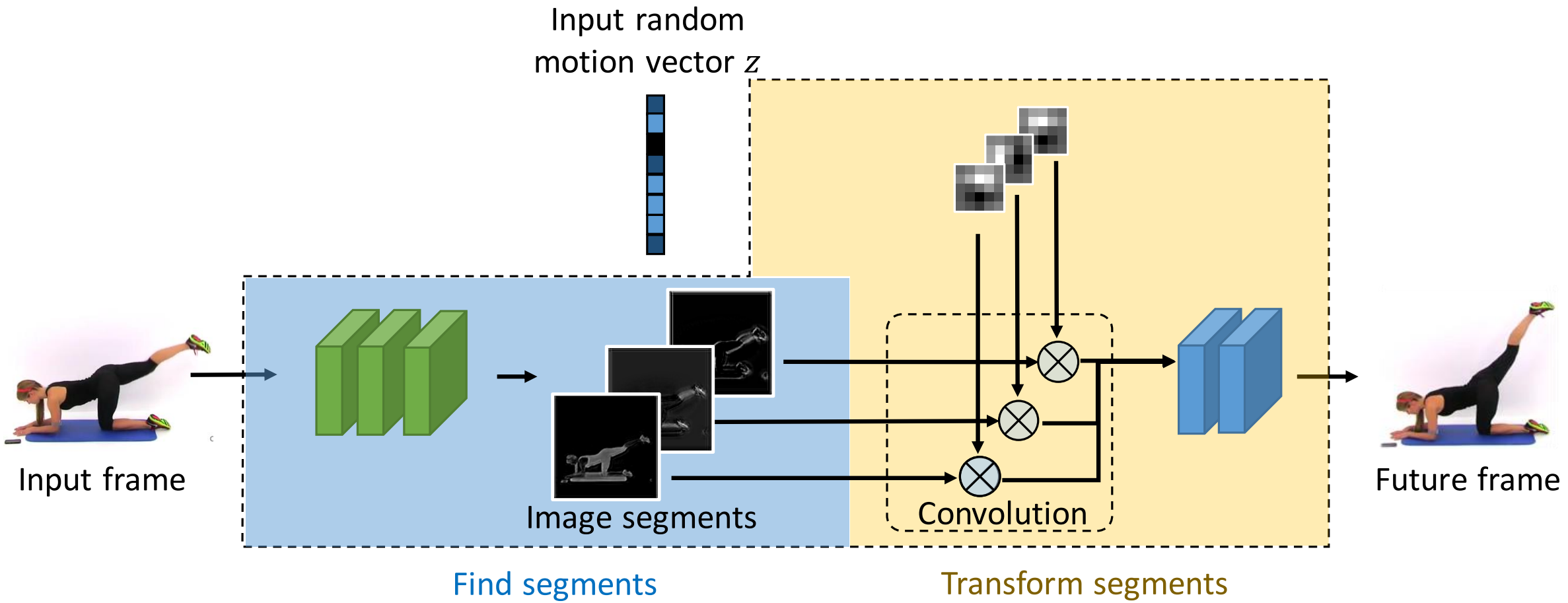
Synthesize by transforming segments



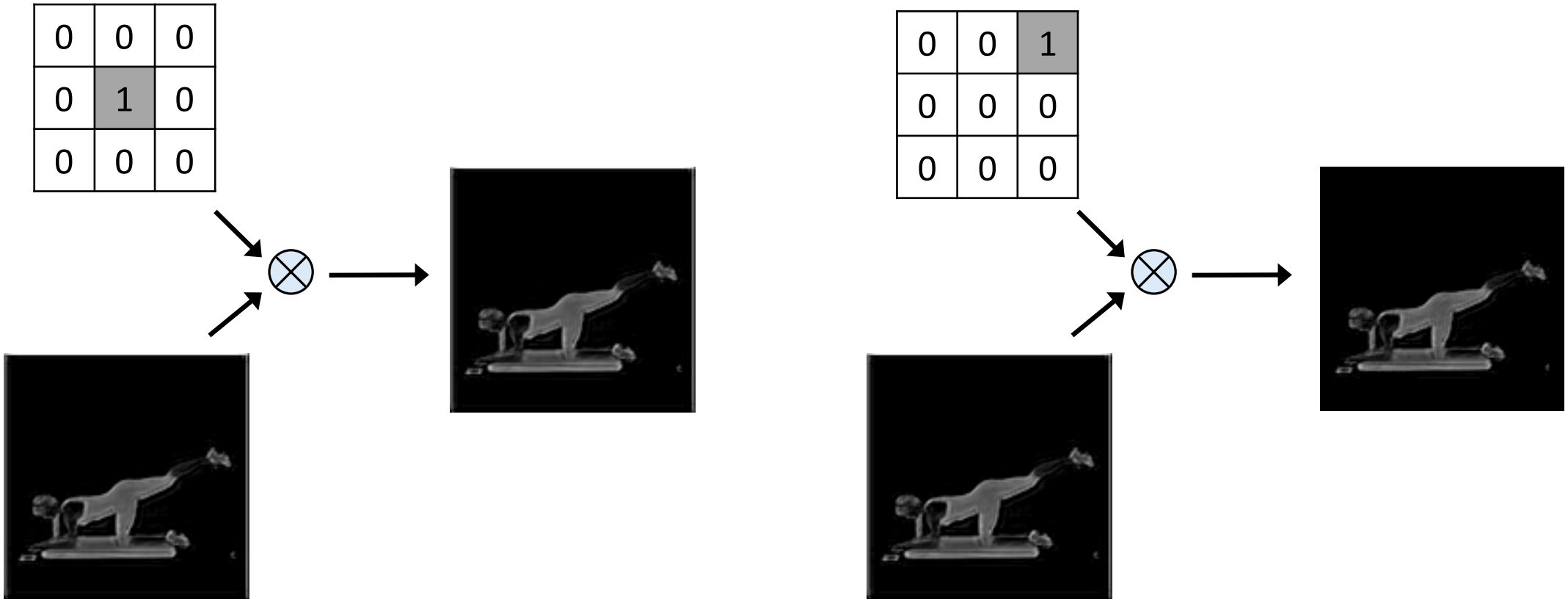
Synthesize by transforming segments



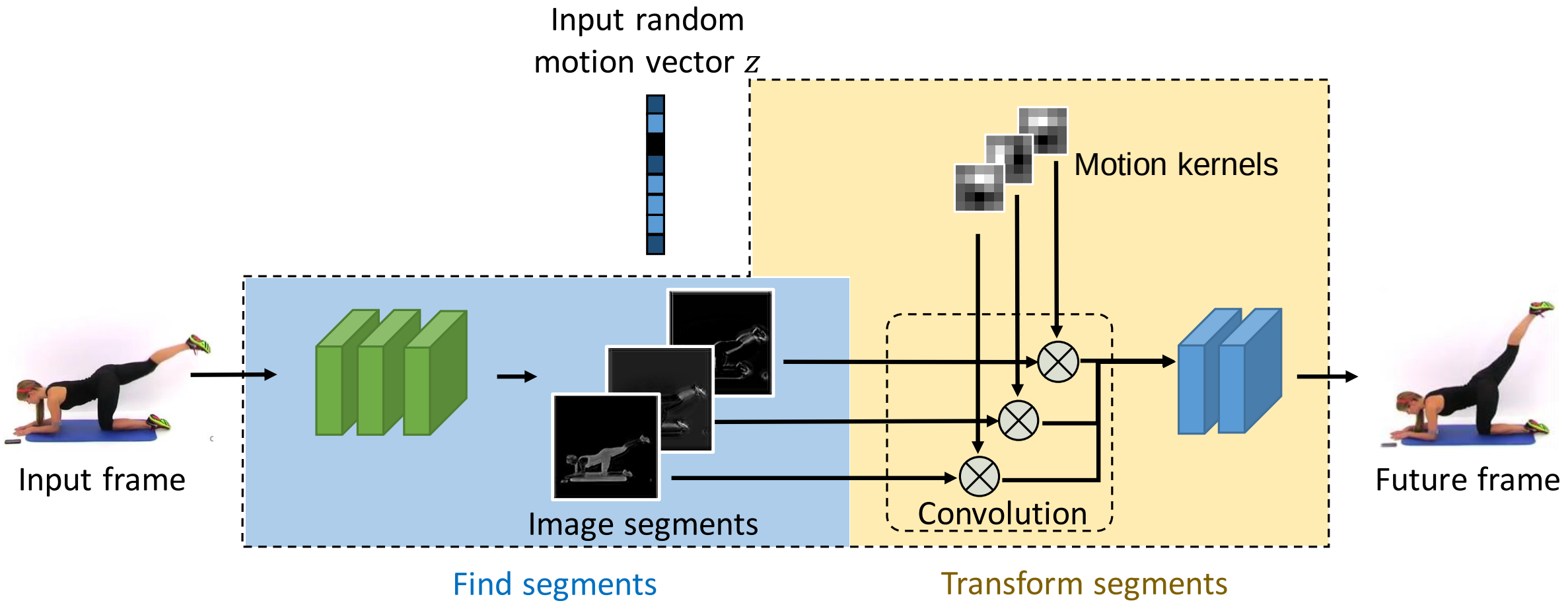
Synthesize by transforming segments



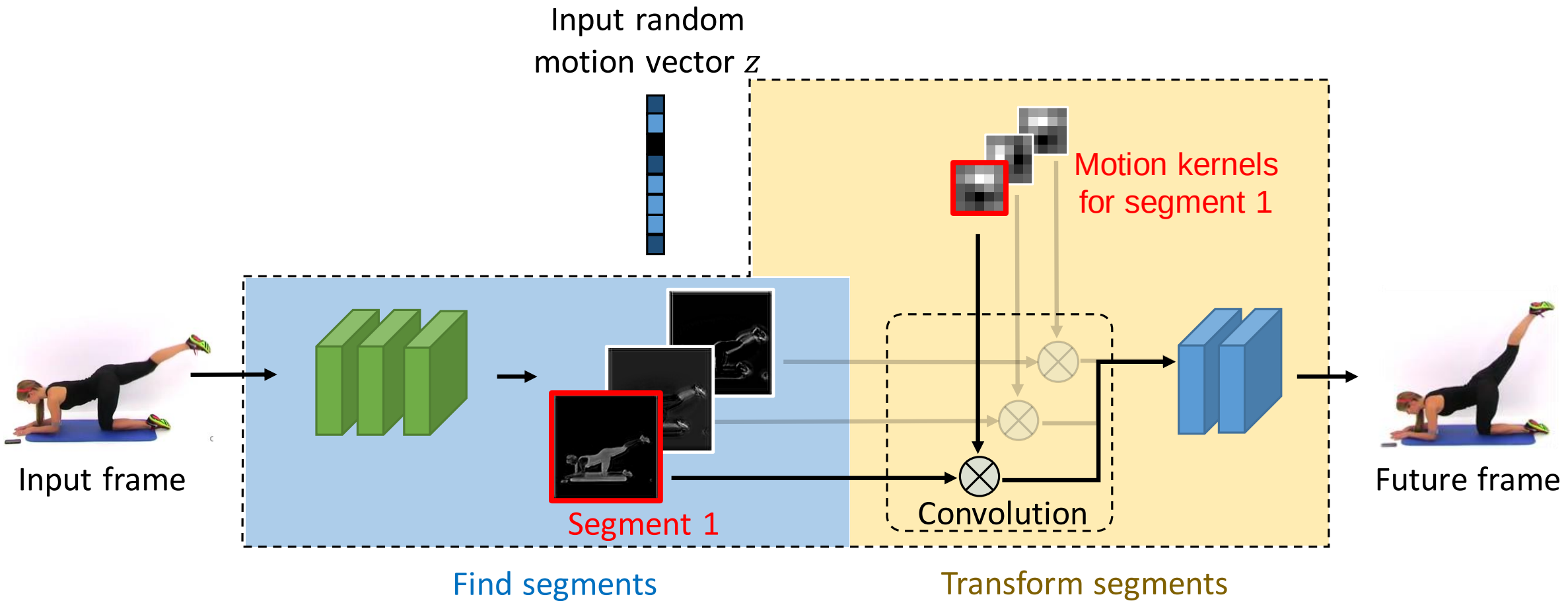
Movement can be synthesized through convolution



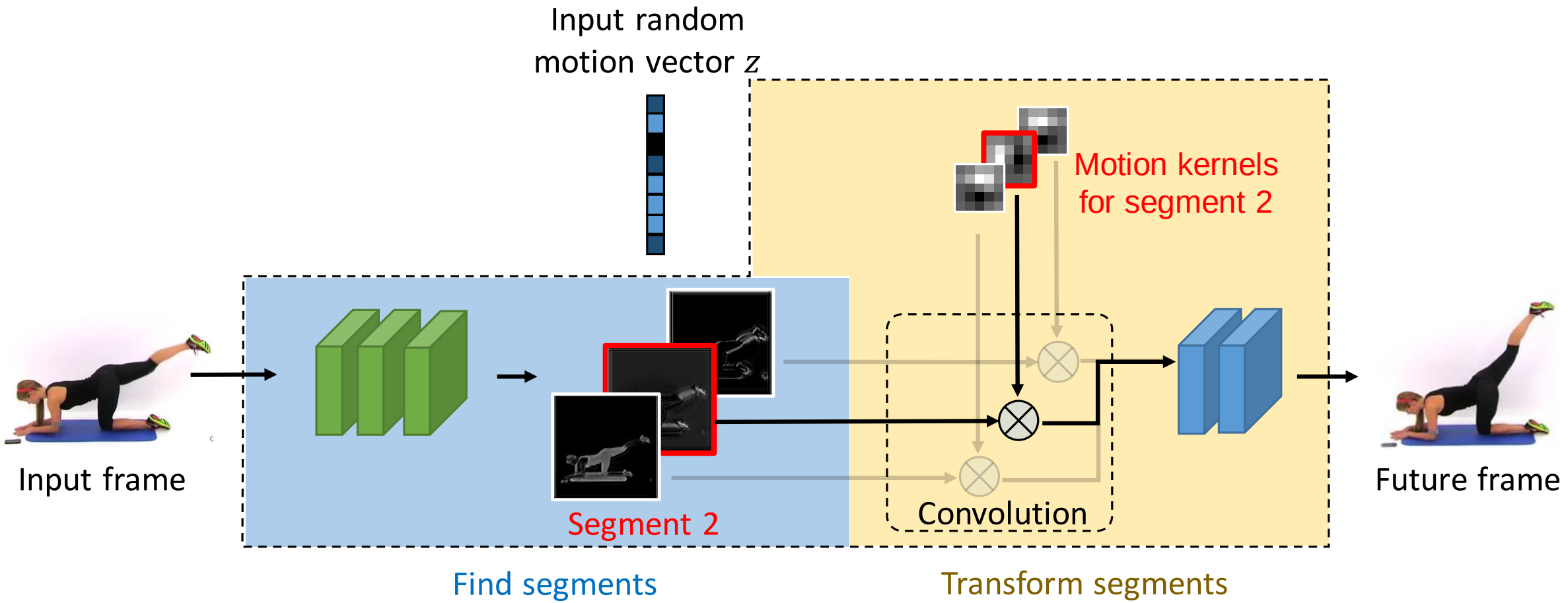
Transforming segments vis Cross-convolution



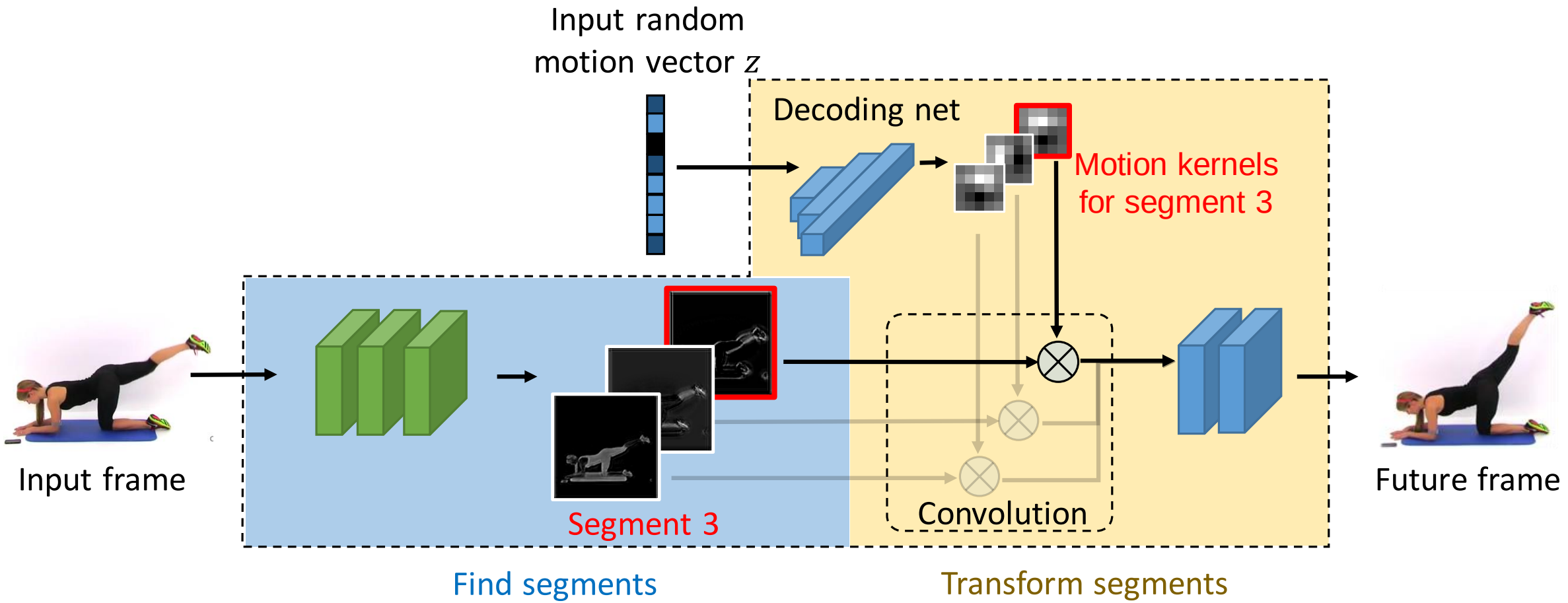
Applying motion to each segment



Applying motion to each segment

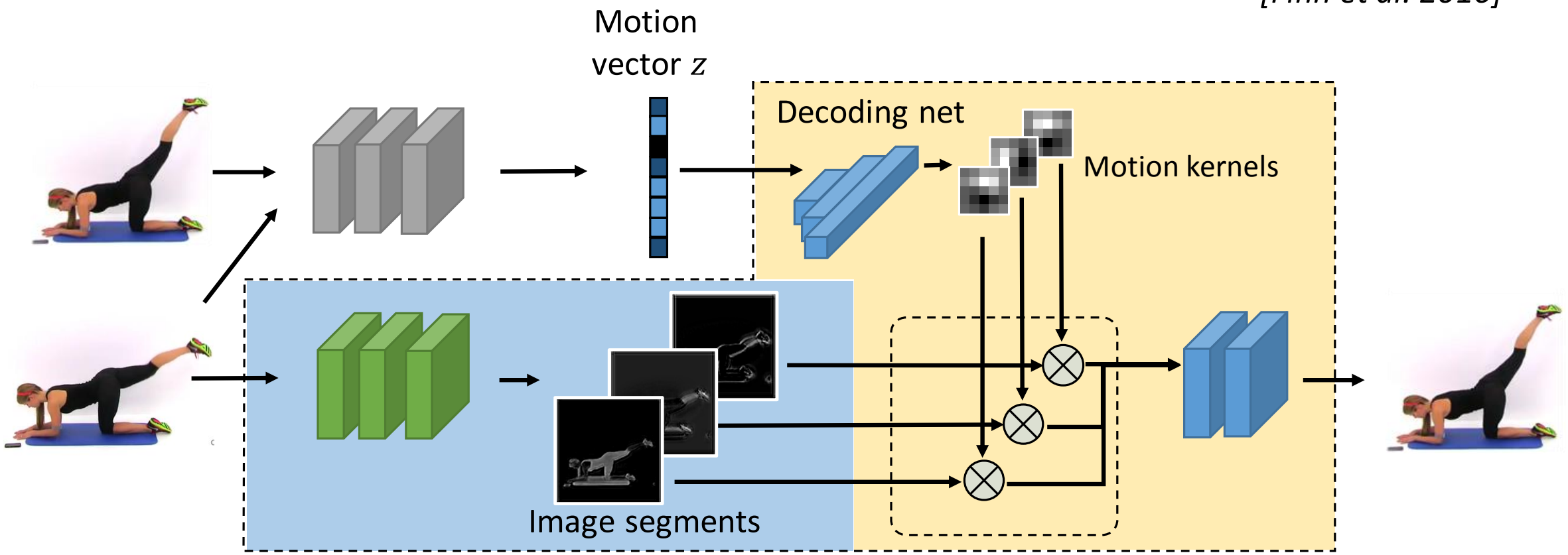


Applying motion to each segment



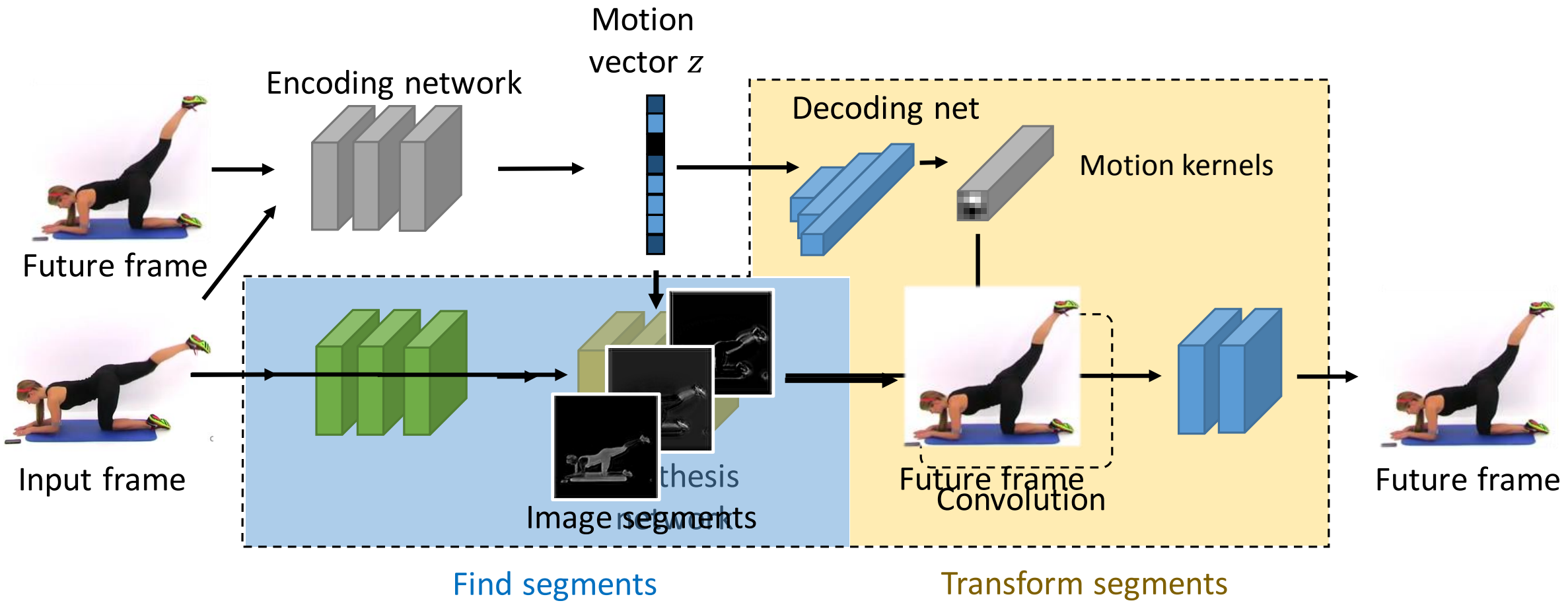
Applying motion to each segment

[Brabandere et al. 2016]
[Finn et al. 2016]

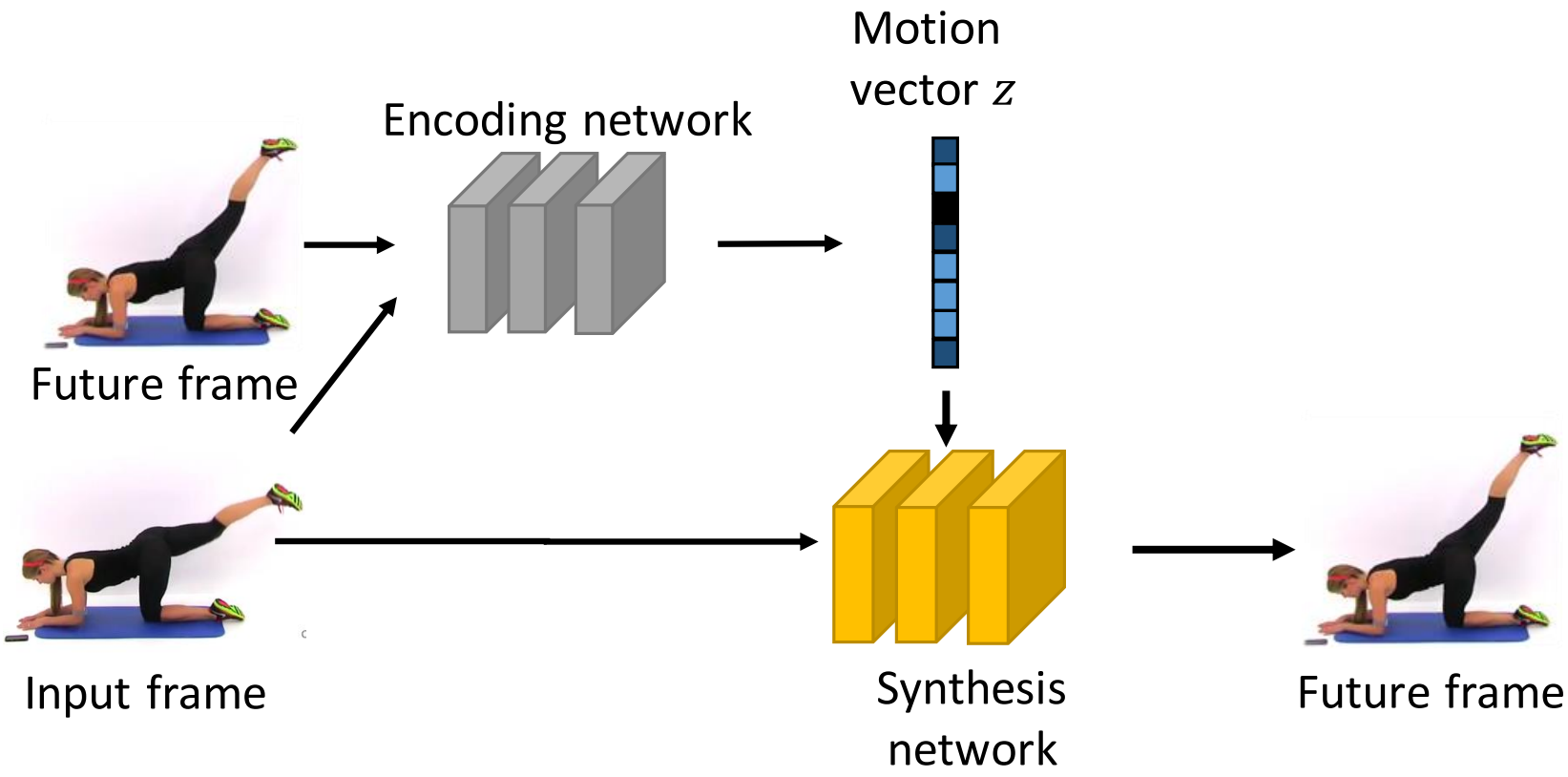


The decoding network generates a motion kernel for each corresponding segment

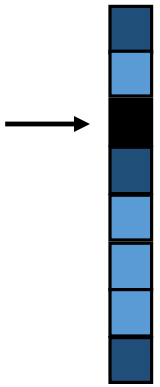
Synthesize by transforming segments



What is encoded in the motion vector?



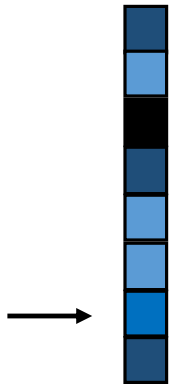
Each dimension encodes a type of motion



Motion vector z

Upward motion when changing this dimension

Each dimension encodes a type of motion

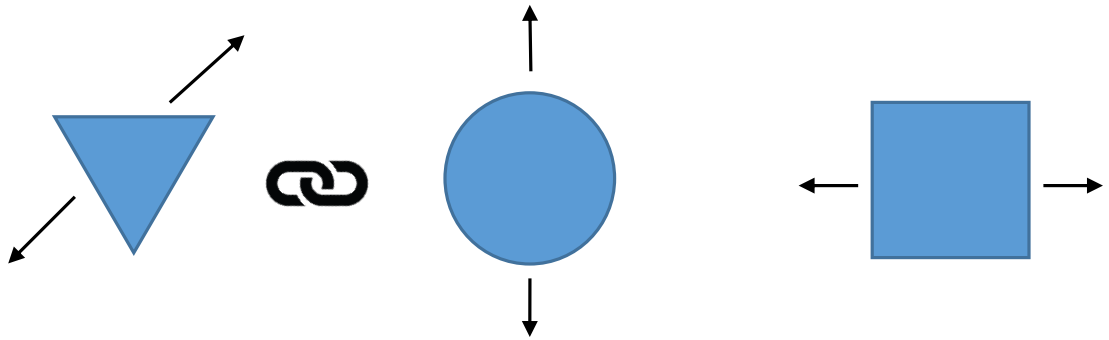


Motion vector z

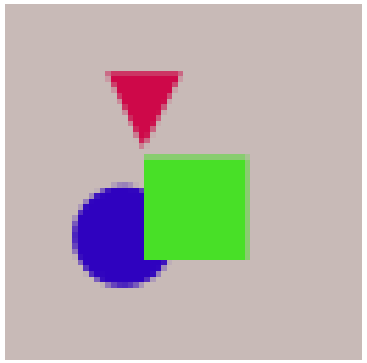
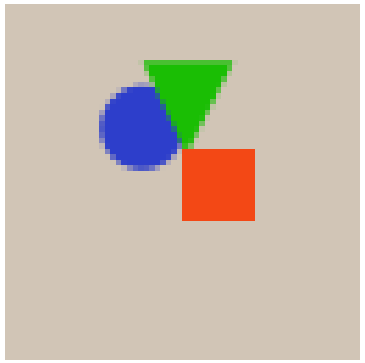
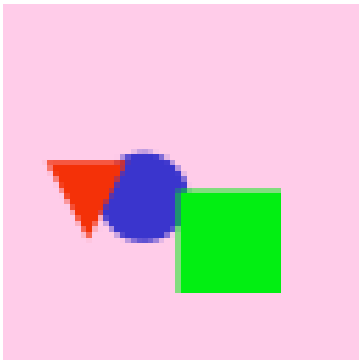
Leg motion when changing this dimension

Results: toy example

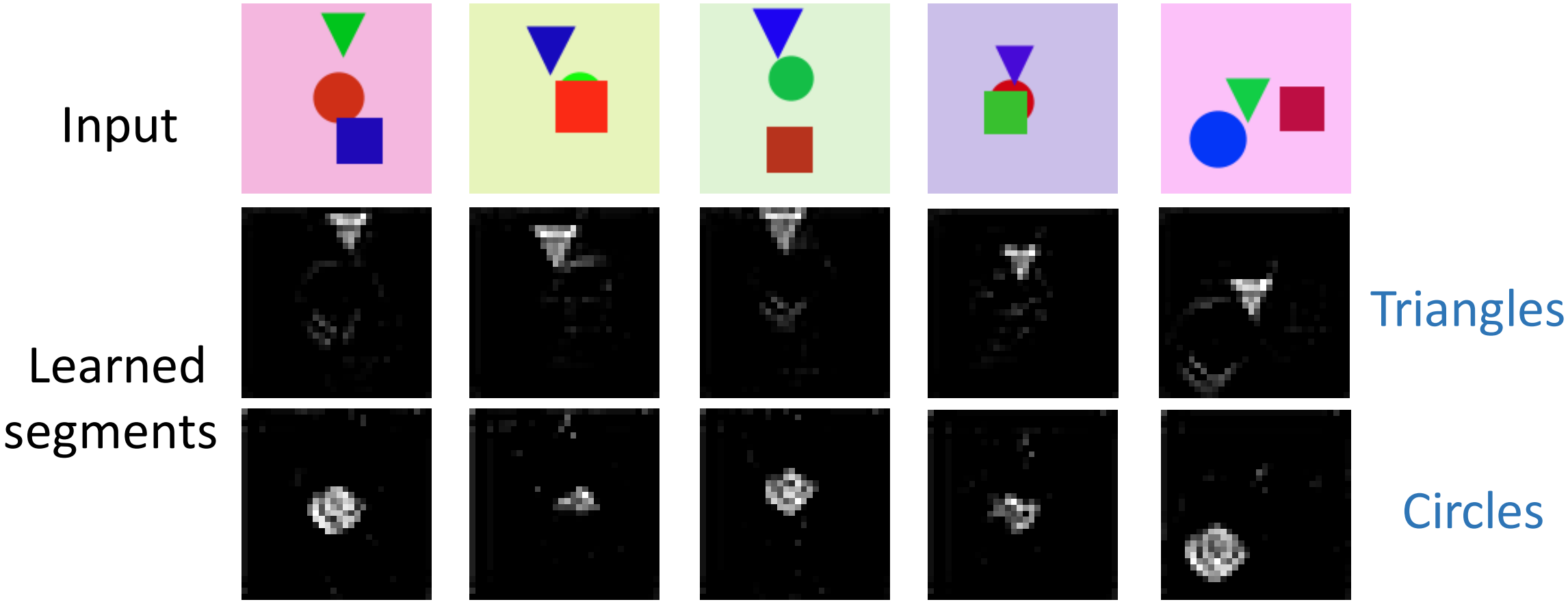
- Simulated shapes



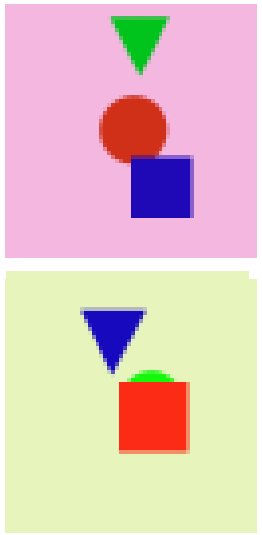
- Training samples

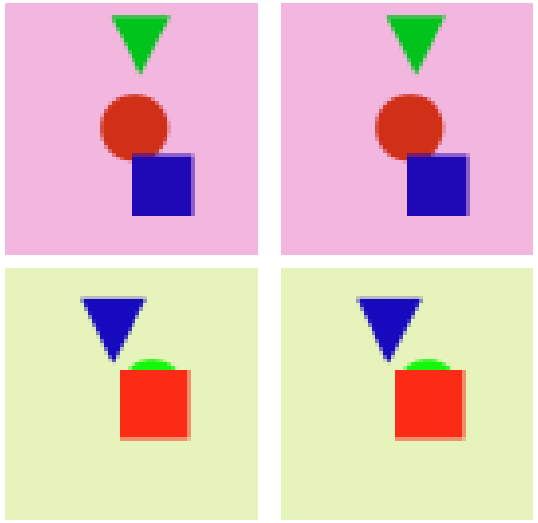
Network automatically detects segments



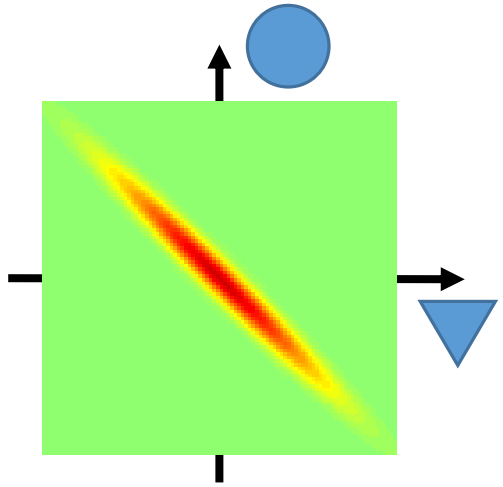
Network learns the correlation between appearance and motion



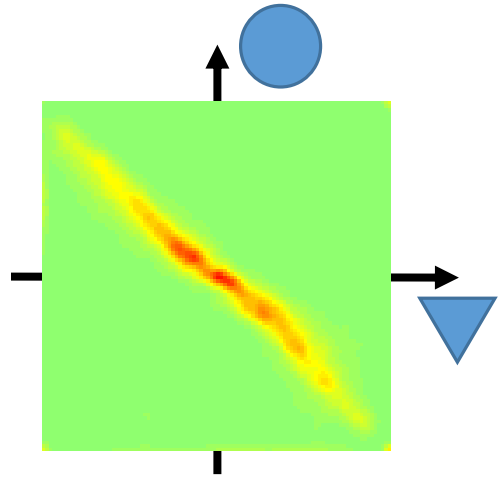
Input



Sampled next frame



Ground truth distribution



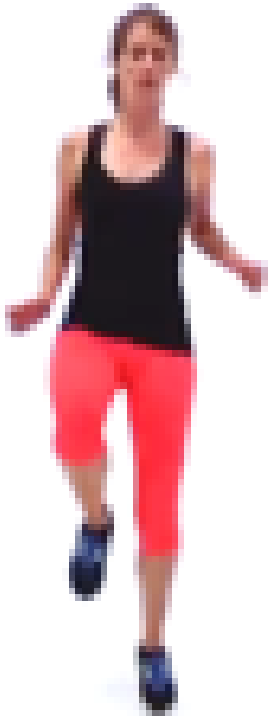
Sample distribution

Results: real-world images

Input

Sampled future frames

Challenge: large motion

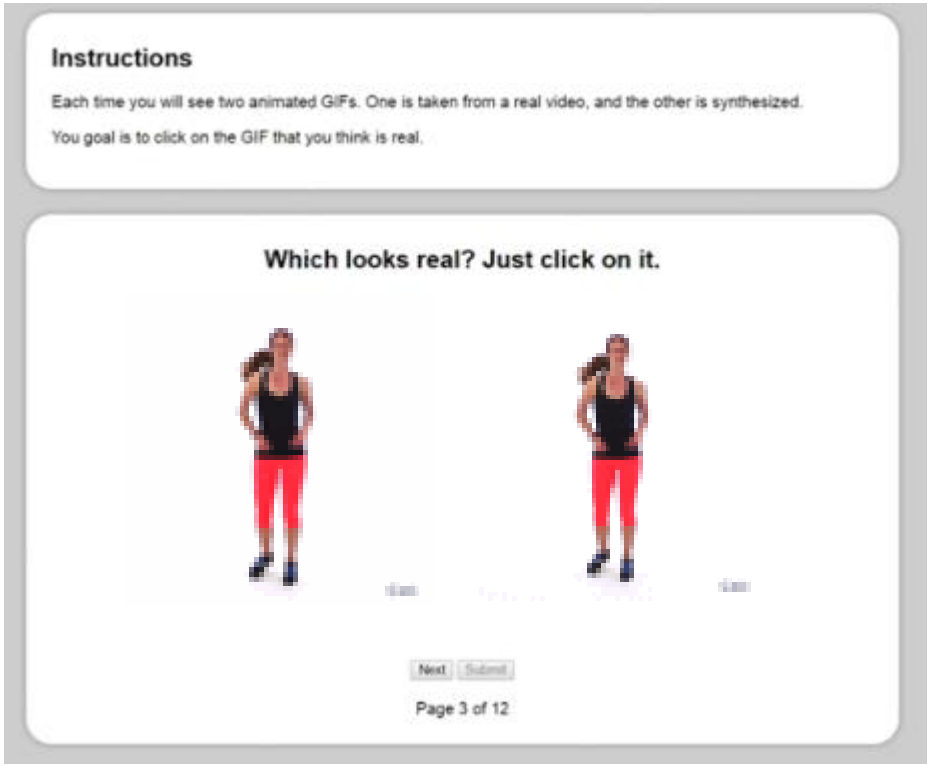


Input

Two sampled future frames

Artifacts appear when motion is large

Mechanical Turk study to assess synthesis quality



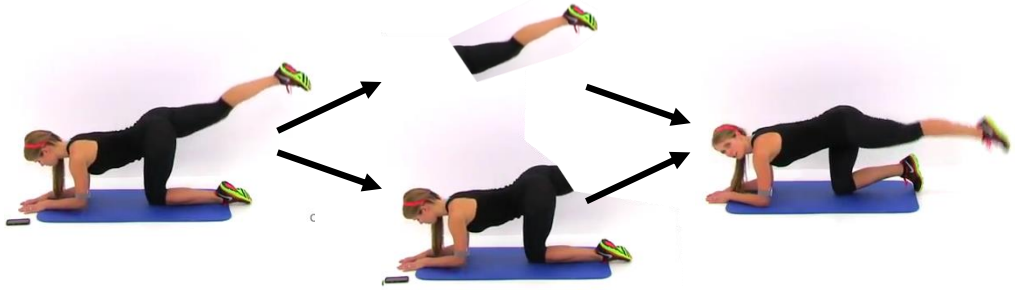
Labeled as real	
Baseline: Transfer flow	25.5 %
Our method	**31.3 %**

Ideal synthesis algorithm achieves 50%

Contributions

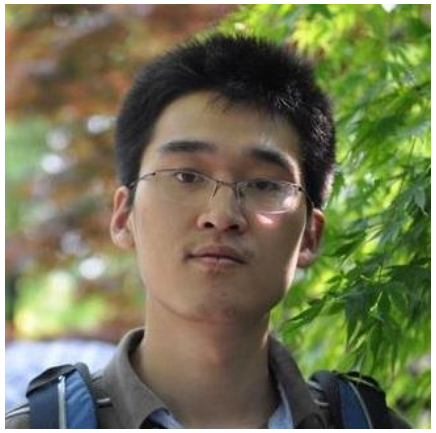
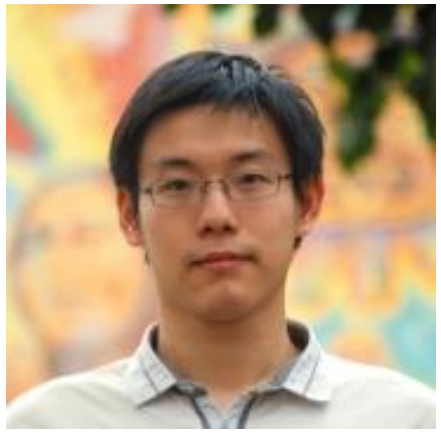
- Sample multiple future frames that are consistent with the input

- Synthesize frames by transforming segments



- Learn a motion representation without supervision

Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks

Tianfan Xue* Jiajun Wu* Katie Bouman Bill Freeman

http://visualdynamics.csail.mit.edu