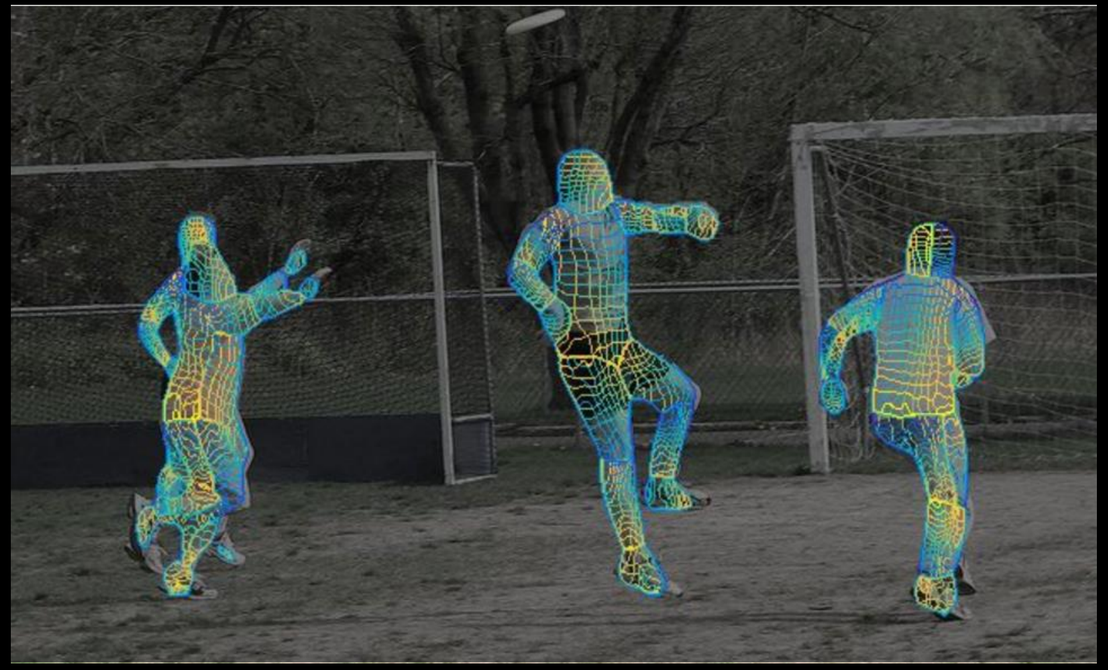
Learning Graph Representations for Video Understanding

Xiaolong Wang
Carnegie Mellon University

Computer Vision

Dog

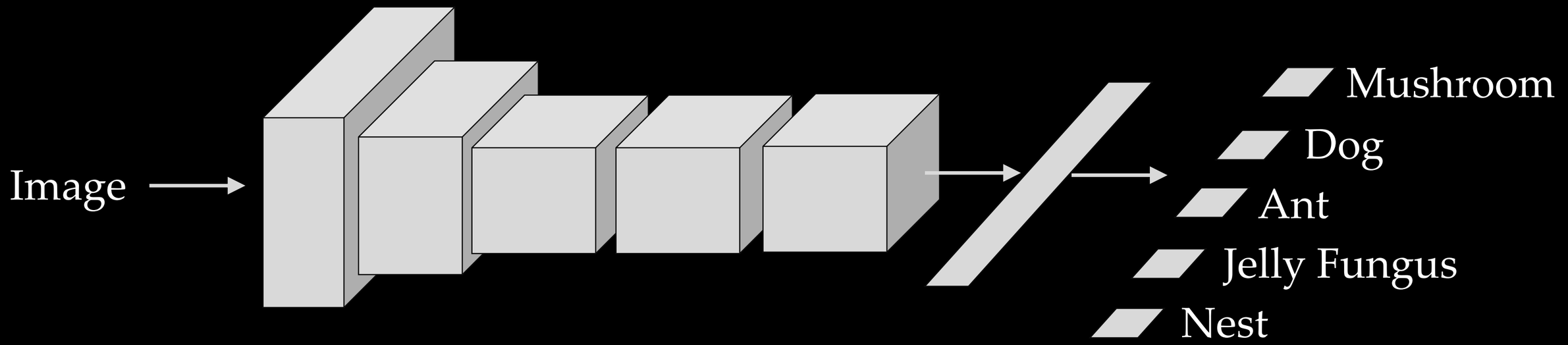


He et al. *Mask R-CNN*. ICCV 2017.
Güler et al. *DensePose: Dense Human Pose Estimation In The Wild*. CVPR 2018.

Deep Learning

ImageNet

Train a Convolutional Neural Network



Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge.* 2014.

Convolutional Neural Networks

- Convolution is local
- Long-range Pairwise relations are not modeled

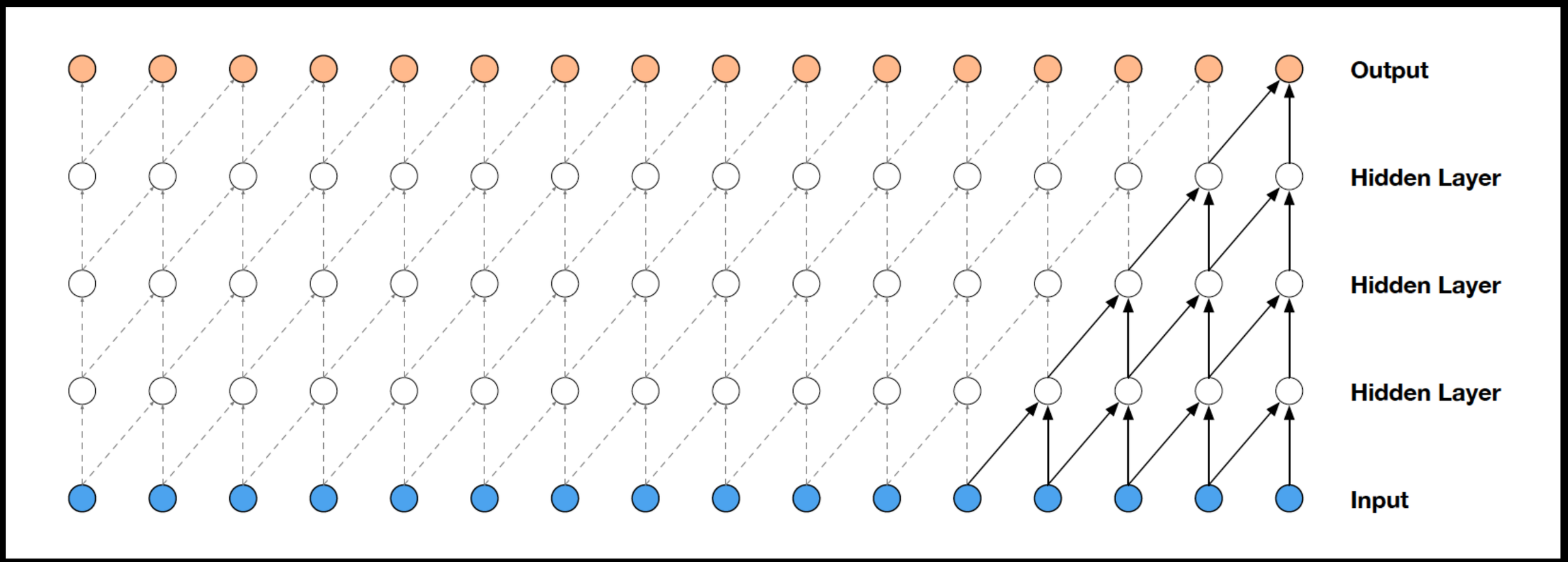
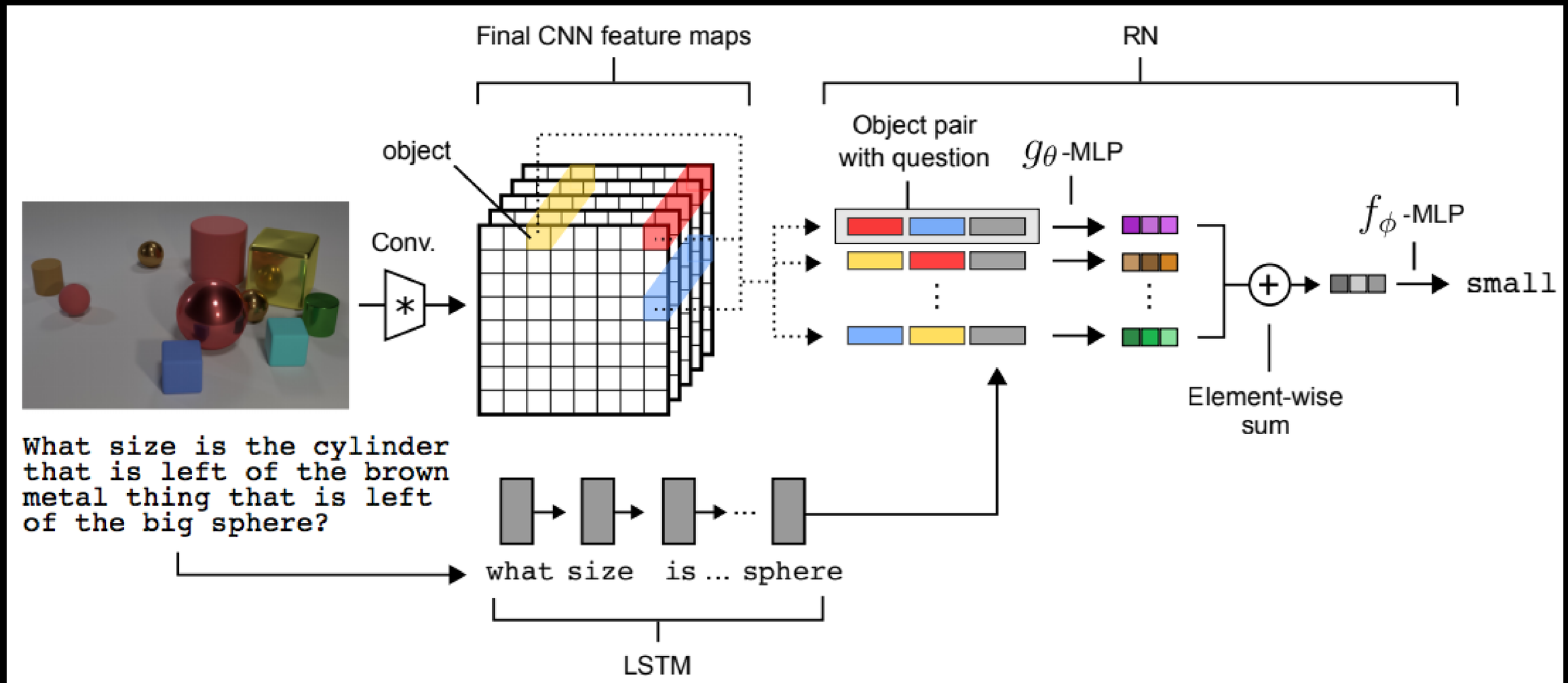


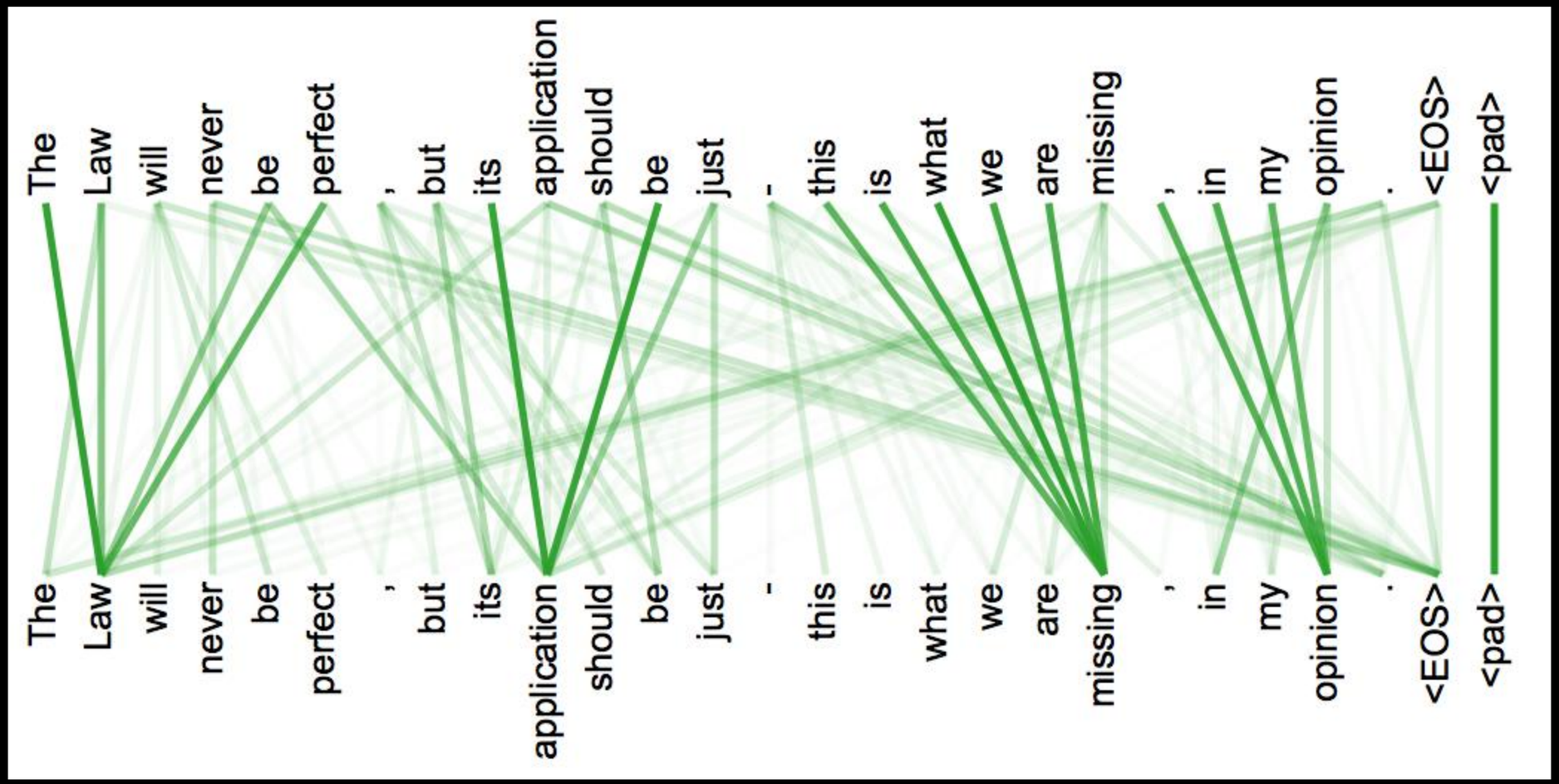
Figure credit: Van Den Oord et al.

Related Work: Relation Networks



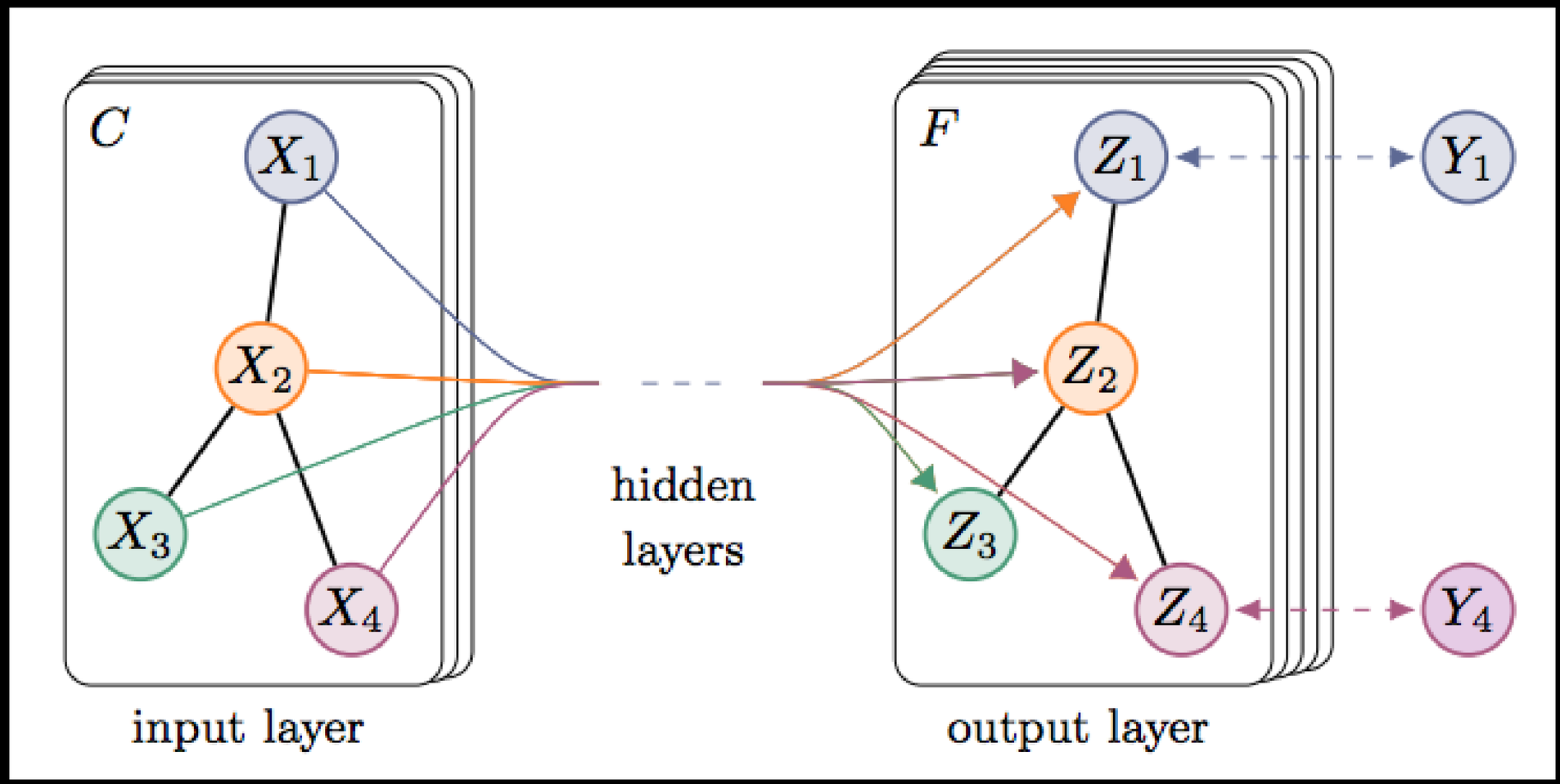
[Santoro et al, 2017]

Related Work: Self-Attention



[Vaswani et al, 2017]

Related Work: Graph Convolution Networks

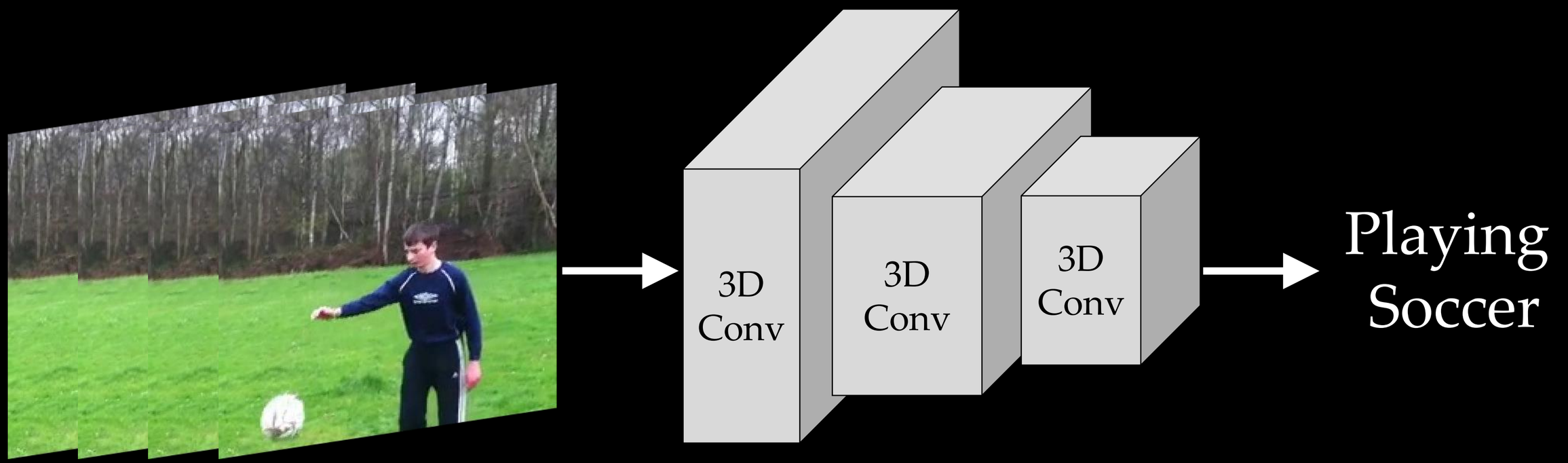


[Kipf et al, 2017]

This Tutorial

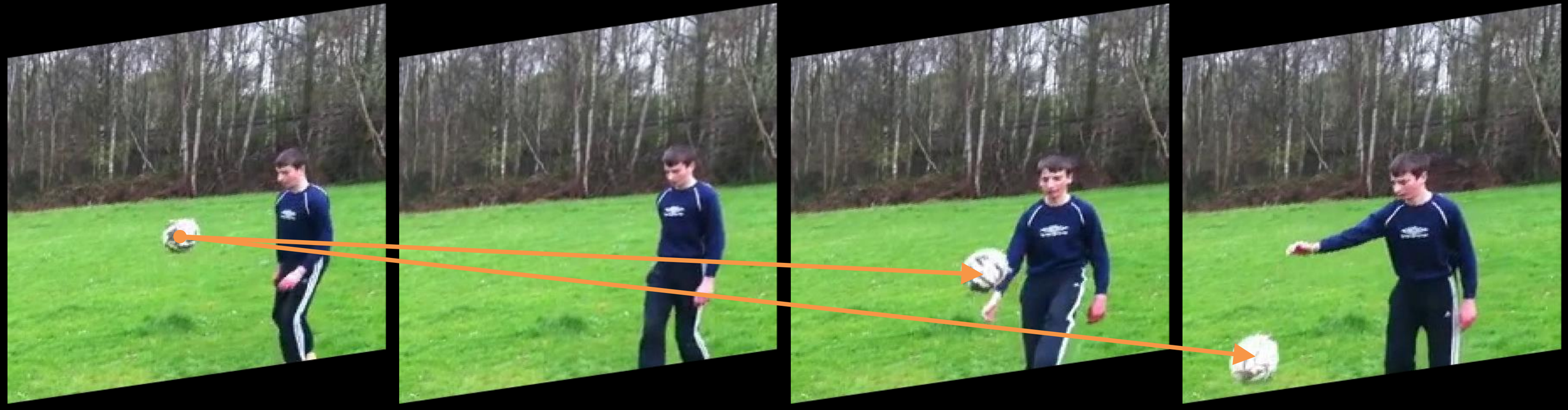
- Perform connections on different graph/relation networks
- Under the application of video understanding
- Both supervised and self-supervised methods

Video Recognition



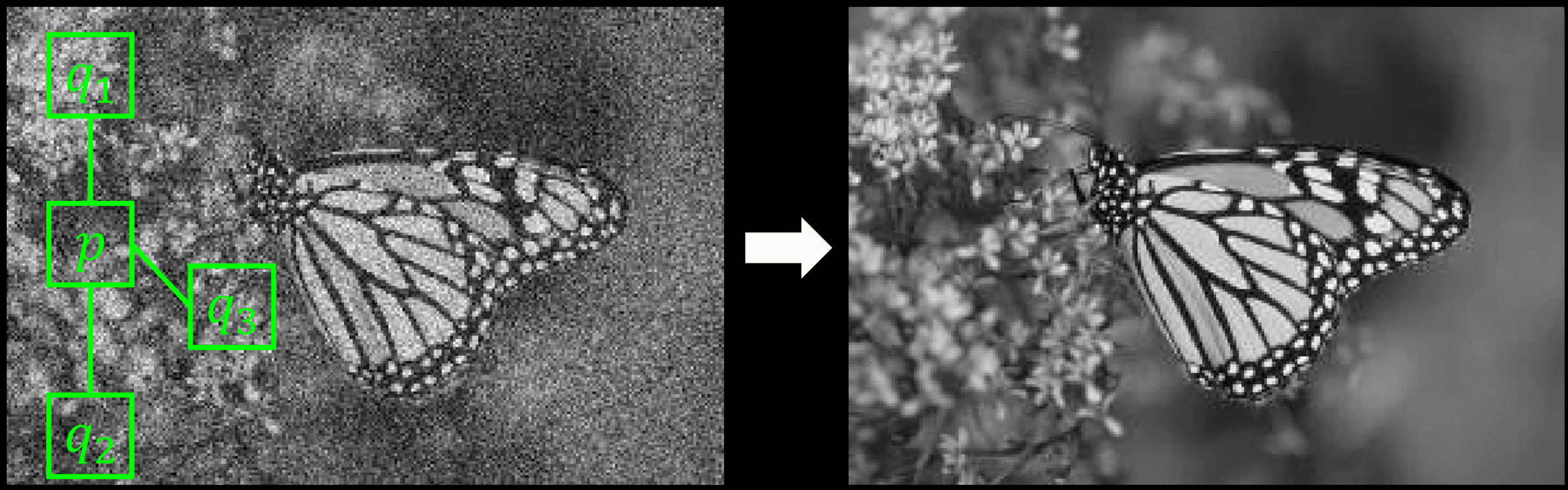
Reasoning for Action Recognition

Long-rang explicit reasoning



X. Wang , R. Girshick , A. Gupta, and K. He. *Non-local Neural Networks*. CVPR 2018.

Non-local Means

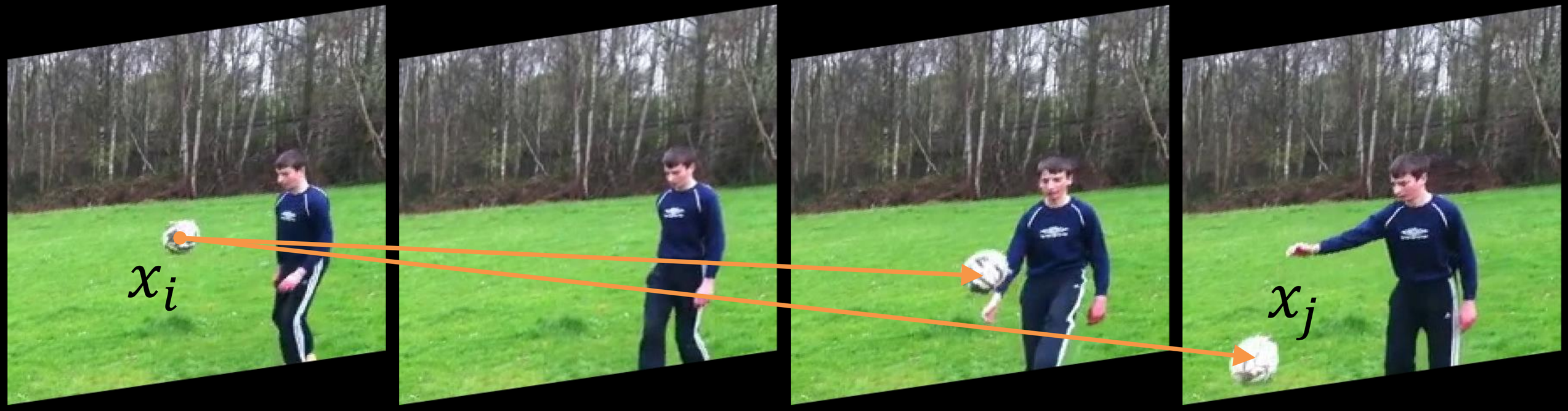


Buades et al. *A non-local algorithm for image denoising*. CVPR, 2005.

Non-local Operator

Operation in feature space

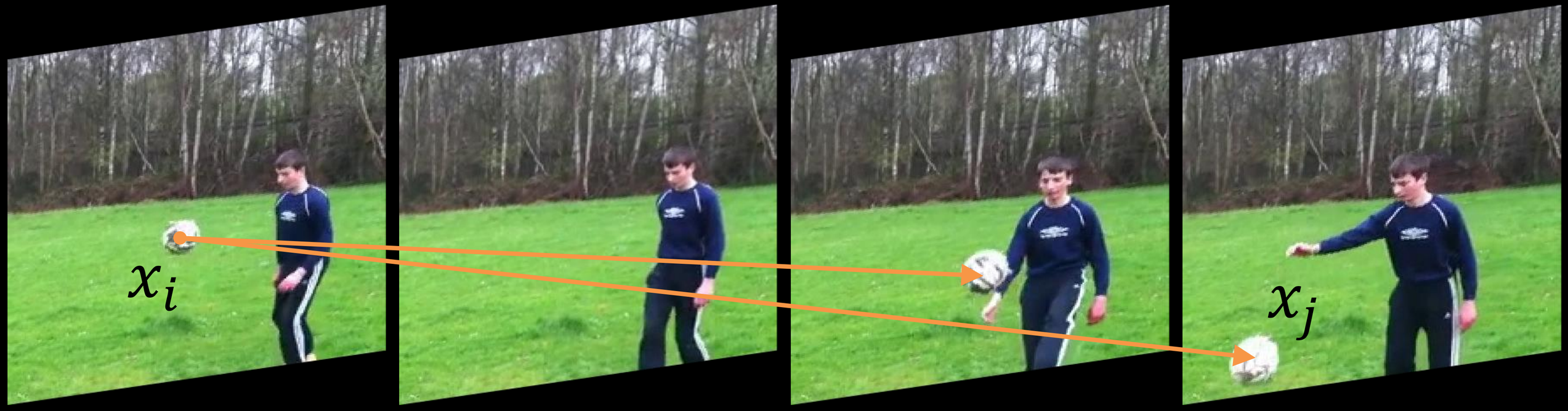
Can be embedded into any ConvNets



Non-local Operator

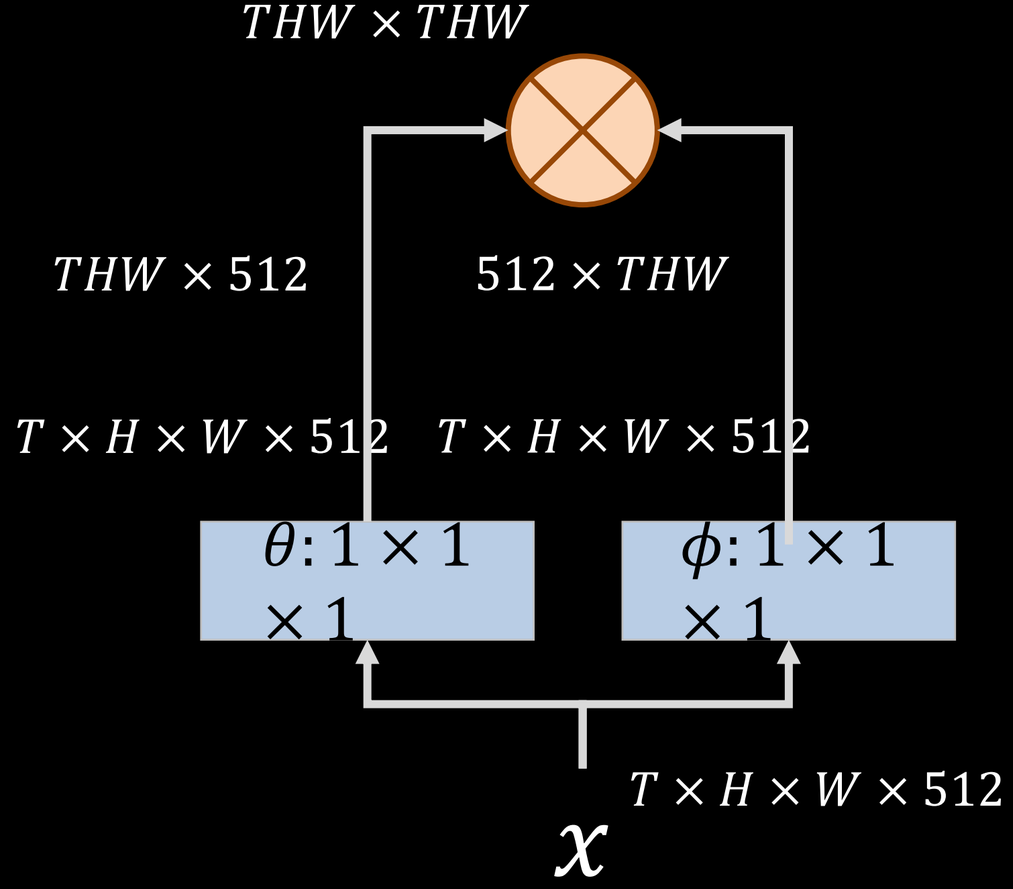
$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) \; g(x_j)$$

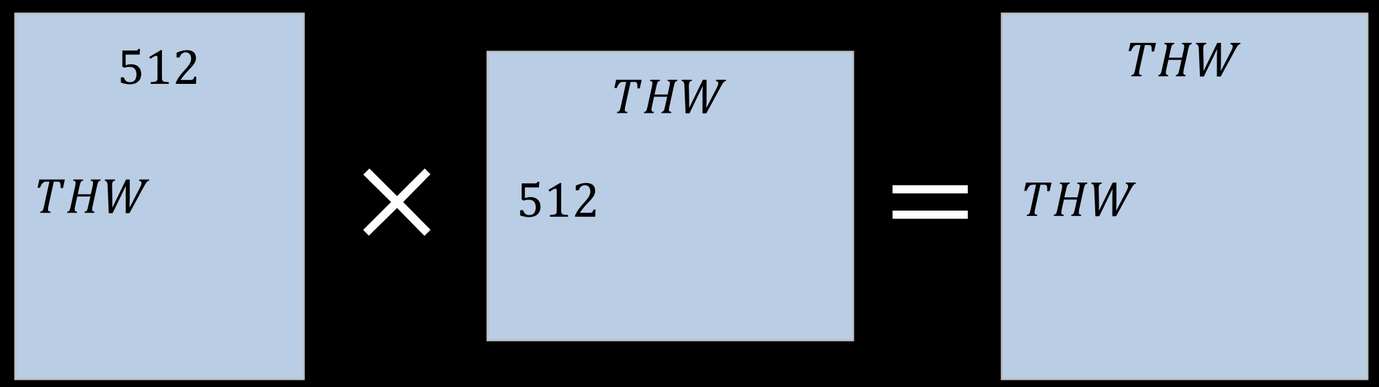
Affinity Features



Non-local Operator

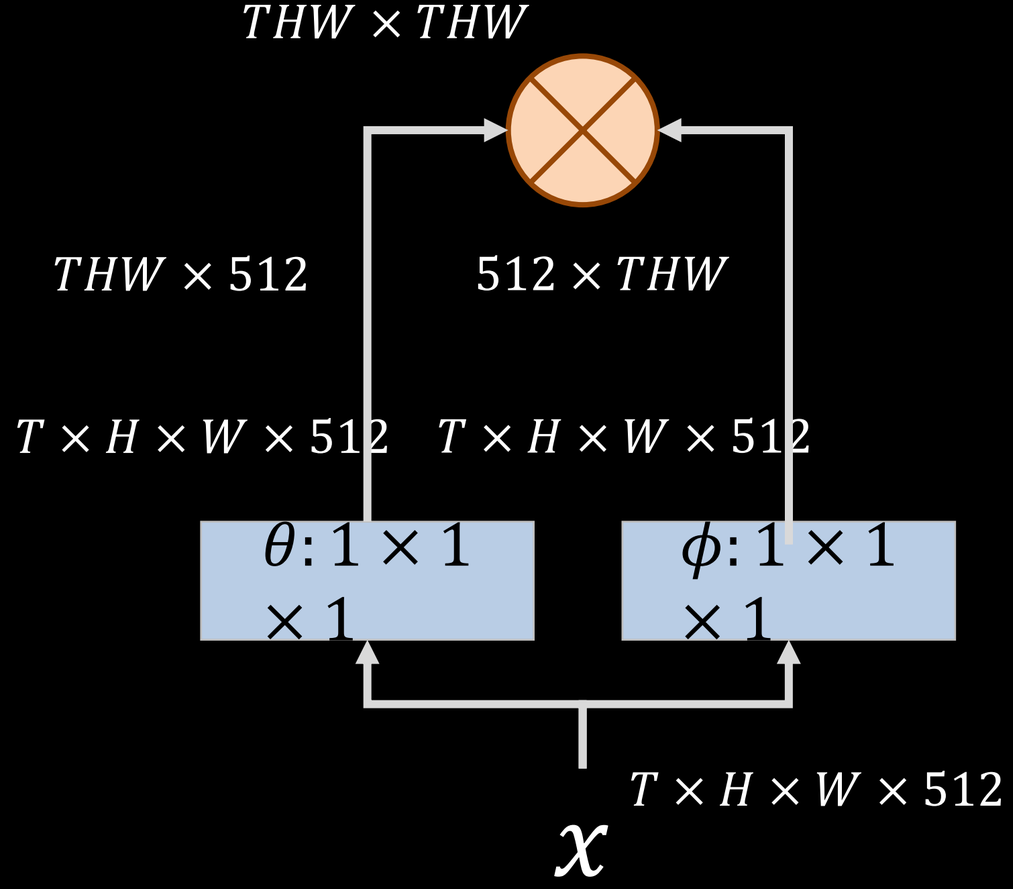
$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\ g(x_j)$$

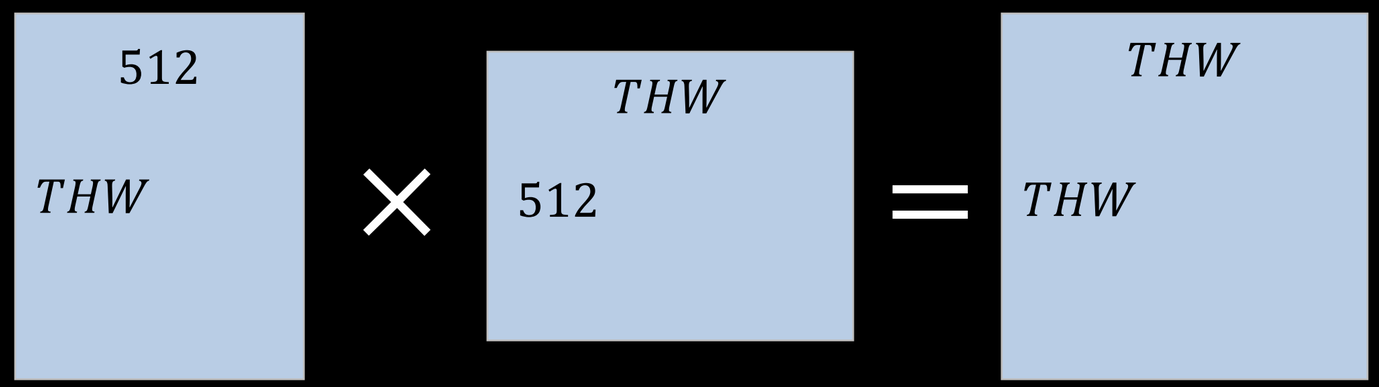




Non-local Operator

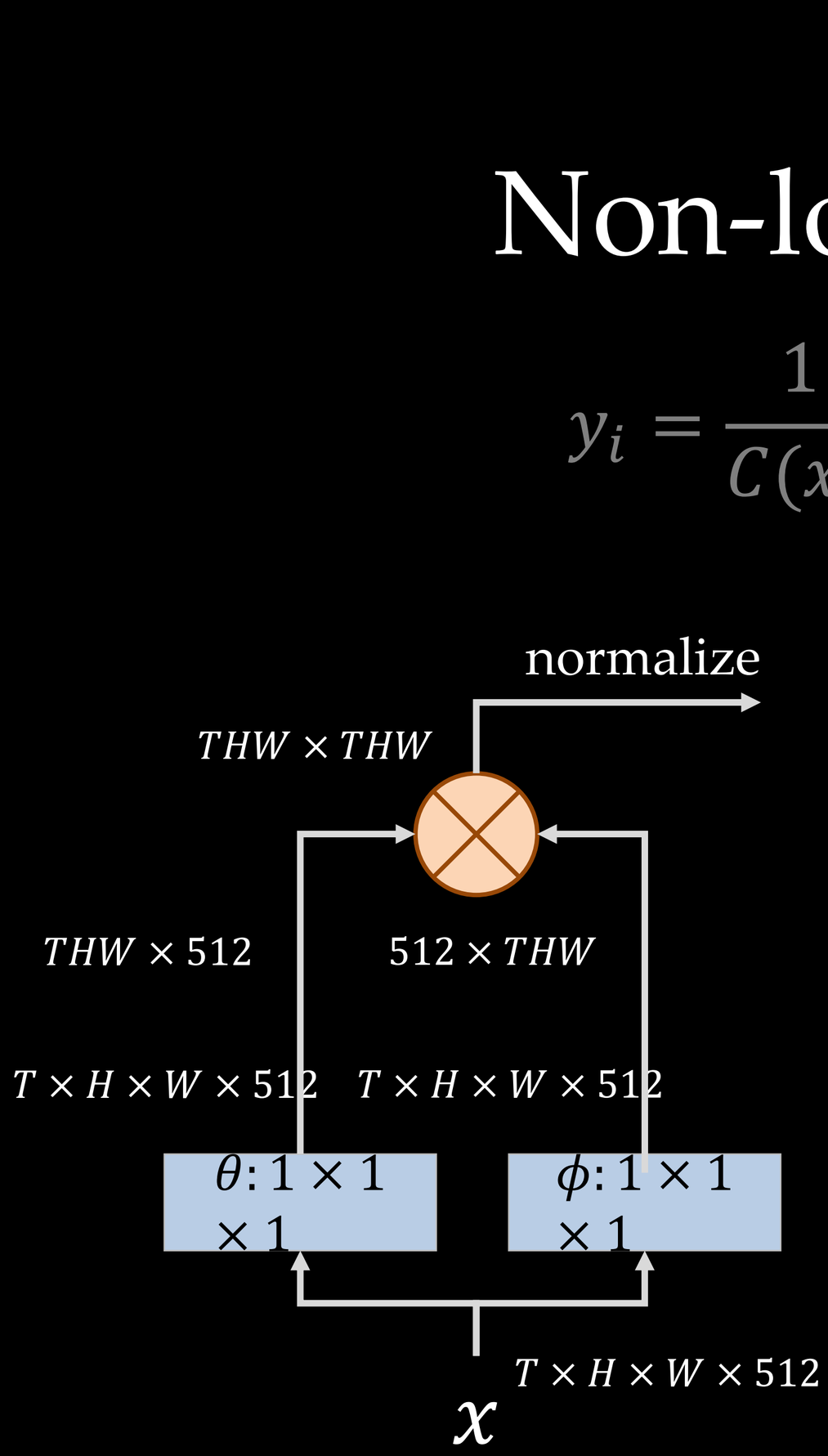
$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\ g(x_j)$$





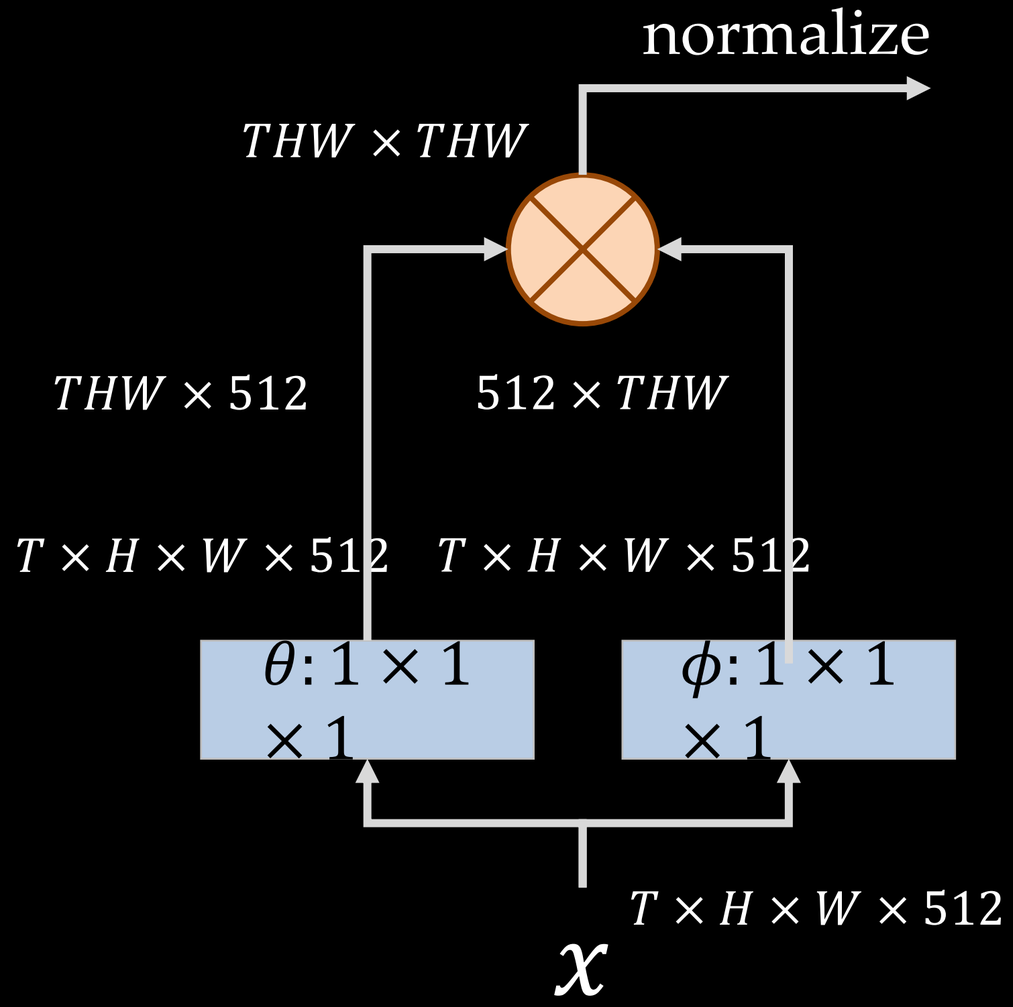
Non-local Operator

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\ g(x_j)$$



Non-local Operator

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) \; g(x_j)$$



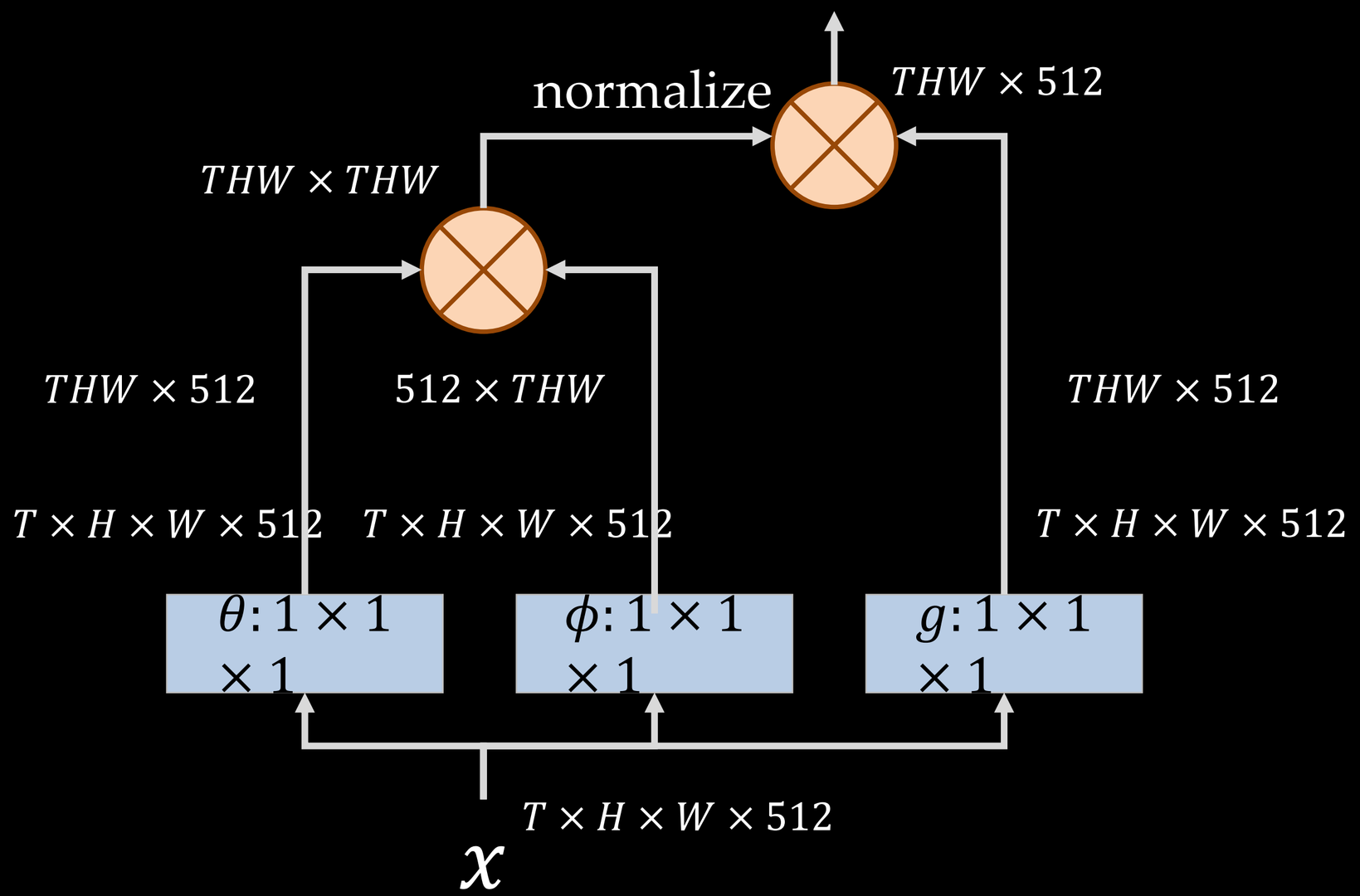
$$f(x_i, x_j) = \exp(x_i^T x_j)$$

$$C(x) = \sum_{\forall j} f(x_i, x_j)$$

$$\frac{f(x_i, x_j)}{C(x)} = \frac{\exp(x_i^T x_j)}{\sum_{\forall j} \exp(x_i^T x_j)}$$

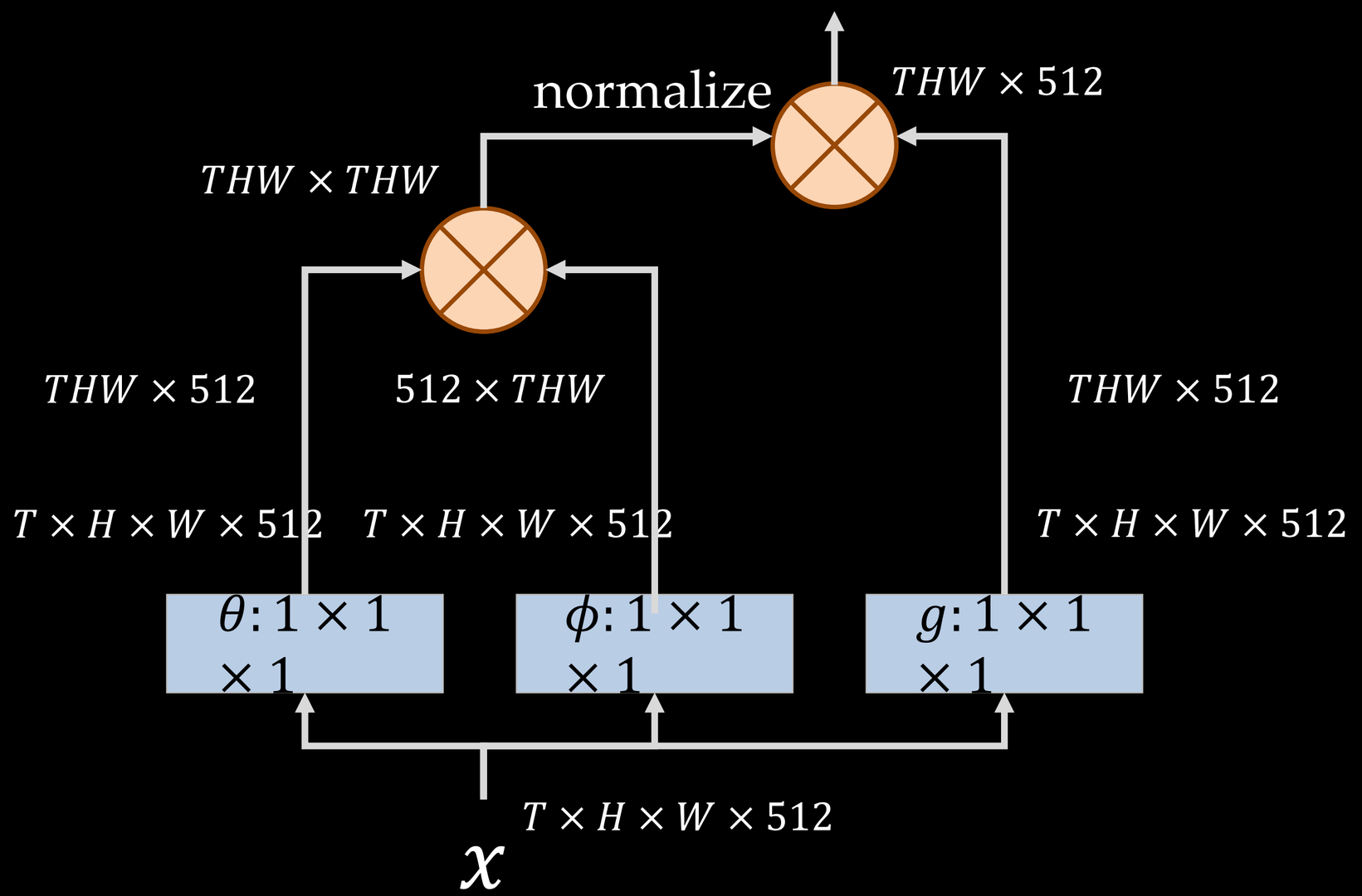
Non-local Operator

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\ g(x_j)$$



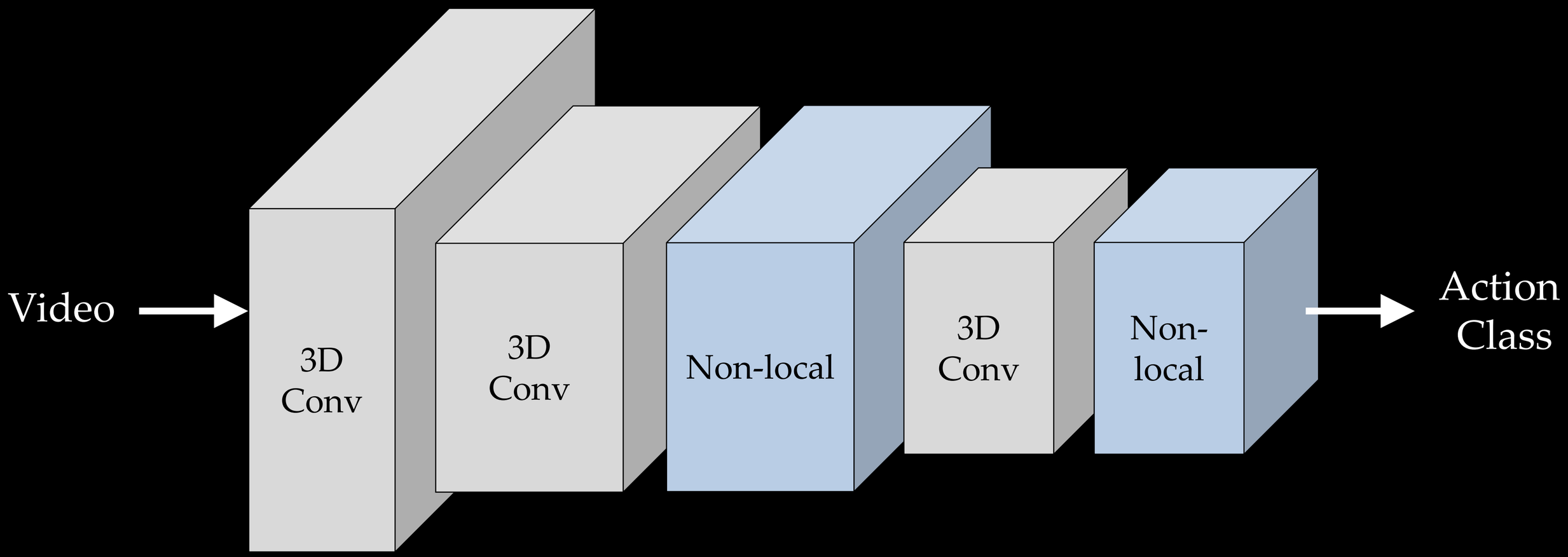
Non-local Operator

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\ g(x_j)$$

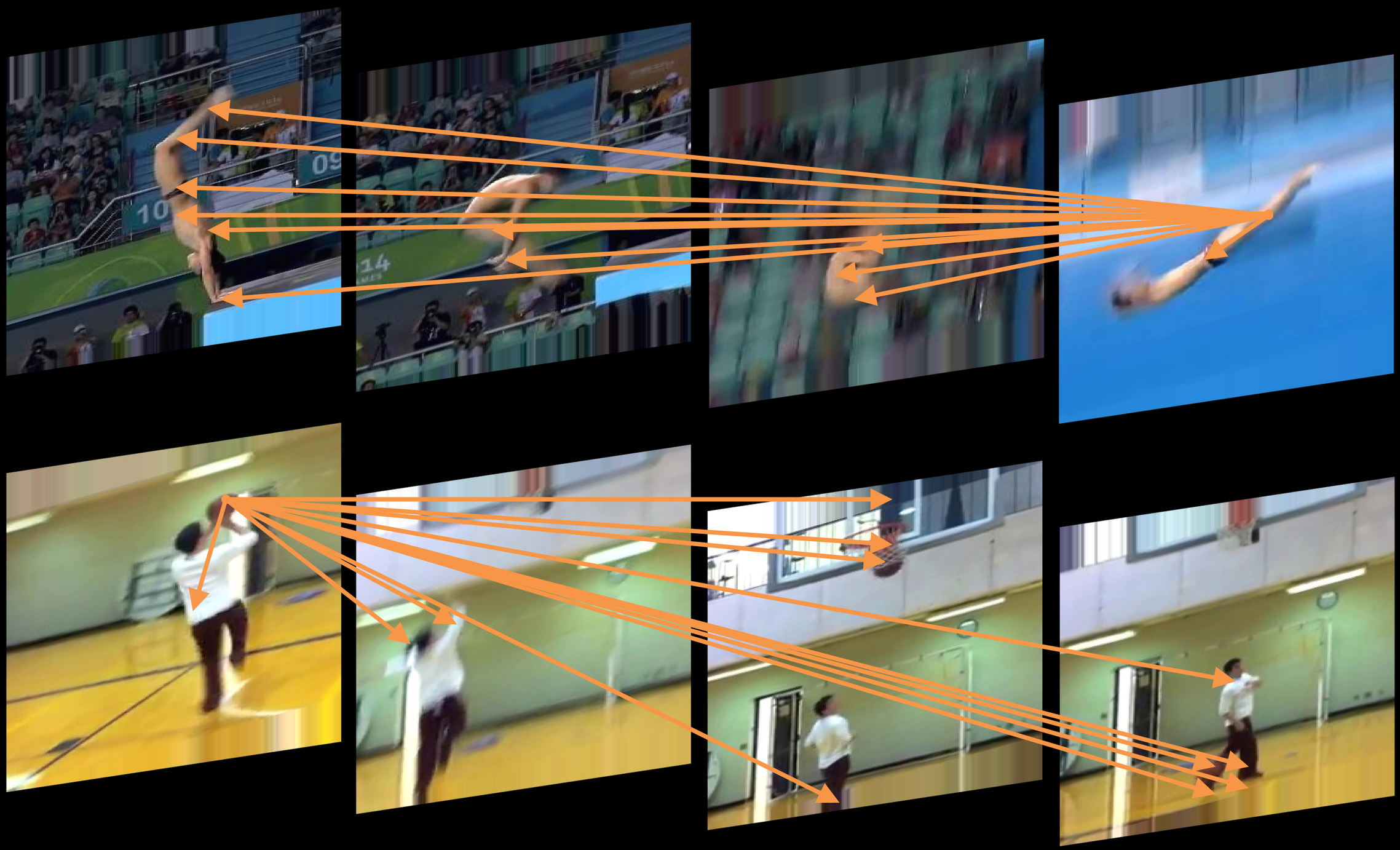


Non-local Operator as A Residual Block

$$z_i = y_i W + x_i$$



Examples



Action Recognition in Daily Lives

We let the people upload their own videos!

Charades Dataset: 157 classes, 9.8k videos, 30s per video

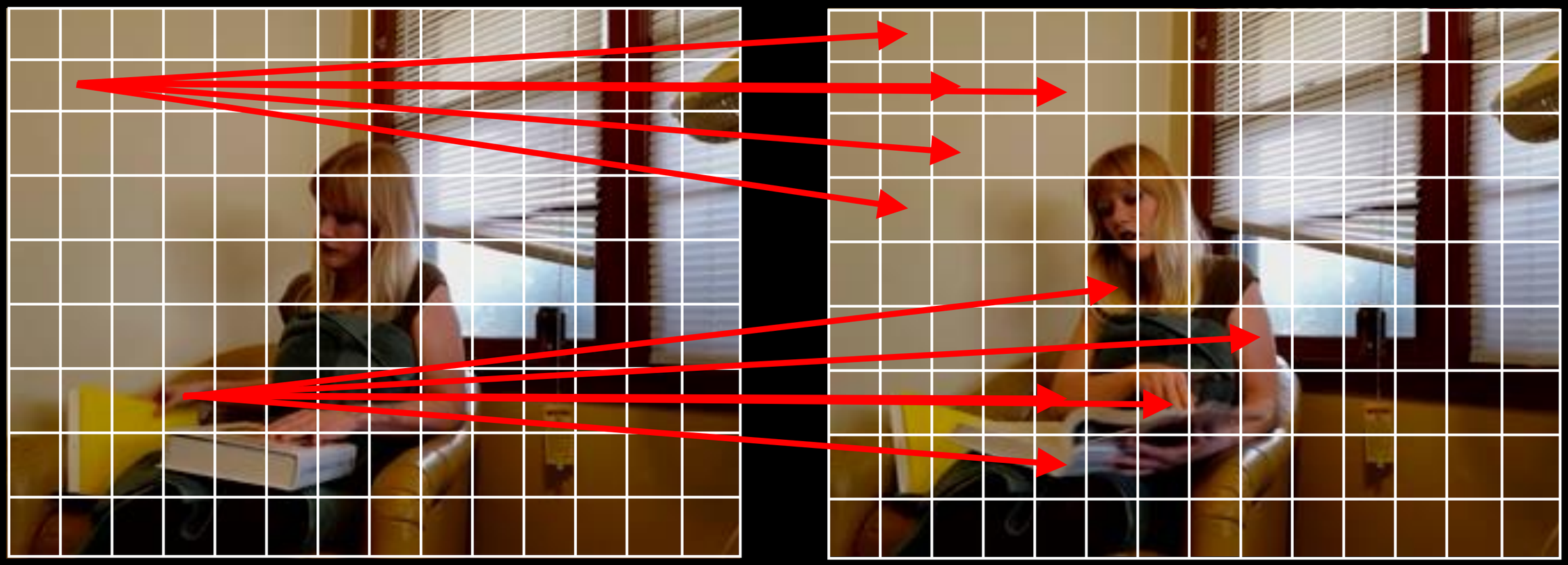
Gunnar A. Sigurdsson, Gül Varol, **Xiaolong Wang**, Ivan Laptev, Ali Farhadi, Abhinav Gupta.
Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV 2016.

Action Recognition on Charades

Method	mAP
3D Conv	31.8%
3D Conv + Non-local	33.5%

Opening A Book

Opening A Book



The Non-local Block

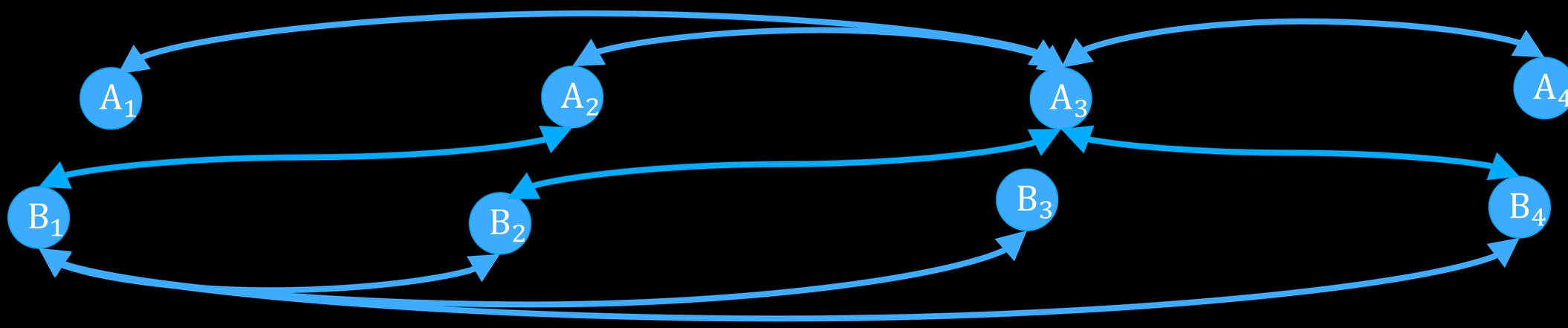
Opening A Book

Object states changes over time

Human-object, object-object interactions

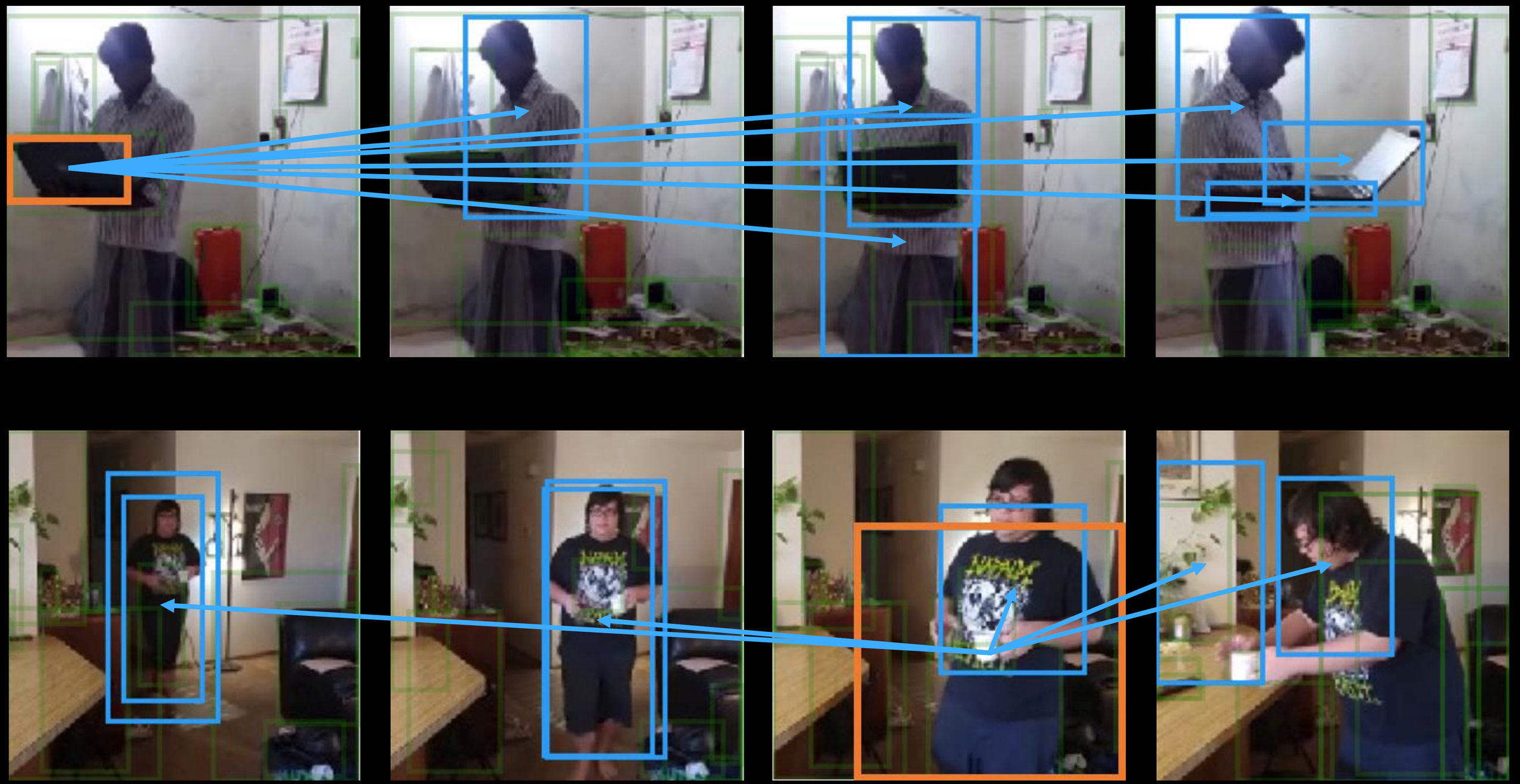
X. Wang and A. Gupta. *Video as Space-Time Region Graphs*. ECCV 2018.

Opening A Book



Highly Correlated

Relations between Regions



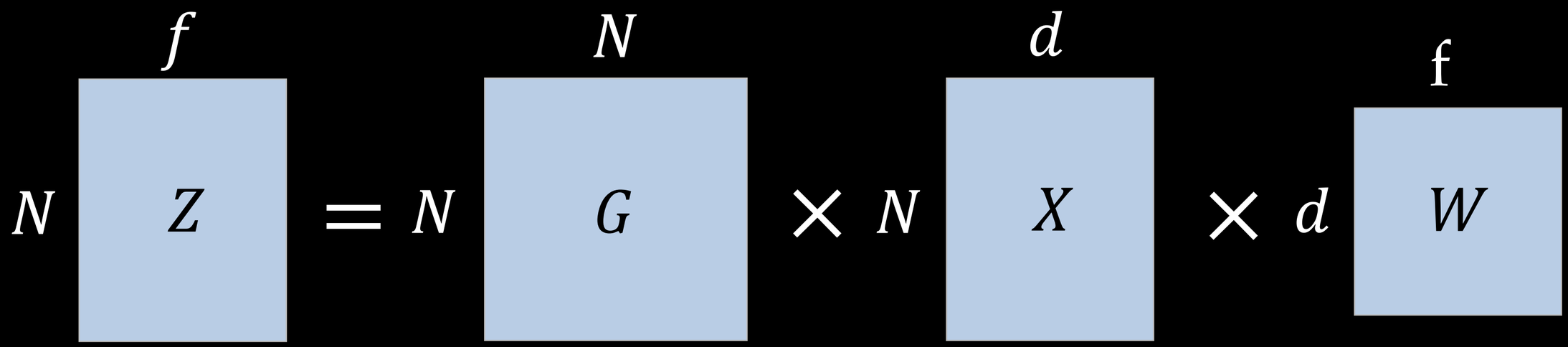
Relations between Regions

$$f(x_i, x_j) = \phi(x_i)^T \, \phi'(x_j)$$

$$G_{ij} = \frac{\exp \, f(x_i, x_j)}{\sum_{\forall j} \exp \, f(x_i, x_j)}$$

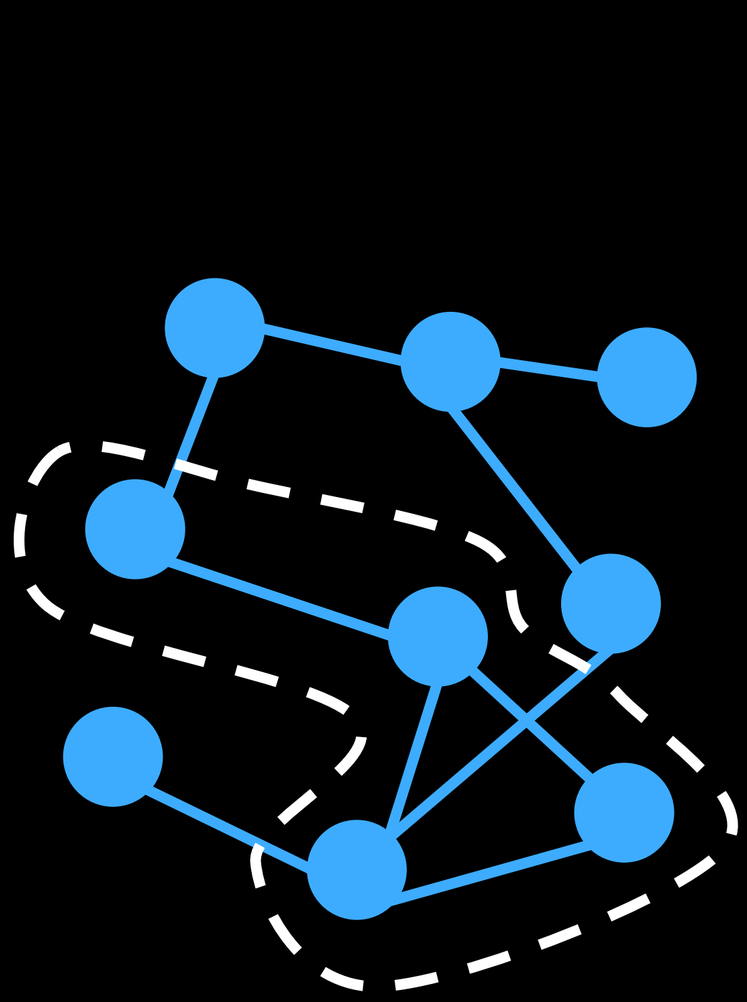
Graph Convolutional Network

$$Z = GXW$$

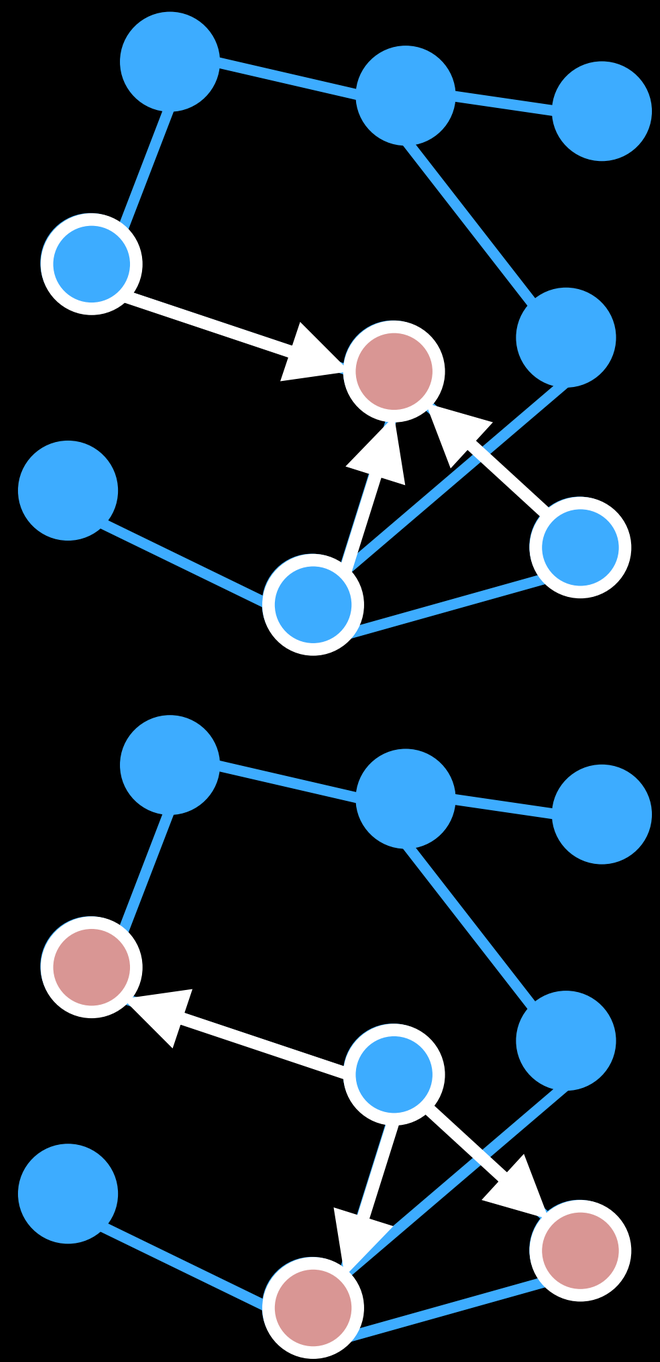


Kipf. Semi-Supervised Classification with Graph Convolutional Networks. 2017

Graph Convolutional Network



Propagation



Connecting Non-local and GCN

The Non-local Operator:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j)\ g(x_j) \qquad z_i = y_i W + x_i$$

$$= \sum_{\forall j} \frac{f(x_i, x_j)}{\sum_{\forall j} f(x_i, x_j)} g(x_j)$$

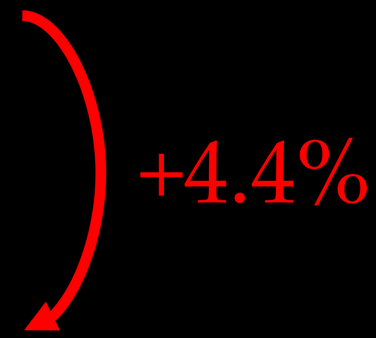
$$= \sum_{\forall j} G_{ij}\ g(x_j) \qquad = \sum_{\forall j} G_{ij}\ g(x_j)\, W + x_i$$

$$\boxed{Z = G\ g(X)\ W} + X$$

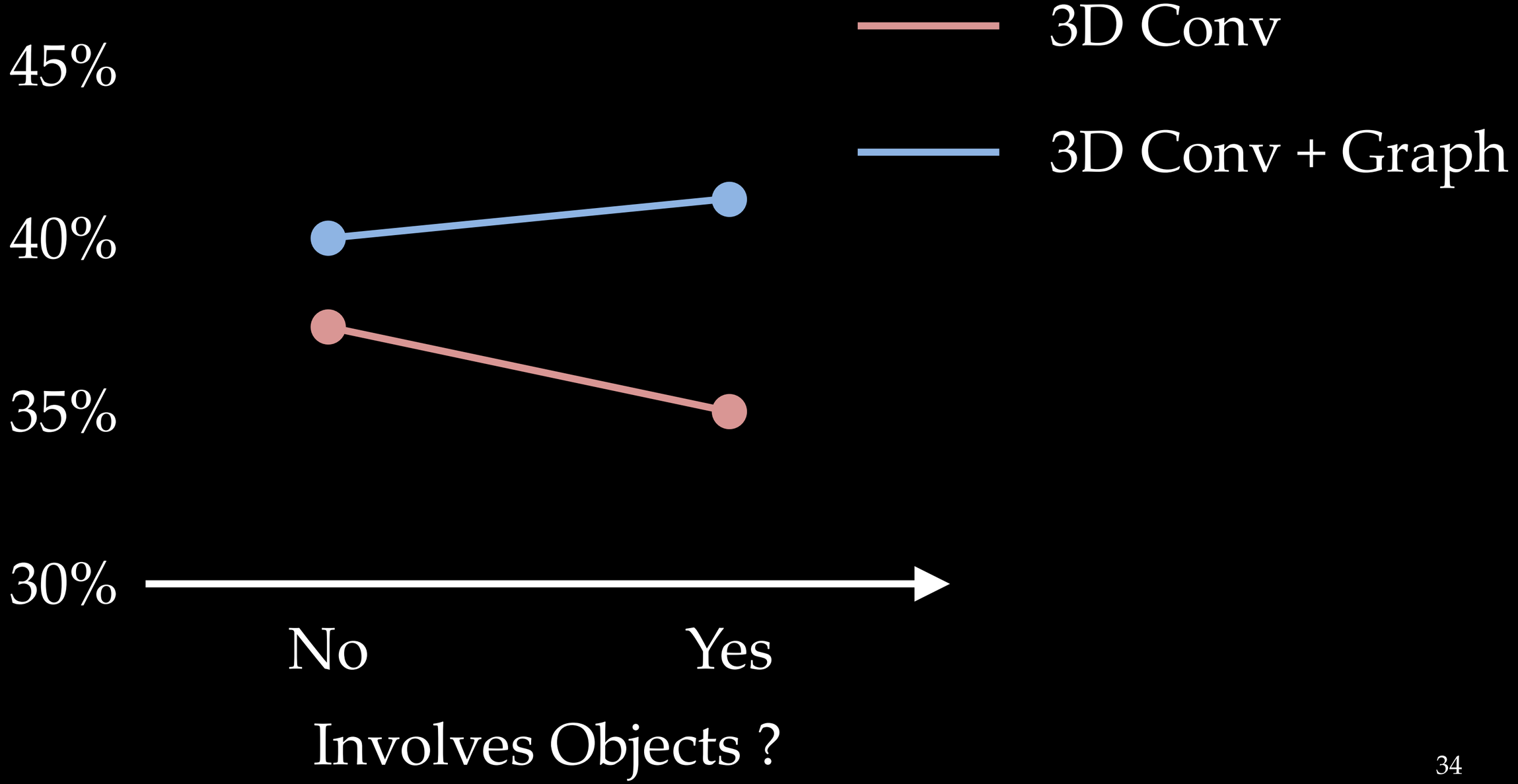
The Graph Convolution

Action Recognition on Charades

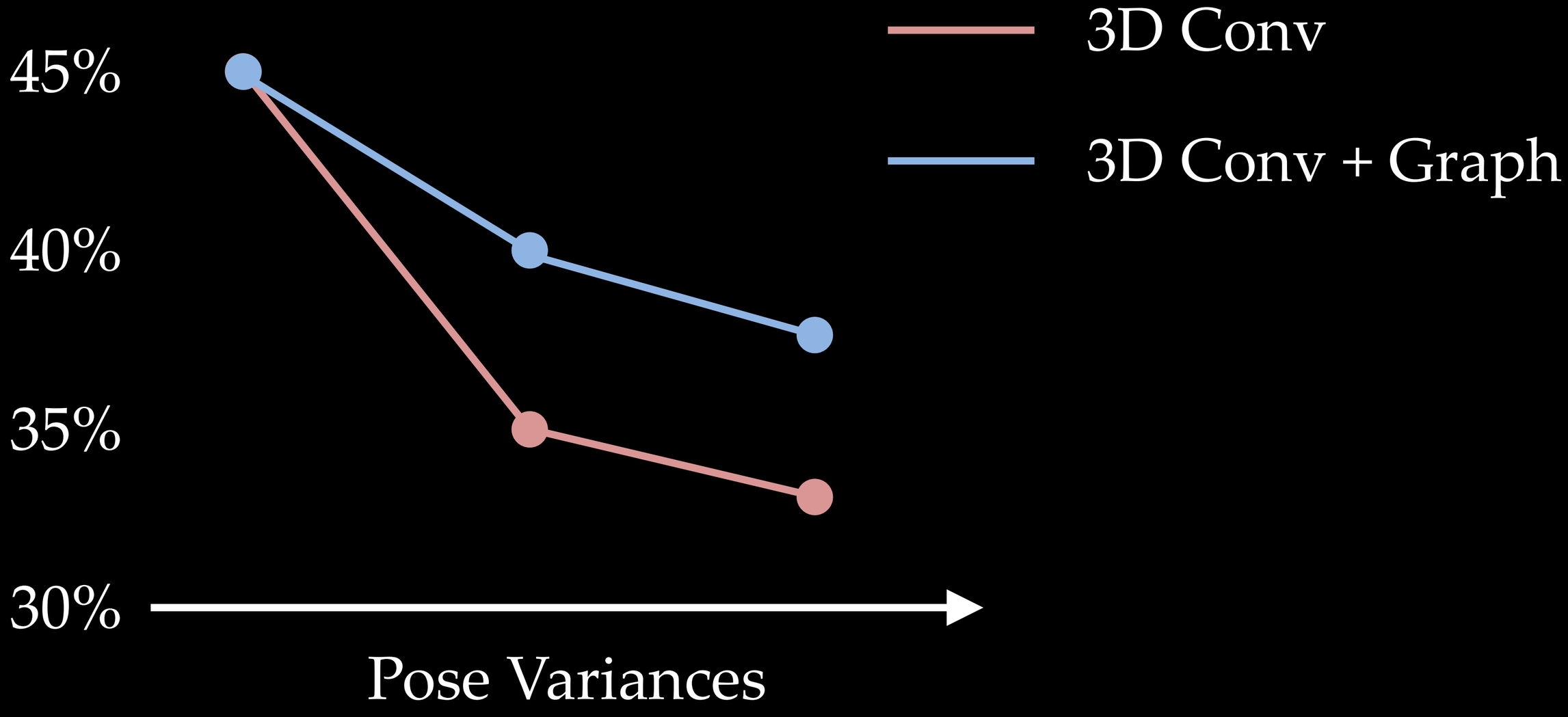
Method	mean AP
3D Conv	31.8%
3D Conv + Non-local	33.5%
3D Conv + Region Graph	36.2%



Action Recognition on Charades



Action Recognition on Charades



Connection to Mean-Shift

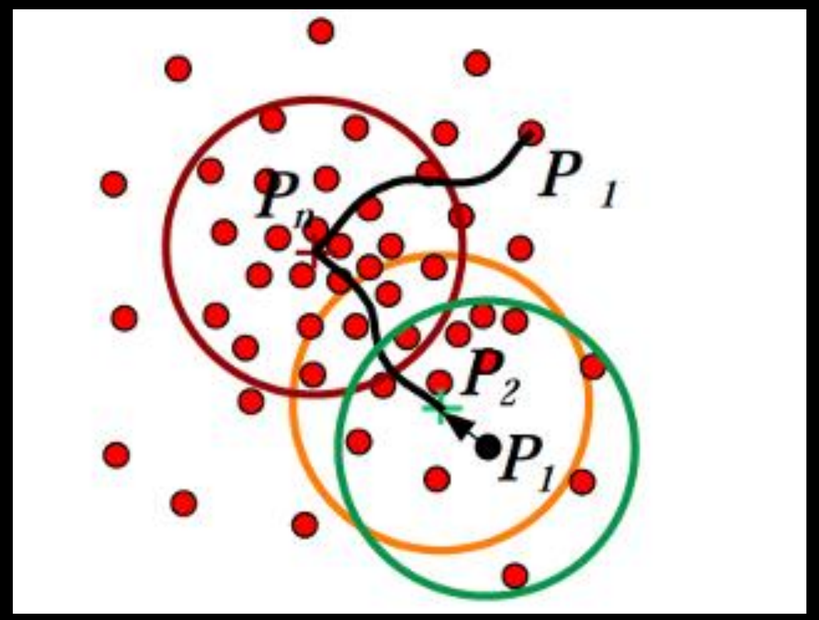
The Non-local Operator:

$$y_i = \sum_{\forall j} \frac{f(x_i, x_j)}{\sum_{\forall j} f(x_i, x_j)} g(x_j)$$

The Mean-Shift Clustering:

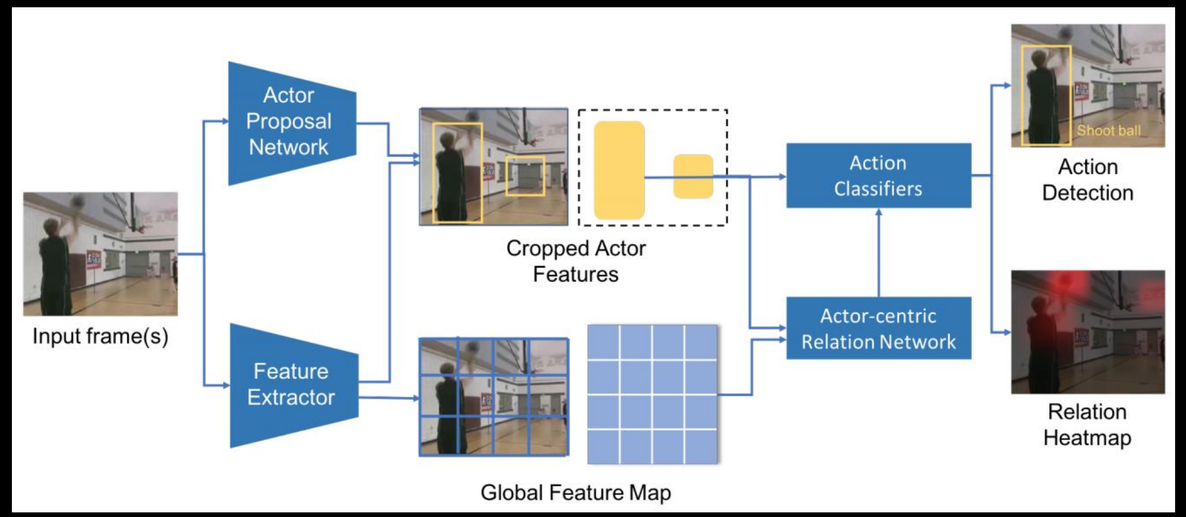
$$m(x) = \sum_{x_j \in N(x)} \frac{K(x, x_j)}{\sum_{x_j \in N(x)} K(x, x_j)} x_j$$

Converging to the same mean?

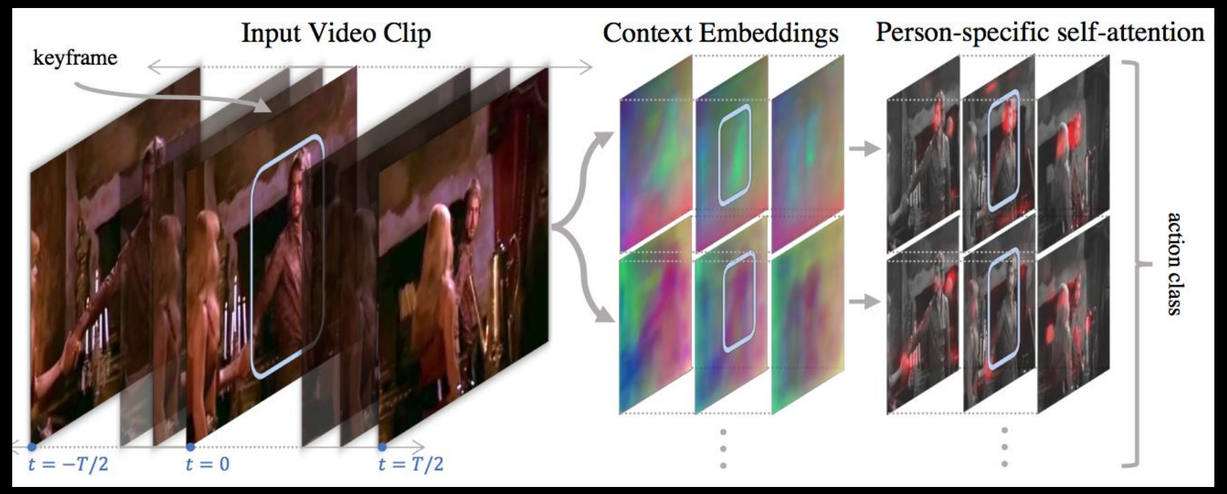


https://tw.rpi.edu/web/project/JeffersonProjectAtLakeGeorge/Clustering

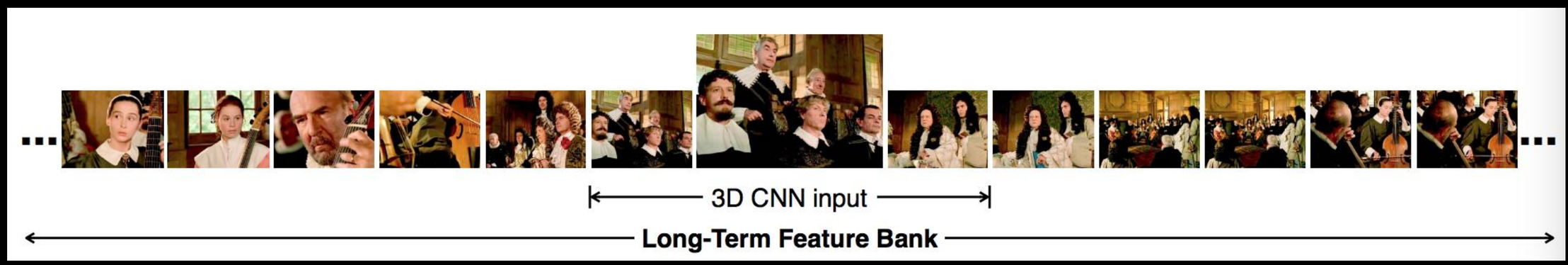
Recent Related Work



Actor-Centric Relation Network
[Sun et al, 2018]

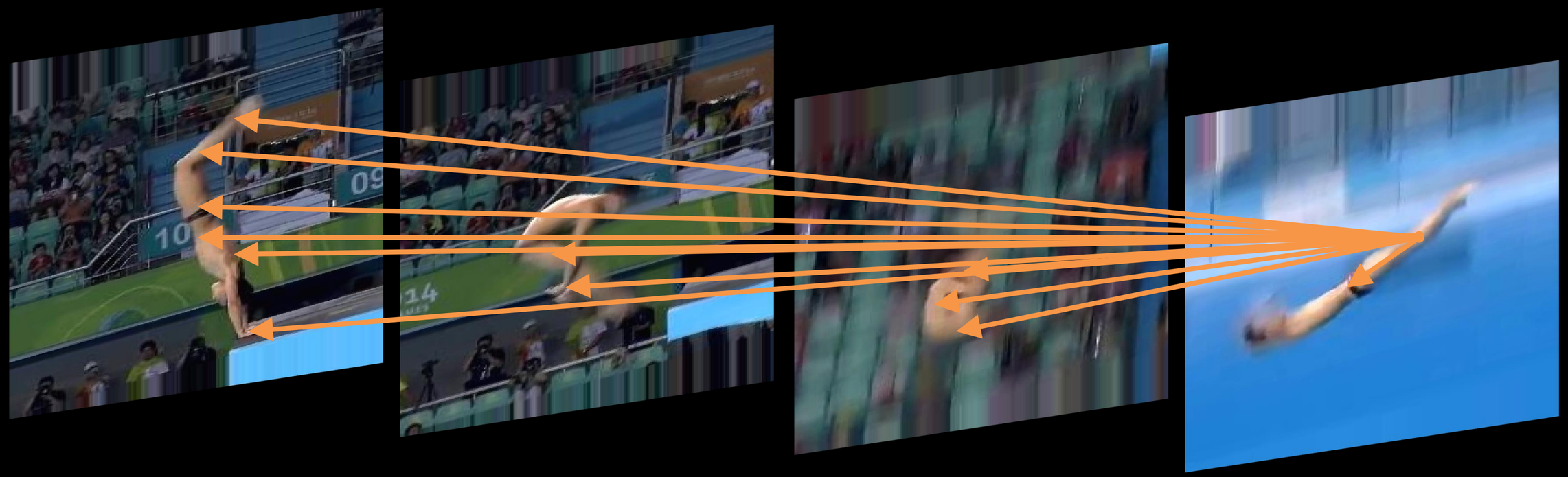


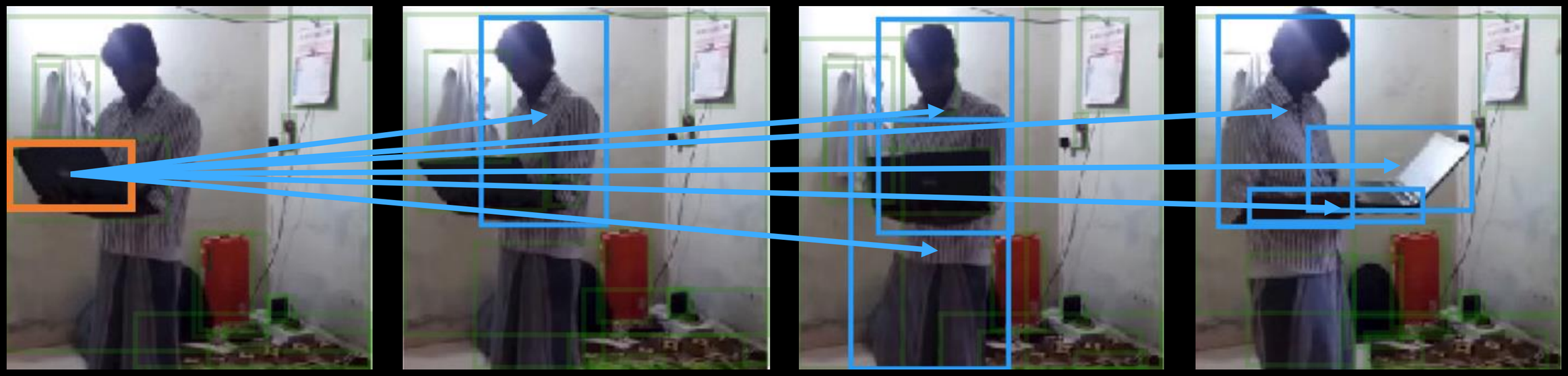
Video Action Transformer Network
[Girdhar et al, 2019]



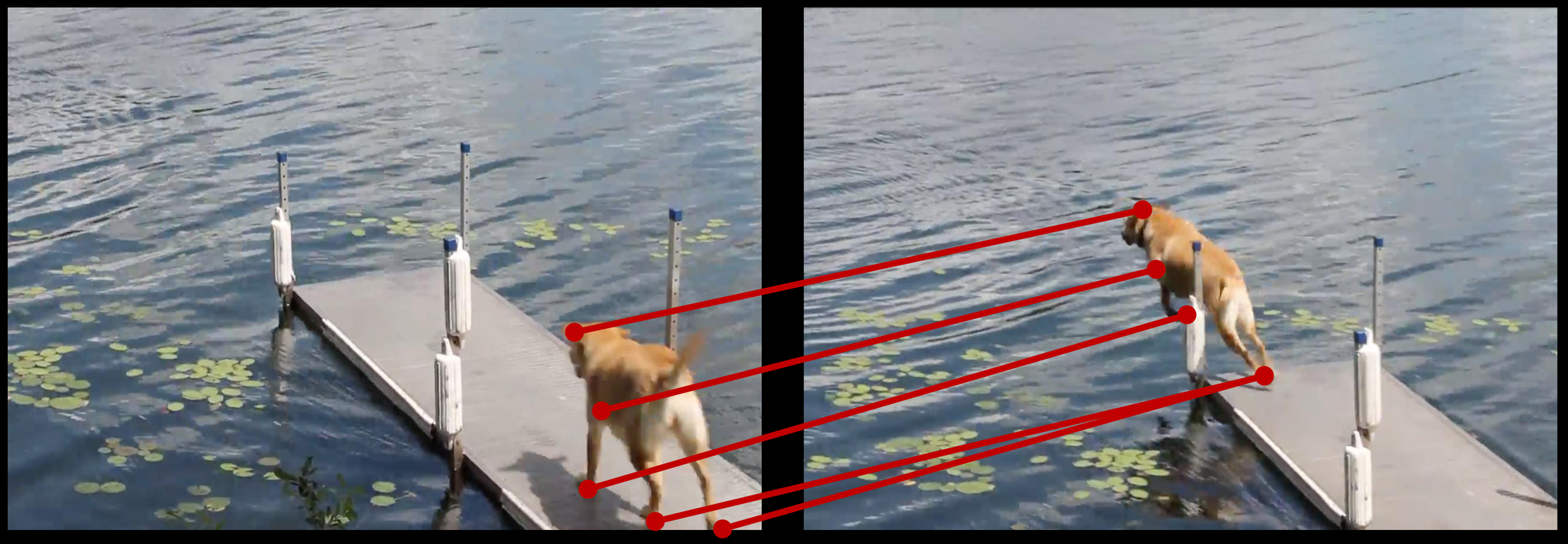
Long-Term Feature Banks for Detailed Video Understanding
[Wu et al, 2019]

Learning Affinity with Semantic Supervision

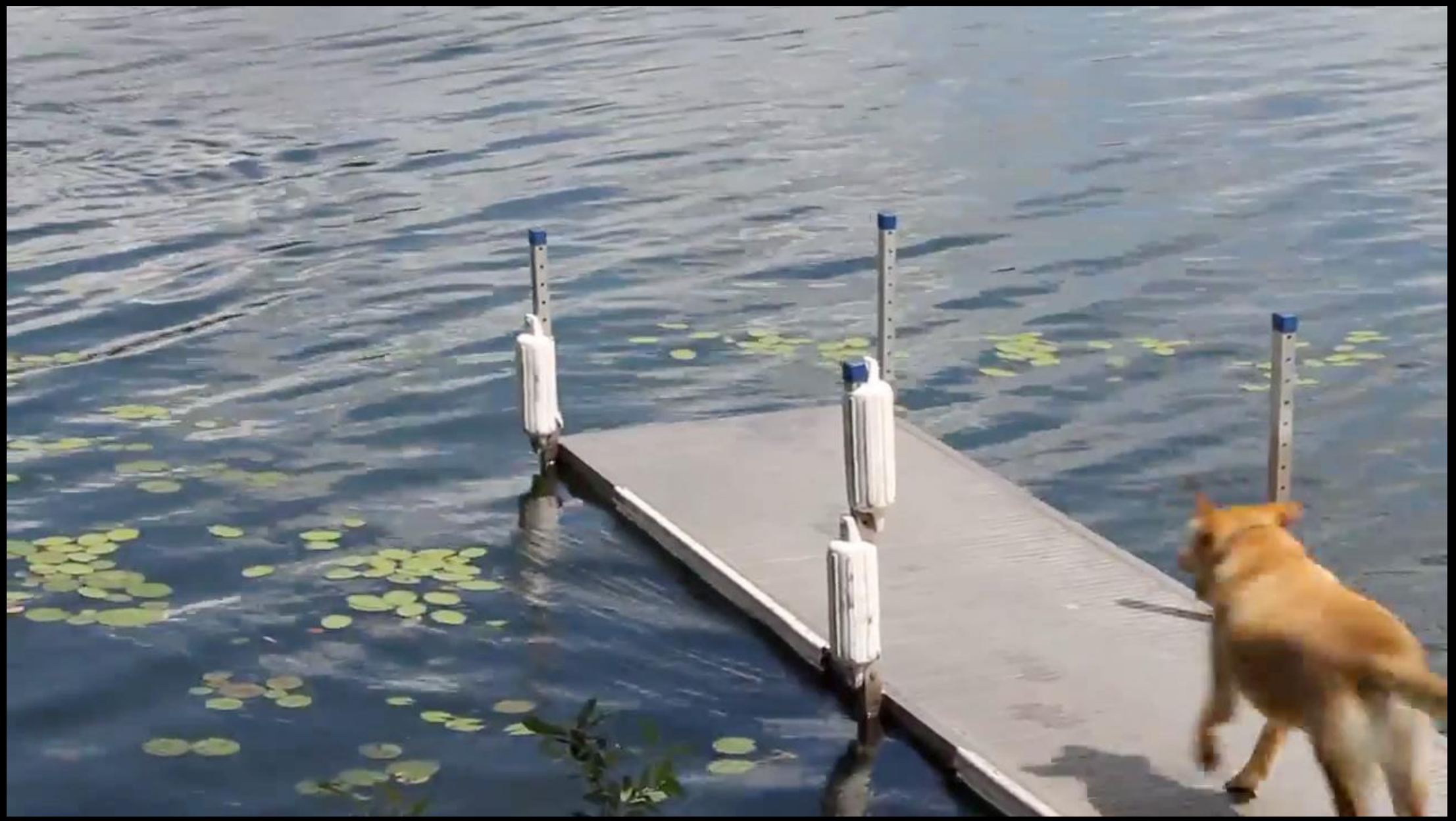




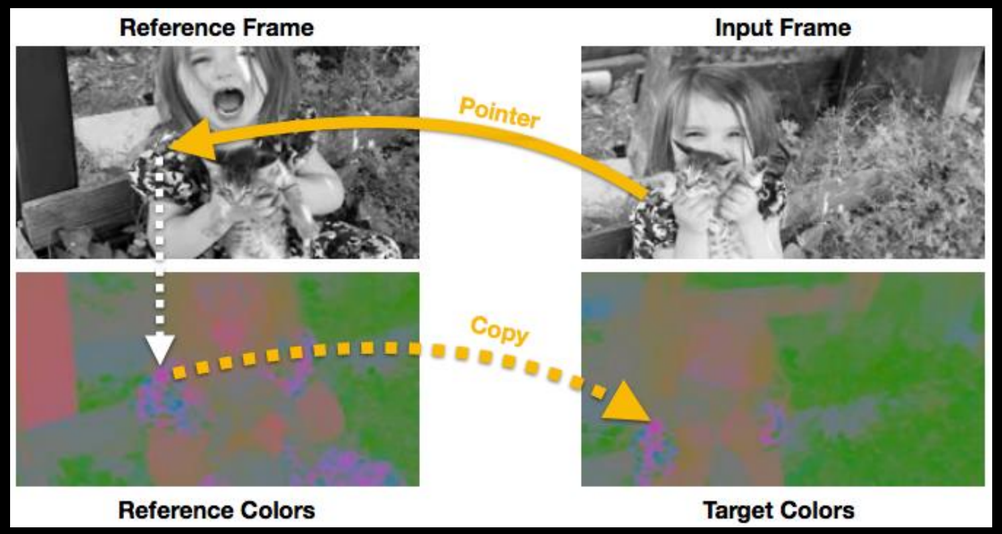
Goal: Learn Correspondence without Human Supervision



The visual world exhibits continuity



Prior Work: Learning from Time



Predict Color in Time
[Vondrick et al, 2018]

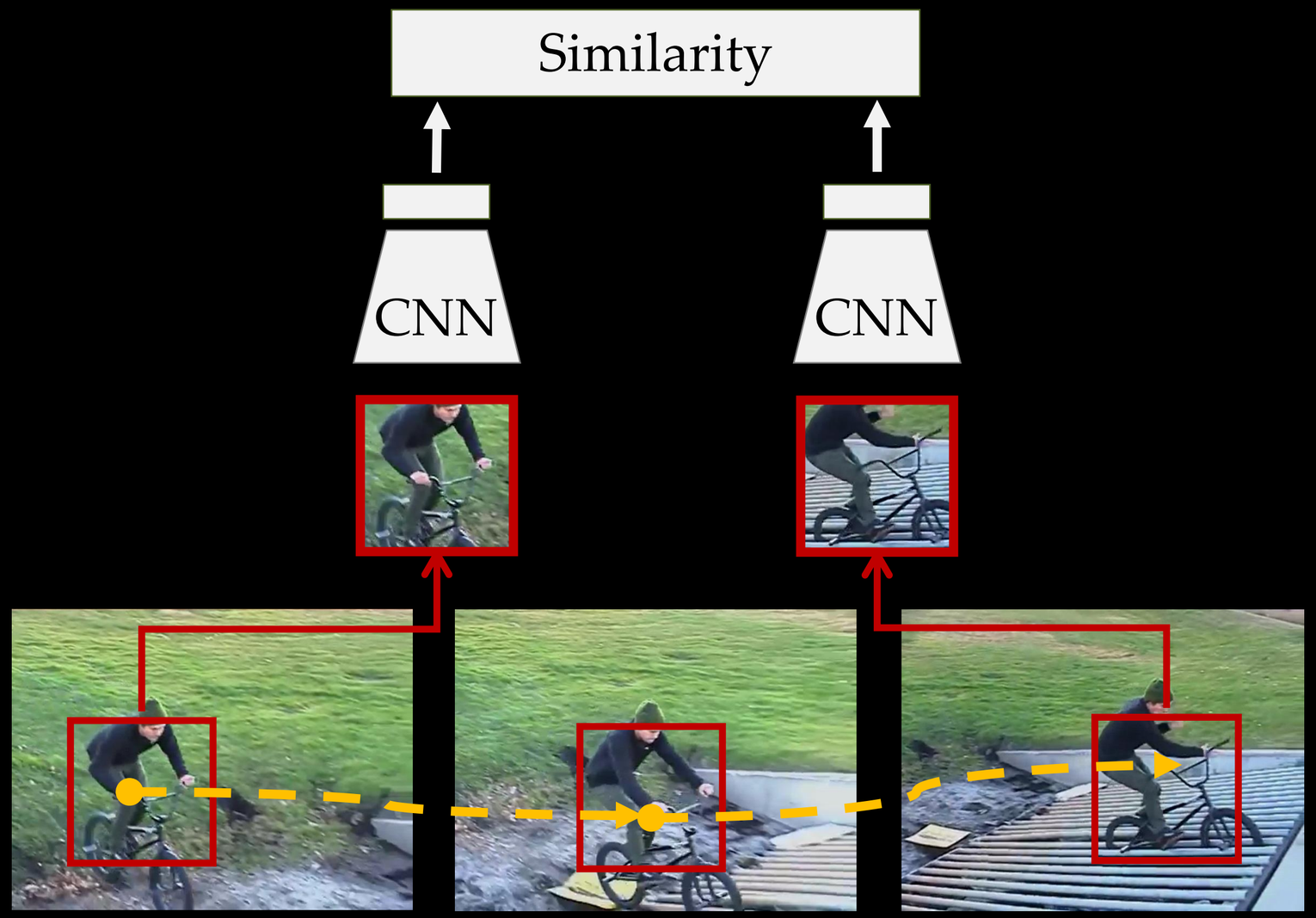
Inputs

Outputs

Predict Pixel in Time
[Mathieu et al, 2015]

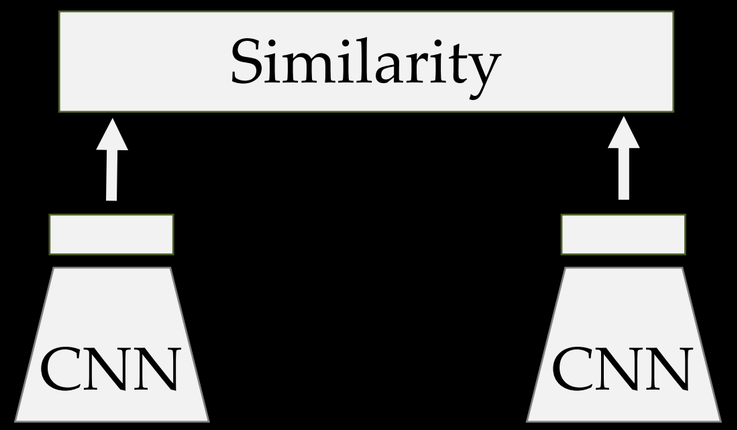
Predict Arrow of Time [Wei et al, 2018]

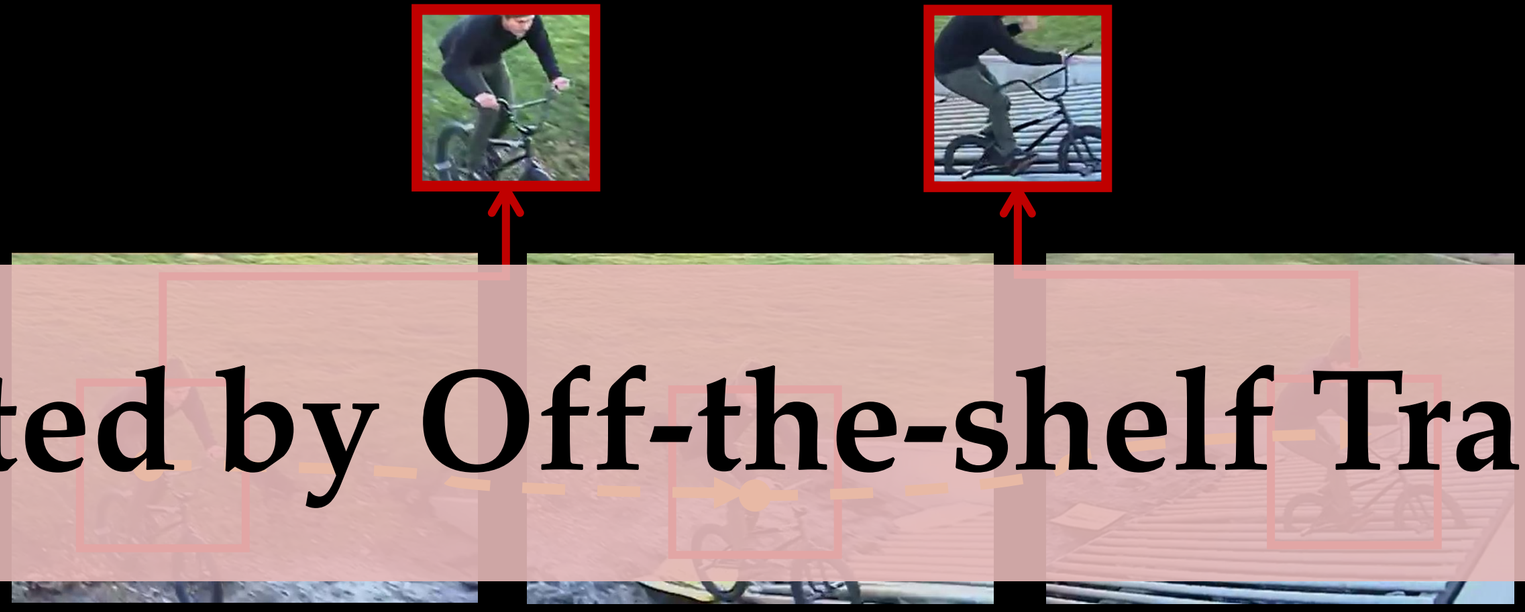
Using Tracking to Learn Features



Tracking → Similarity
[Wang et al, 2015]

Using Tracking to Learn Features





Limited by Off-the-shelf Trackers

Tracking → Similarity
[Wang et al, 2015]

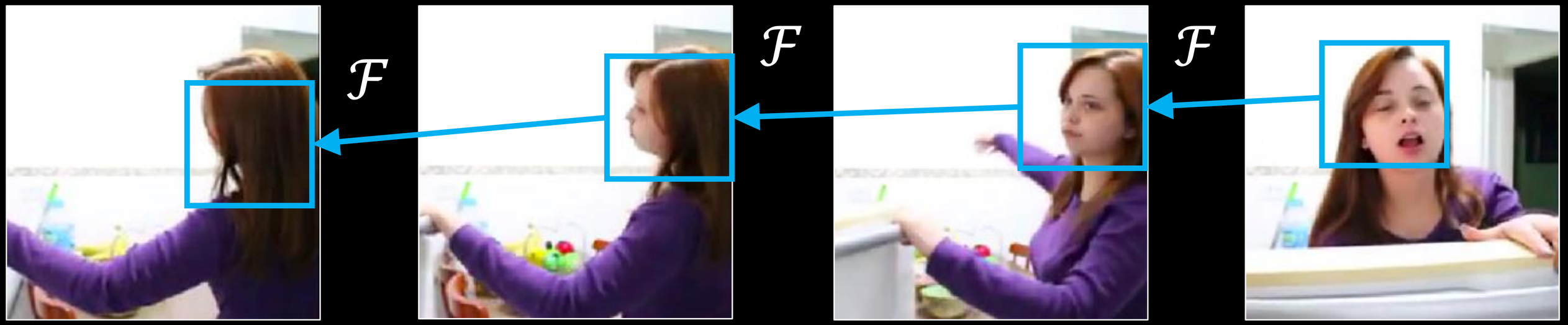
Similarity requires tracking

Tracking requires similarity

Let’s jointly learn both!

Learning to Track

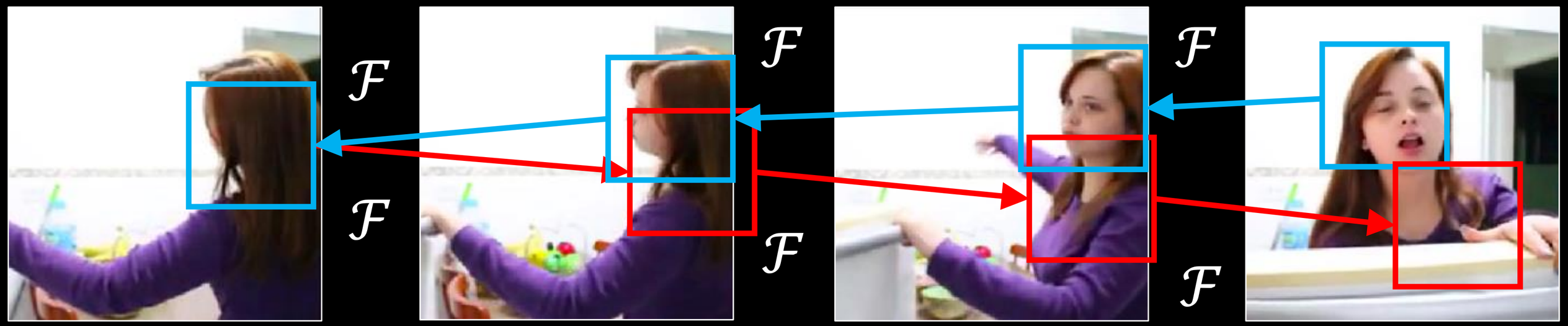
$\mathcal{F}$: a deep tracker



How to obtain supervision?

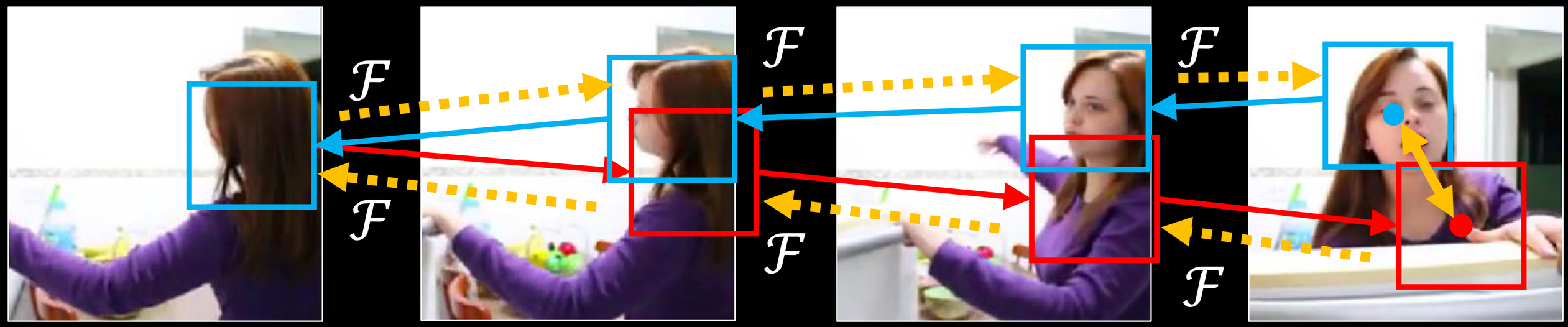
Supervision: Cycle-Consistency in Time

Track backwards



Track forwards, back to the future

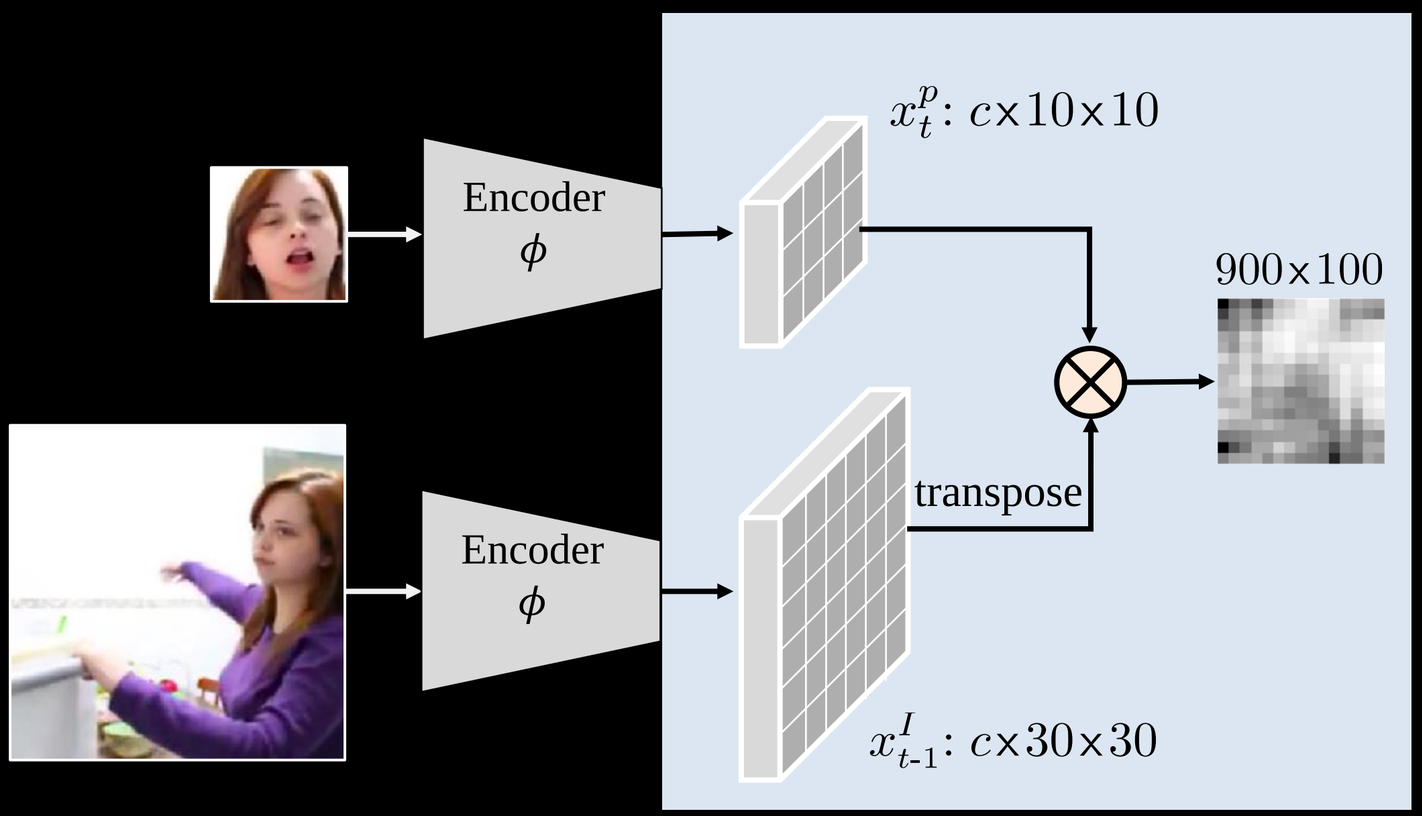
Supervision: Cycle-Consistency in Time

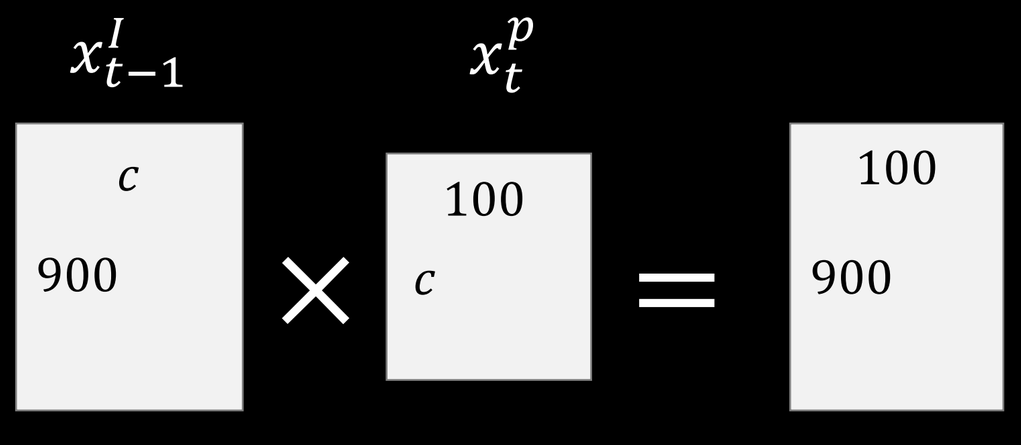


Backpropagation through time along the cycle

Differentiable Tracking

Patch feature in time t: x_t^p





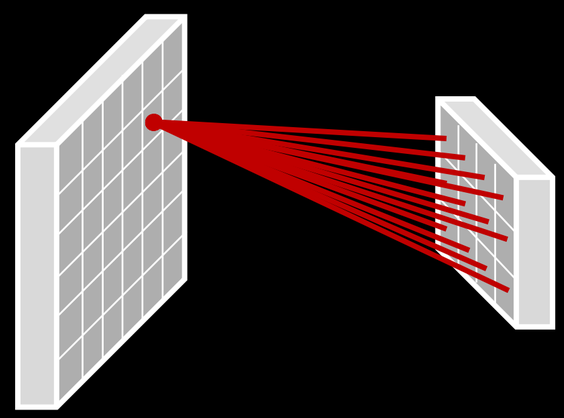


Image feature in time $t-1$: x_{t-1}^I

Differentiable Tracking

Patch feature in time t: x_t^p

Patch feature in time $t-1$: x_{t-1}^p

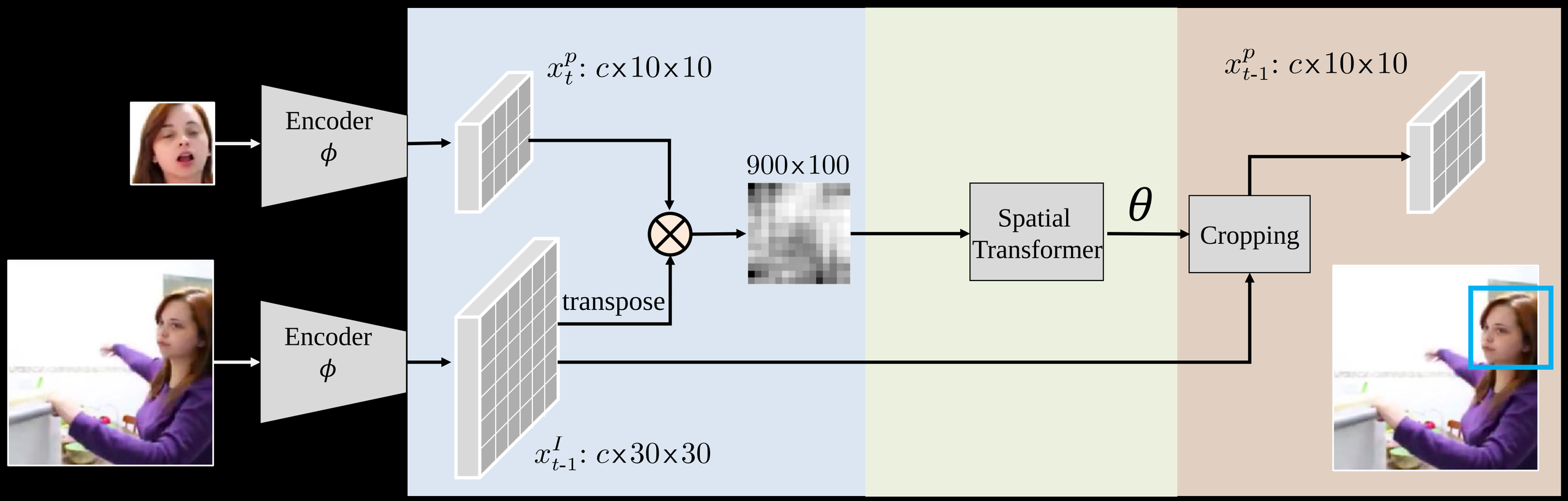
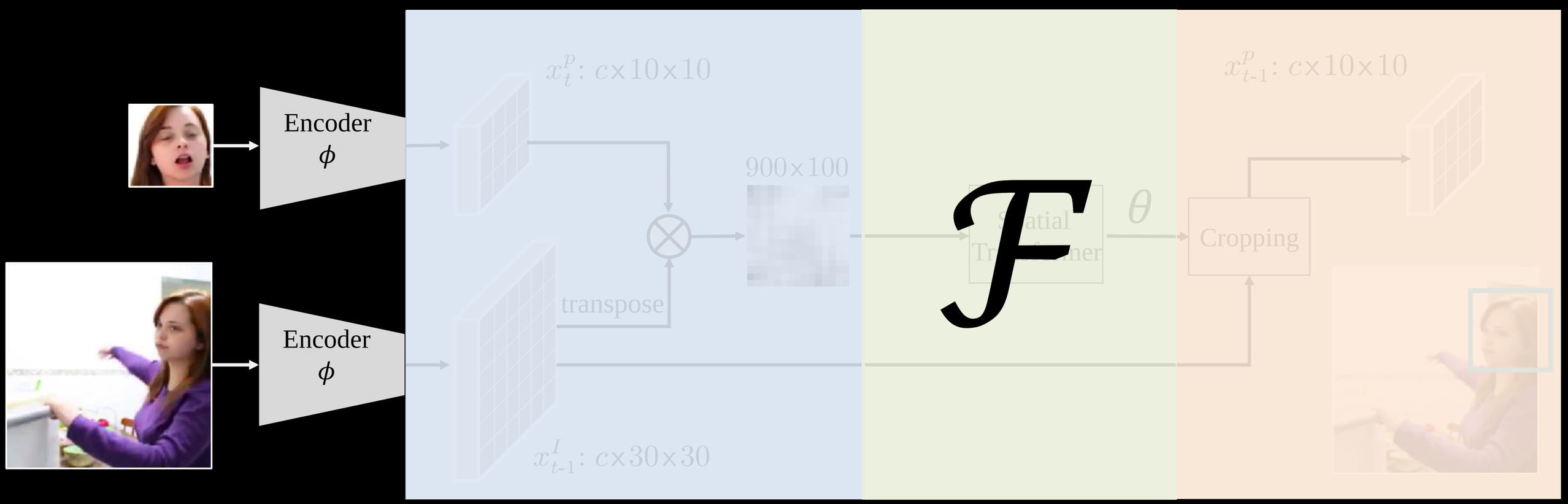


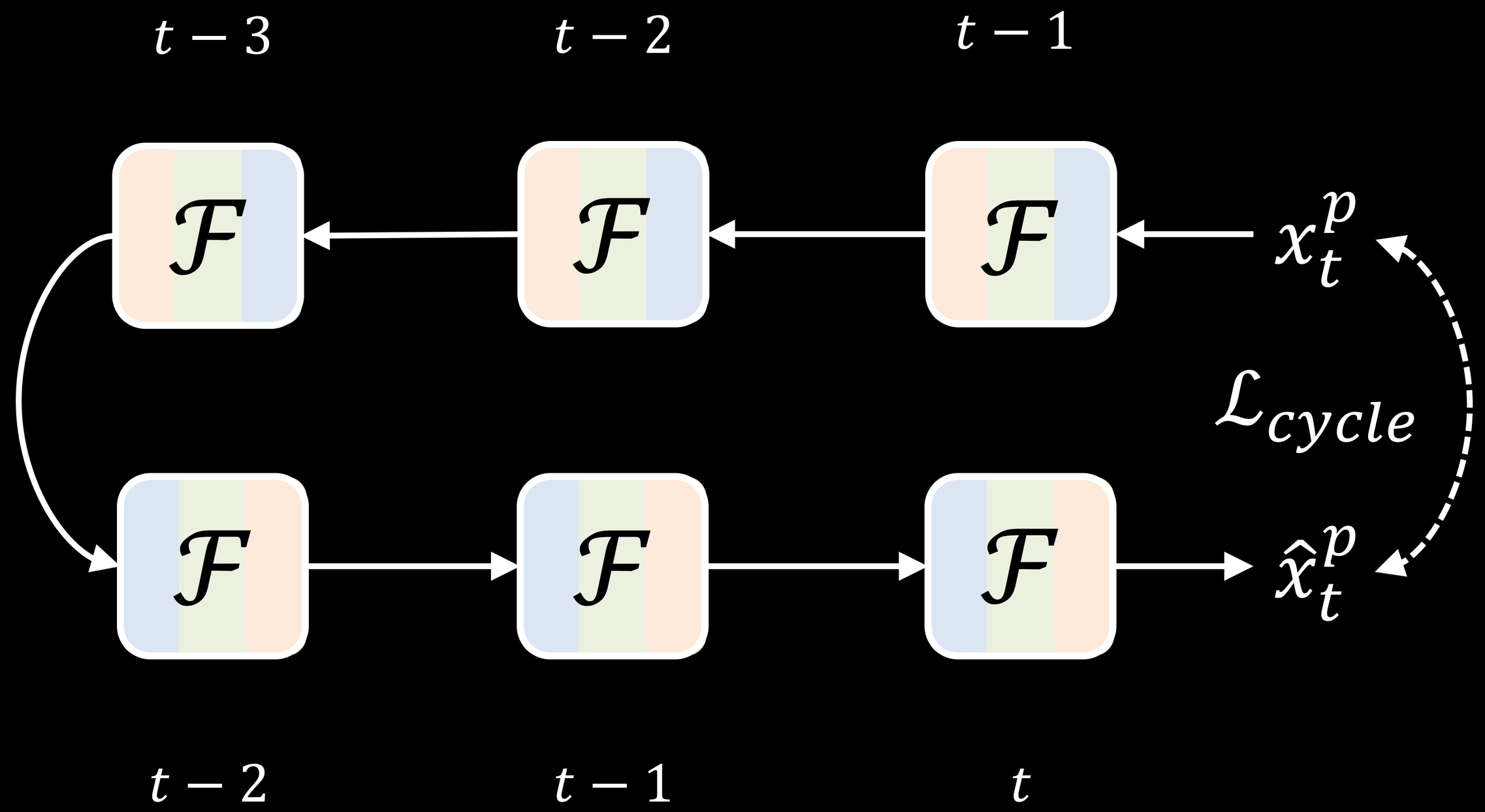
Image feature in time $t-1$: x_{t-1}^I

Differentiable Tracking

$$x_{t-1}^{p} = \mathcal{F}(x_{t-1}^{I}, x_{t}^{p})$$

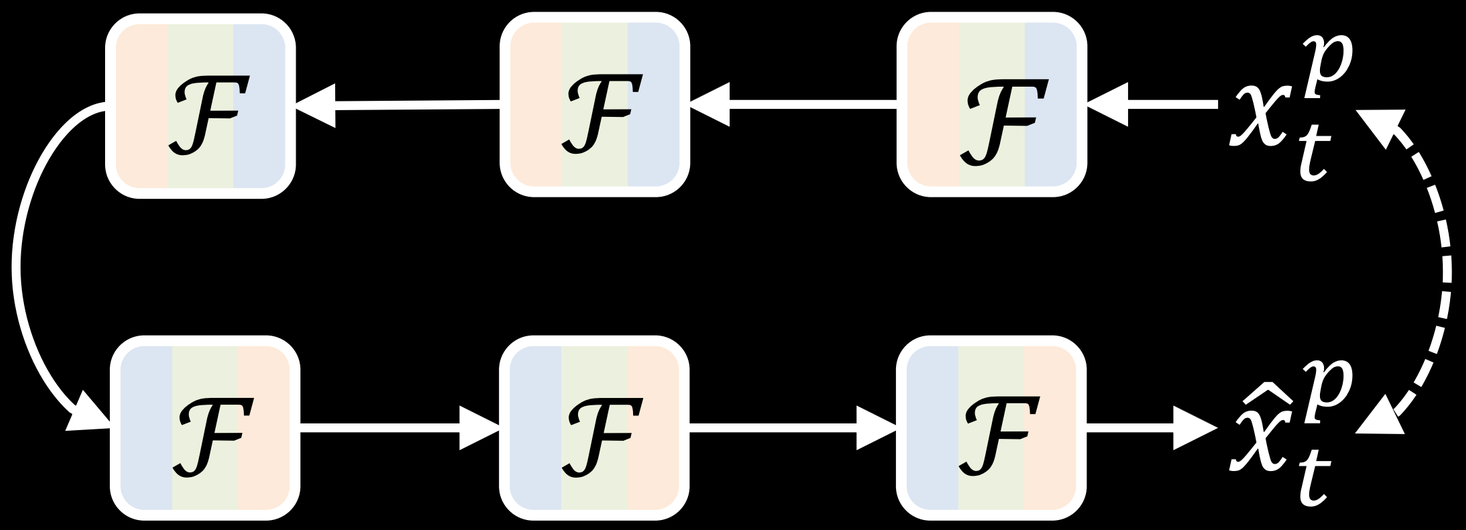


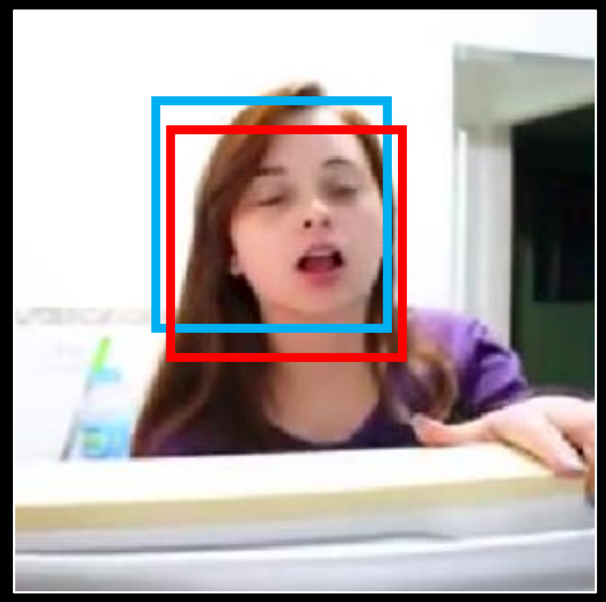
Recurrent Tracking



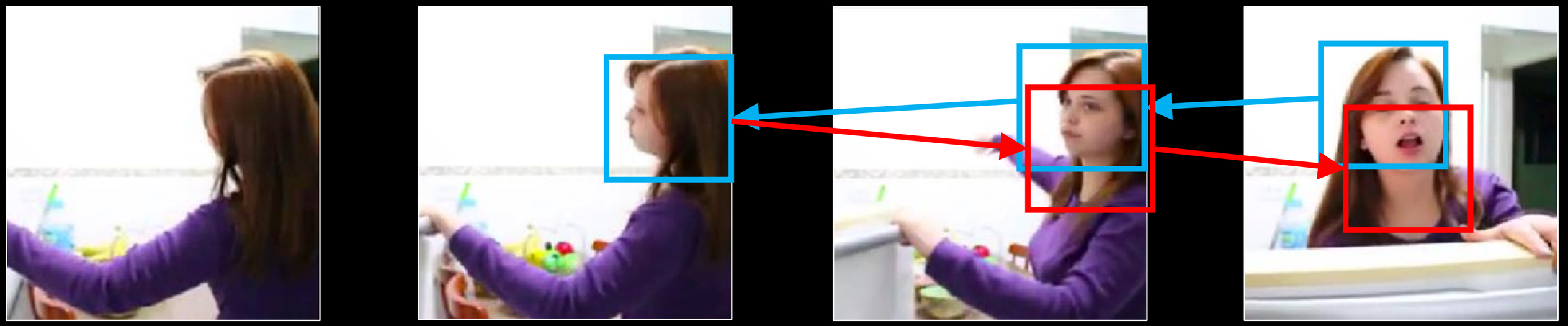
Cycle-Consistency Loss Function

$$\mathcal{L}_{cycle} = ||Loc(\hat{x}_t^p) - Loc(x_t^p)||_2^2$$



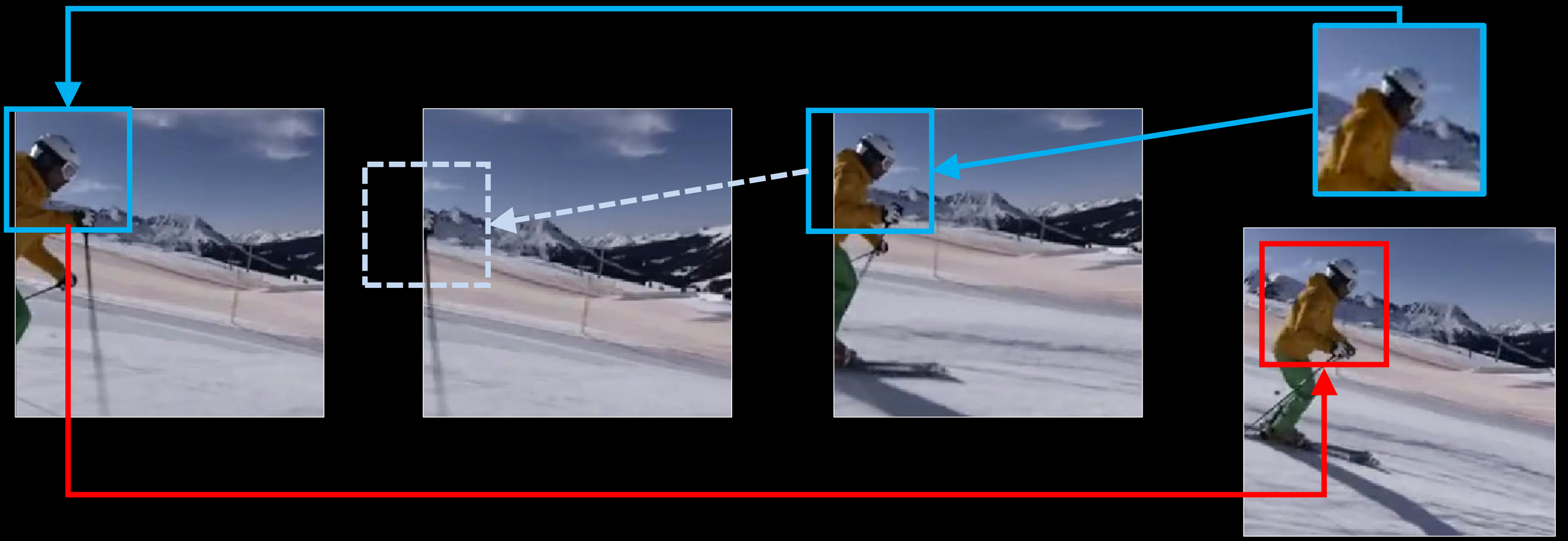


Multiple Cycles



Sub-cycles: a natural curriculum

Skip Cycles

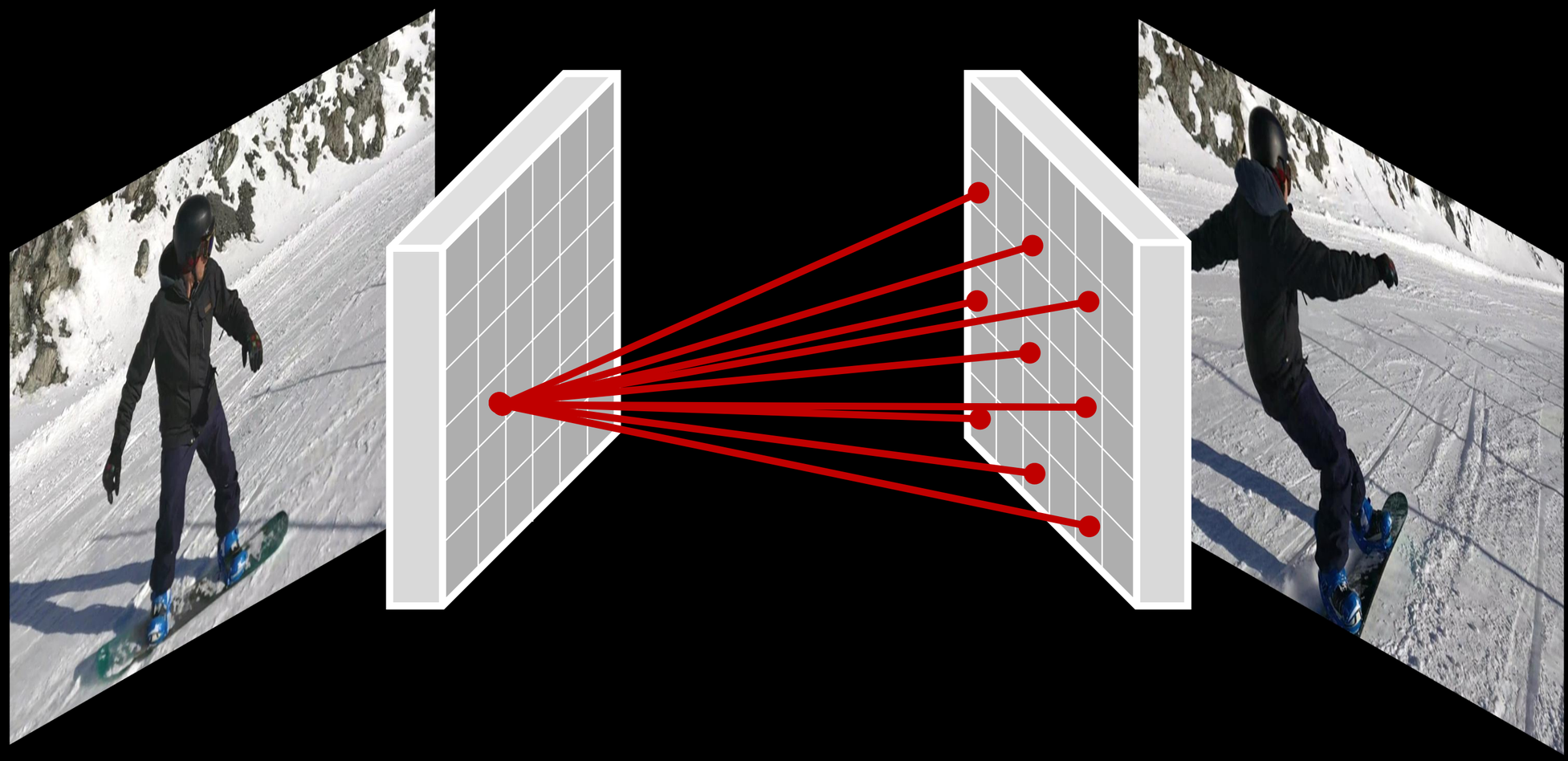


Skip-cycles: skipping occlusions

Visualization of Training

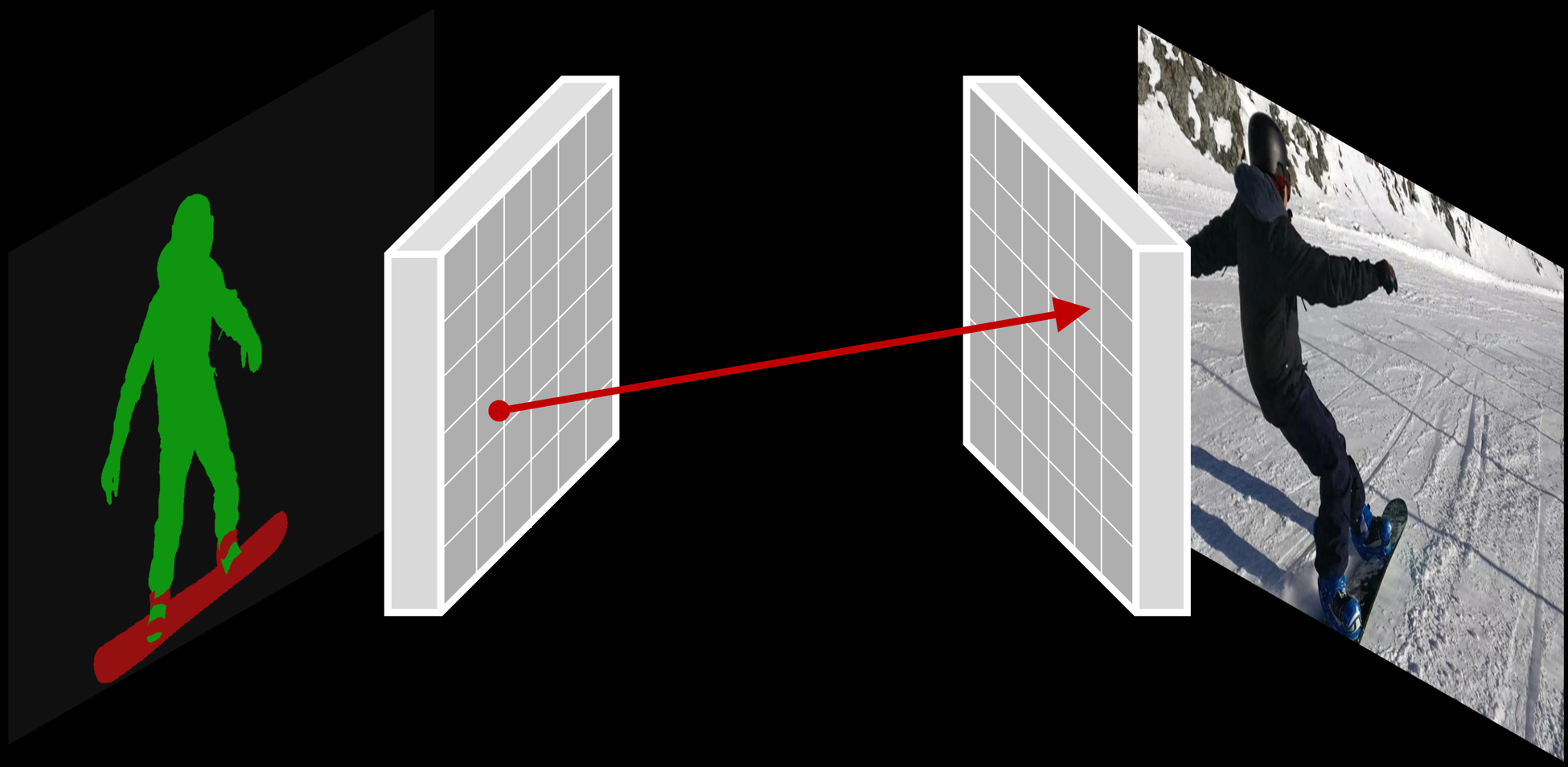
Iteration: 1200

Test Time: Nearest Neighbors in Feature Space ϕ



$t-1$ t

Test Time: Nearest Neighbors in Feature Space ϕ



$t-1$ t

Instance Mask Tracking

DAVIS Dataset

DAVIS Dataset: Pont-Tuset et al. *The 2017 DAVIS Challenge on Video Object Segmentation*. 2017.

Pose Keypoint Tracking

JHMDB Dataset

Comparison

Our Correspondence

Optical Flow

Pose Keypoint Tracking

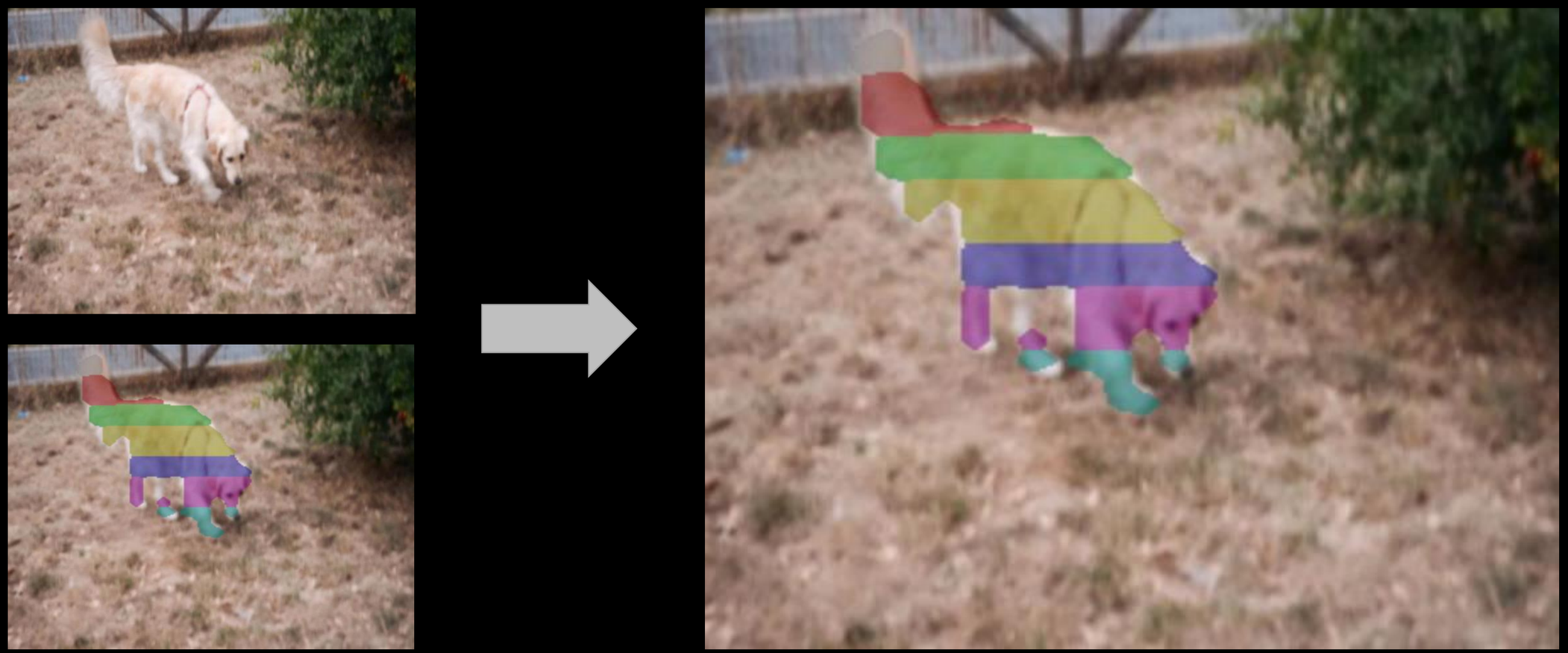
JHMDB Dataset

Method	PCK @.1
Optical Flow	45%
Vondrick et al.	45%
Ours	58%

Vondrick et al. Tracking Emerges by Colorizing Videos. ECCV 2018.

Texture Tracking

DAVIS Dataset



DAVIS Dataset: Pont-Tuset et al. *The 2017 DAVIS Challenge on Video Object Segmentation*. 2017.

Semantic Masks Tracking

Video Instance Parsing Dataset

Zhou et al. *Adaptive Temporal Encoding Network for Video Instance-level Human Parsing*. ACM MM 2018.

Questions?