Lecture 25: Recent directions in motion estimation


30

What can we learn from motion?

Correspondence

"Common Fate"

Wehrtheimer (1938)

We could try learning it from single images…

input

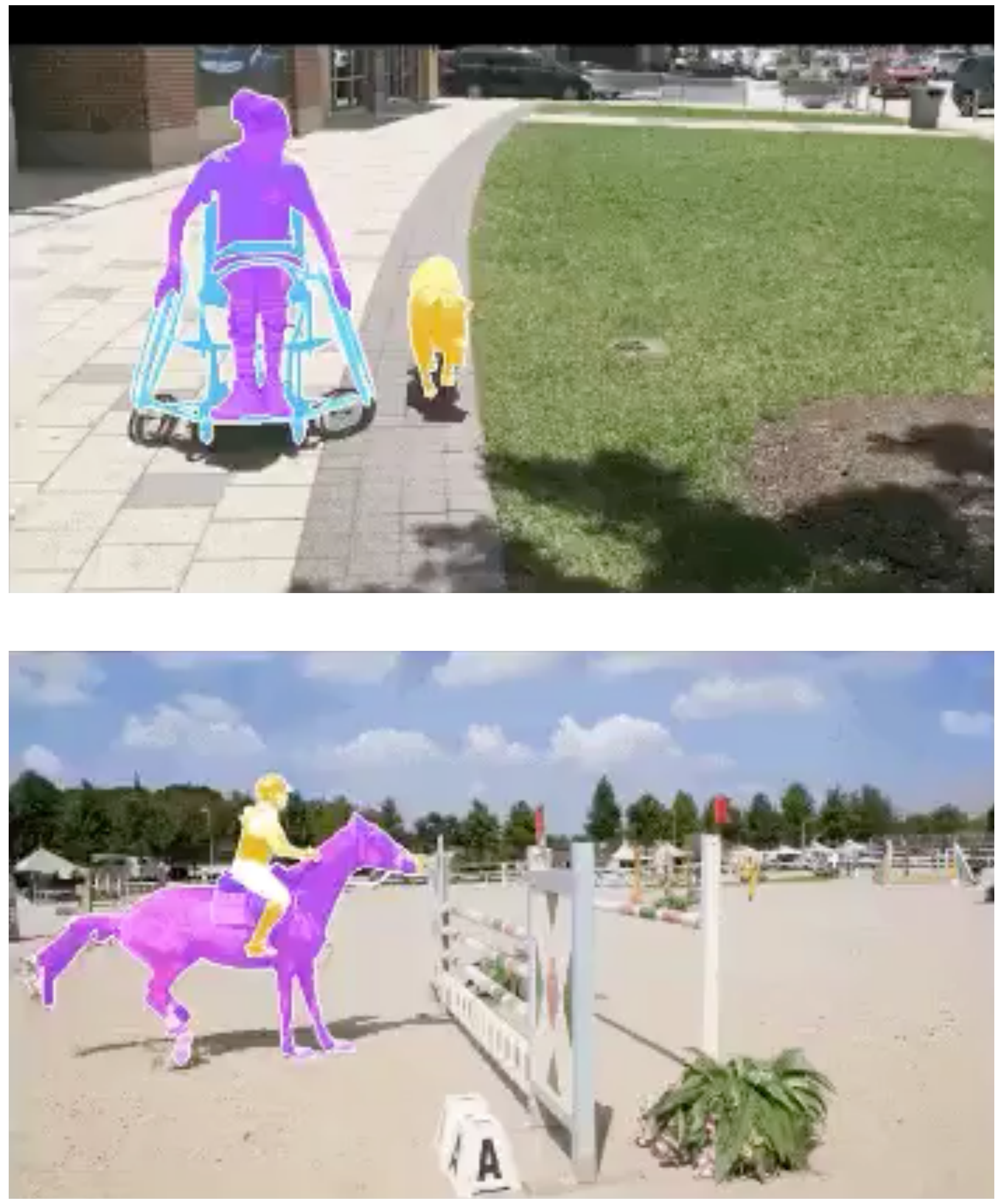
color
crop
flip
blur

SimCLR augmentations (Chen et al., 2020)

…but the video gives us these for free!

Correspondence in computer vision

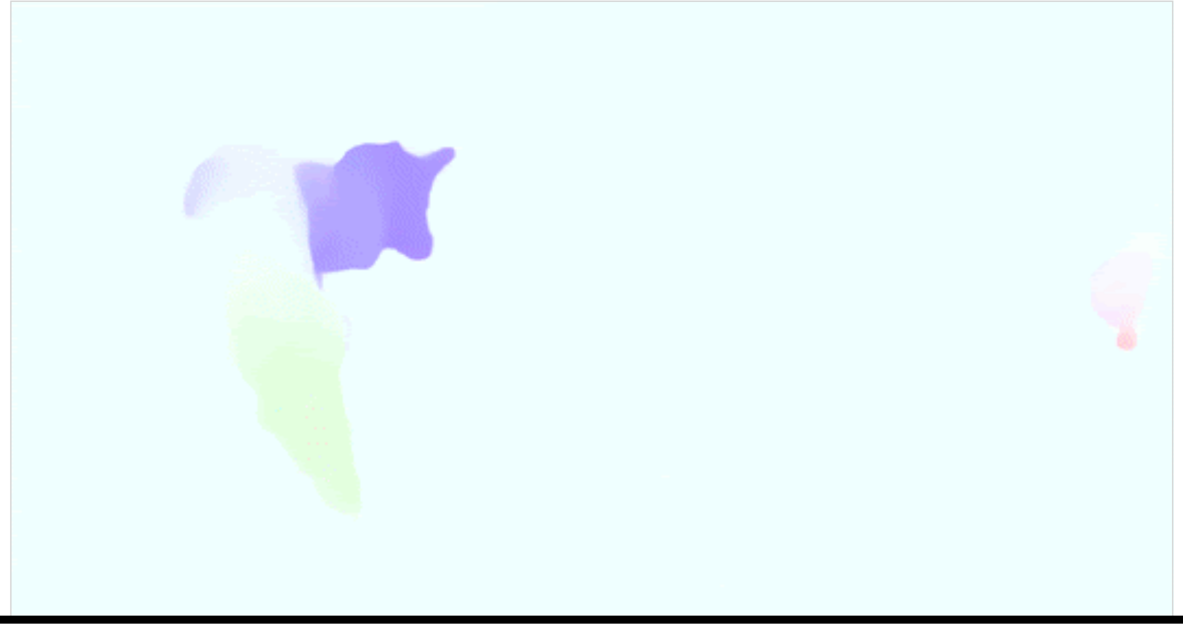
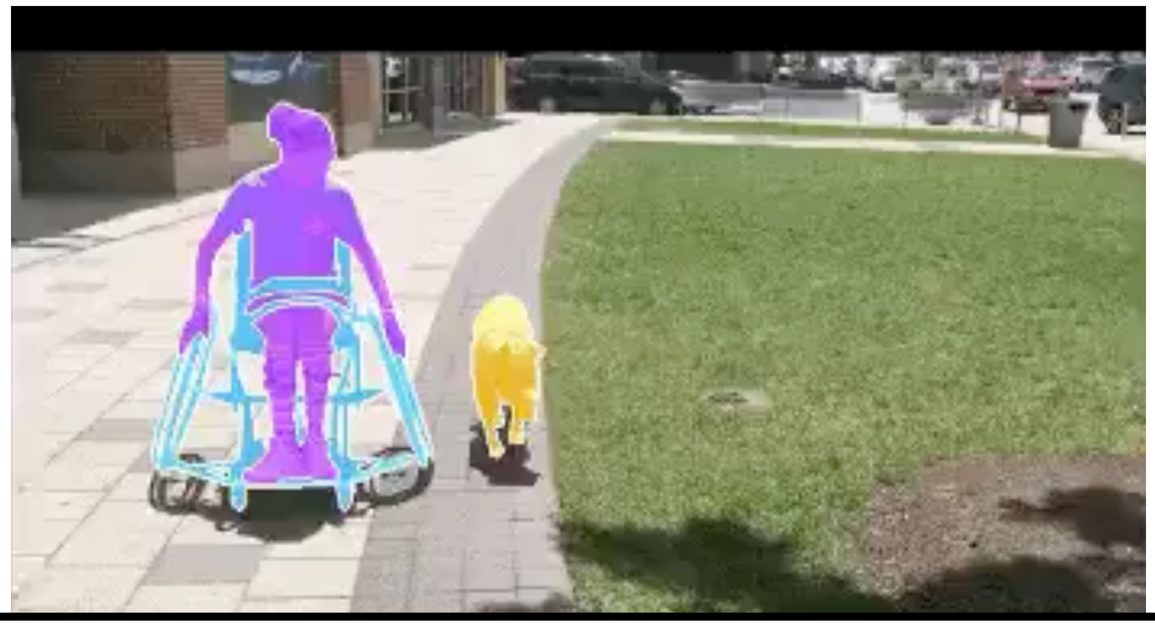
optical flow



segment tracking

pose tracking

Correspondence in computer vision


Each task needs specialized data and model.


optical flow

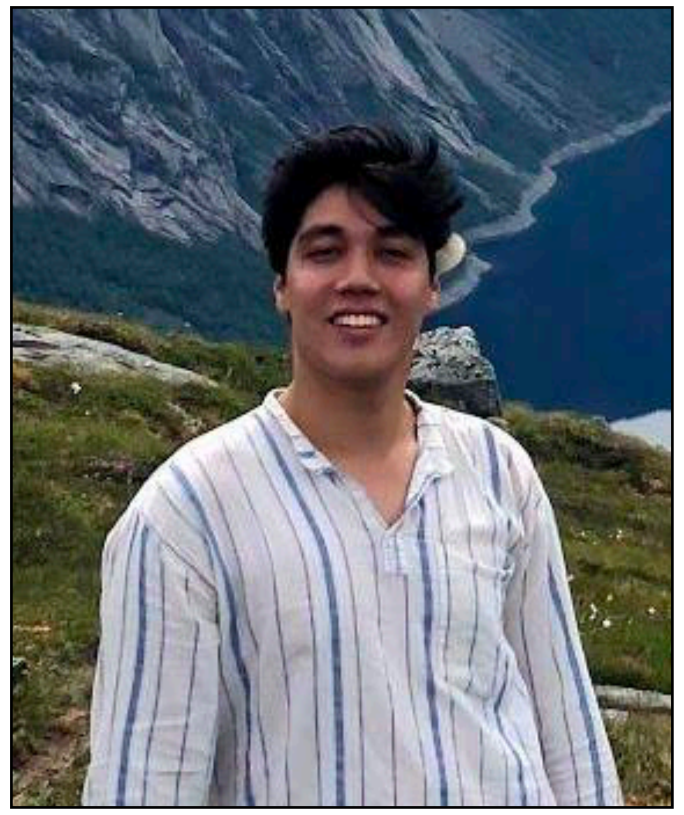
segment tracking

pose tracking

Dense but short-range

Long-range but sparse

Space-Time Correspondence as a Contrastive Random Walk



Allan Jabri

Andrew Owens

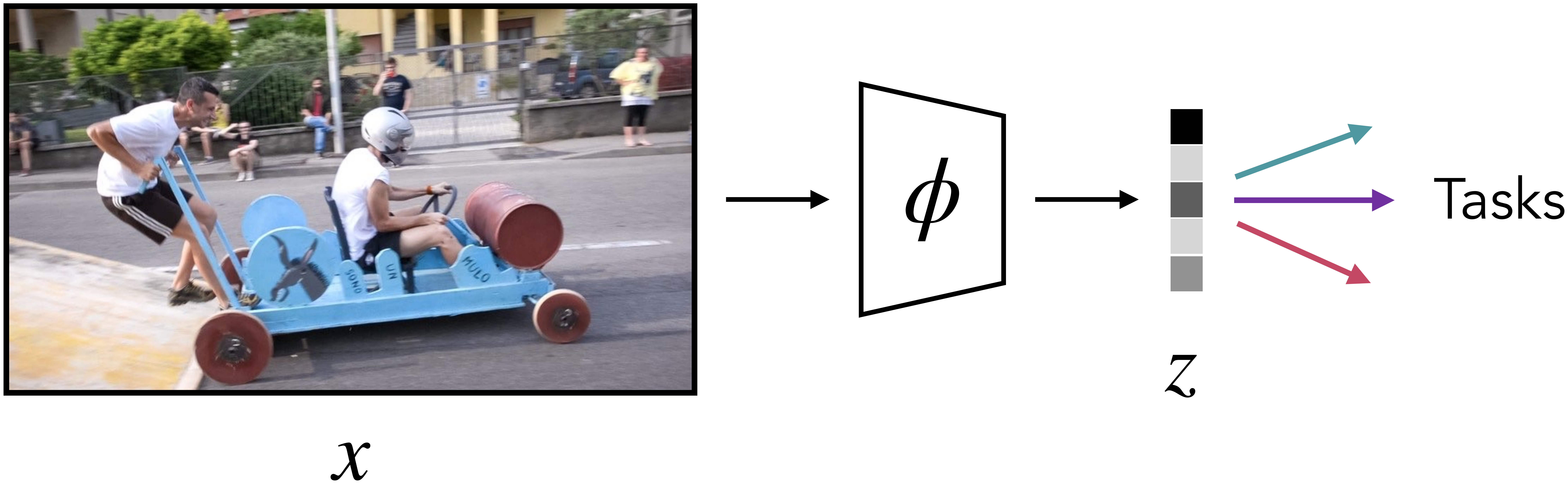
Alexei Efros

NeurIPS 2020 (Oral)

Funding: NSF and Soros Fellowship

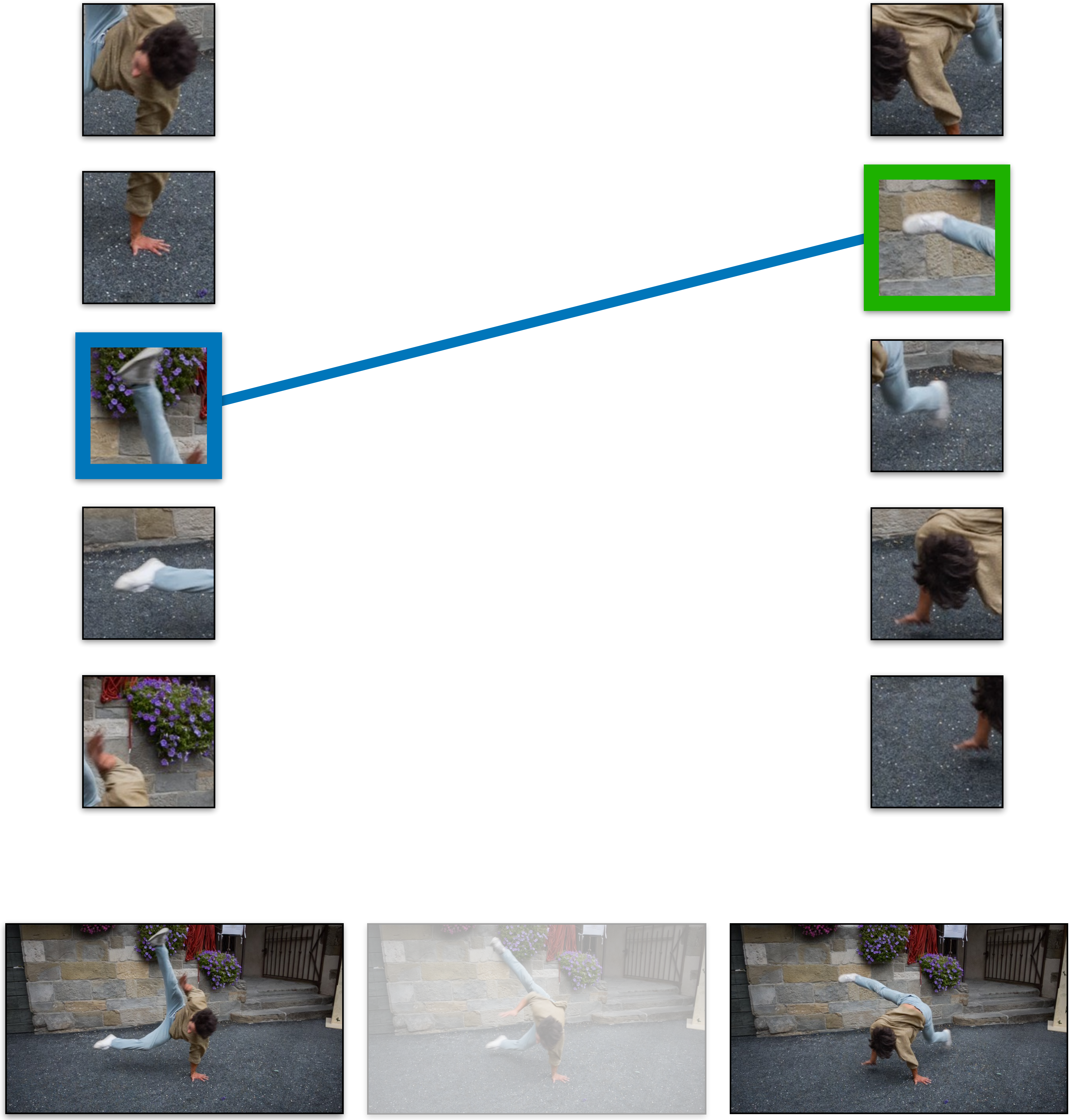
Toward "universal" methods for correspondence

Representation learning for correspondence

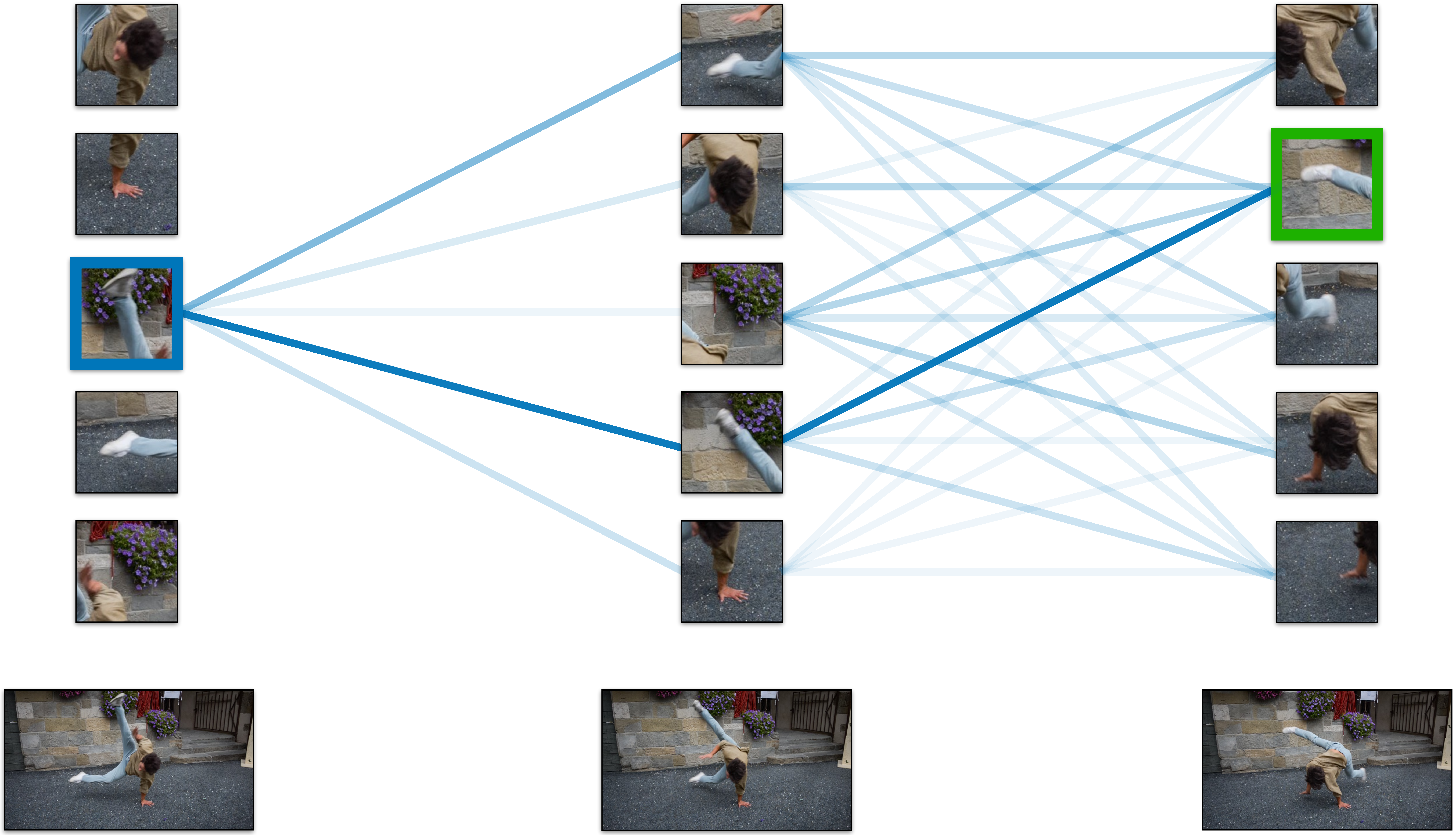


Supervised learning for tracking

Supervised Learning

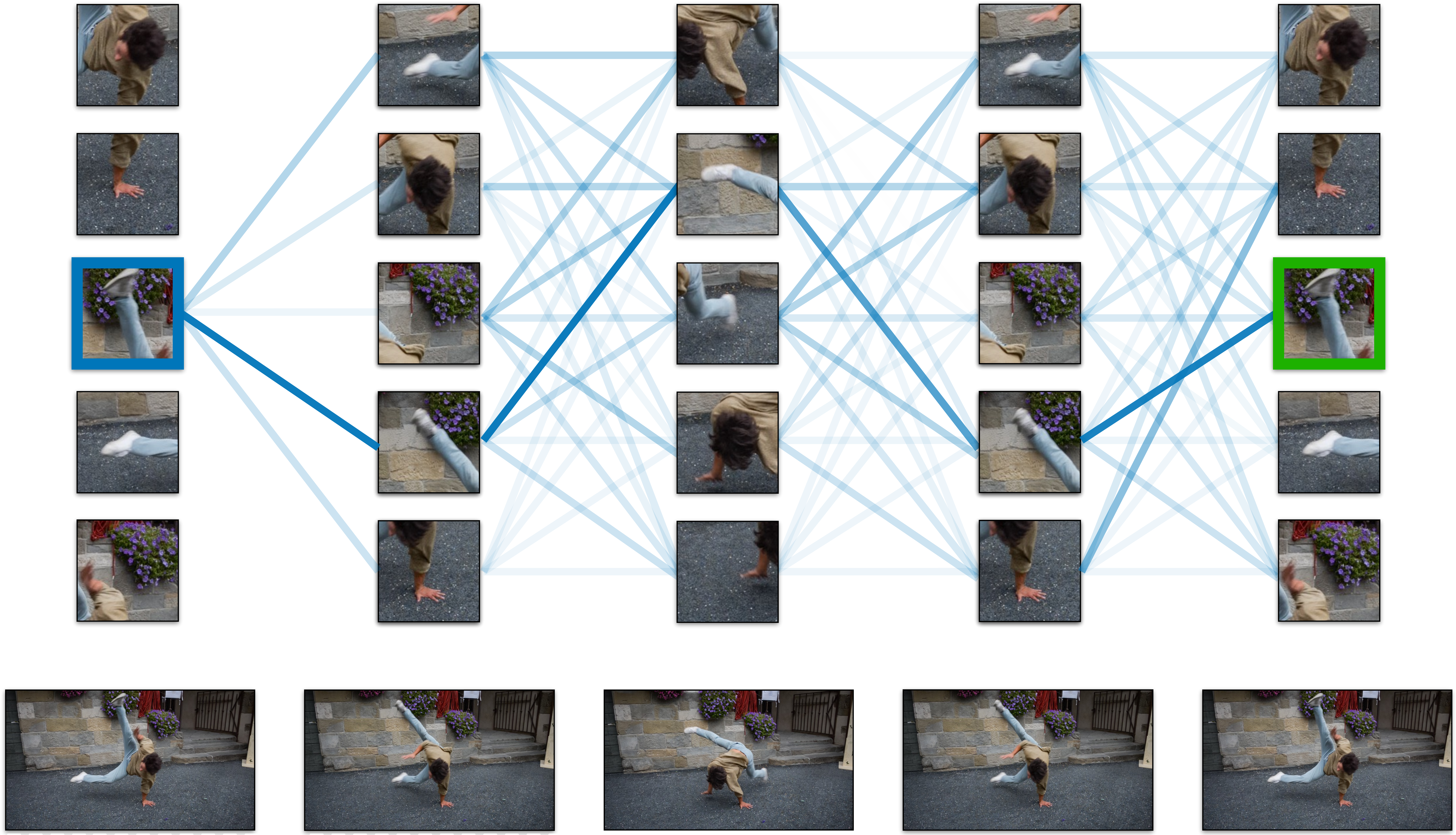


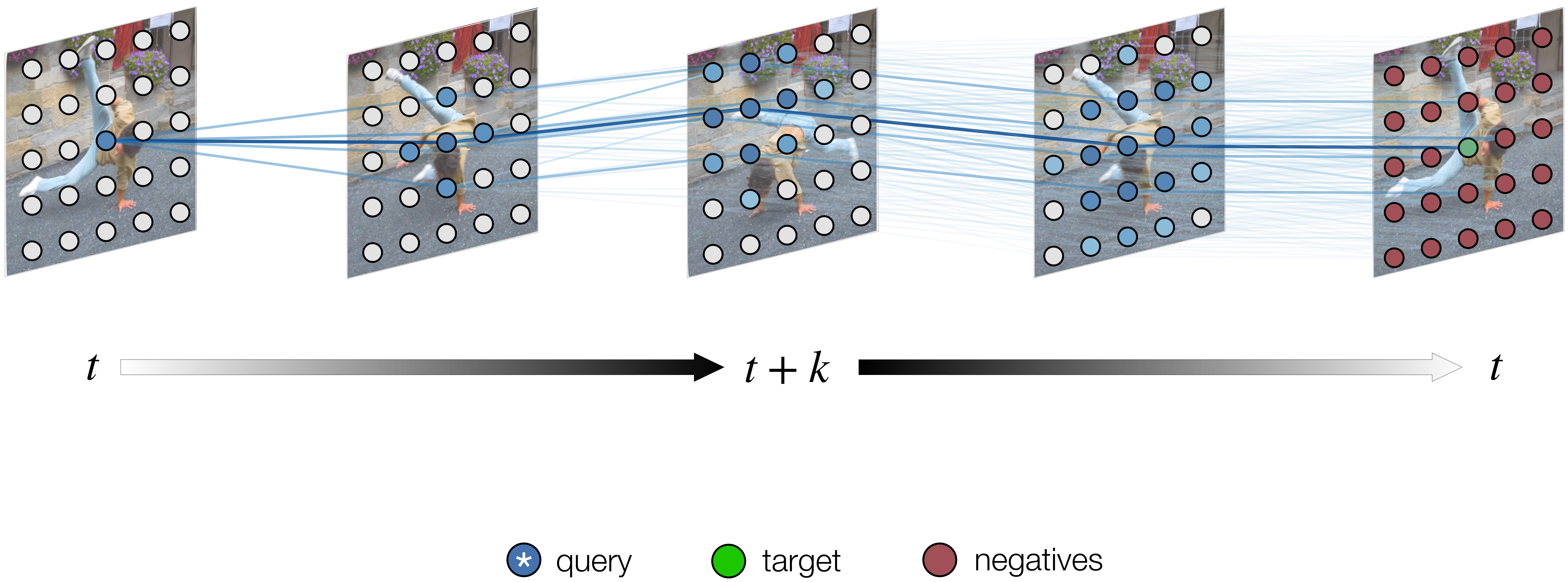
Latent views



Palindromes

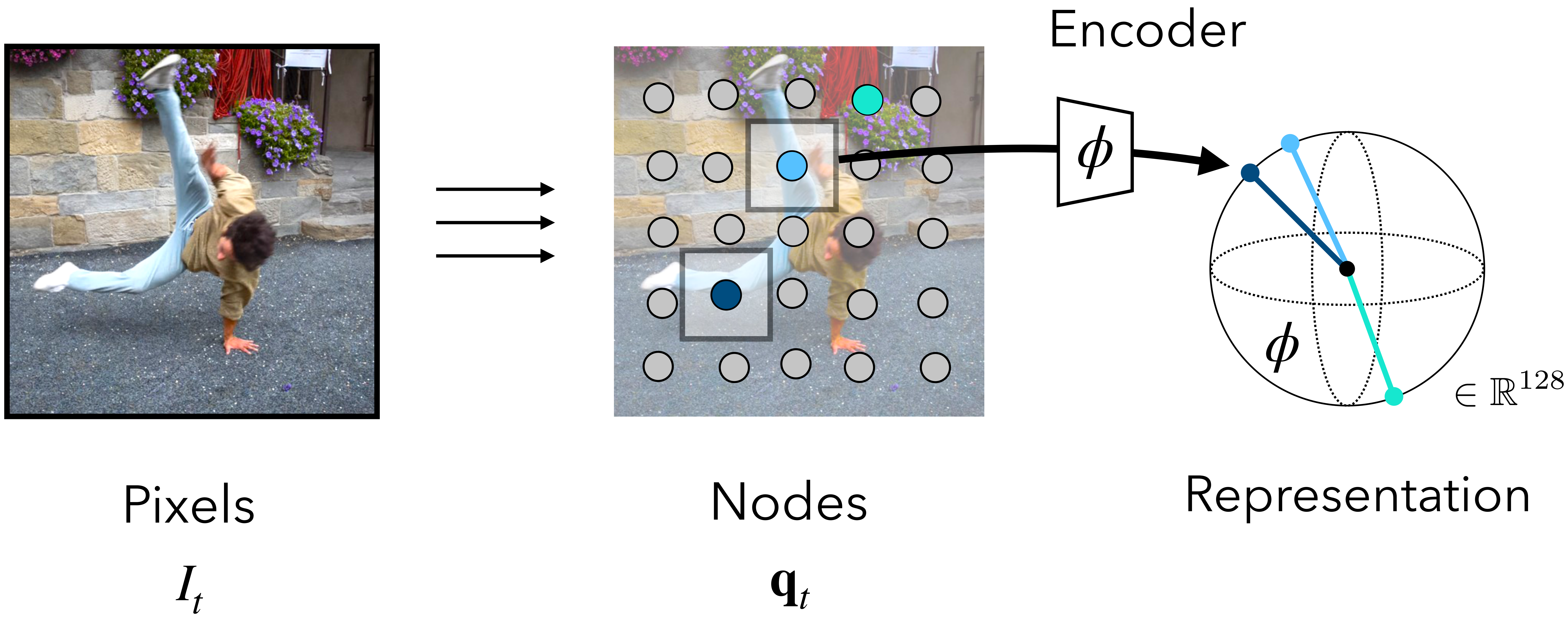
Self-supervised Learning



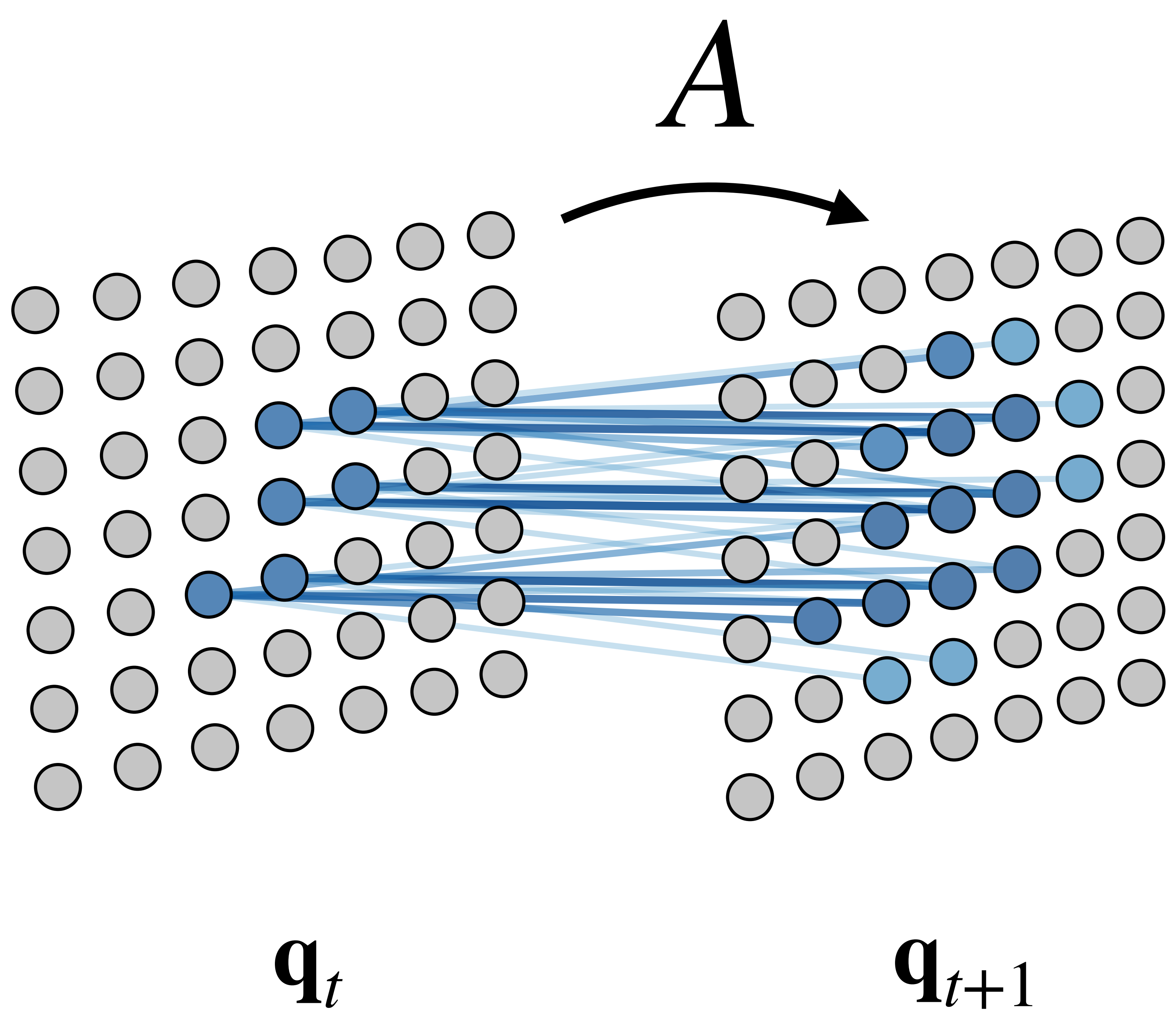


t
$t + k$
t
query
target
negatives

Video as a Graph



Video as a Graph

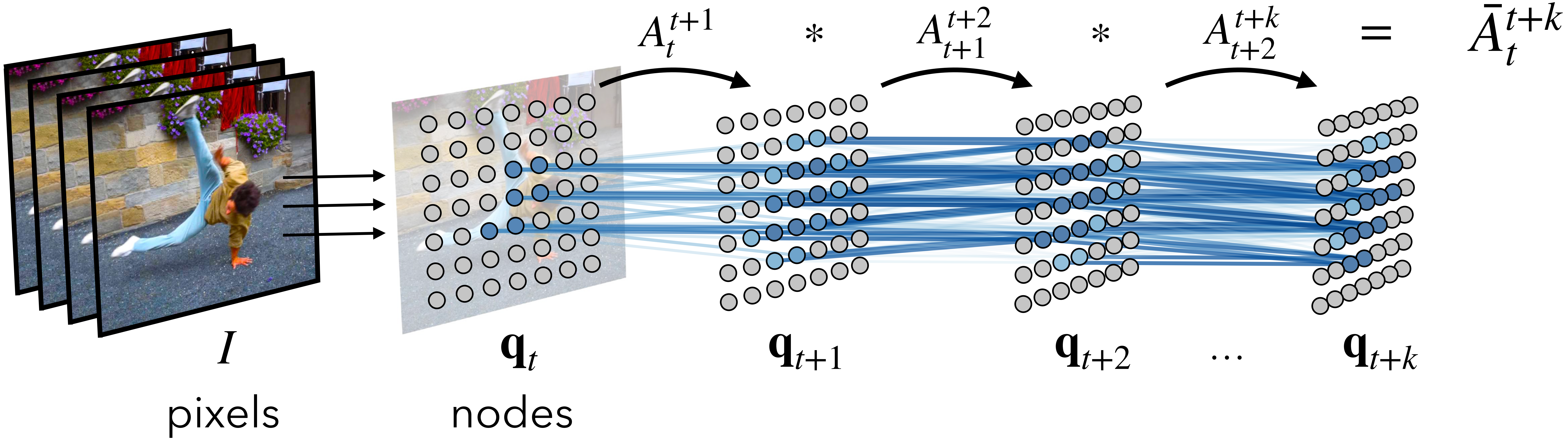


$$A_{ij} = \frac{e^{d_\phi(q_t^i,\ q_{t+1}^j)/\tau}}{\sum_l e^{d_\phi(q_t^i,\ q_{t+1}^l)/\tau}}$$

$$= P(X_{t+1} = j \mid X_t = i)$$

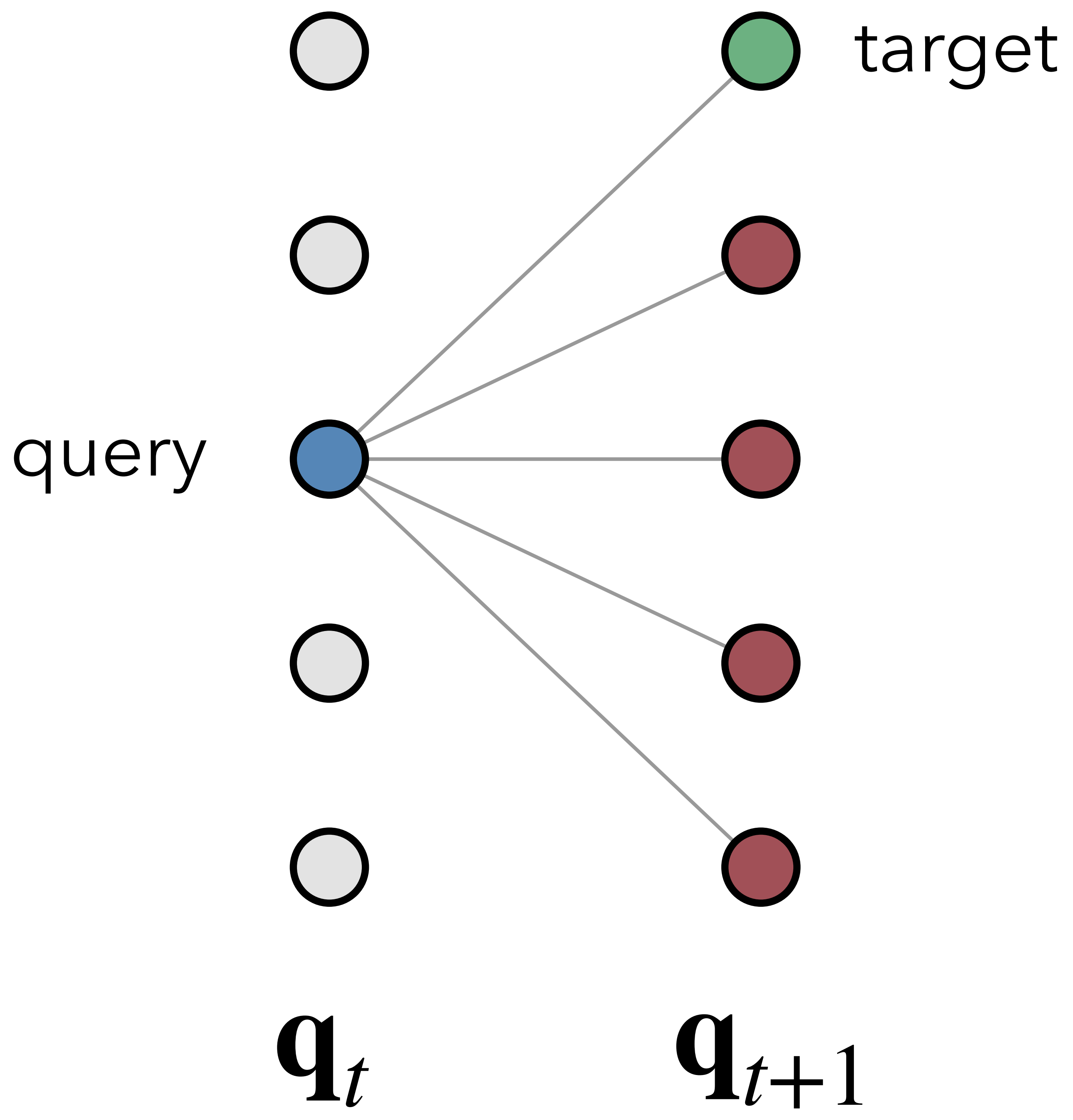
where $d_\phi(x, y) = \phi(x)^\top \phi(y)$

X_t is the position of walker at time t



Learn representation ϕ = Fit transition probabilities $\bar{A}_t^{t+k}$

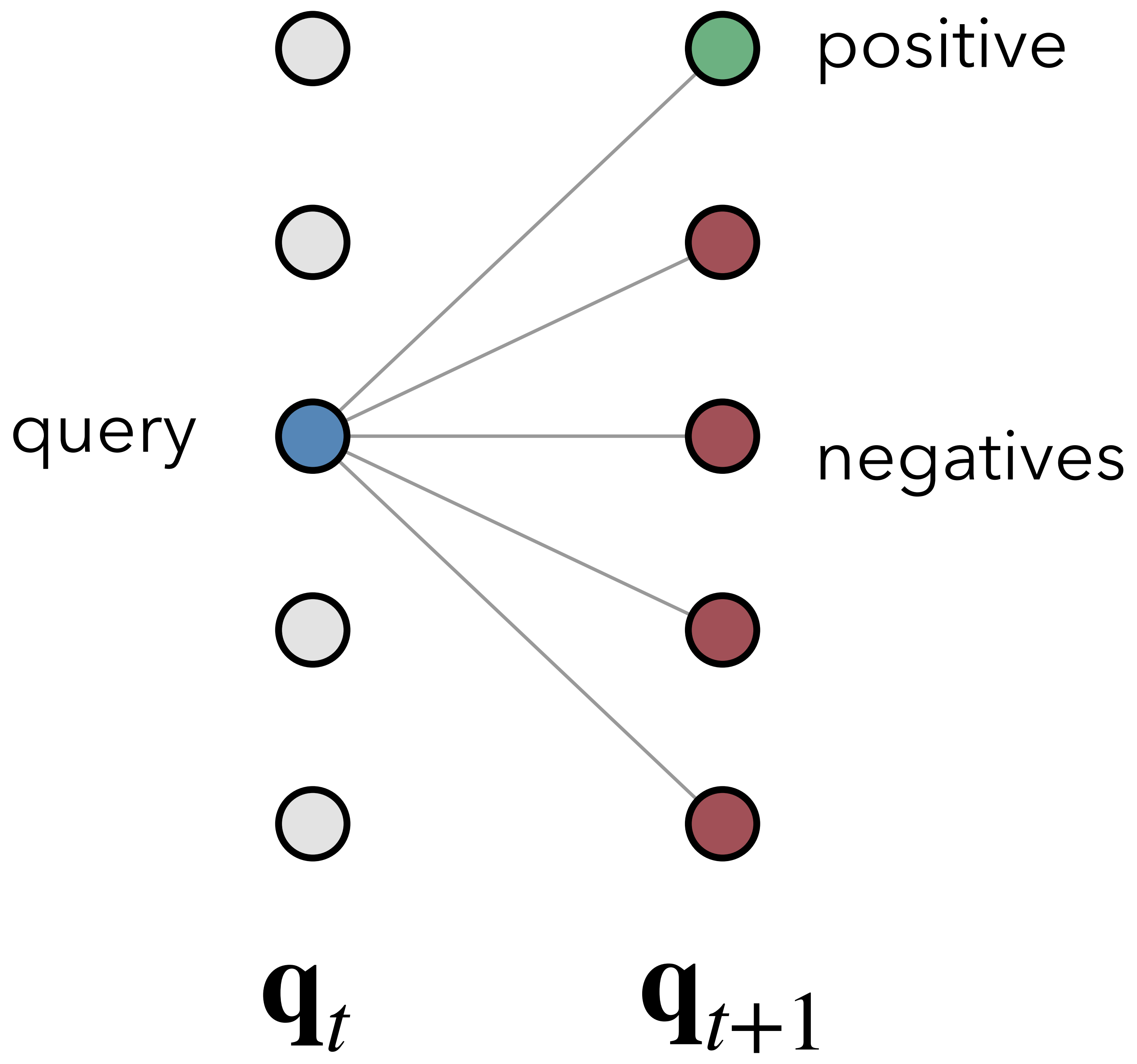
One Step



Maximize

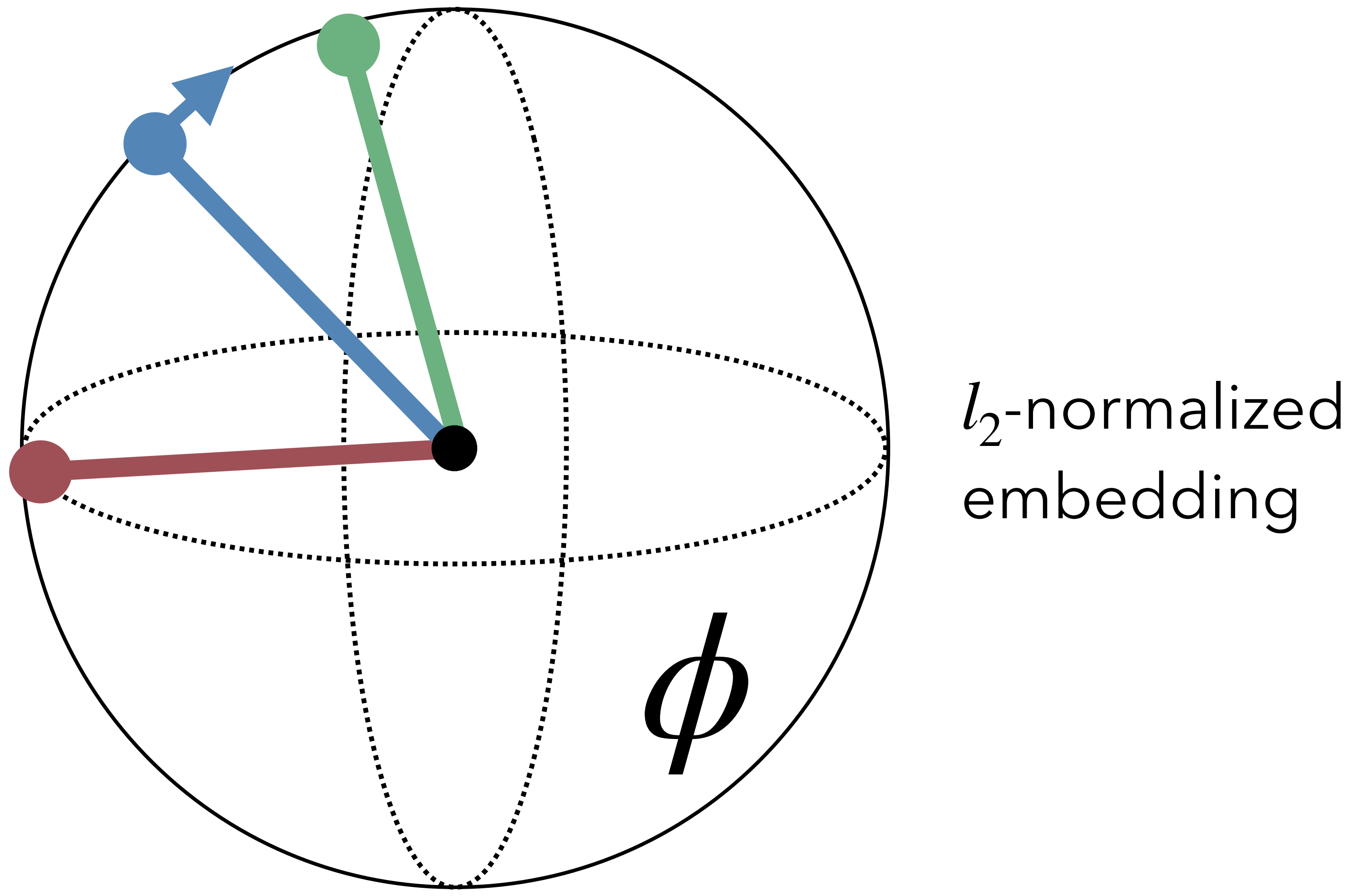
$$P(X_{t+1} = target \,|\, X_t = query)$$

One Step

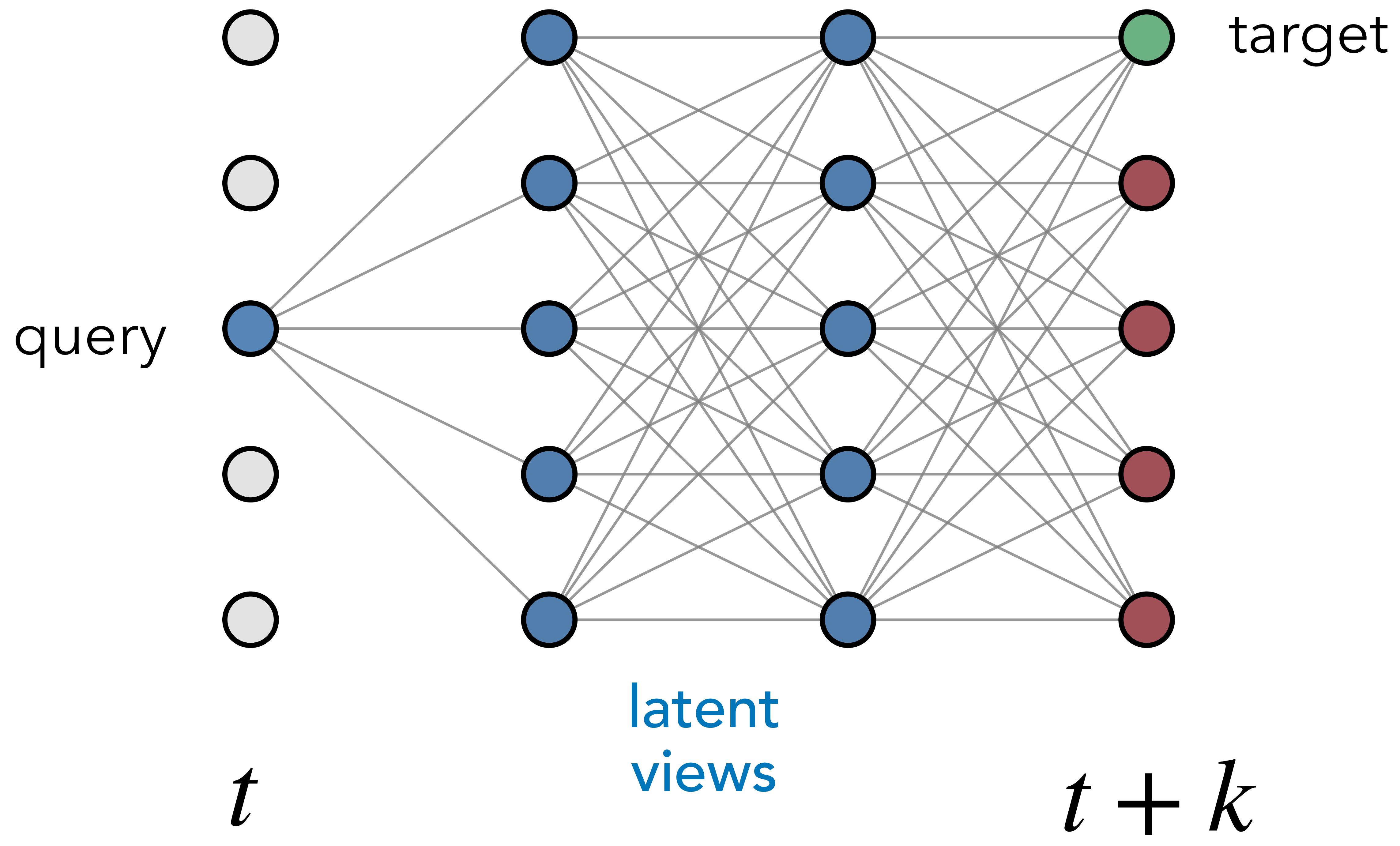


Maximize

$$P(X_{t+1} = \text{pos} \mid X_t = \text{query}) = \frac{e^{\phi(\text{query})^{\top}\phi(\text{pos})}}{\sum_{q_l} e^{\phi(\text{query})^{\top}\phi(q_l)}}$$



Chaining Correspondences



k-step Transition Matrix

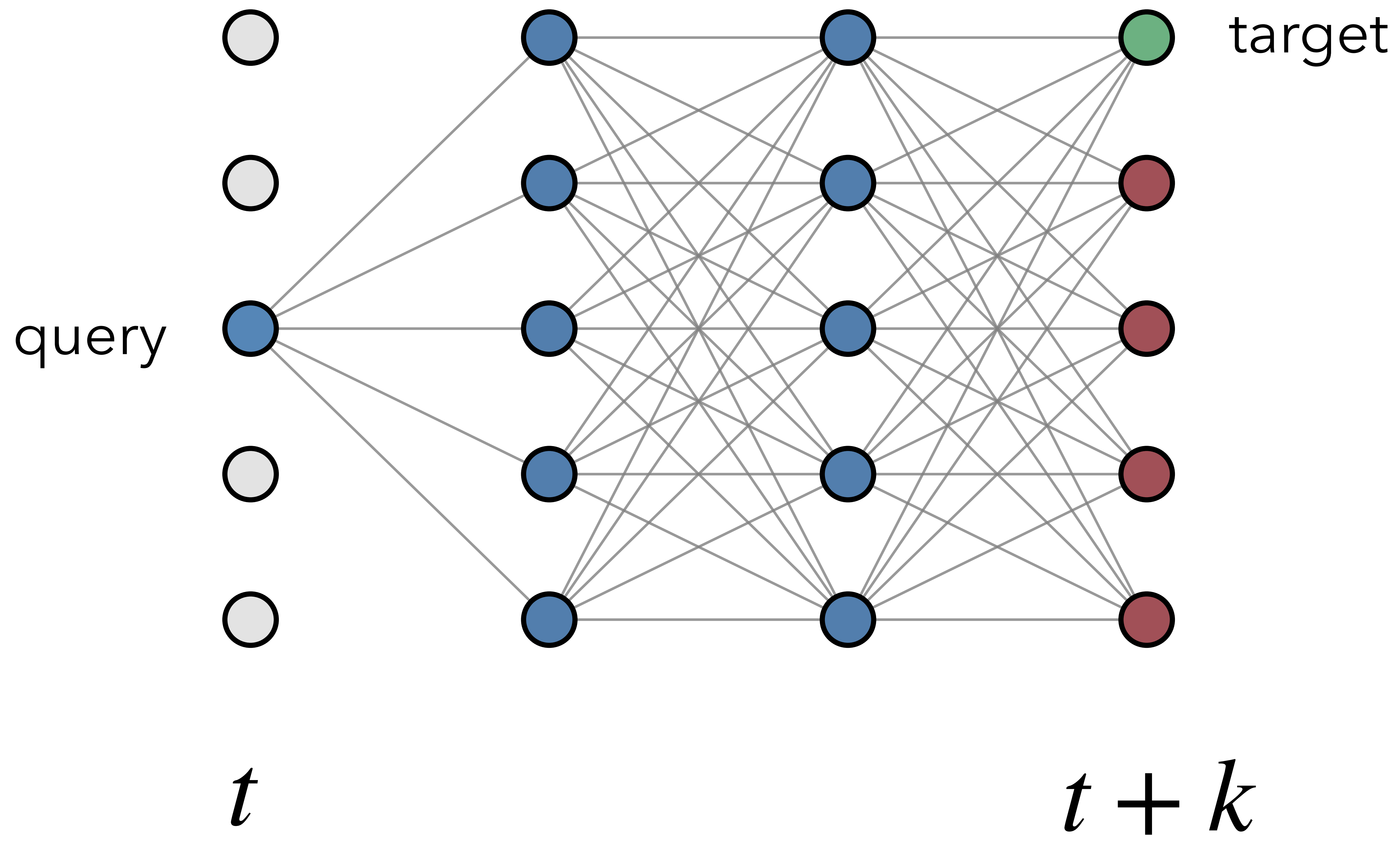
$$\bar{A}_t^{t+k} = \prod_{i=0}^{k-1} A_{t+i}^{t+i+1}$$

$$= P(X_{t+k} \mid X_t)$$

Sum over intermediate time steps

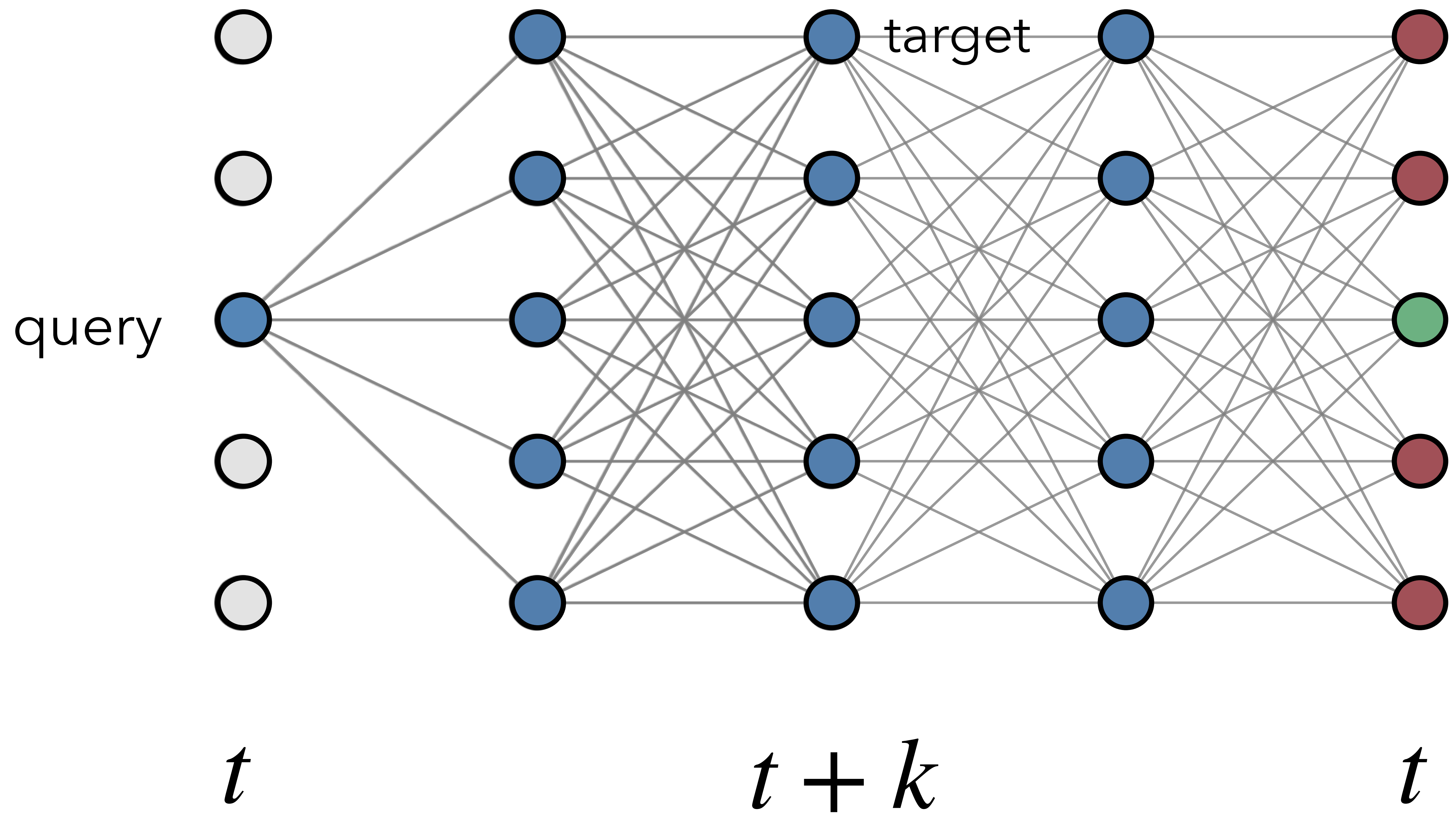
Fitting $\bar{A}_t^{t+k}$ provides supervision for A_t^{t+1}, A_{t+1}^{t+2} …, A_{t+k-1}^{t+k}

Chaining Correspondences



A single target supervises chains of learning problems

Supervised → Self-Supervised



Train on Palindromes

$$\mathcal{L}_{cyc}^{k} = \mathrm{tr}(\log(\bar{A}_t^{t+k}\bar{A}_{t+k}^{t}))$$

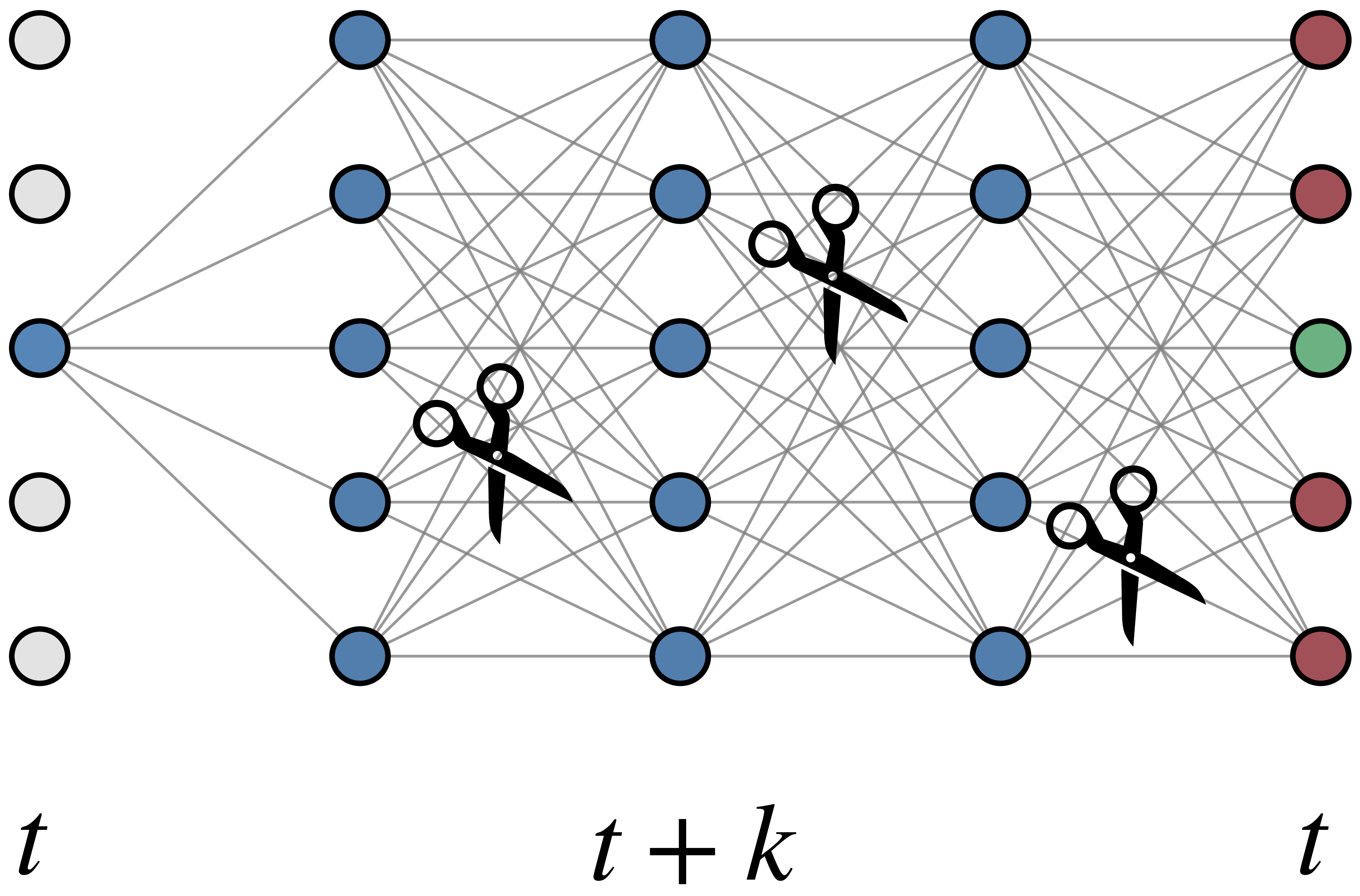
See also [Wang, Jabri, Efros, "Learning correspondence from the cycle-consistency of time", 2019]

Image
w/ query

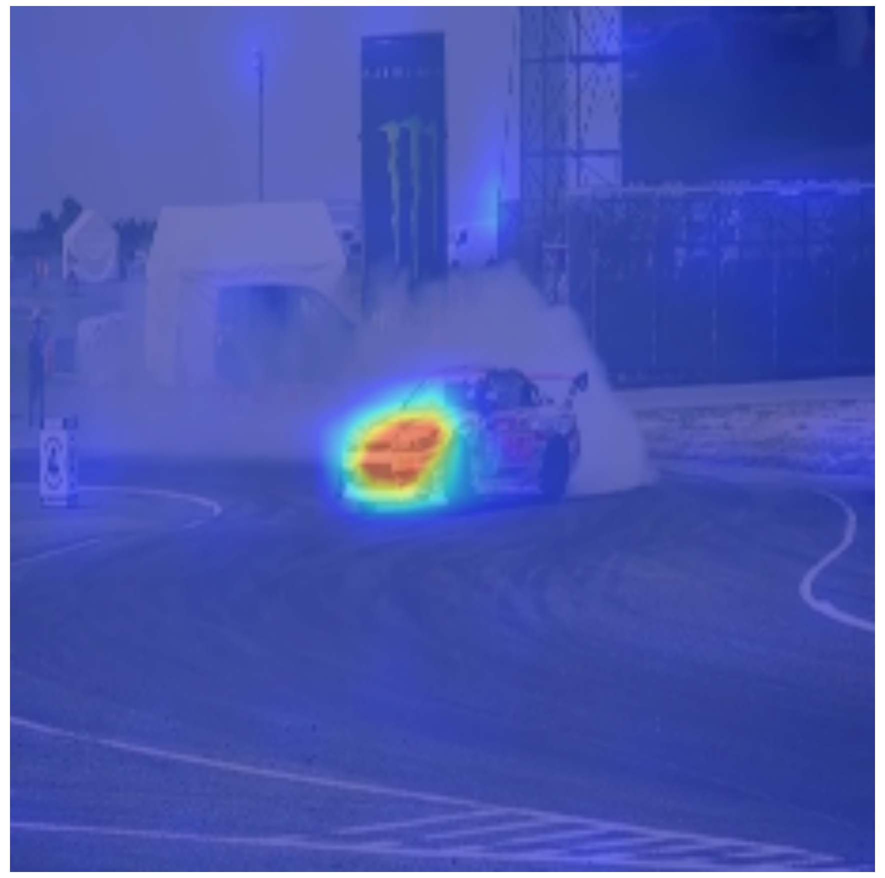
Transition
to future frame

Higher-level
Correspondence?

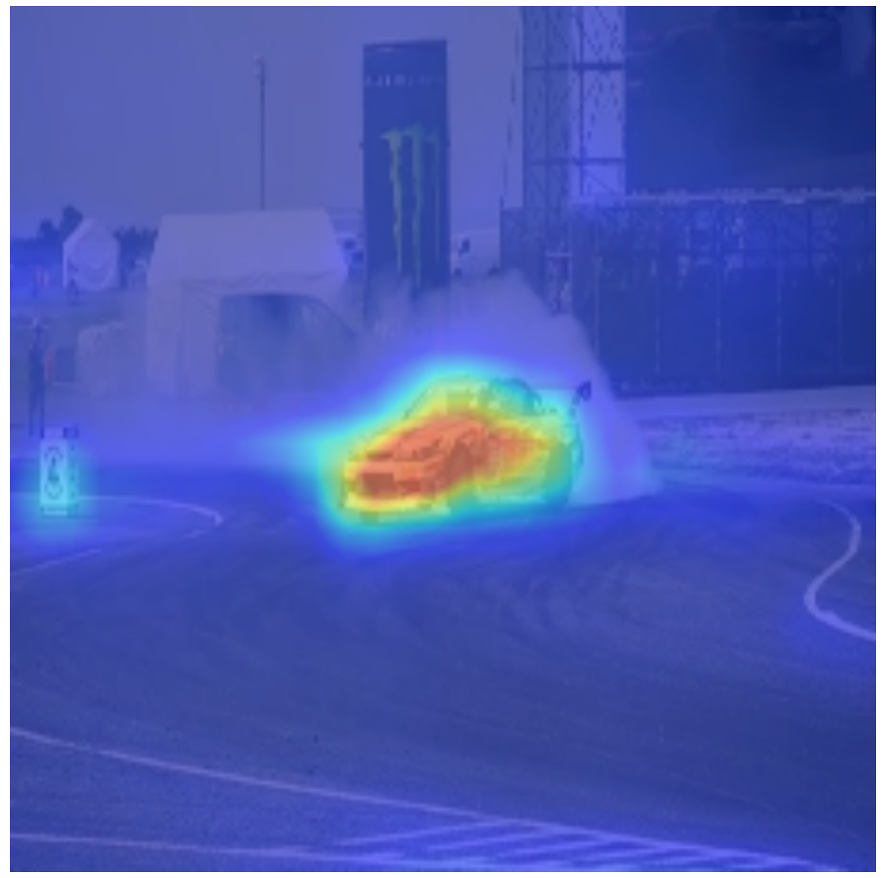
Edge Dropout



Force alternate, **context** paths



No Edge Dropout



Edge Dropout
p = 0.1

Evaluation: Using ϕ for Label Propagation

Object Propagation 1-4 Objects **DAVIS Benchmark**

Semantic Part Propagation 20 Parts **VIP Benchmark**

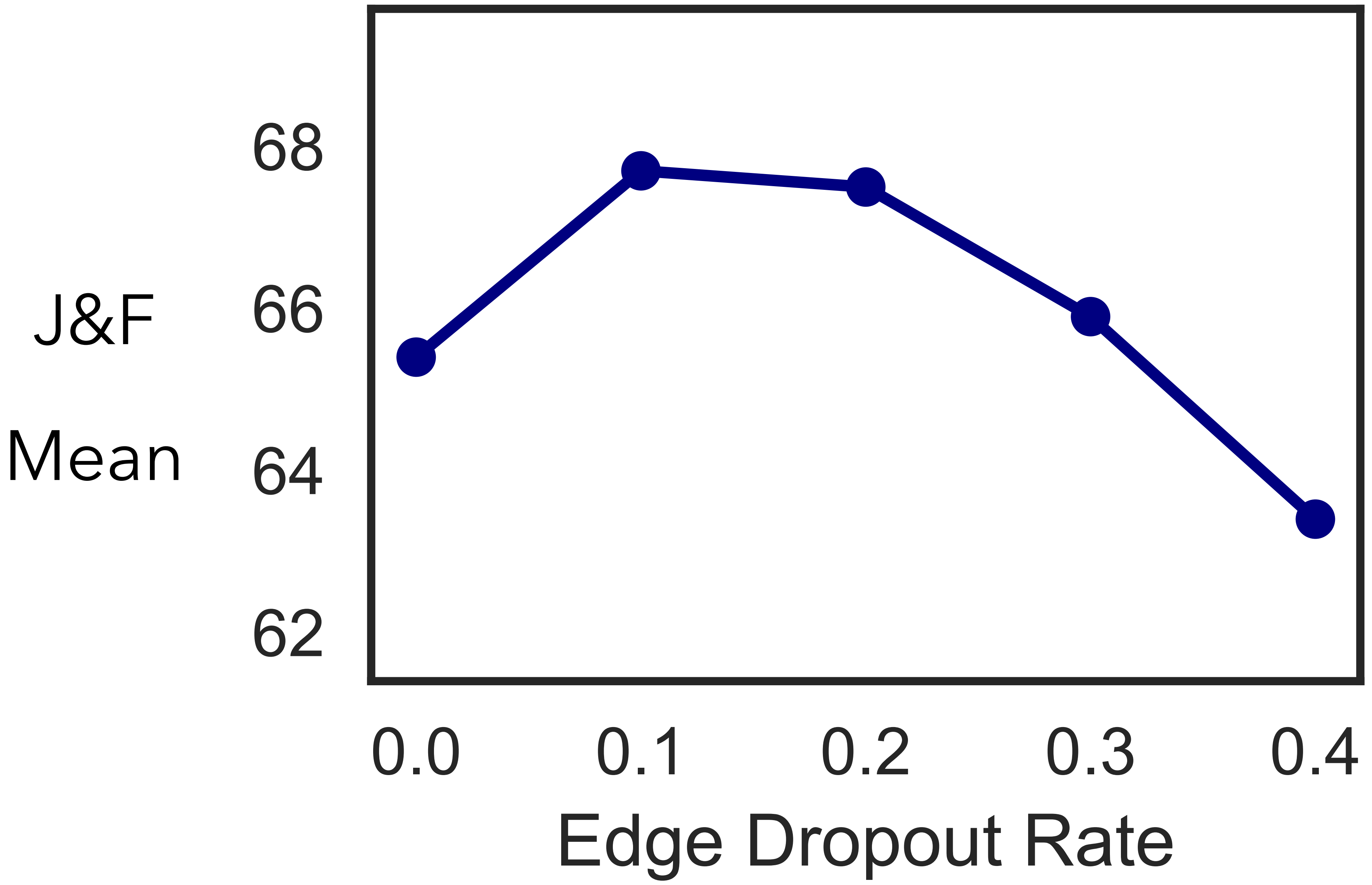
Pose Propagation 15 Keypoints **JHMDB Benchmark**

Qualitative Results: Video Object Propagation (DAVIS)

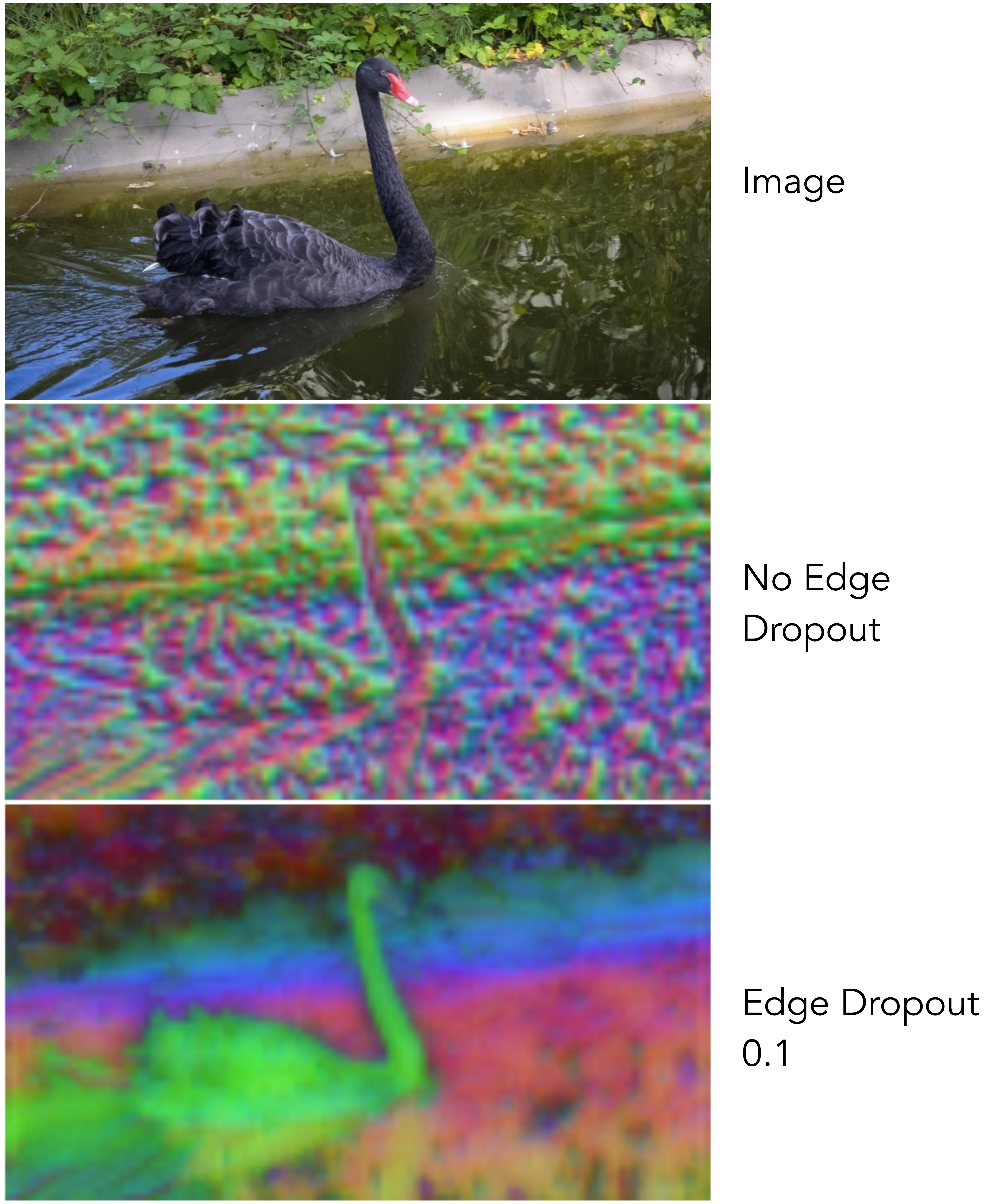
UVC
Li et al. (2019)

Ours

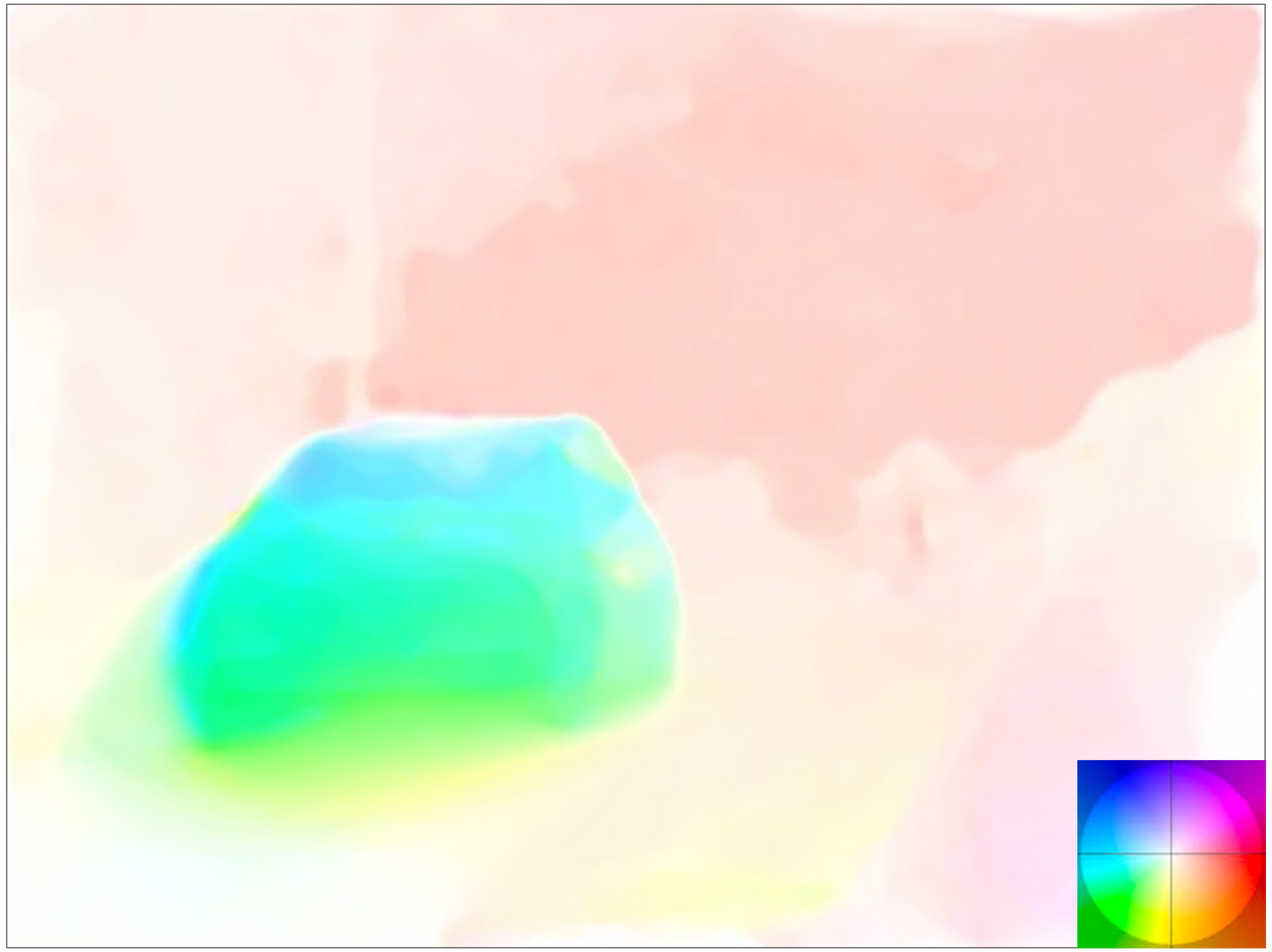
Effect of Edge Dropout



PCA Feature Visualization



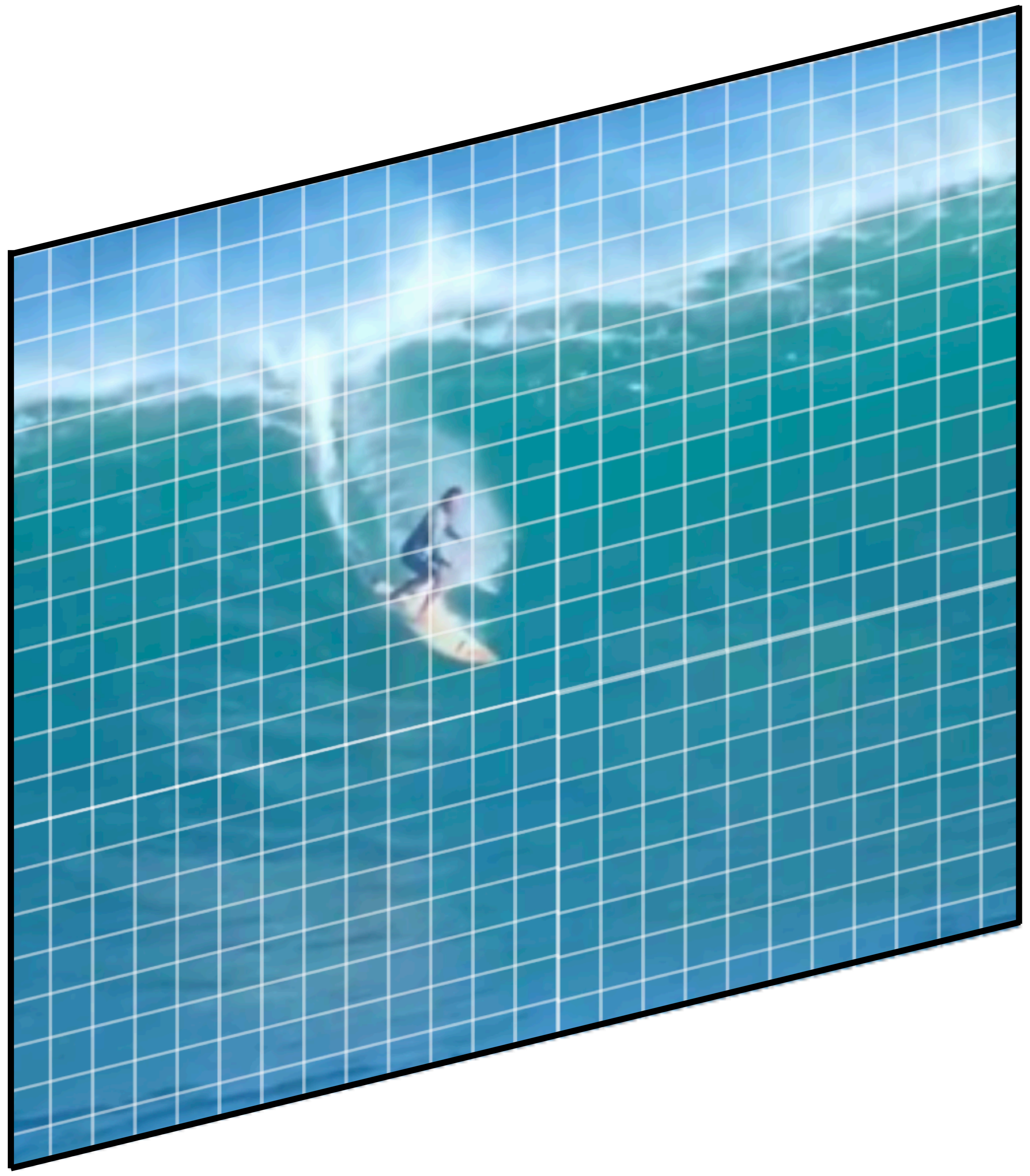
Can we track pixels?



Optical flow

Example source: Ce Liu

Can we track pixels?



large attention matrix

$HW \times HW$

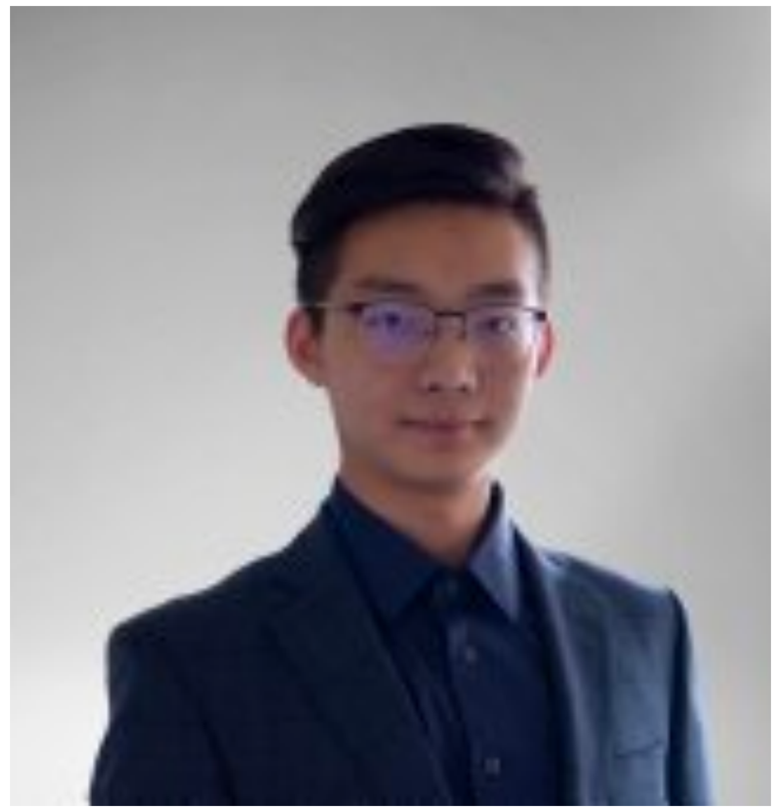
$$A = \mathrm{softmax}\left(Q_t Q_{t+1}^\top\right) = P(X_{t+1}|X_t)$$

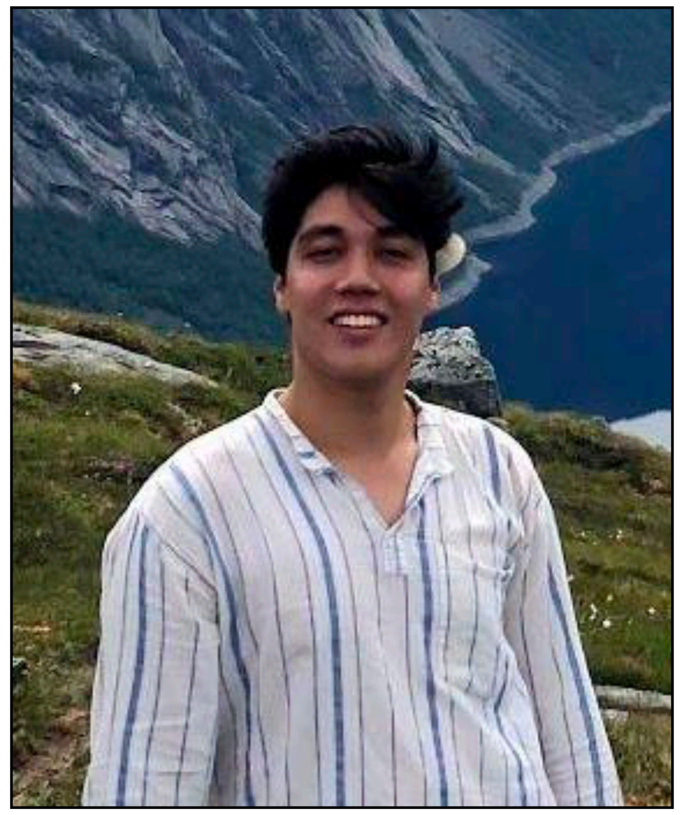
$$\bar{A}_t^{t+k} = \prod_{i=0}^{k-1} A_{t+i}^{t+i+1}$$

large matrix multiplies

$O(H^4W^4)$

Learning Pixel Trajectories with Multiscale Contrastive Random Walks


Zhangxing Bian

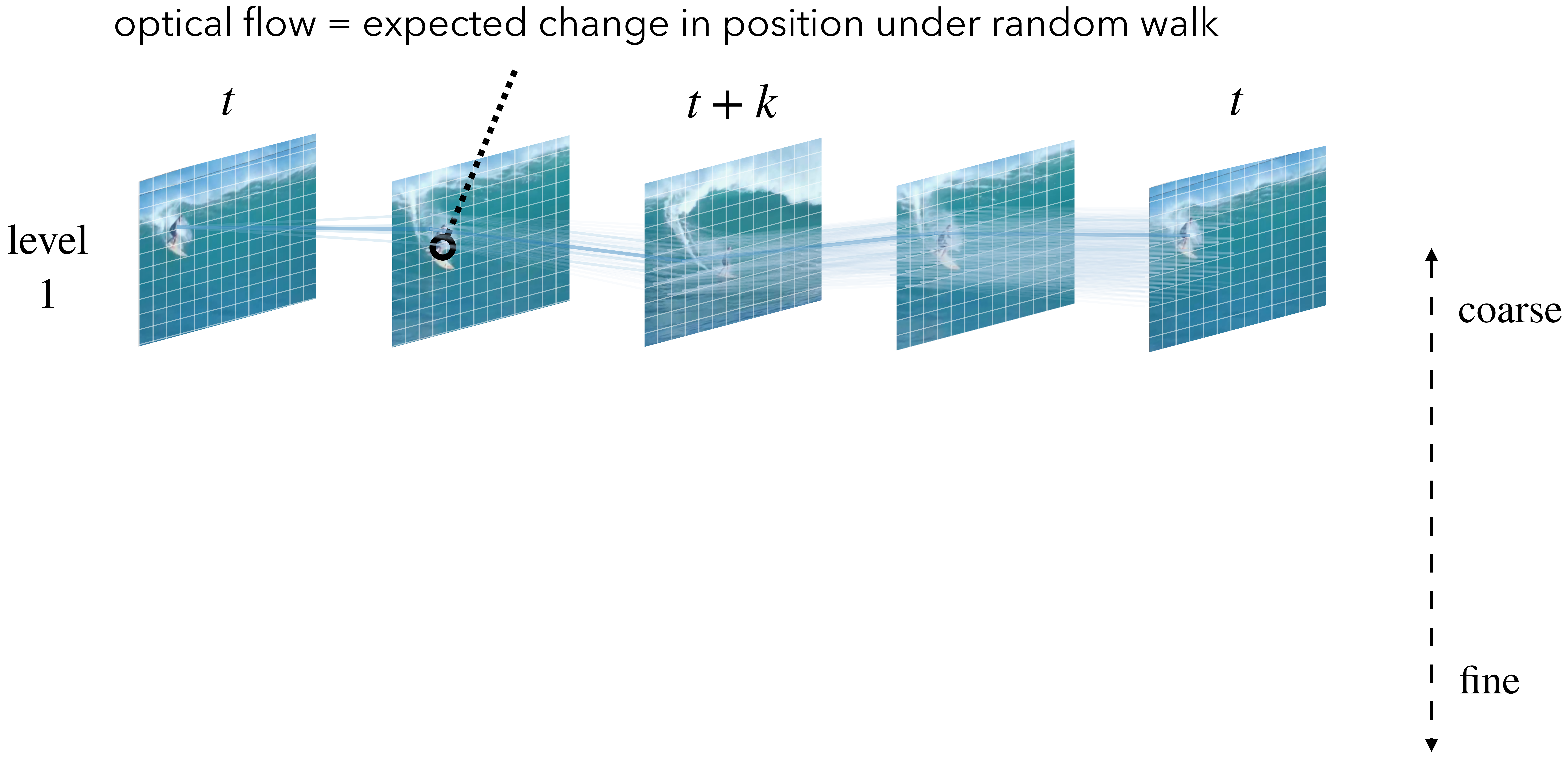

Allan Jabri

Andrew Owens


Alexei Efros

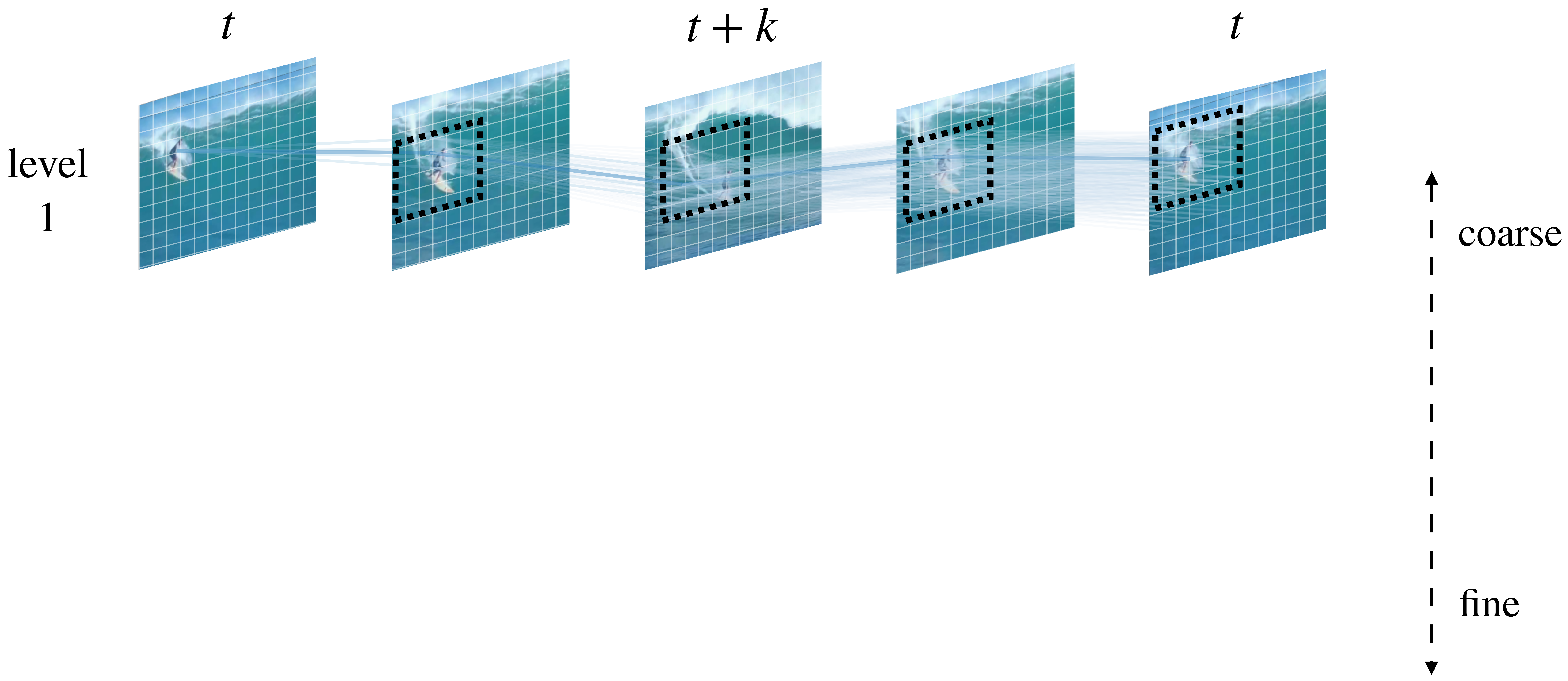
Funding: TRI, Cisco, and Berkeley Deep Drive

Multiscale walk

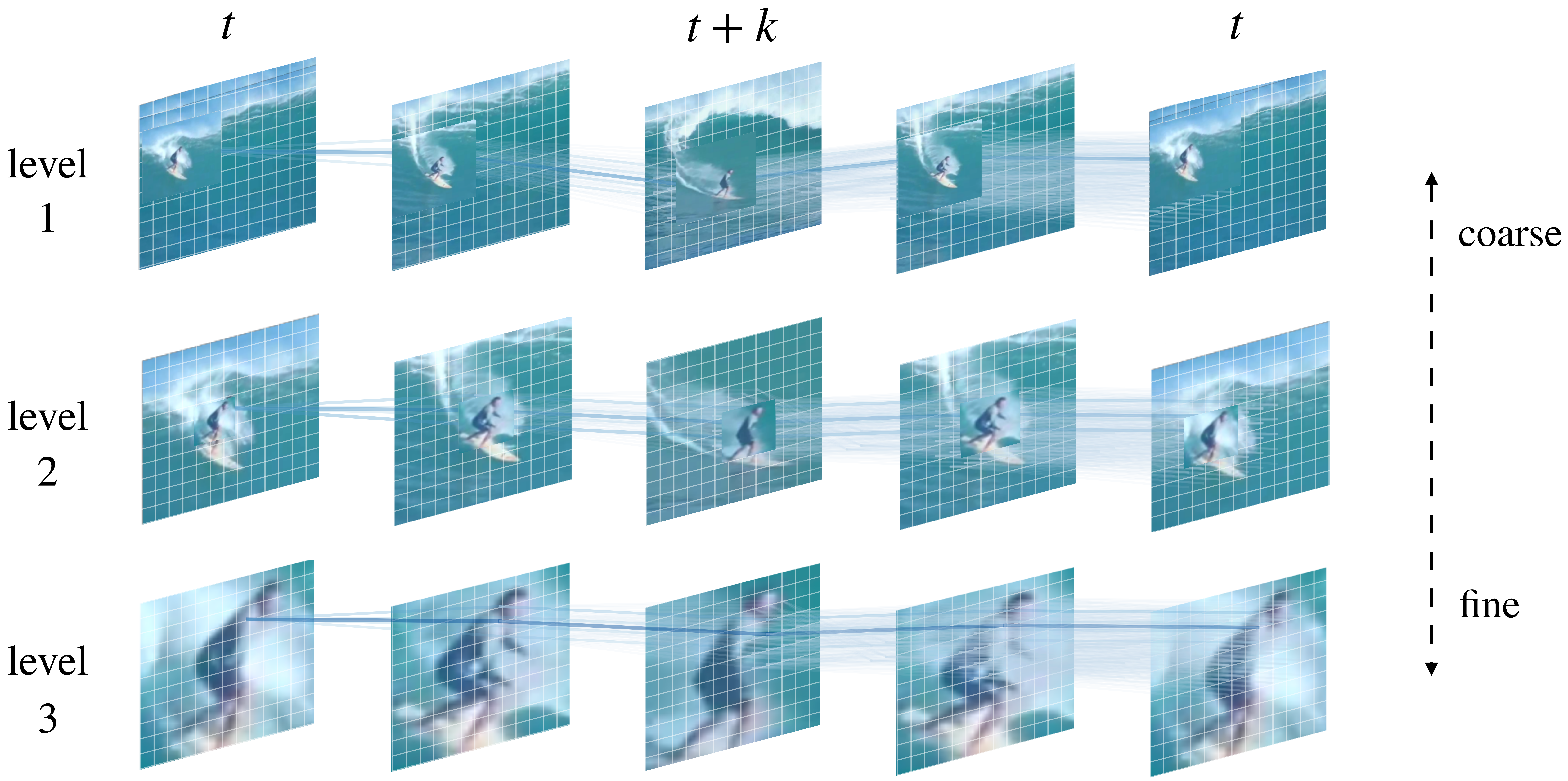


Multiscale walk

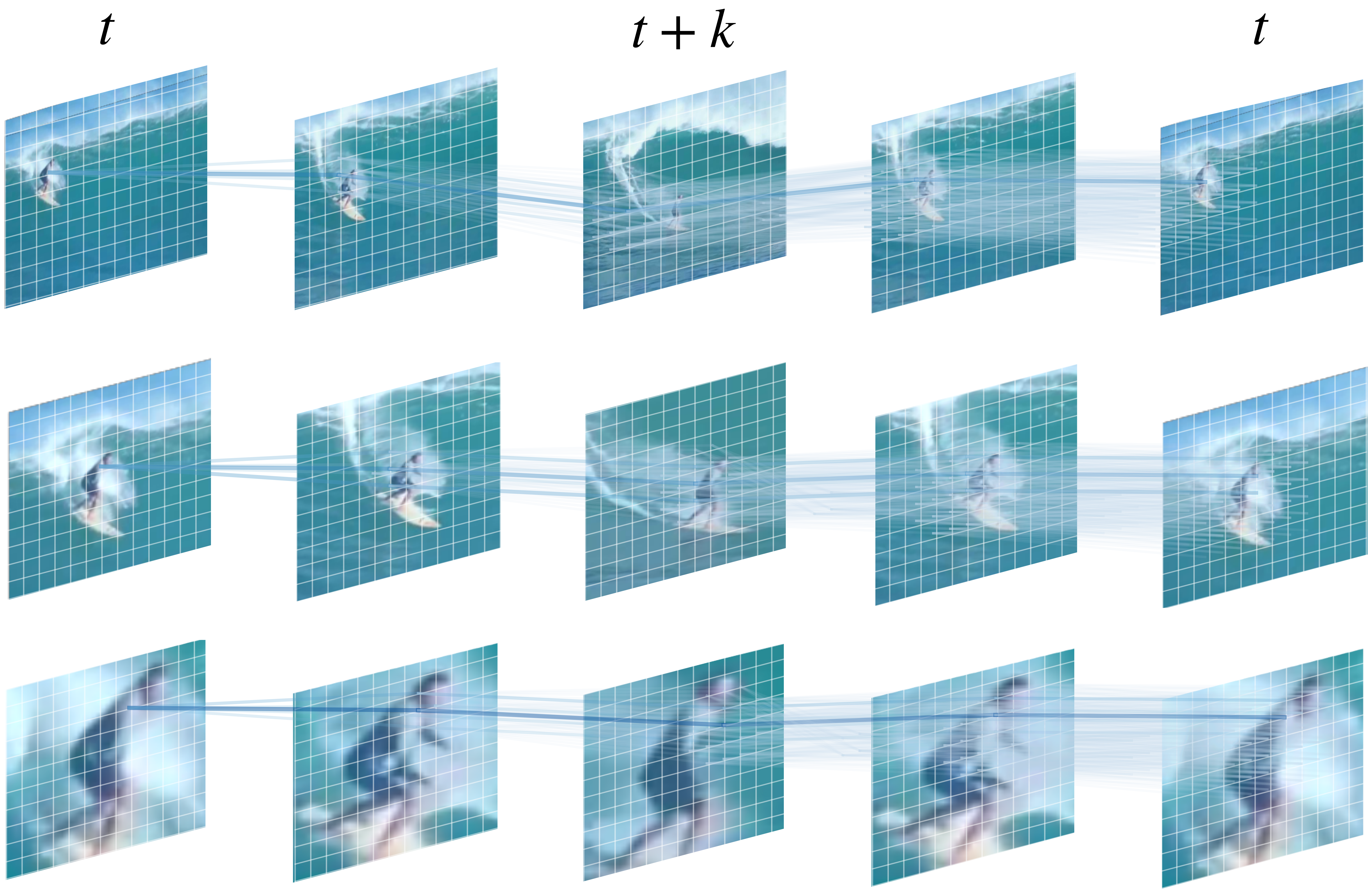
recenter walk and repeat using higher-res features



Multiscale walk



Multiscale walk



Multiscale walk loss:

$$\mathcal{L}_{\text{msCRW}} = -\sum_{l=1}^{L} \text{tr}(\log(\bar{A}^l_{t,t+k}\bar{A}^l_{t+k,t}))$$

Smoothness Regularization (Jonschkowski et al., 2020):

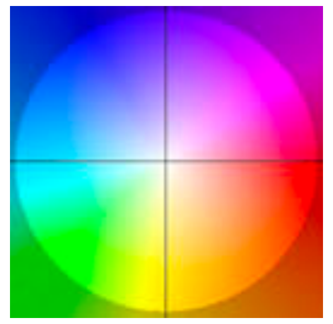
$$\mathcal{L}_{\text{smooth}} = \mathbb{E}_p \sum_{d \in \{x,y\}} \exp(-\lambda_c I_d(p)) \left|\frac{\partial^2 \mathbf{f}_{s,t}(p)}{\partial d^2}\right|$$

Network backbone from PWC-Net (Sun et al., 2017).

Optical flow

Ground truth | Ours

Contrastive random walk + smoothness

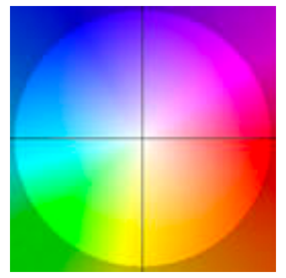


Optical flow

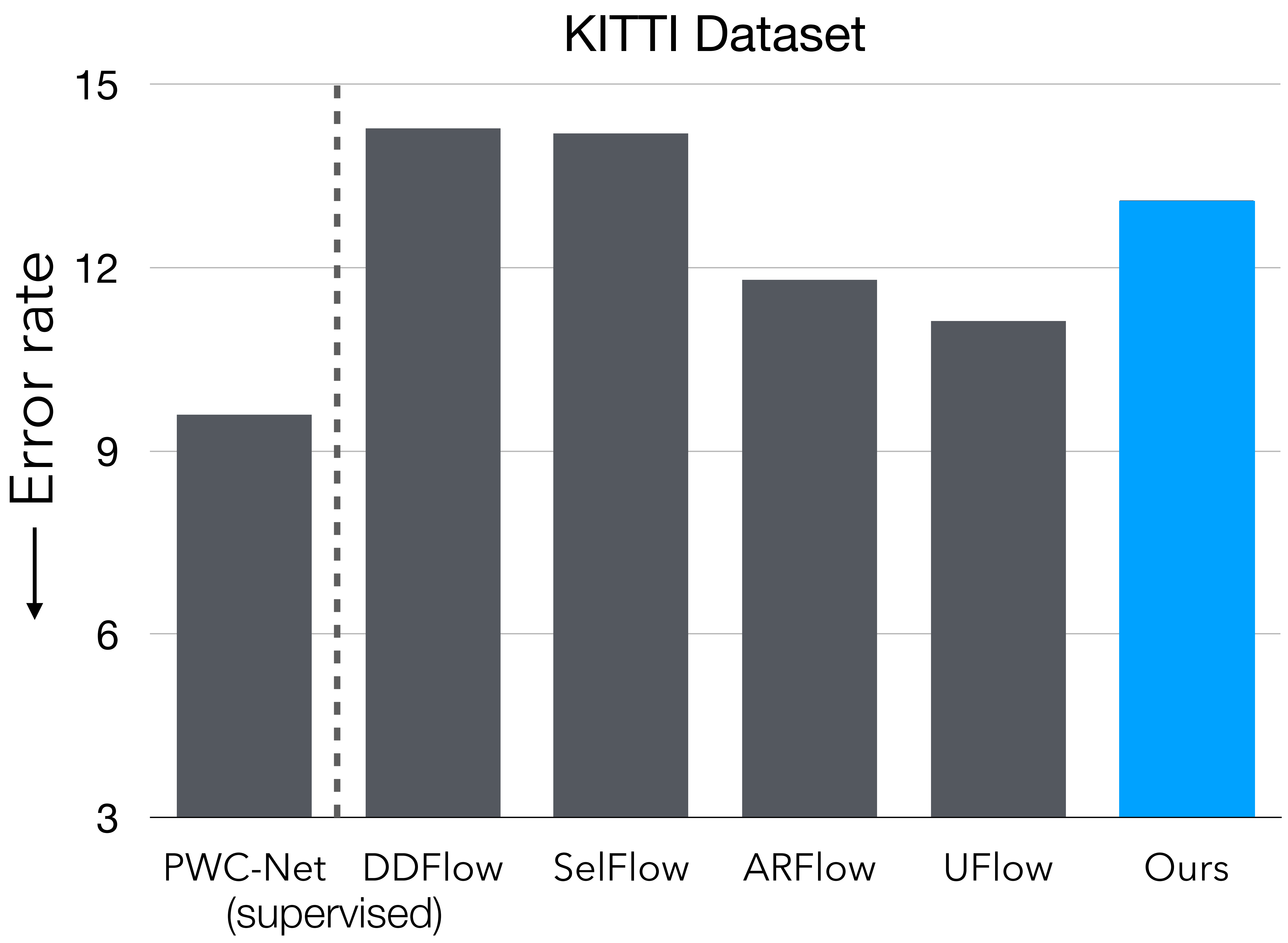
Ground truth | Ours (nonparametric) | Ours + regression

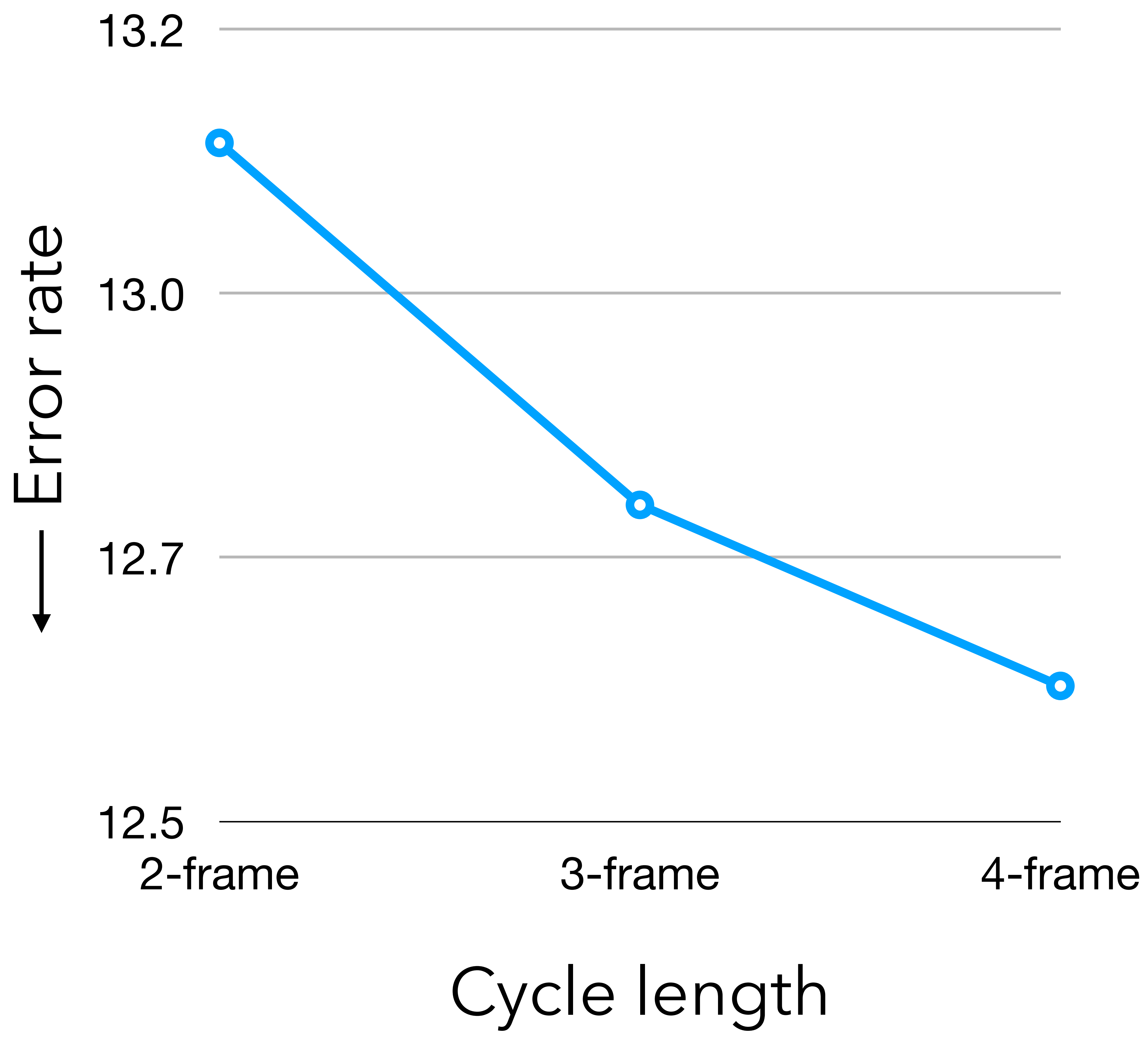
After adding flow regression module + occlusion handling.

Implementation follows UFlow (Jonschkowski et al., 2020)



Optical flow results





Toward a unified model

Label

Propagation

Optical Flow

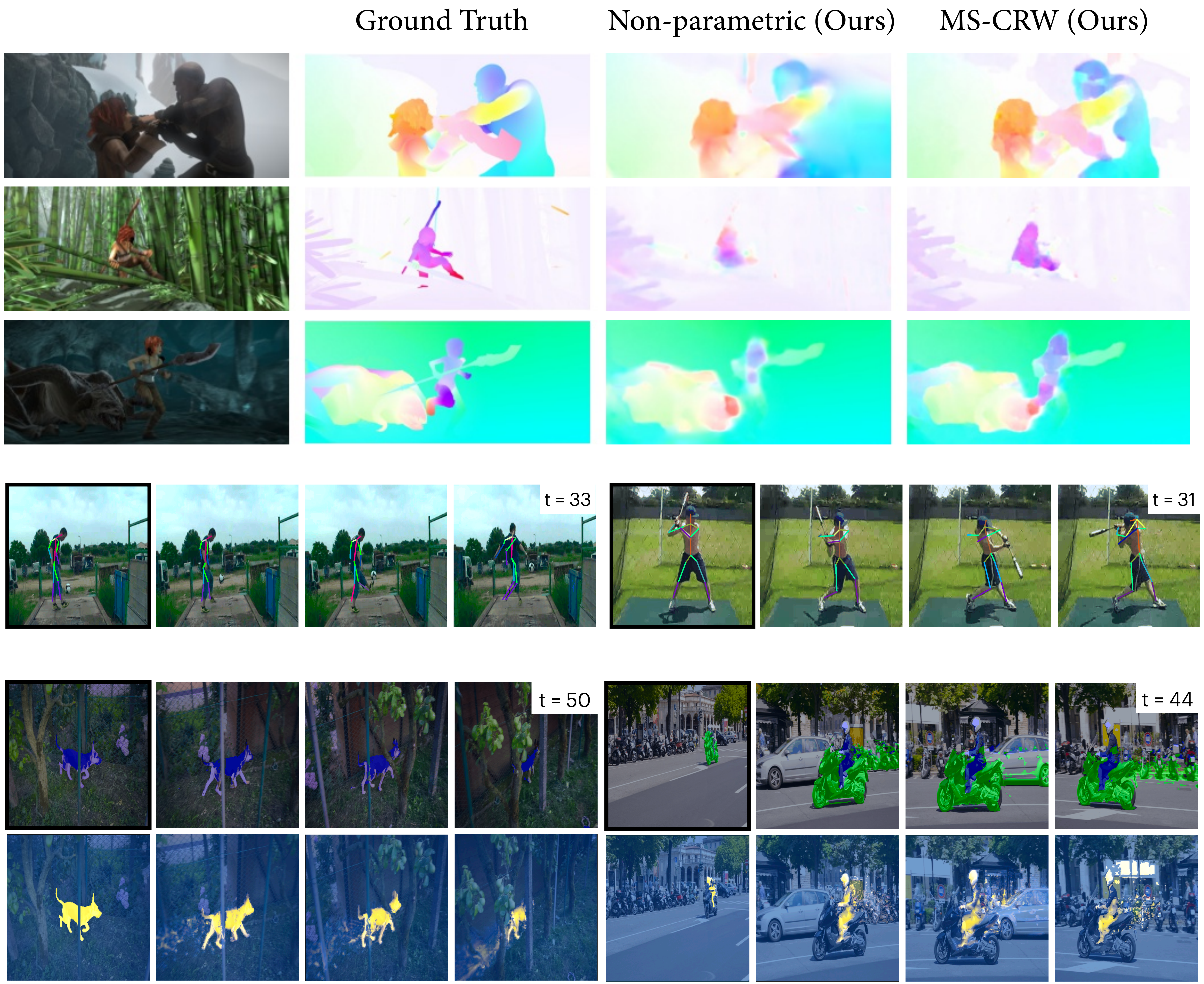
Label: $(\Delta x, \Delta y)$

Pose Tracking

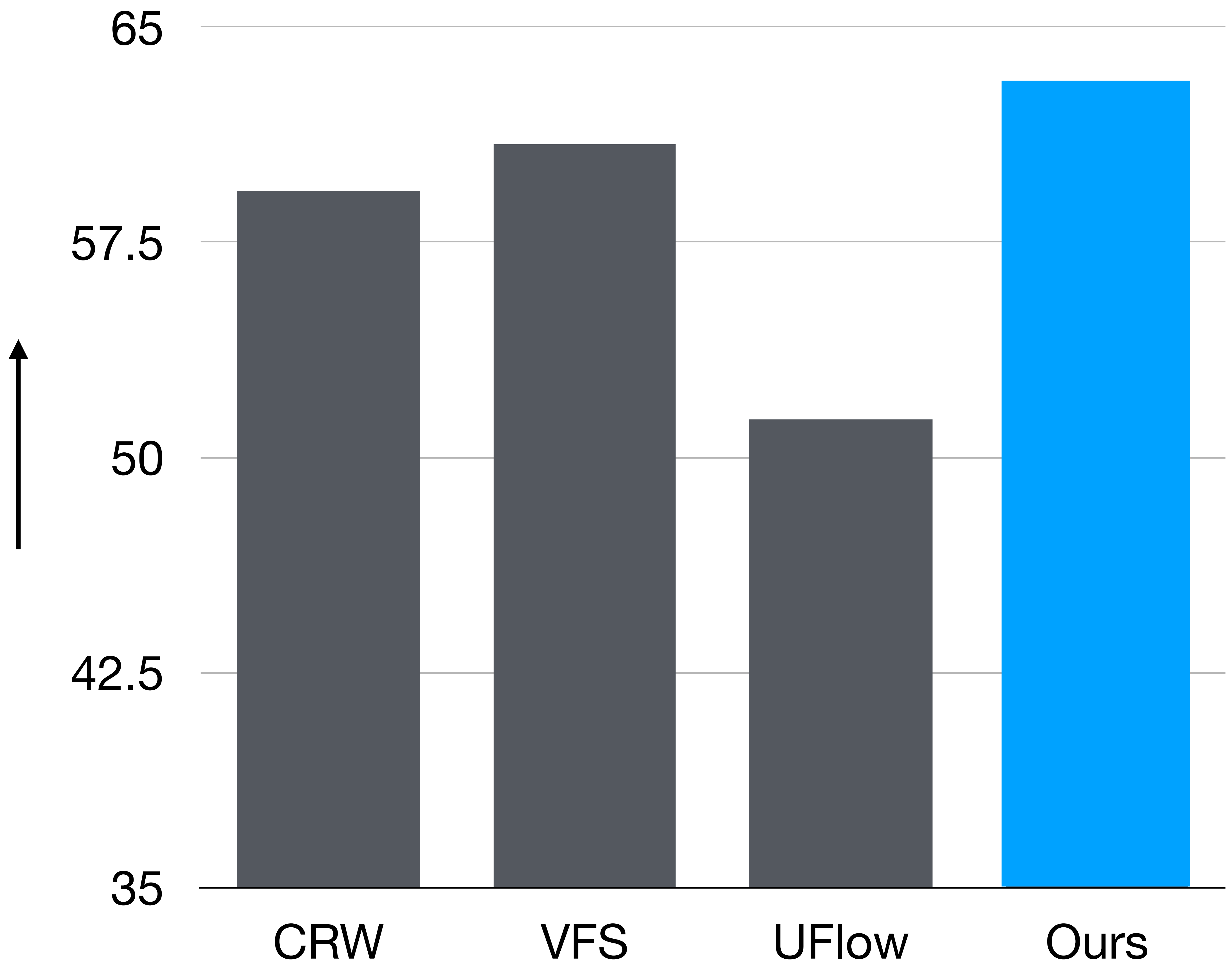
Label: 1-Of-K

Objects Seg.

Label: 1-Of-K

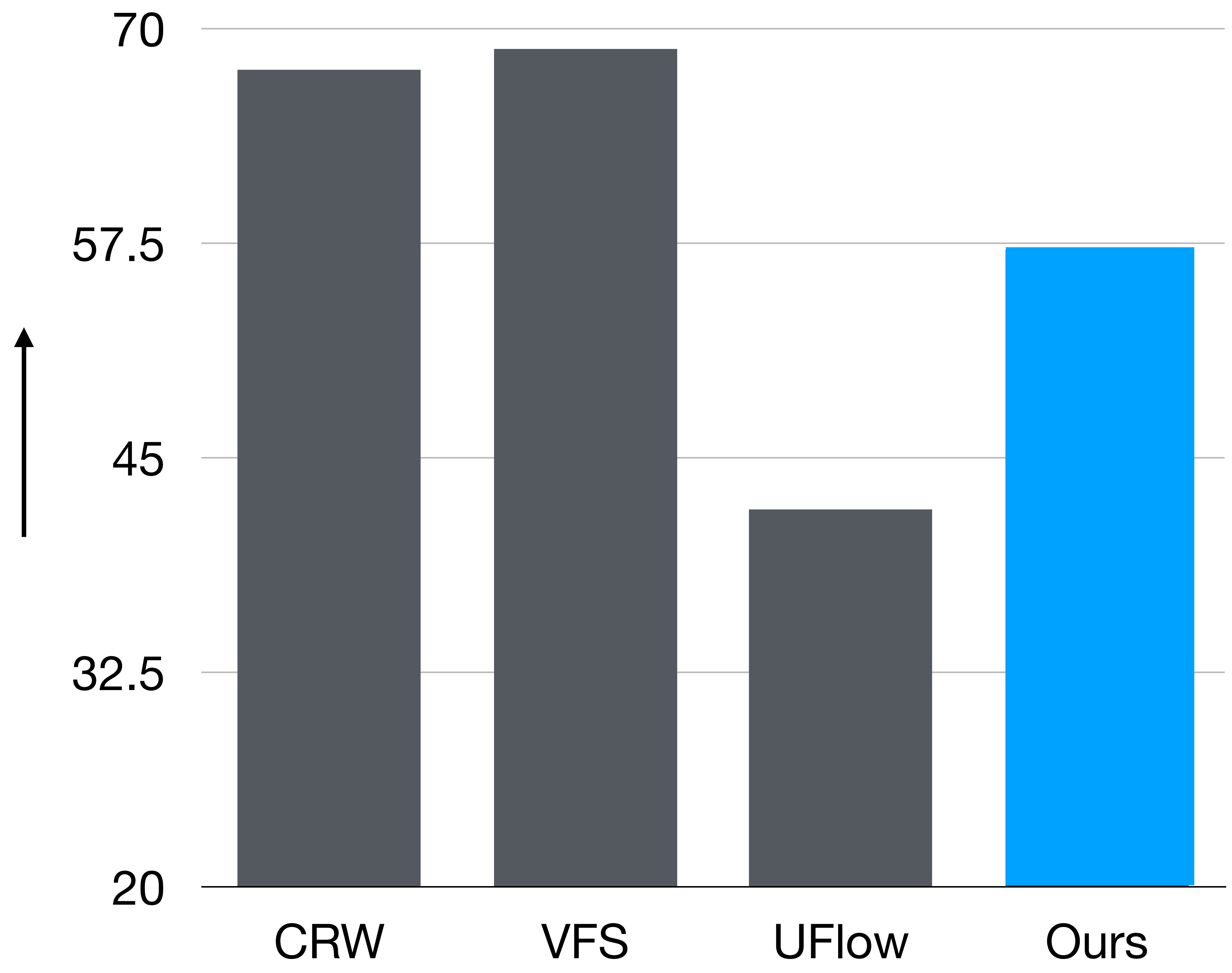


Pose and Segment Tracking



Pose tracking (@0.1)

[JHMDB Dataset]



Segment tracking (J&F)

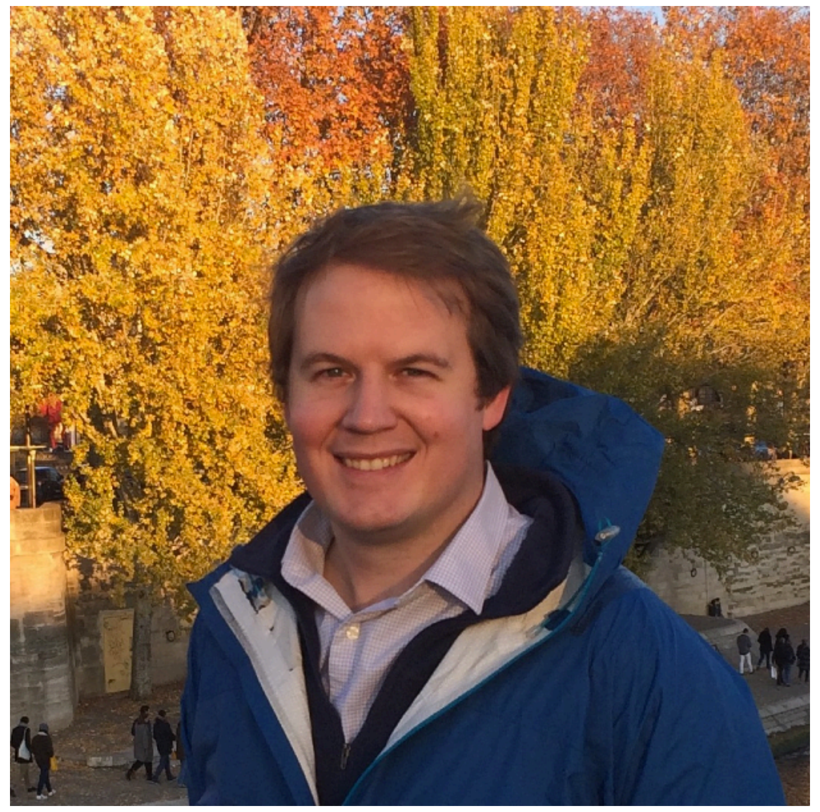
[DAVIS Dataset]

UNIVERSITY OF MICHIGAN

Sound Localization by Self-Supervised Time Delay Estimation

ECCV 2022

Ziyang Chen

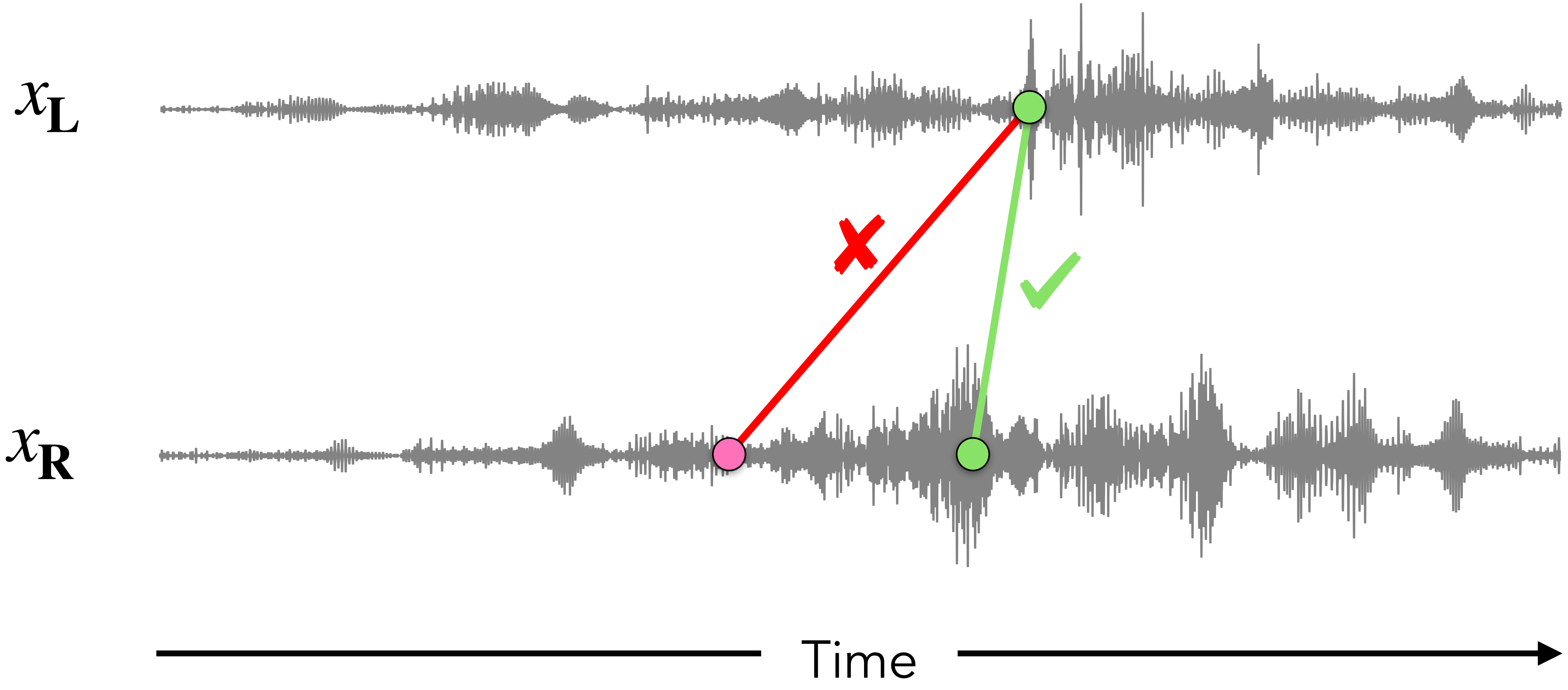


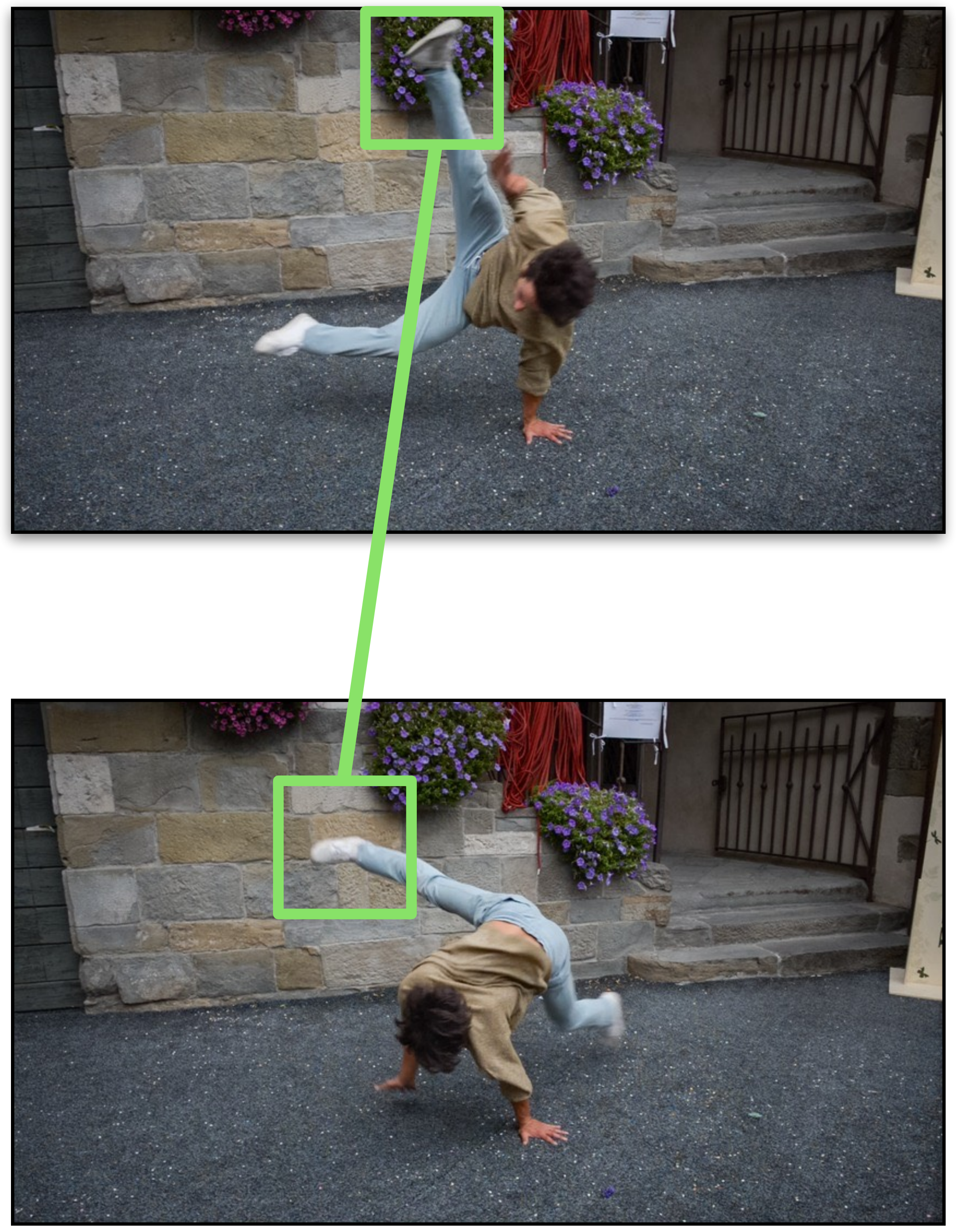
David F. Fouhey

Andrew Owens

Interaural time difference cues

Interaural Correspondence





Visual correspondence

Vondrick et al., 2018
Wang et al., 2019
Jabri et al., 2020

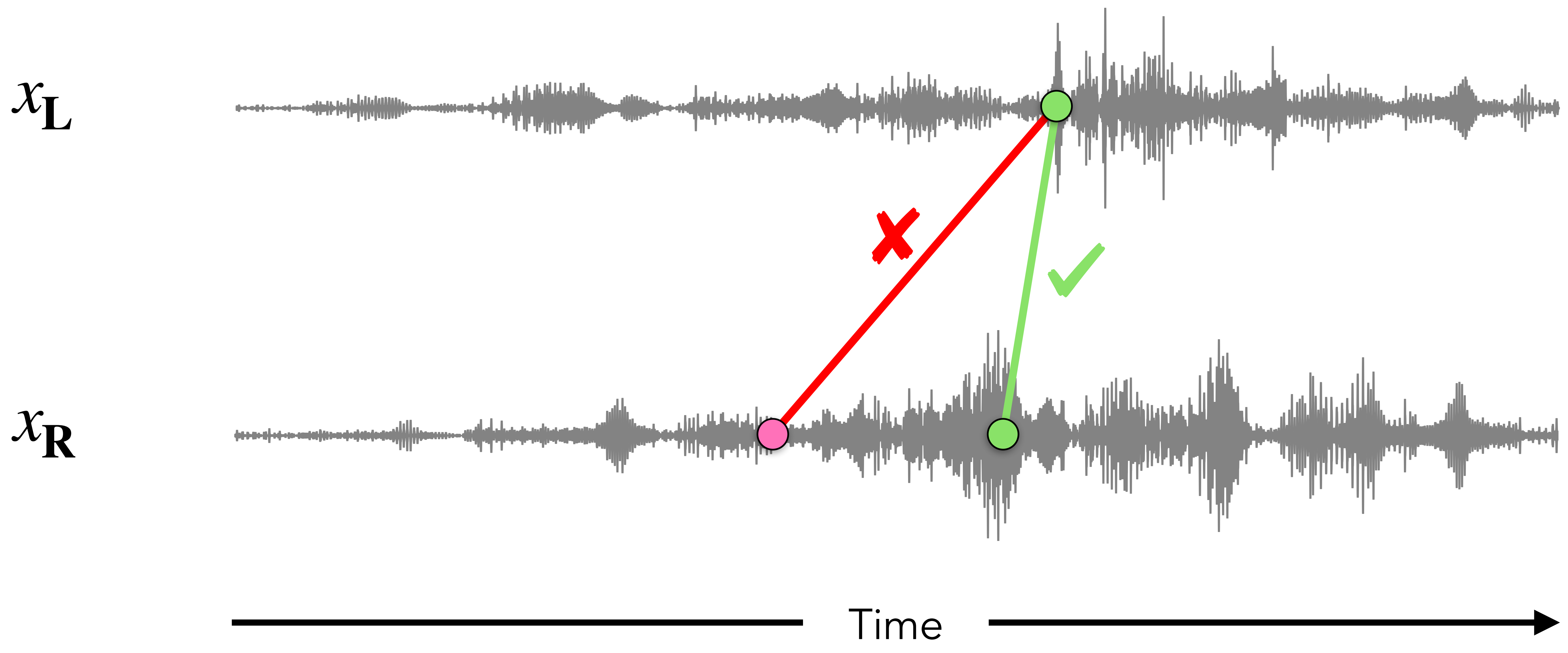
Interaural Correspondence

Objective: find a time delay τ that maximizes **generalized cross-correlation**:

$$R_{\mathbf{x}_1,\mathbf{x}_2}(\tau) = \mathbb{E}_t\left[\mathbf{h}_1(t) \cdot \mathbf{h}_2(t-\tau)\right]$$

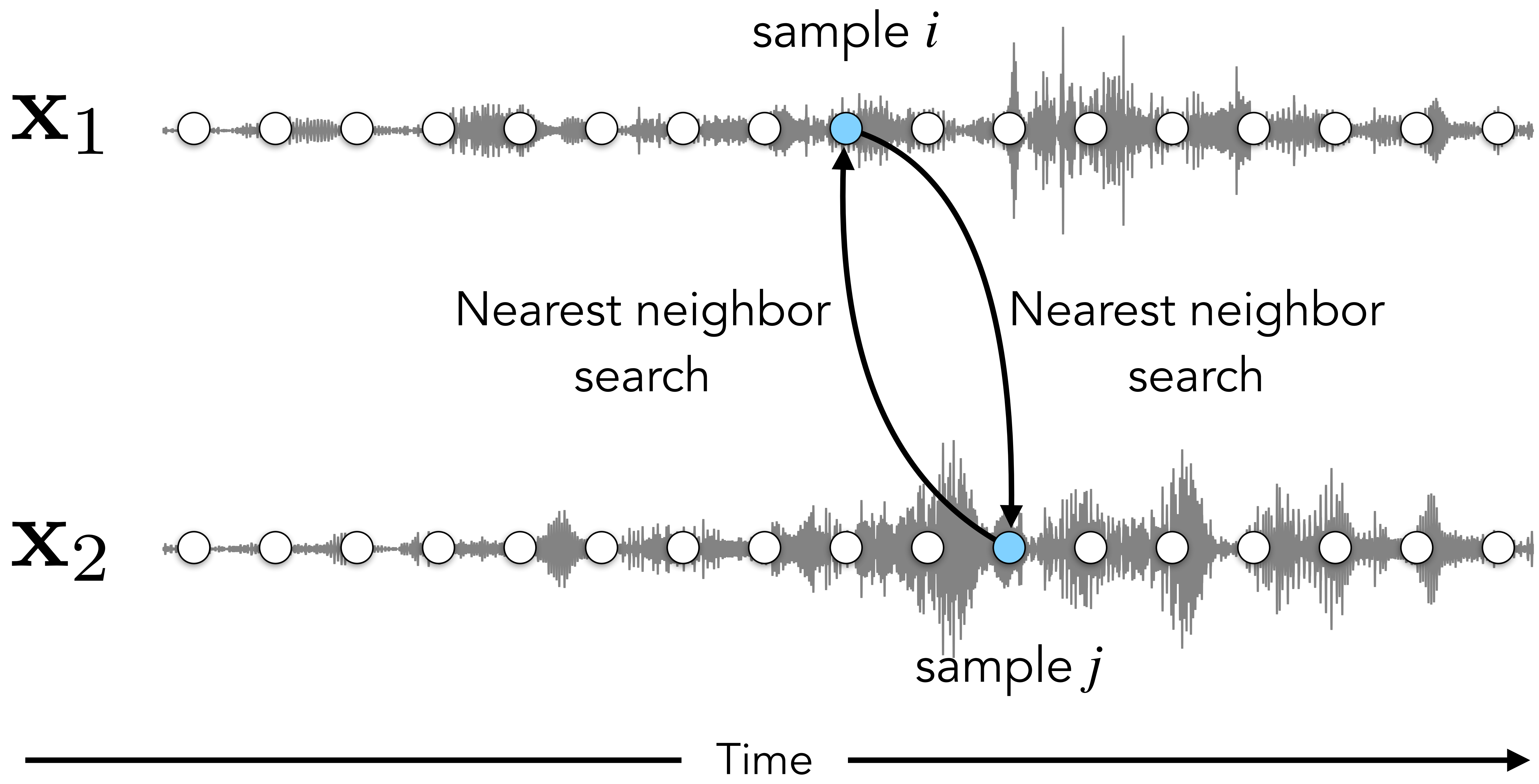
$$\downarrow$$

Learn audio embedding $\mathbf{h}$



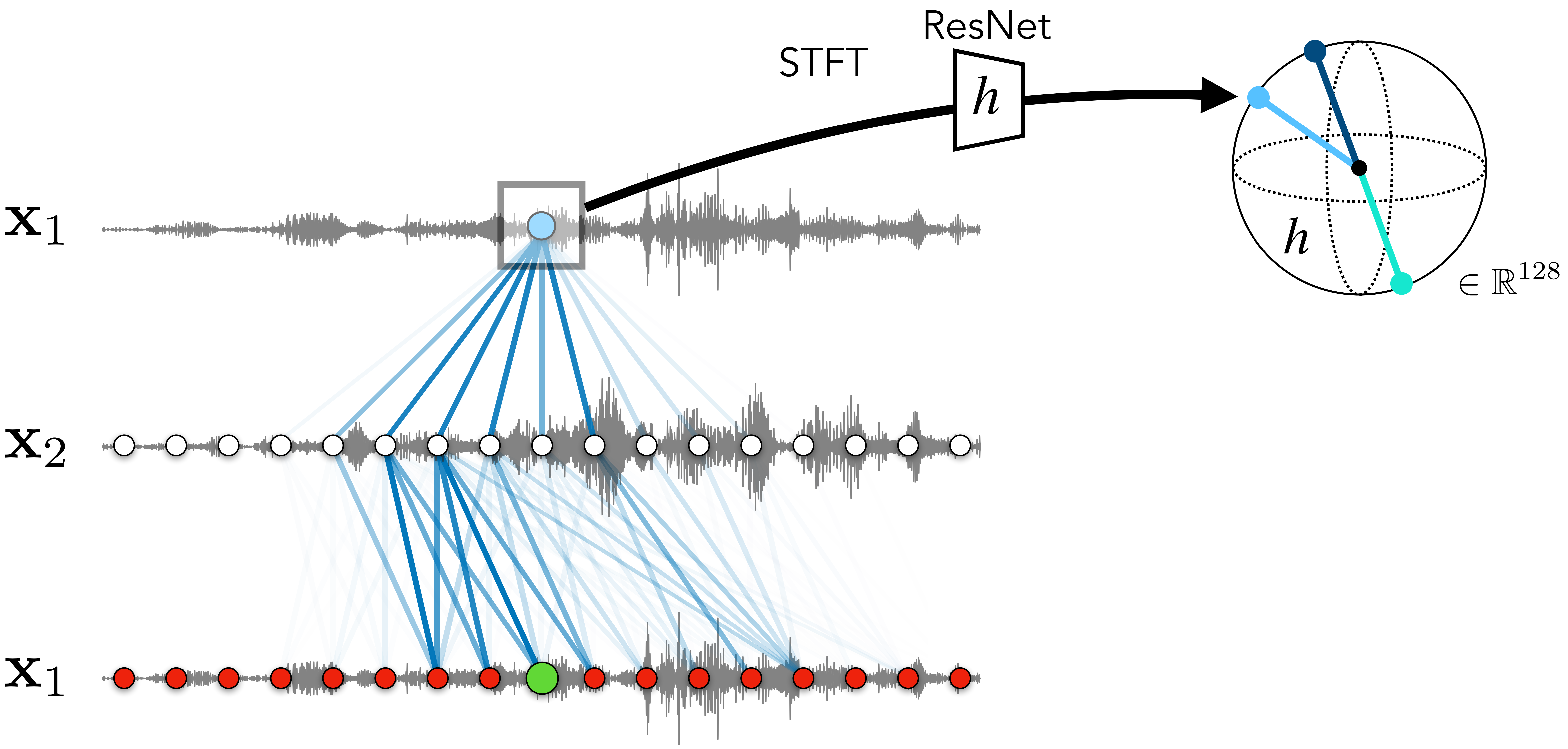
[Knapp and Carter, 1976]

Cycle-Consistency as Supervision



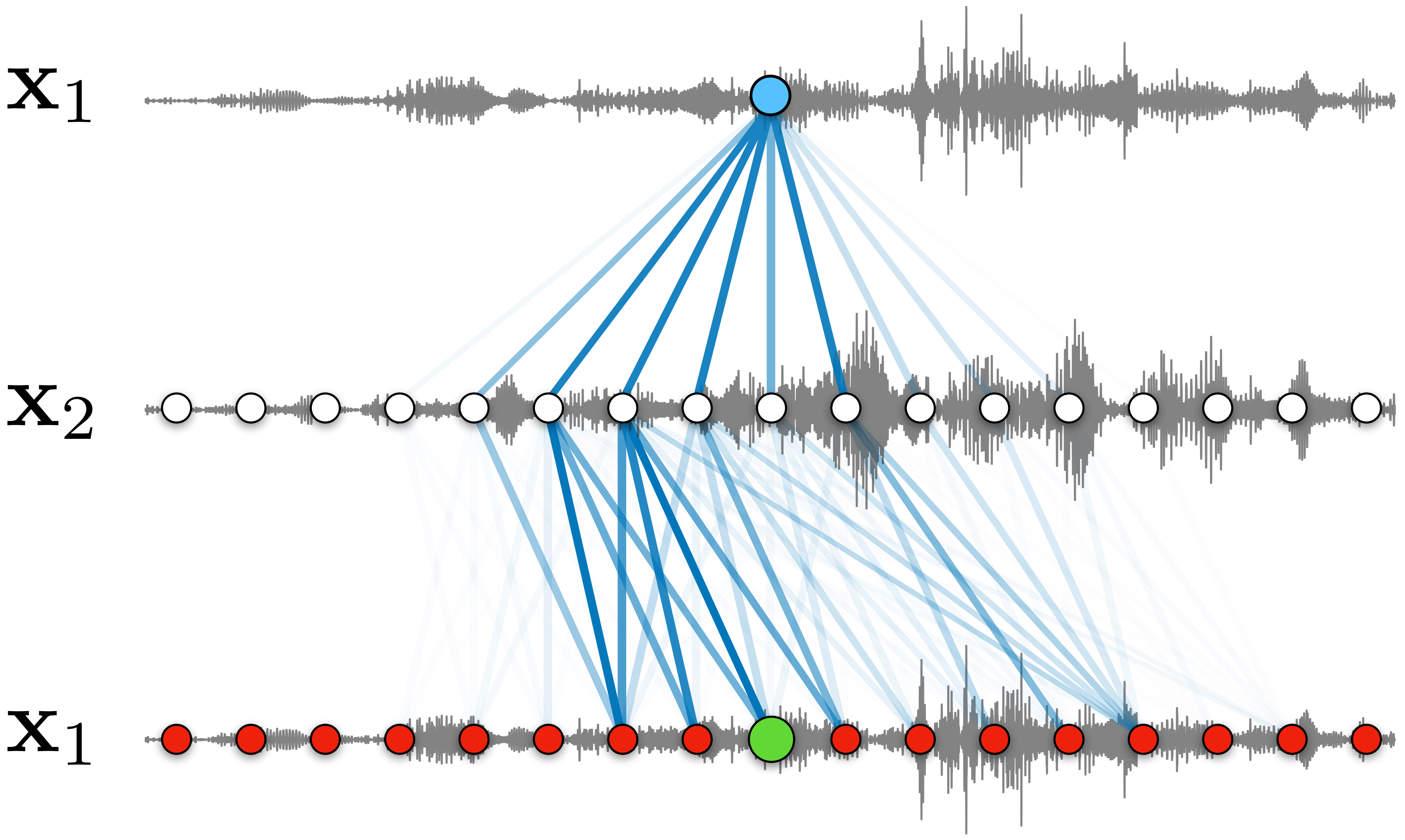
Wang et al., 2019
Jabri et al., 2020

Learning Interaural Correspondence



Interaural contrastive random walk
(StereoCRW)

Learning Interaural Correspondence



Interaural contrastive random walk (StereoCRW)

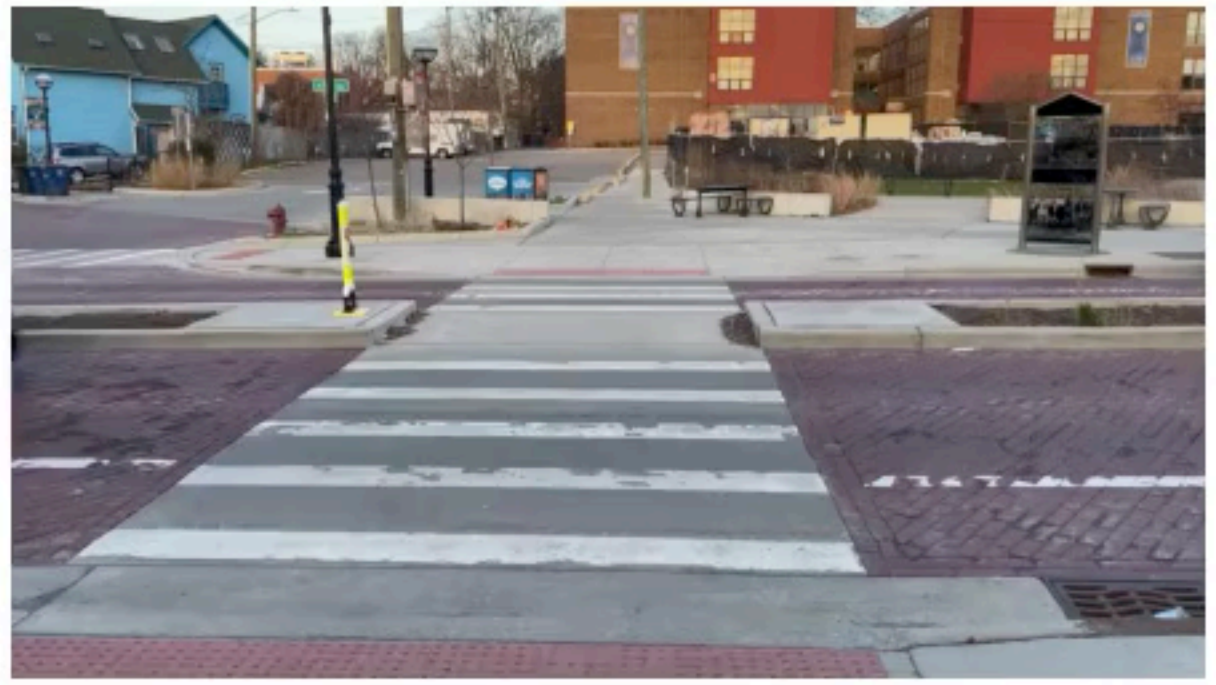
Transition probability from sample s in x_i to sample t in x_j :

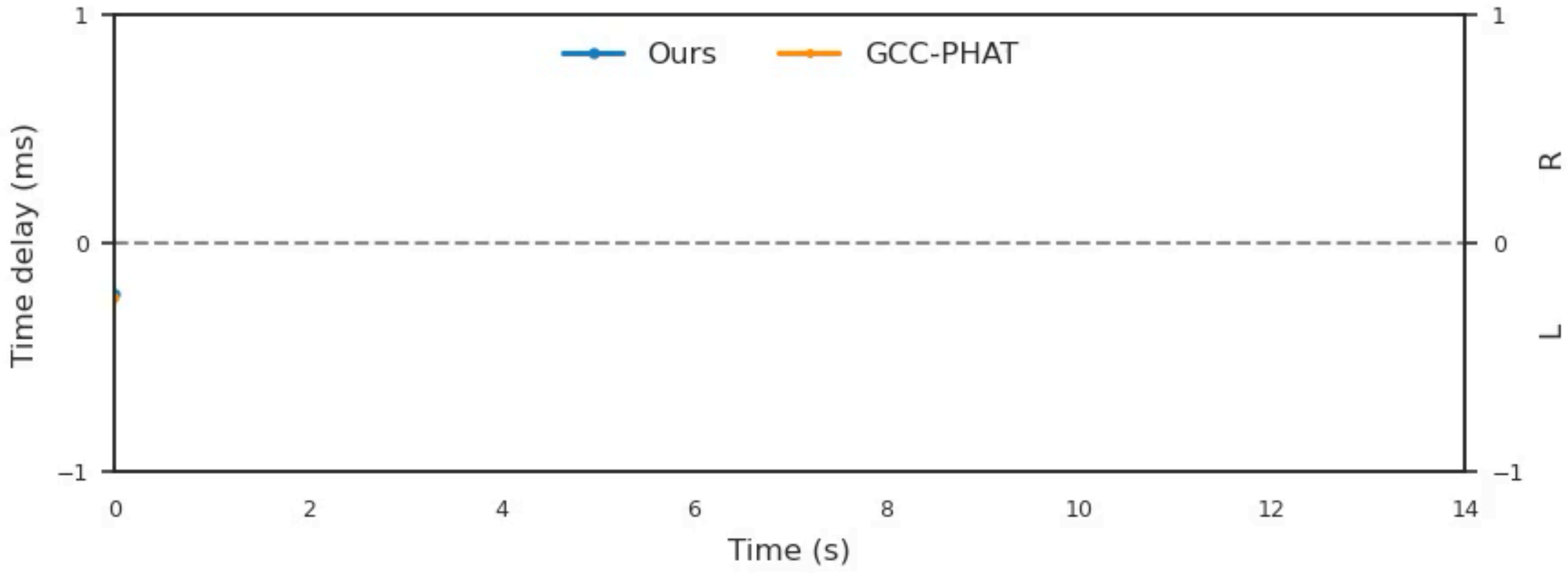
$$A_{ij}(s,t) = \frac{\exp(\mathbf{h}_i(s) \cdot \mathbf{h}_j(t)/c)}{\sum_{k=1}^{n} \exp(\mathbf{h}_i(s) \cdot \mathbf{h}_j(k)/c)}$$

Maximize the return probability of a walk moves between channels:

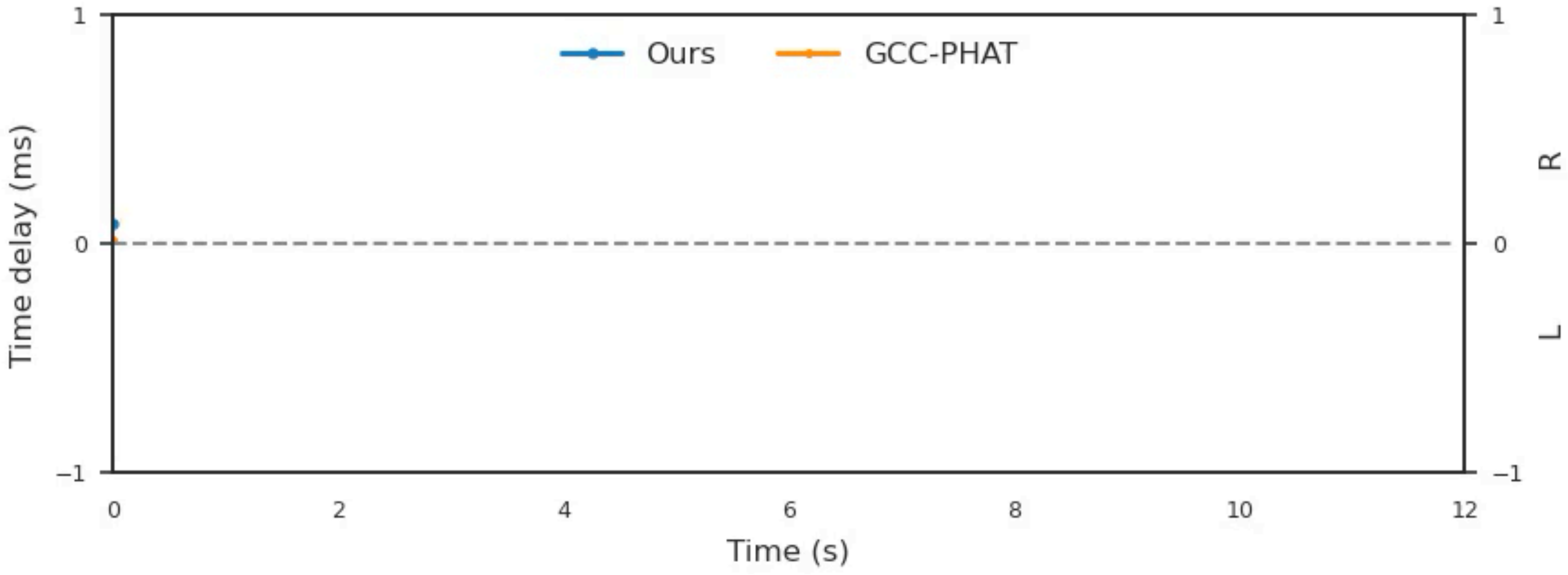
$$\mathrm{tr}(\log(A_{12}A_{21}))$$

Application to in-the-wild phone recordings





Self-recorded video using iPhone 12



iPhone 12 video from Flickr

Tracking

?

Lots of different formulations…

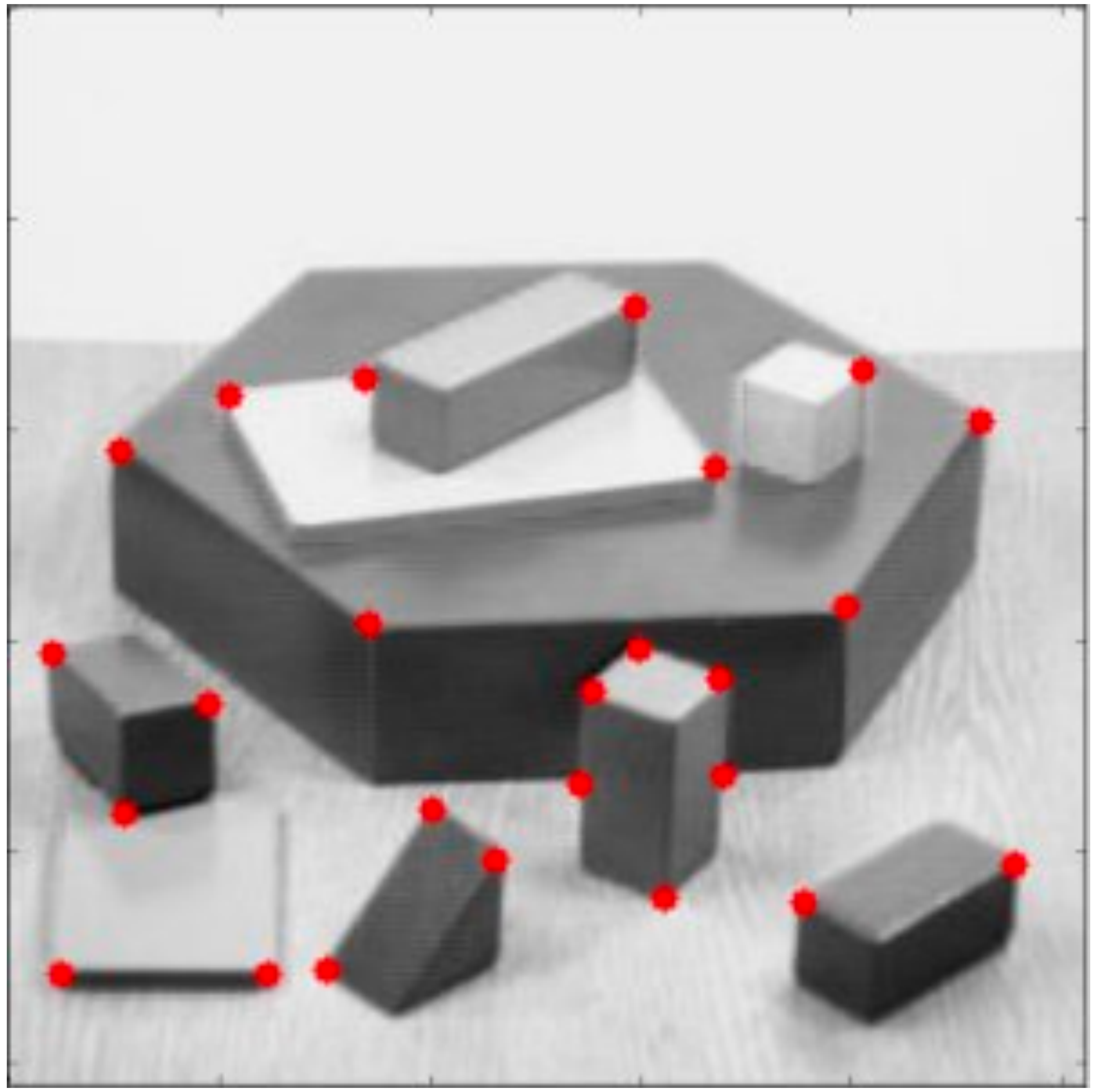
Other formulations of motion estimation

- Point tracking
- Object tracking
- Unsupervised optical flow

Some challenges in tracking

- Track scene structures over long periods of time
- Deal with occlusion and disocclusion
- Often framed in terms of tracking *objects*
- Keep instances distinct

Sparse feature tracking

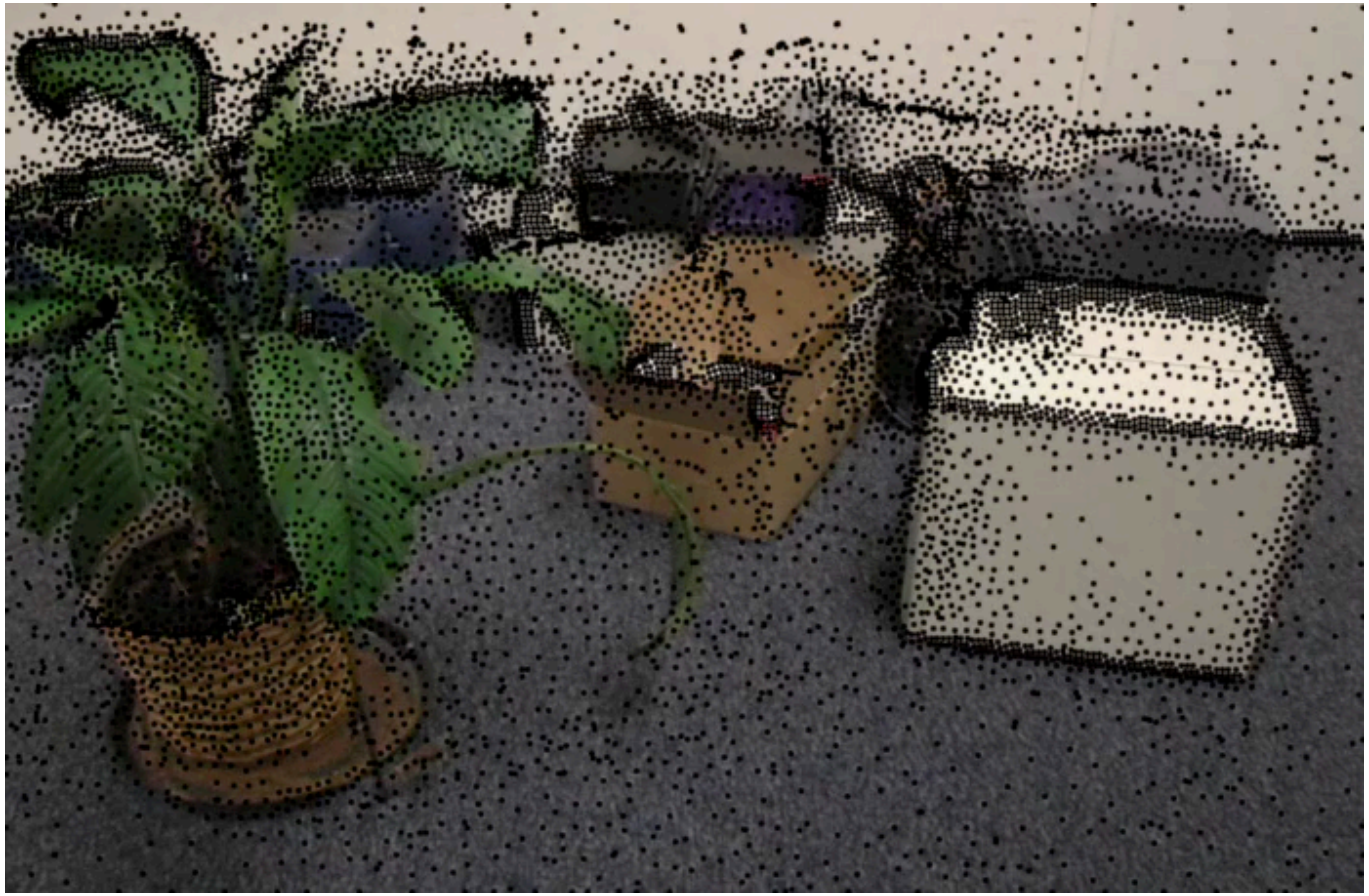


- Detect sparse features (like Harris corners)
- Compute a descriptor at each one
- Match in the next frame

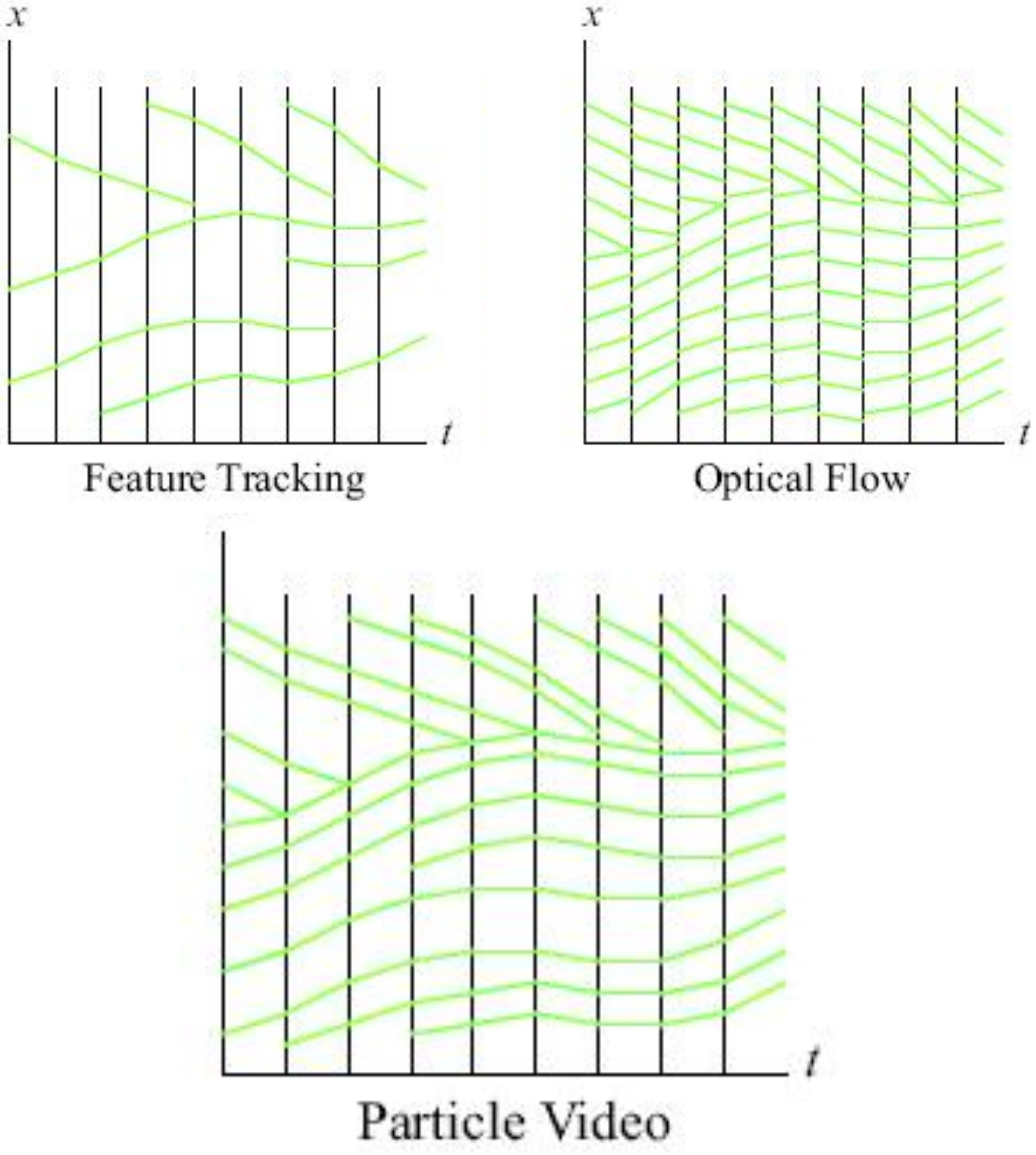
[Shi & Tomasi, “Good Features to Track”, 1994]

Image source: <u>OpenCV</u>

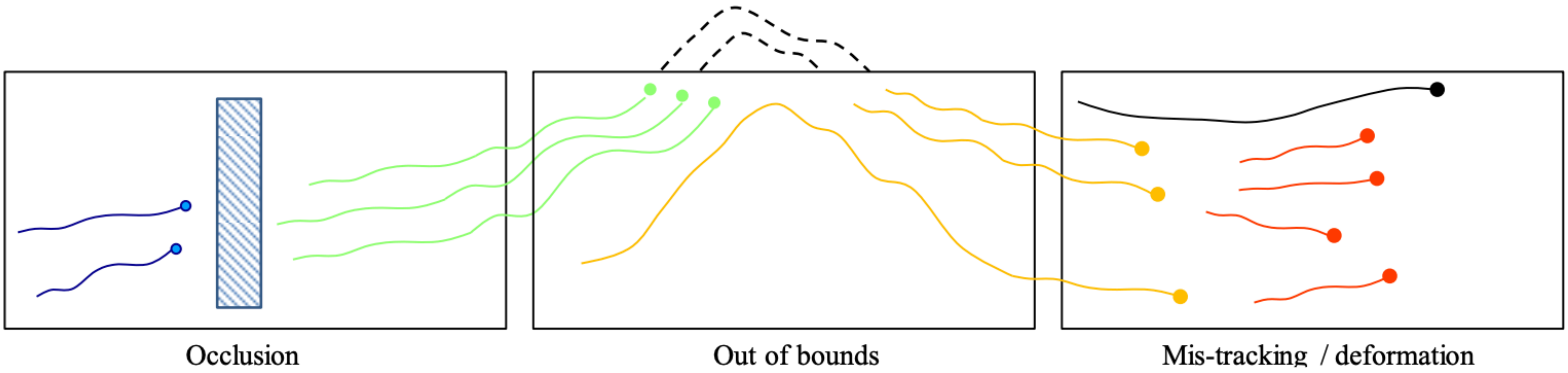
Semi-dense tracking



[Sand & Teller, “Particle Video”, 2006]

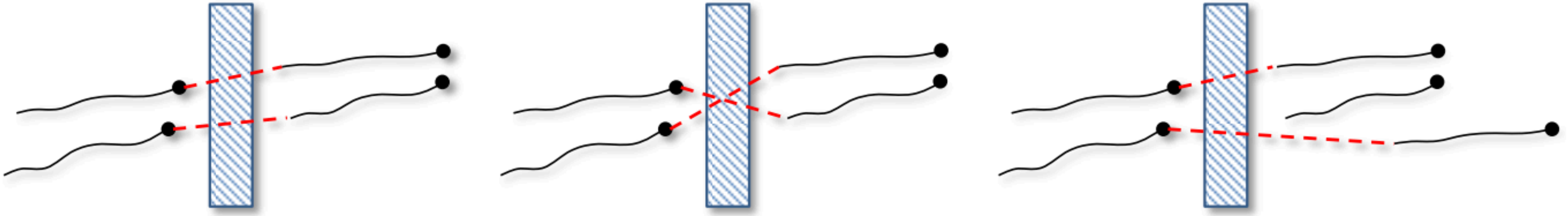


Challenges



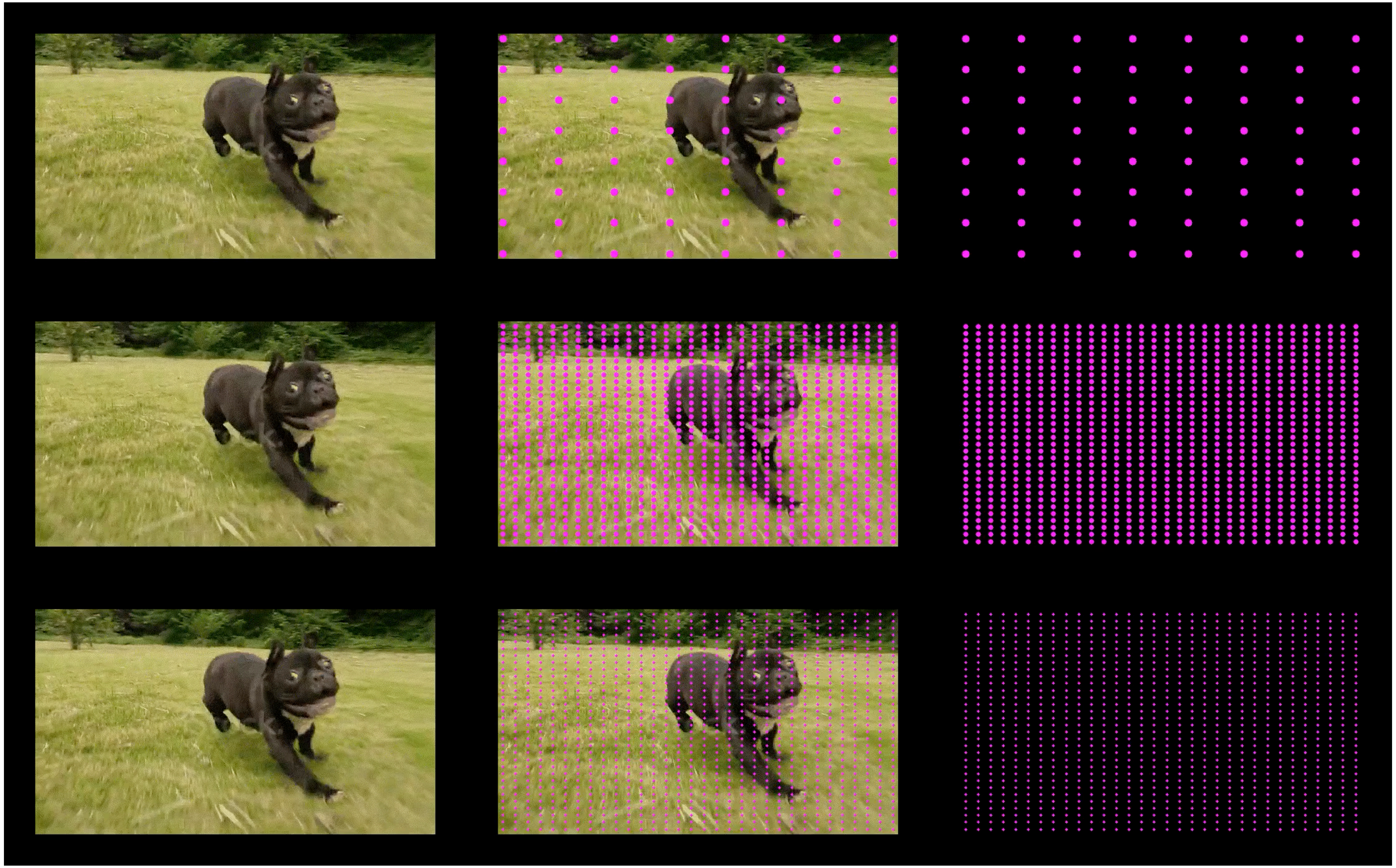
[Rubinstein et al., “Towards Longer Long-Range Motion Trajectories”, 2012]

Association problem



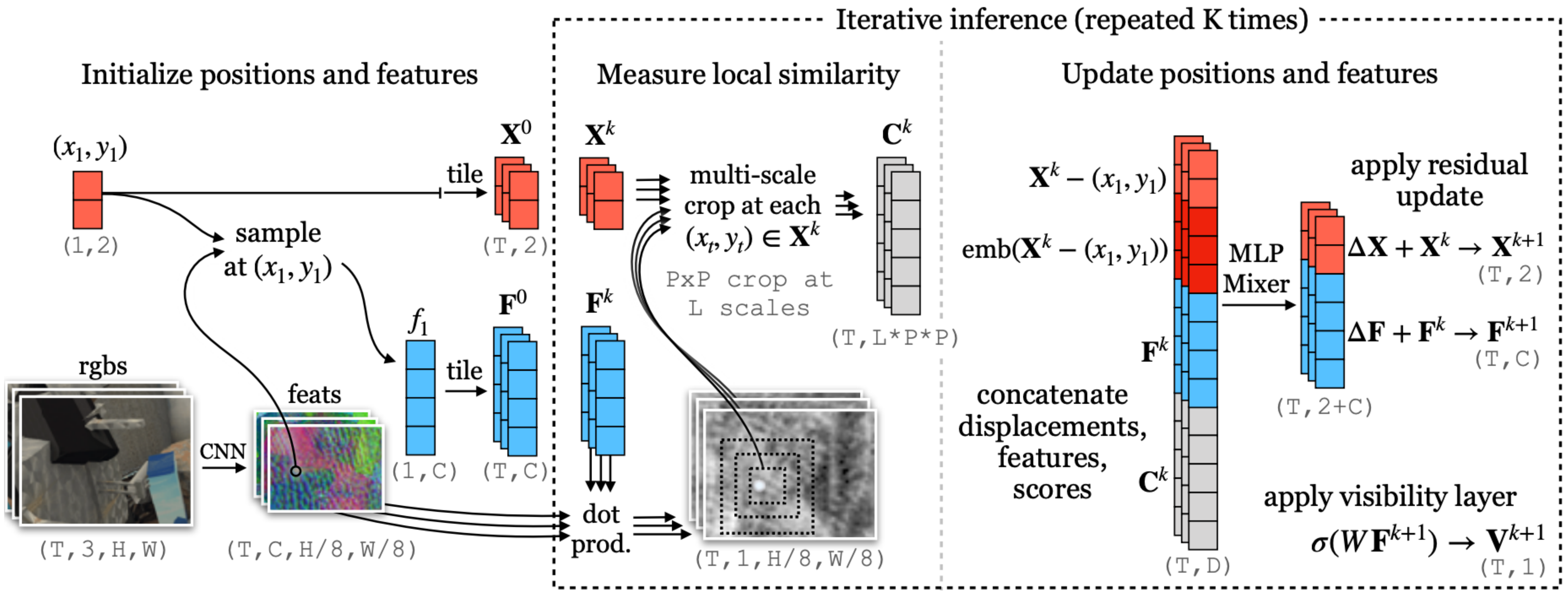
[Rubinstein et al., “Towards Longer Long-Range Motion Trajectories”, 2012]

Persistent independent particles



[Harley et al., “Particle Video Revisited”, 2022]

Persistent independent particles



Other formulations of motion estimation

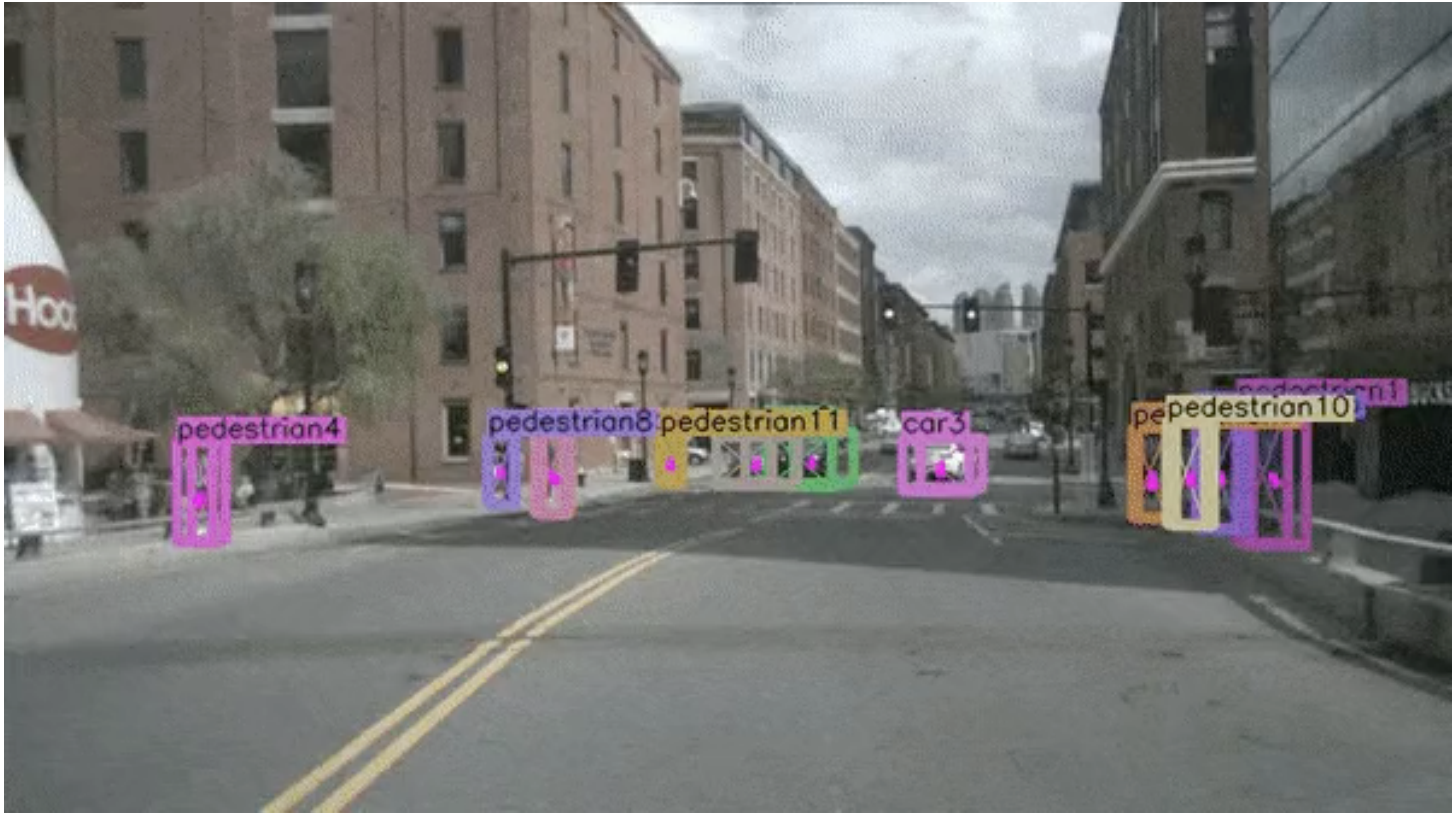
- Point tracking
- **Object tracking**
- Unsupervised optical flow

Tracking as repeated detection

- Detect an objects at time *t*
- Update the detector (optional)
- Detect objects at *t+1* and match them

[Ramanan et al., “Strike a Pose: Tracking People by Finding Stylized Poses”, 2005]

Tracking objects



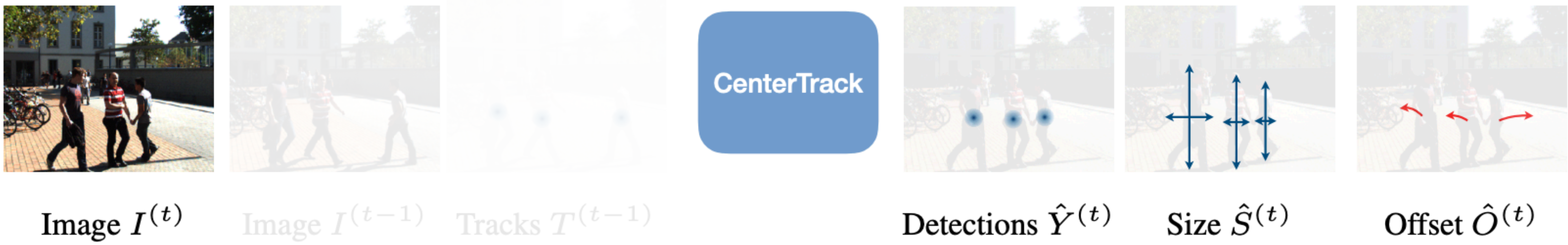
Representative recent example: [Zhou et al., “Tracking objects as points”, 2020]

Tracking objects as points

[Zhou et al., “Tracking objects as points”, 2020]

Inputs and outputs

“Recurrent” predictions



Learning to associate objects

During training:

- Randomly jitter detections from previous frames to simulate prediction errors
- Add false positives near the ground truth.
- Choose “previous” frame from {-2, -1, 0, 1, 2} frames away.

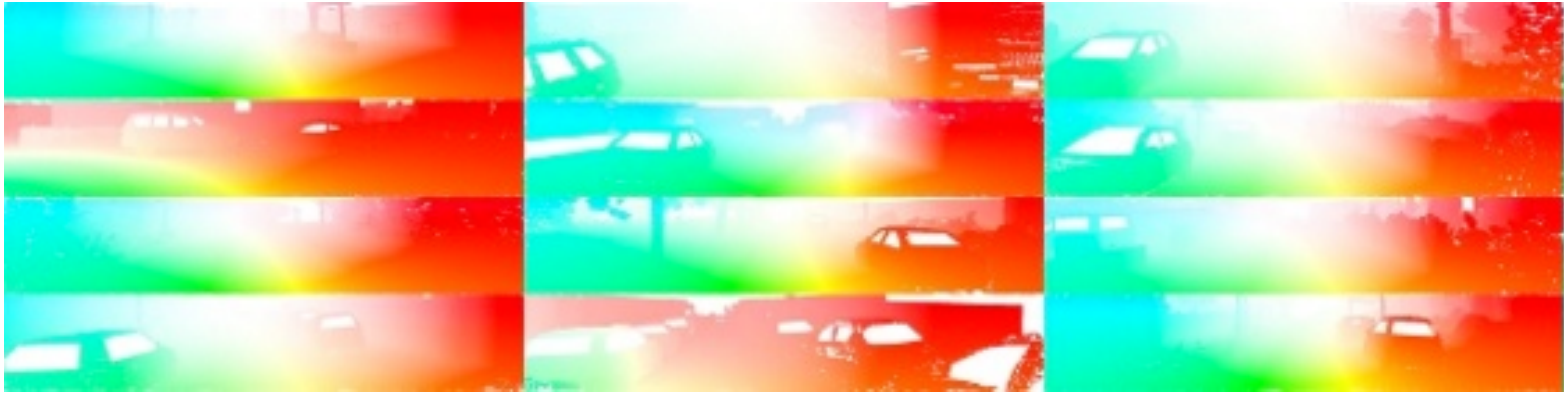
During inference:

- Use a greedy association
- No nearby match? Spawn a new tracklet

Other formulations of motion estimation

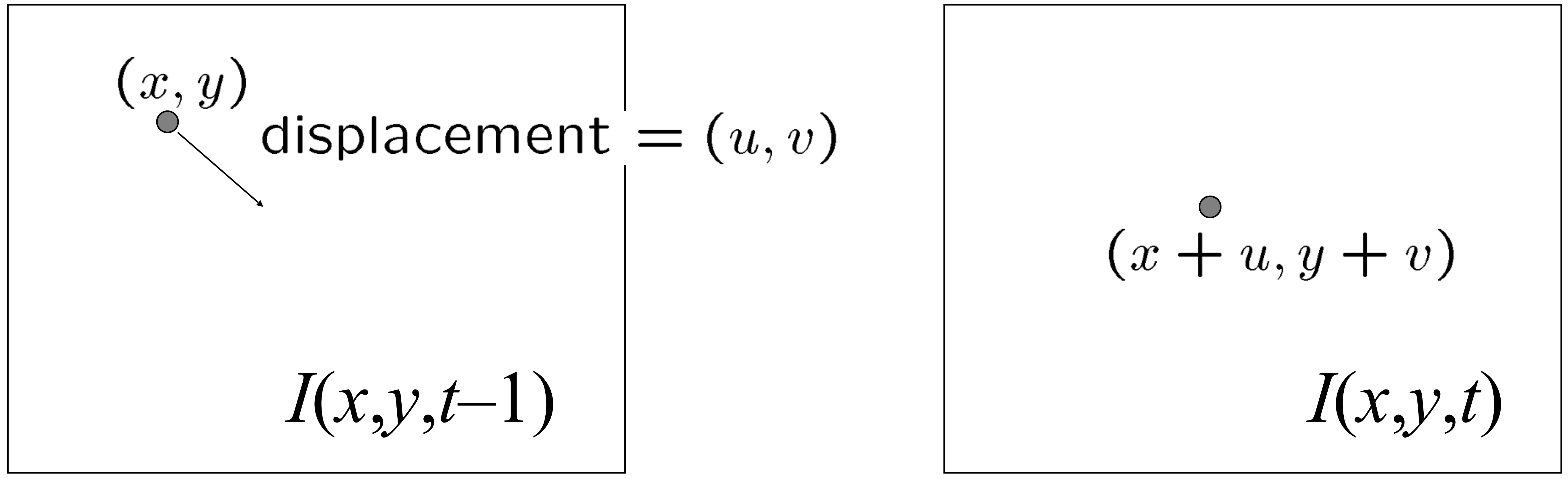
- Point tracking
- Object tracking
- **Unsupervised optical flow**

Problem: hard to get flow supervision



KITTI dataset [Geiger et al.]

Unsupervised optical flow



Recall: minimize matching error + smoothness [Horn and Schunck 1981]

$$\underbrace{\sum_{x,y}[I(x,y,t-1) - I(u(x),v(y),t)]^2}_{E_d(u,v)\ \text{match cost}} + \underbrace{\sum_{p}\sum_{p' \in \mathcal{N}}(u(p)-u(p'))^2 + (v(p)-v(p'))^2}_{E_s(u,v)\ \text{smoothness}}$$

Solution we saw before: optimize using nonlinear least squares.

Unsupervised optical flow

Recall: minimize matching error + smoothness [Horn and Schunck 1981]

$$\underbrace{\sum_{x,y}[I(x,y,t-1) - I(u(x),v(y),t)]^2}_{E_d(u,v) \text{ match cost}} + \underbrace{\sum_{p}\sum_{p' \in \mathcal{N}}(u(p)-u(p'))^2 + (v(p)-v(p'))^2}_{E_s(u,v) \text{ smoothness}}$$

Estimate with neural net instead: $\begin{bmatrix} u(p) \\ v(p) \end{bmatrix} = f(I_t, I_{t+1}, p; \theta)$

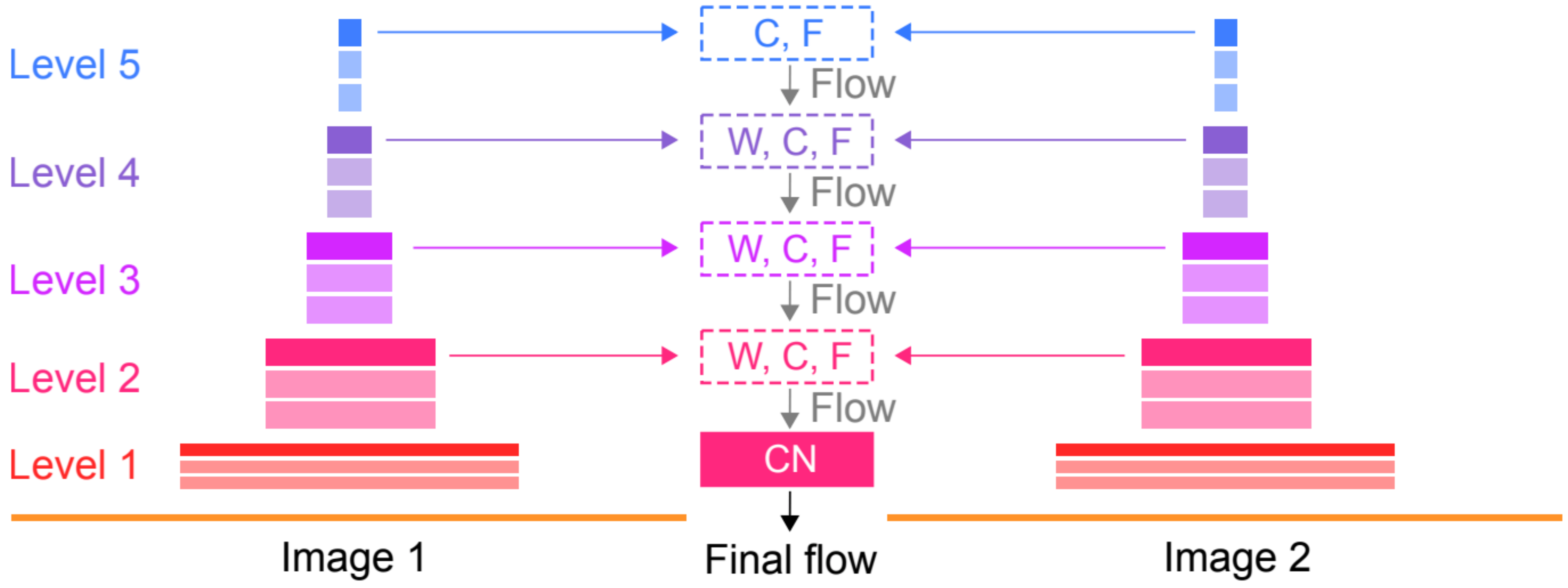
Why might this work better?

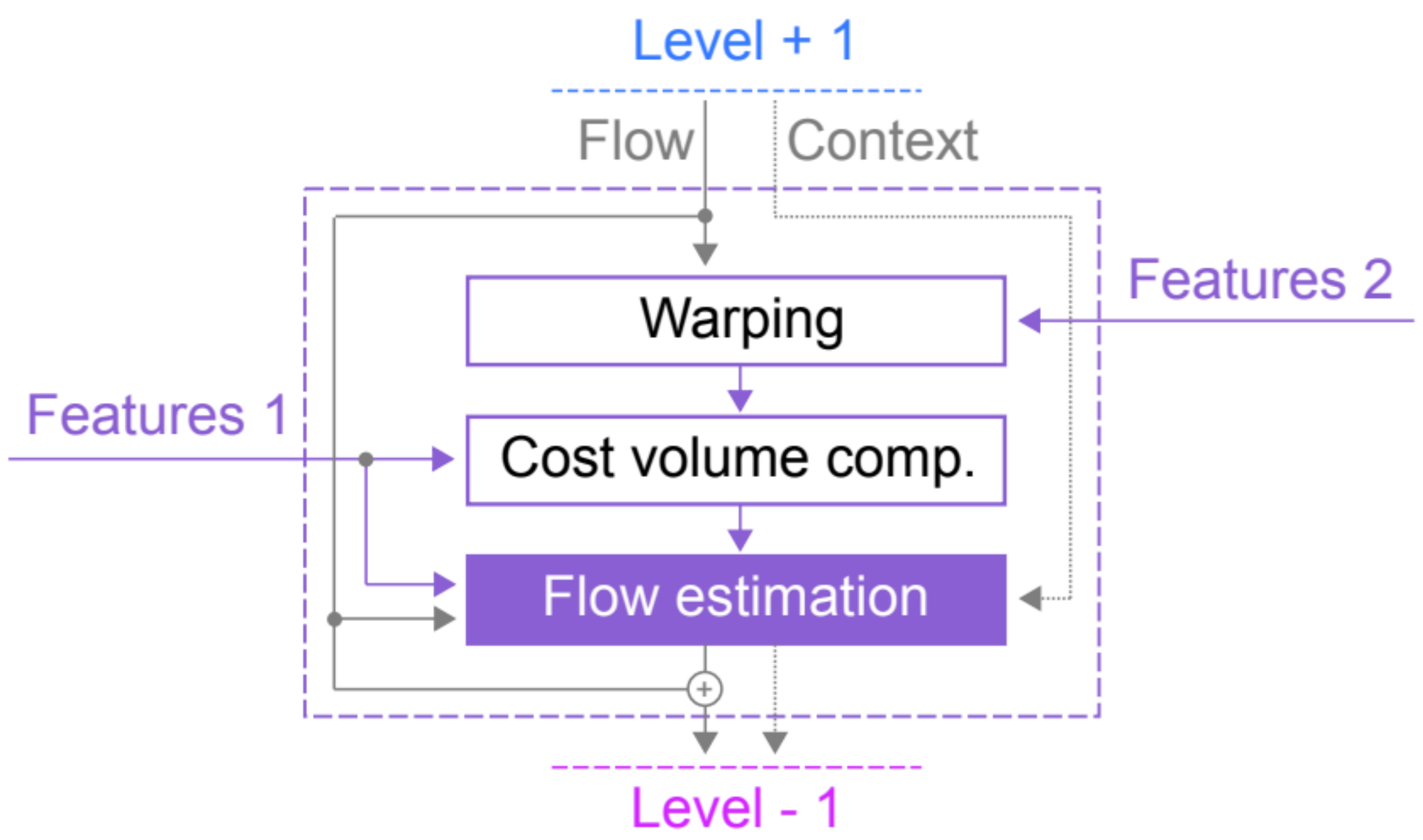
What Matters in Unsupervised Optical Flow

Rico Jonschkowski[1,2], Austin Stone[1,2], Jonathan T. Barron[2],
Ariel Gordon[1,2], Kurt Konolige[1,2], and Anelia Angelova[1,2]

[1]Robotics at Google and [2]Google AI

An optical flow network





[Johnschkowski et al., “What Matters in Unsupervised Optical Flow”, 2020]

Learning with photometric cost + smoothness prior

Create a loss that encourages the following.

The flow you generate should have the following properties:

1. Matched pixels should have similar color.
2. Should have spatial smoothness.
3. Special handling for pixel that don’t have a match (e.g. occlusions)

Qualitative results