Efficient Temporal Consistency for Streaming Video Scene Analysis

Ondrej Miksik **Daniel Munoz**

J. Andrew Bagnell Martial Hebert

Oxford Brookes Vision Group

Carnegie Mellon
THE ROBOTICS INSTITUTE

Scene Analysis/Parsing

Munoz ECCV 2010, Ladicky ECCV 2010, Tighe ECCV 2010, Sorcher ICML 2011, Farabet ICML 2012

Temporal Inconsistency

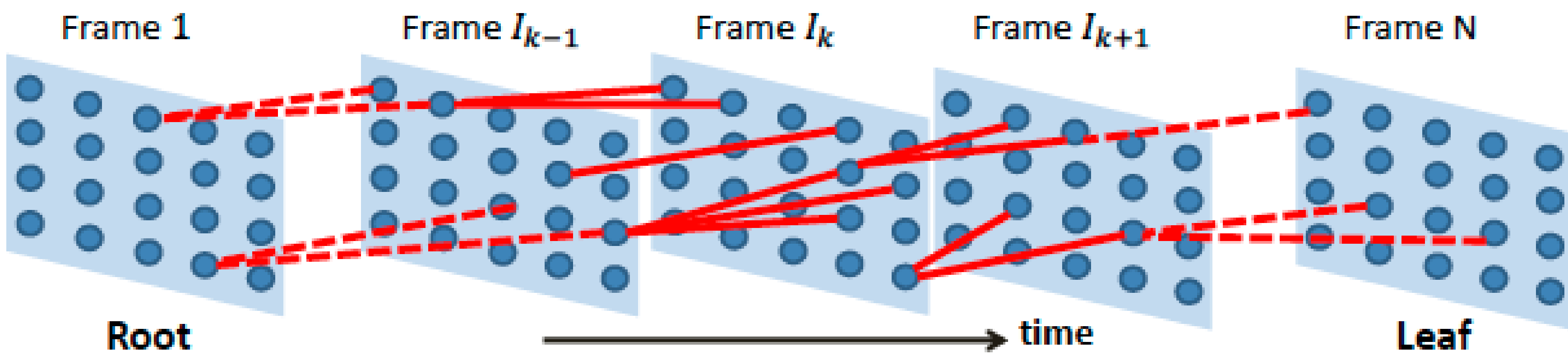
www.youtube.com/user/dmunozcmuri

Per-frame
Munoz ECCV 2010

Temporal
This Work

sky | tree | road | sidewalk | building | car
column_pole | pedestrian | bicycle | fence | sign_symbol

Spatio-temporal Graphical Model



- Approximate inference
- Typically the MAP labeling (no uncertainty)
- Mini-batch processing

Wojek & Schiele ECCV 2008, Ess et al. BMVC 2009, Xiao & Quan ICCV 2009

Spatio-temporal Segmentation

- Hard decisions

Grundmann CVPR 2010, Brendel & Todorovic ICCV 2011, Xu & Corso ECCV 2012

Spatio-temporal Segmentation

- Hard decisions

Grundmann CVPR 2010, Brendel & Todorovic ICCV 2011, Xu & Corso ECCV 2012

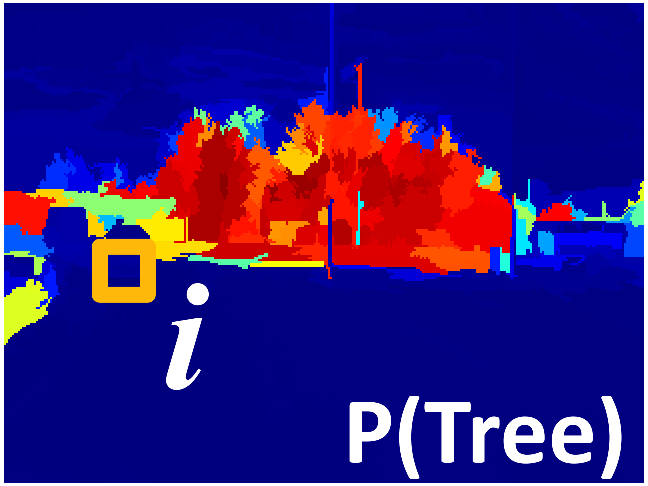
This Work

- Causal/Online/Running-average Algorithm
- Efficient and easy to implement
- Complements your favorite scene parser
- Outputs per-pixel probabilities

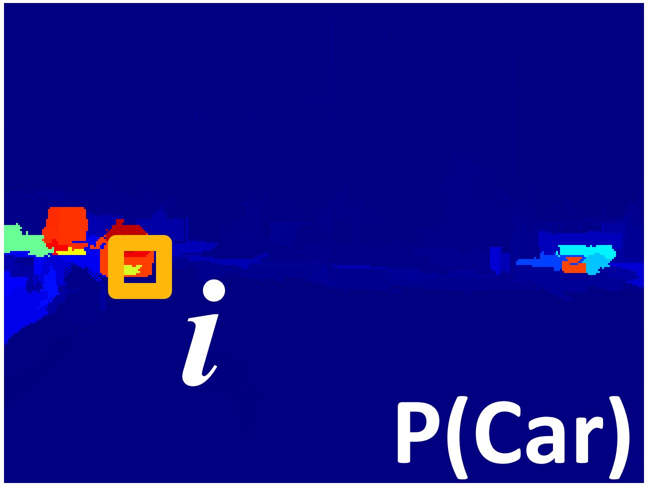
t-1

Temporally smoothed

Per-pixel Probabilities



. . .



$$\mathbf{y}_i = [P_i(\text{Tree}), P_i(\text{Sidewalk}), \ldots, P_i(\text{Car})]^T$$

t-1

t

Scene Parser

Munoz 2010

Farabet 2011

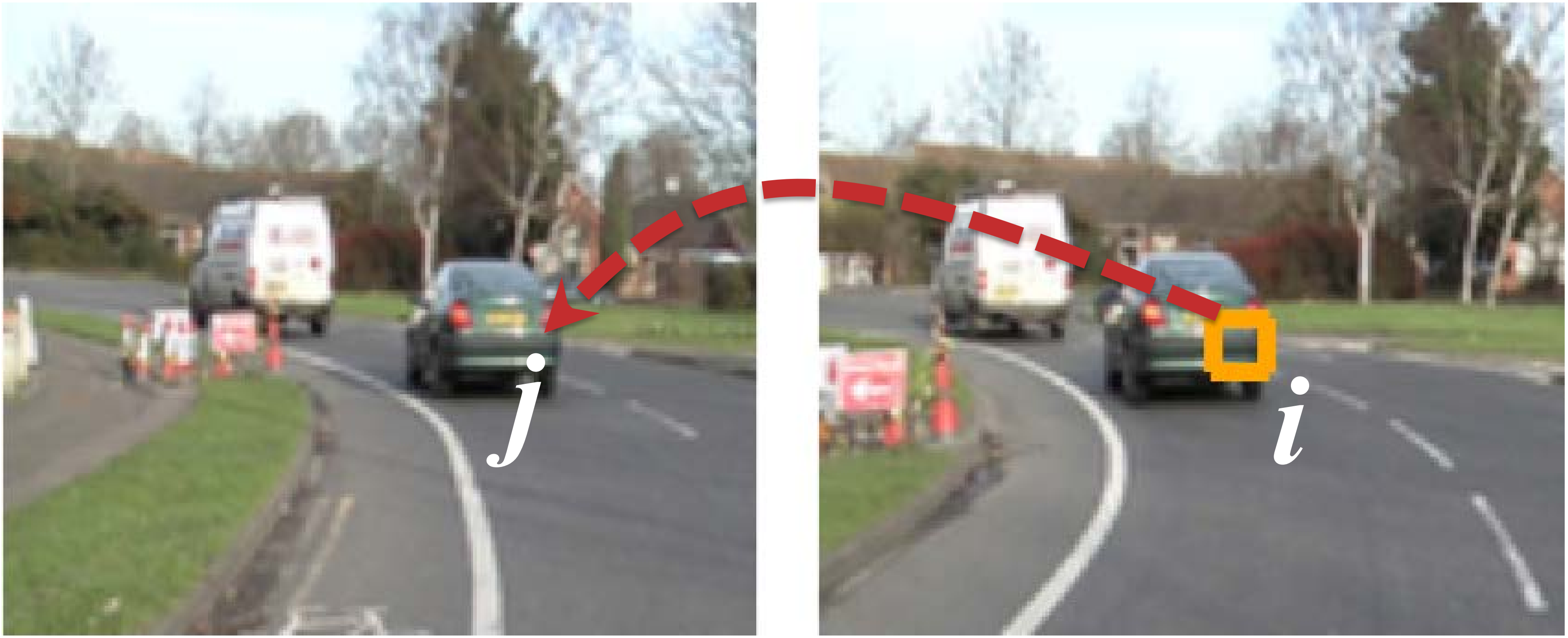
Wojek 2008

Make Consistent!

Temporally smoothed

Per-frame

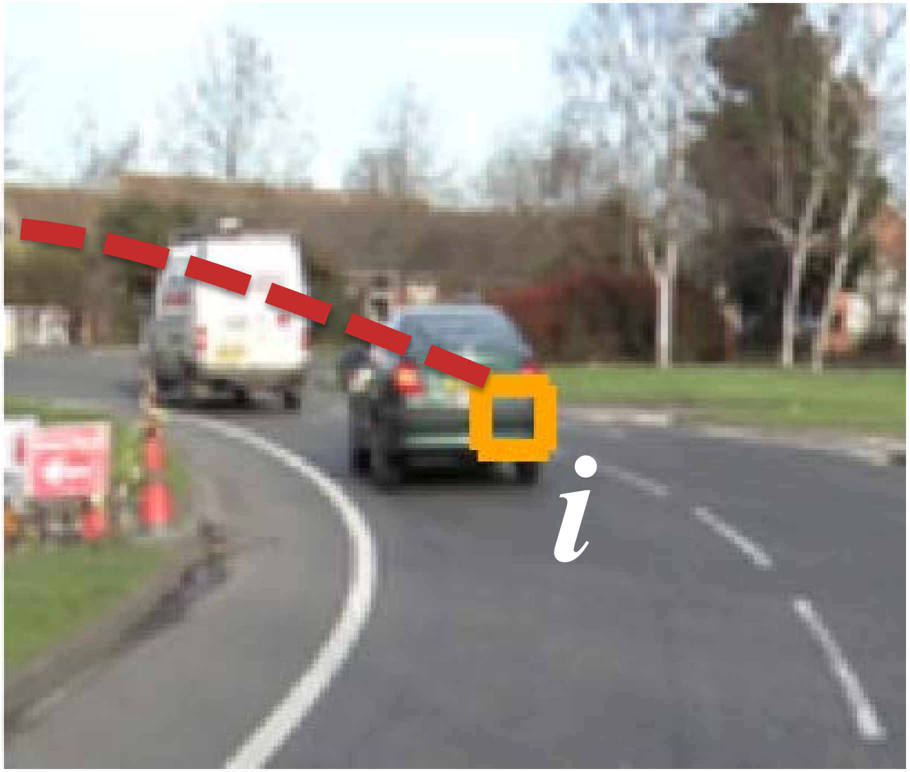
Ideally: Exact Correspondences



t-1 t

$$\mathbf{y}_i^{(t)} \quad \hat{\mathbf{y}}_j^{(t-1)}$$

Ideally: Exact Correspondences

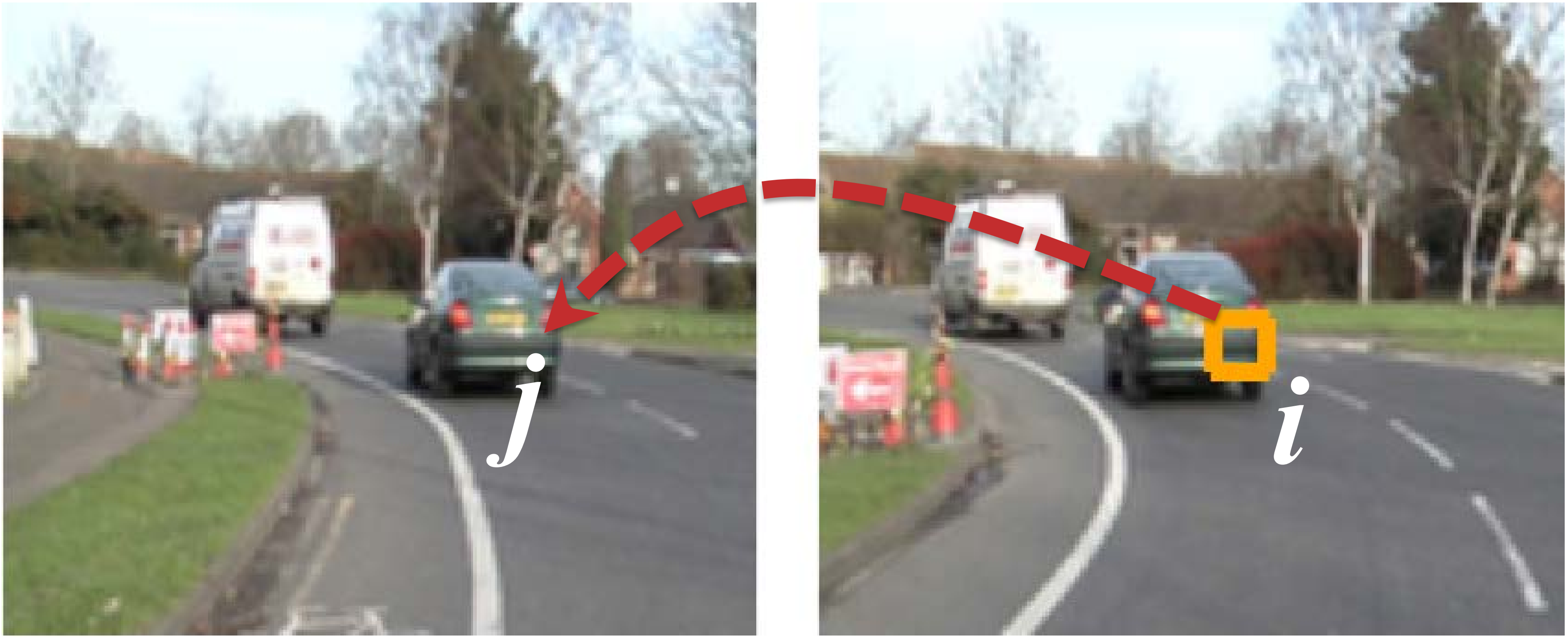


t-1

t

$$\hat{\mathbf{y}}_i^{(t)} \propto \mathbf{y}_i^{(t)} \circ \hat{\mathbf{y}}_j^{(t-1)}$$

Ideally: Exact Correspondences



t-1 t

$$\hat{\mathbf{y}}_i^{(t)} \propto \mathbf{y}_i^{(t)} + \hat{\mathbf{y}}_j^{(t-1)}$$

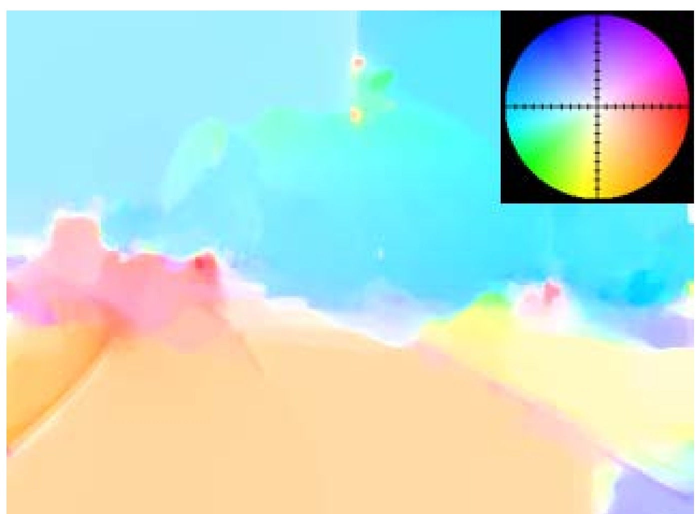
Approx. Correspondences via Optical Flow

t-1

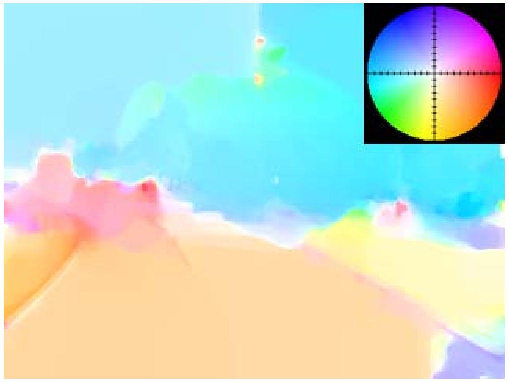
t

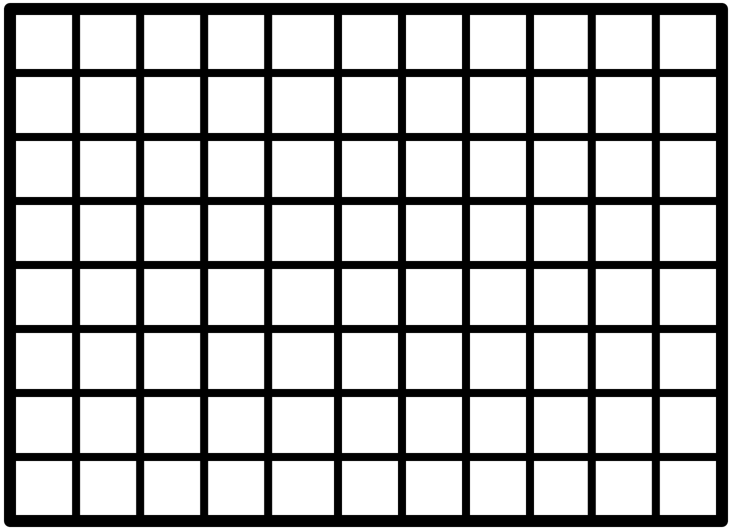
Optical Flow

Werlberger 2009

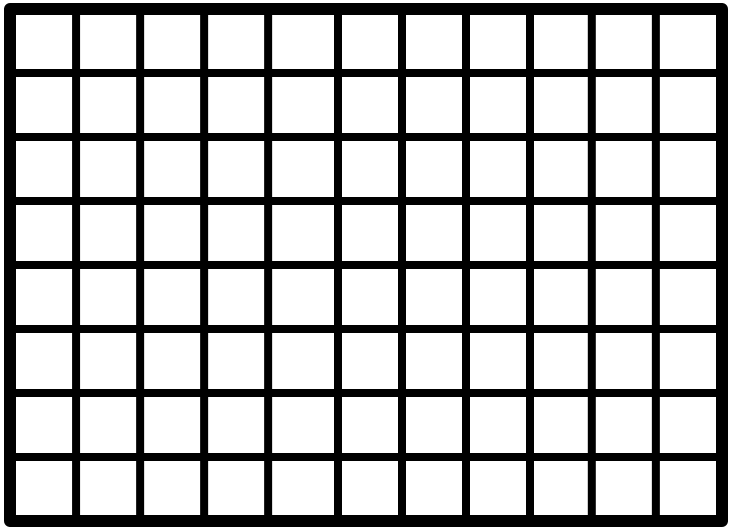


Handling No/Imperfect Correspondences



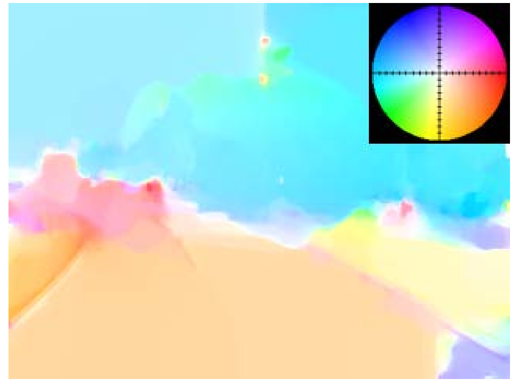


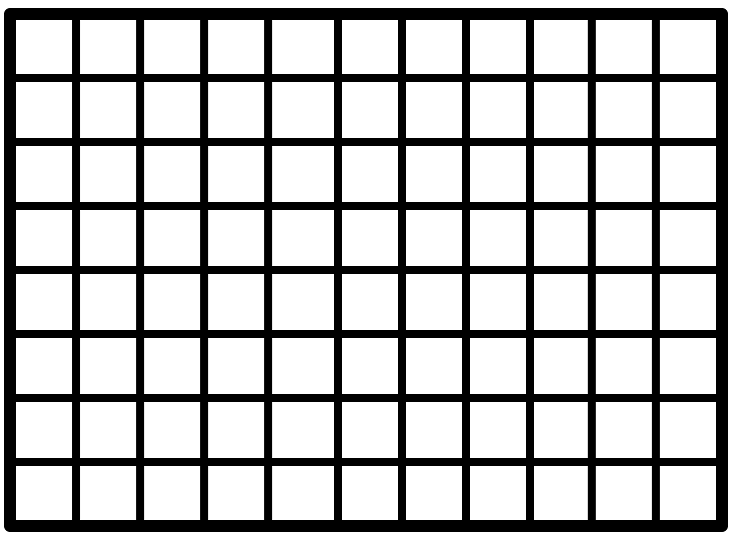
t-1



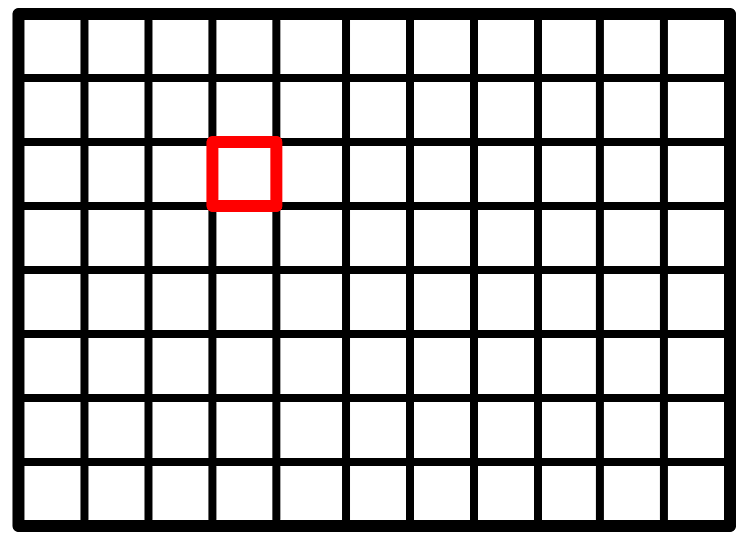
t

Handling No/Imperfect Correspondences



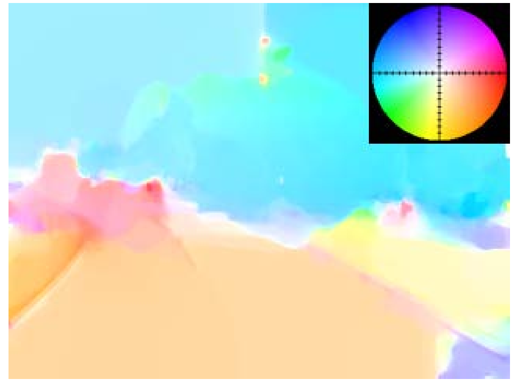


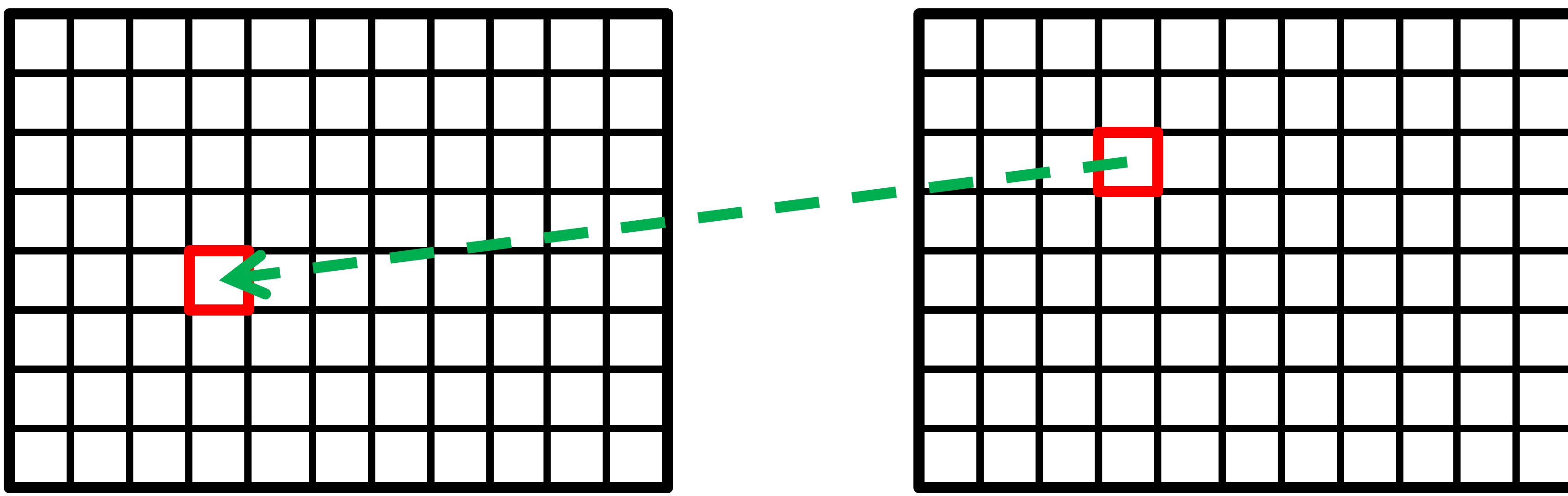
t-1



t

Handling No/Imperfect Correspondences

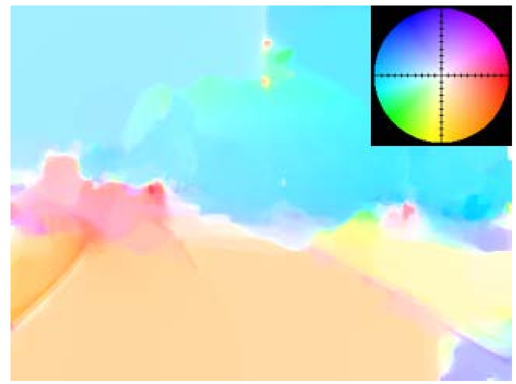


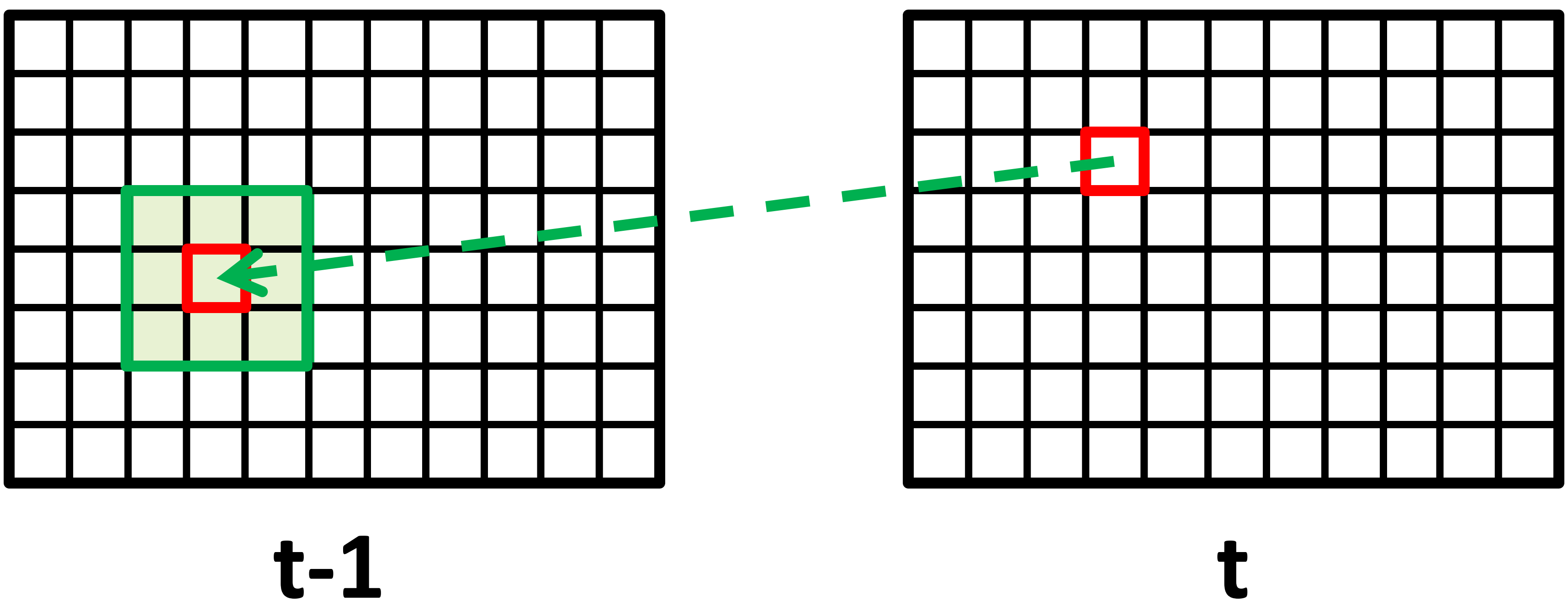


t-1

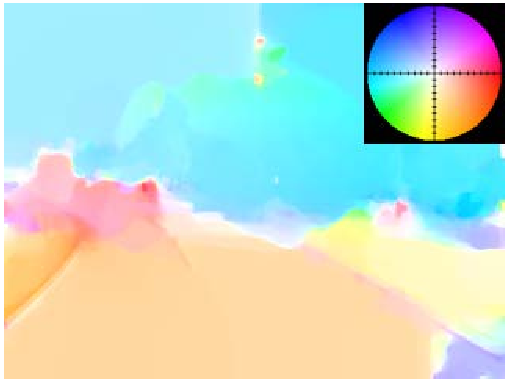
t

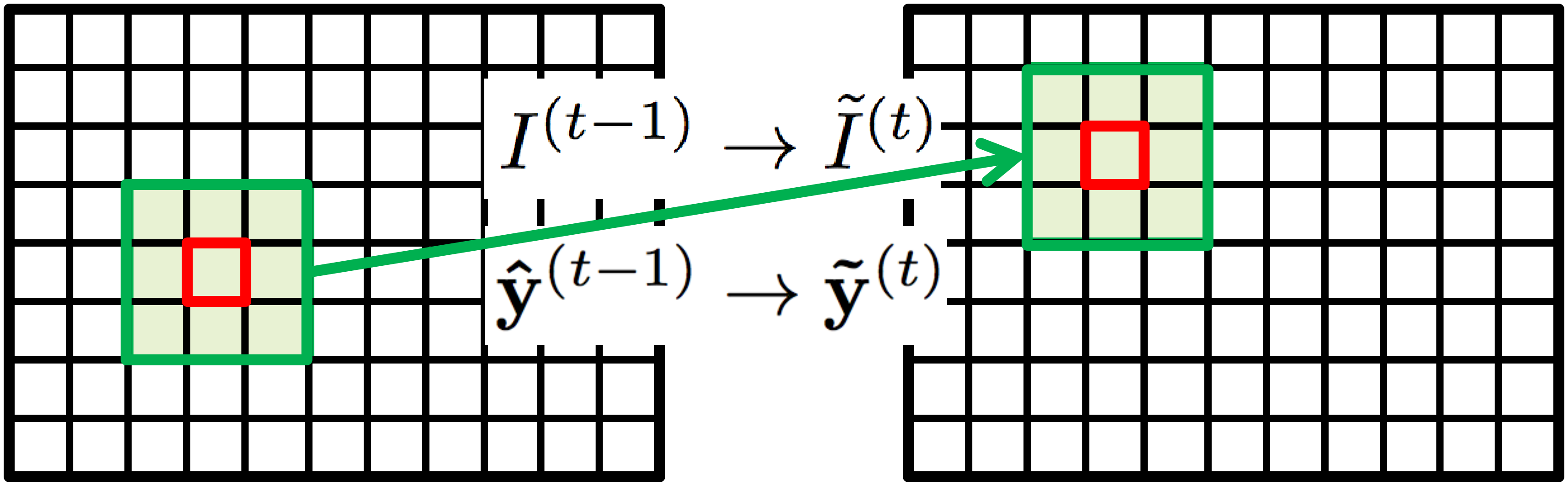
Handling No/Imperfect Correspondences





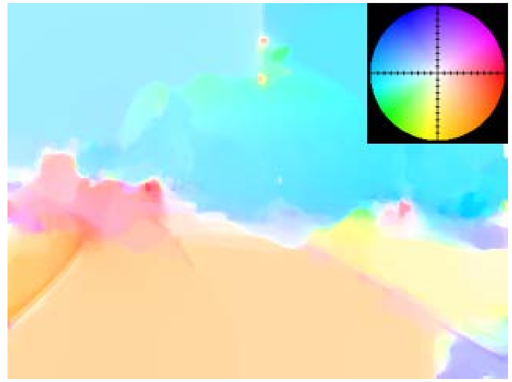
Handling No/Imperfect Correspondences

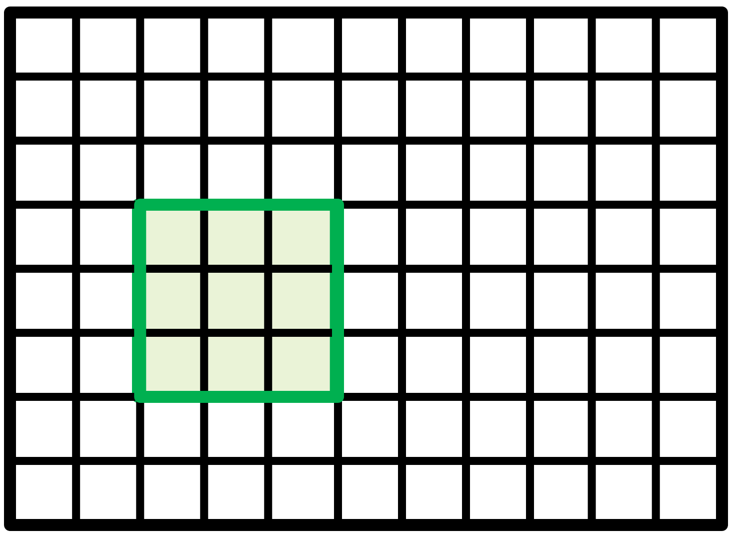




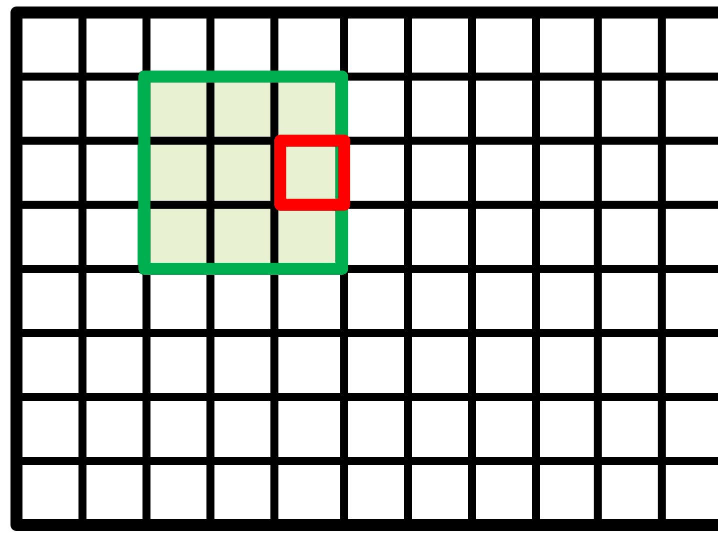
t-1 **t**

Handling No/Imperfect Correspondences



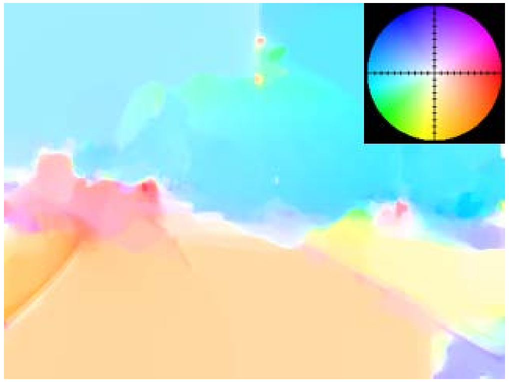


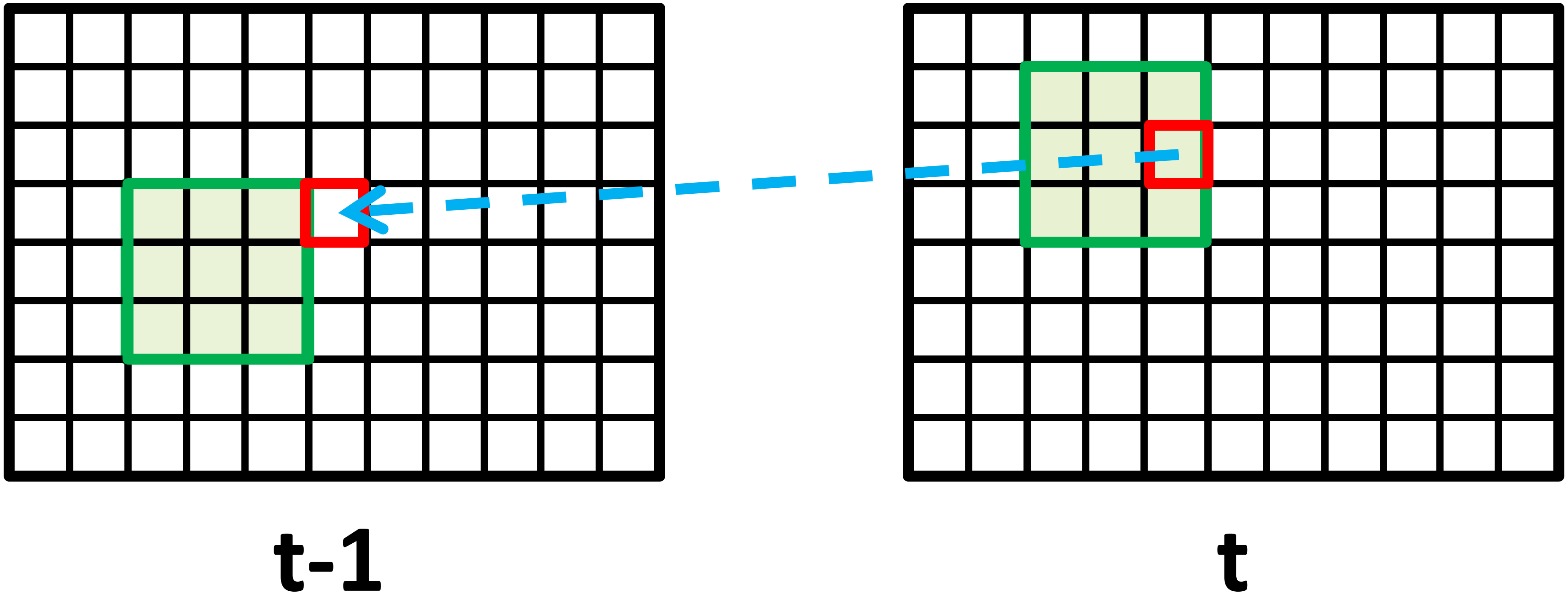
t-1



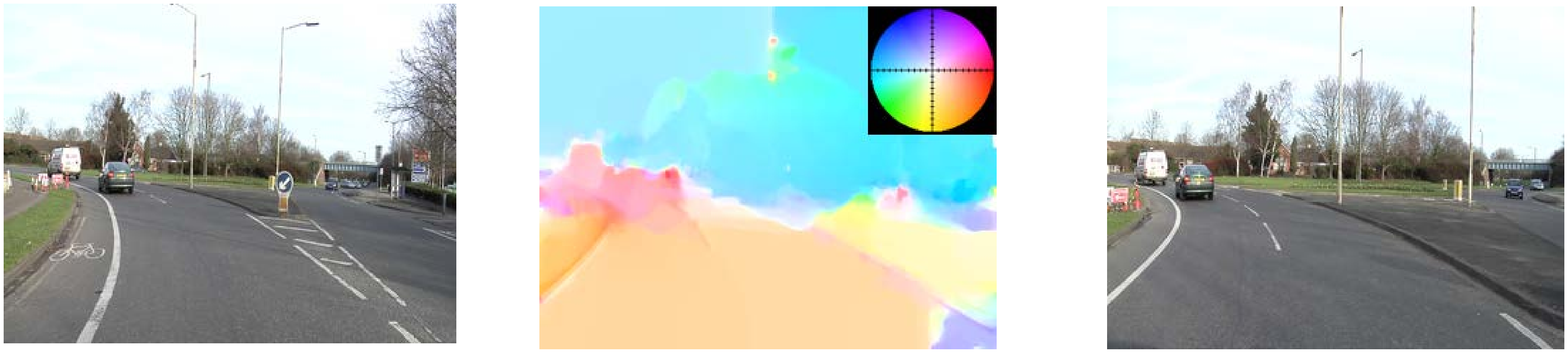
t

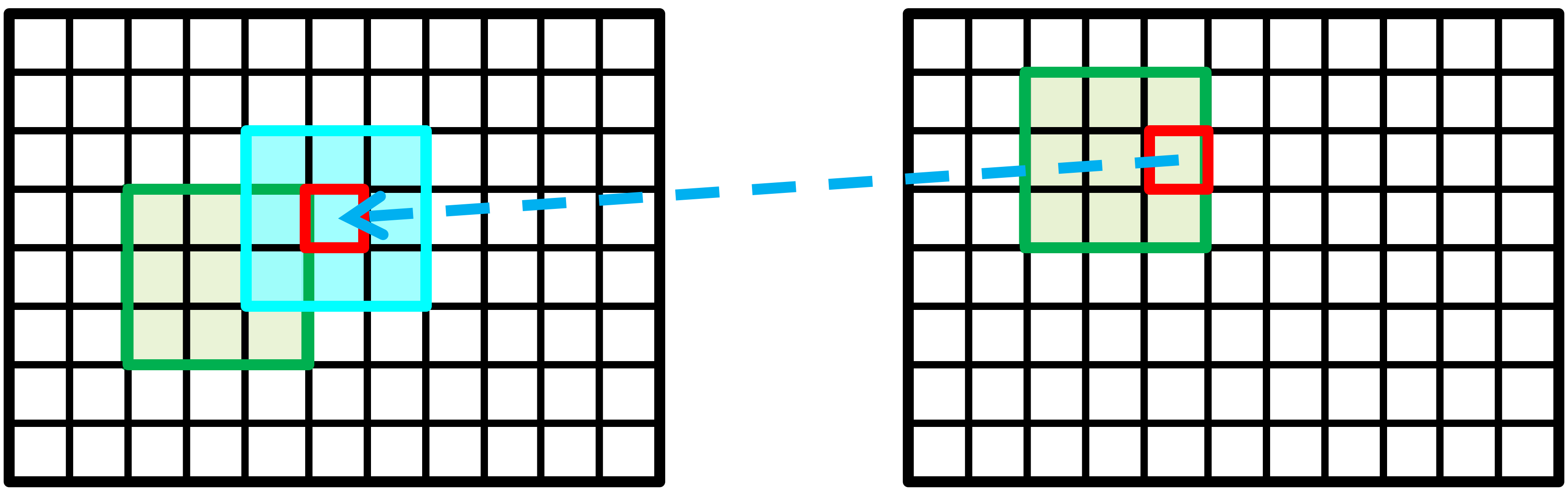
Handling No/Imperfect Correspondences





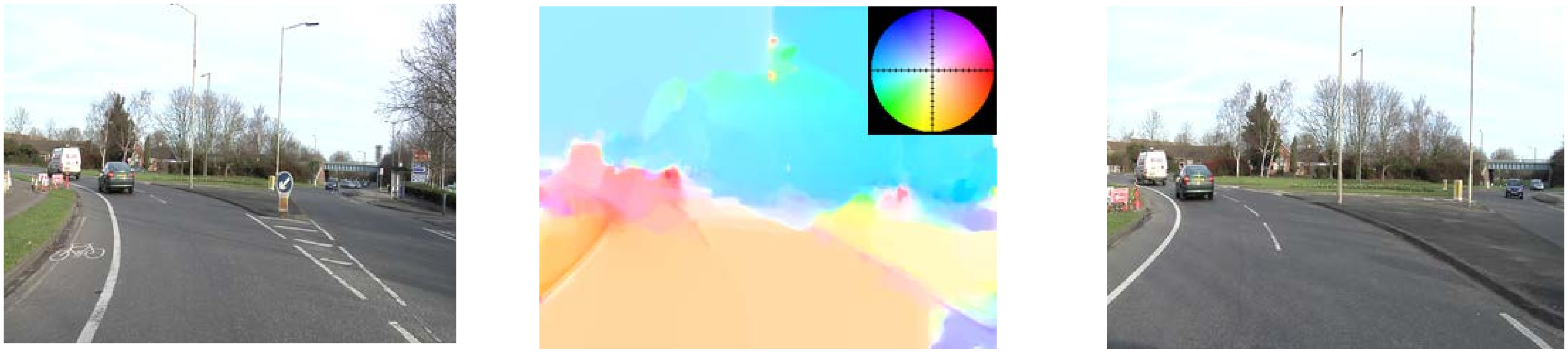
Handling No/Imperfect Correspondences

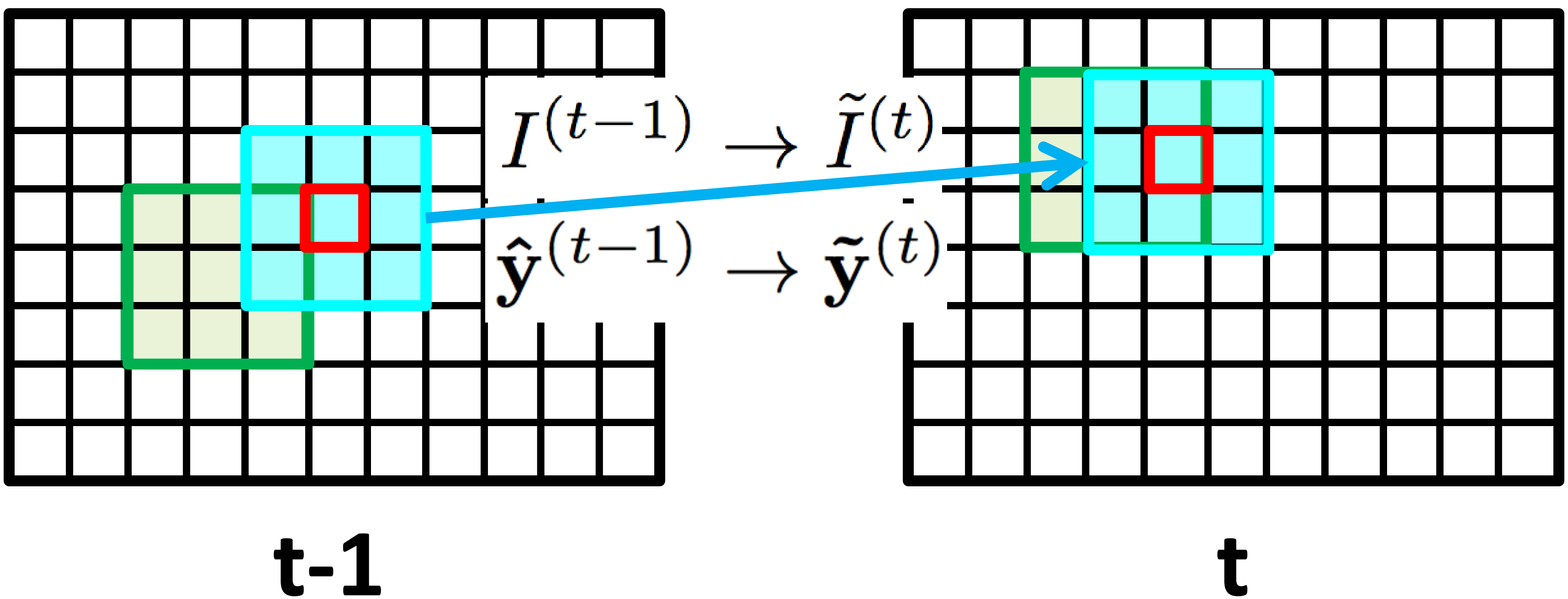




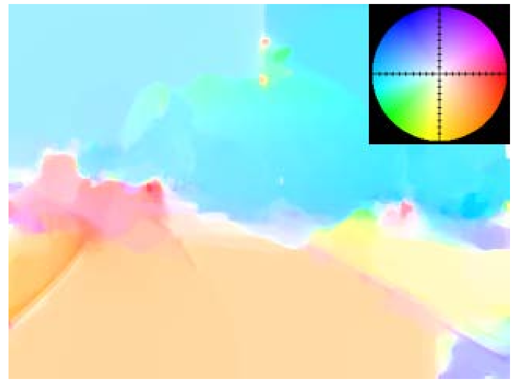
t-1 t

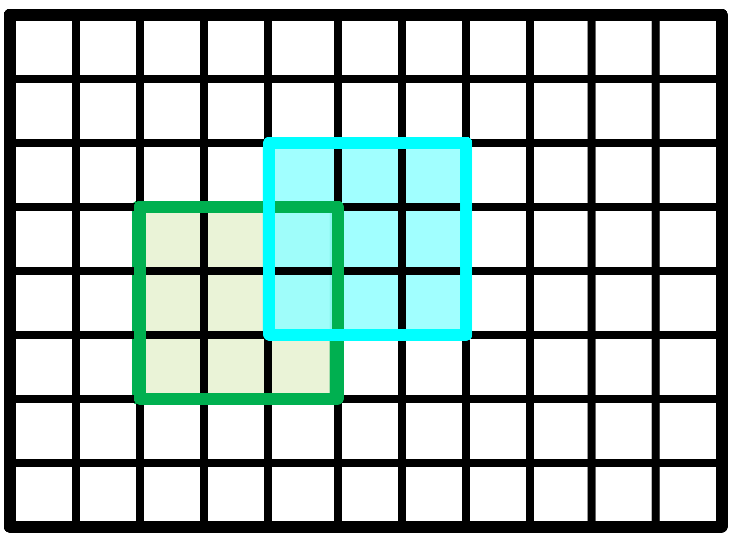
Handling No/Imperfect Correspondences



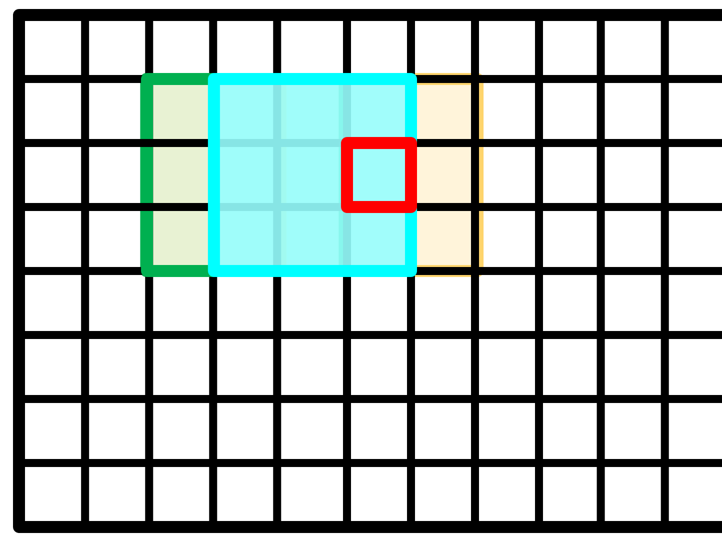


Handling No/Imperfect Correspondences



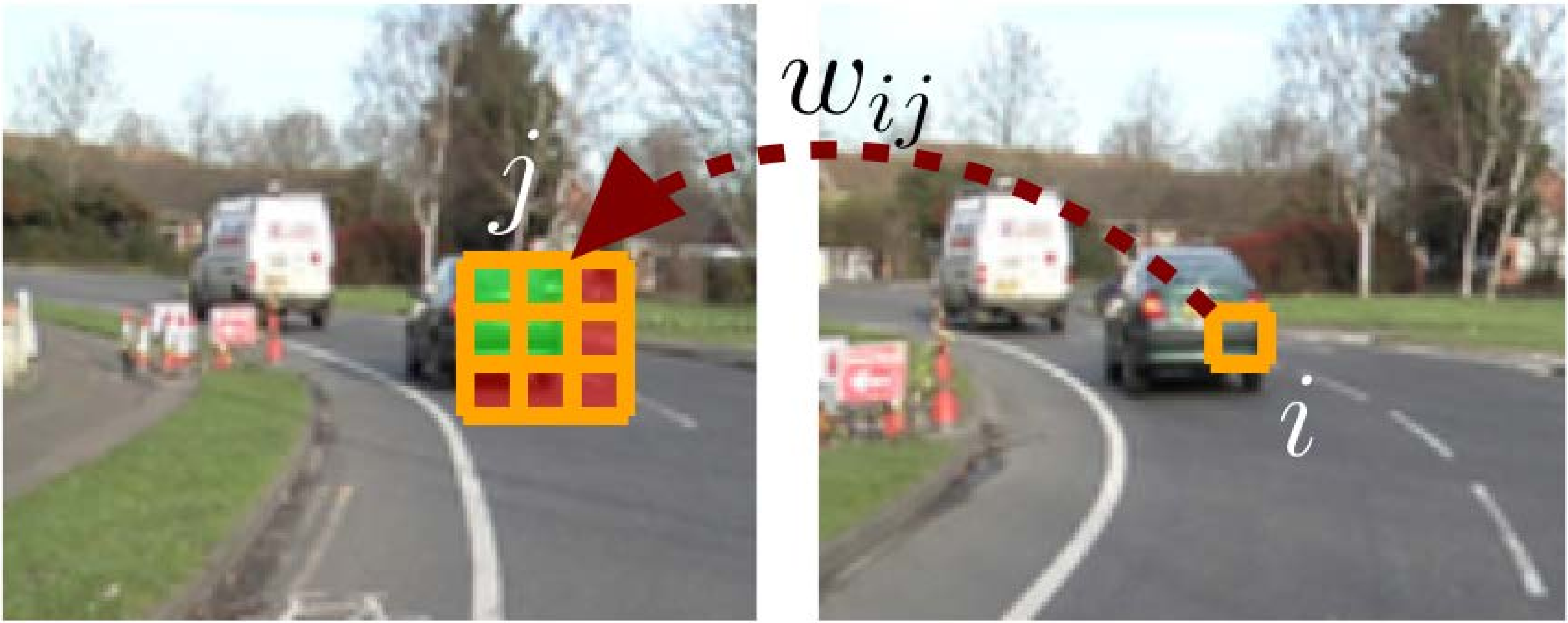


t-1



t

Causal/Running Update



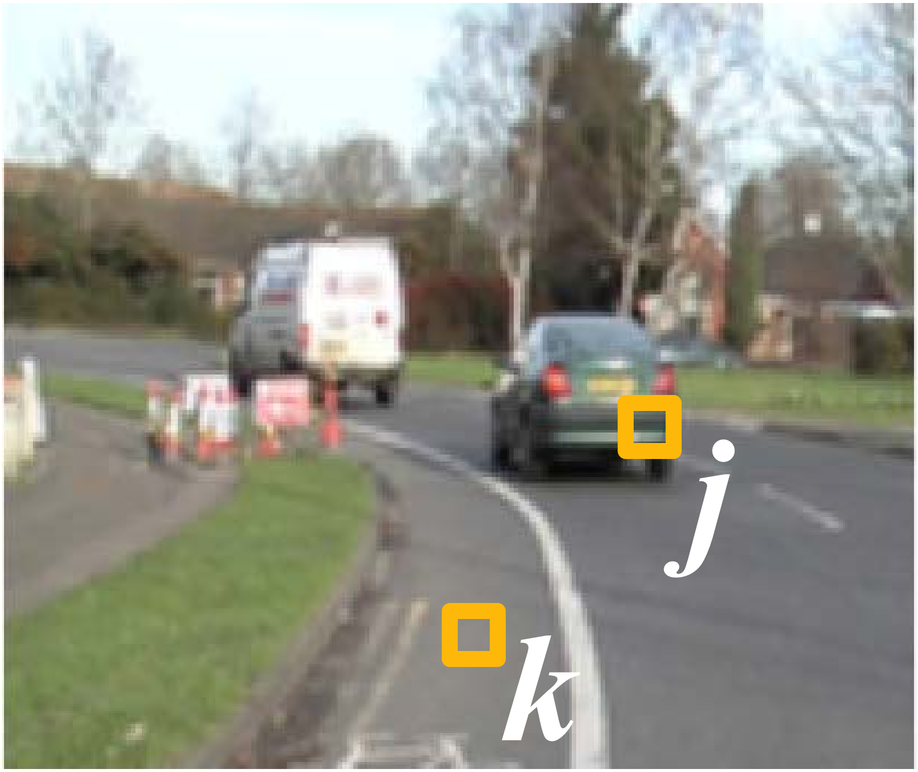
t-1 t

$$\hat{\mathbf{y}}_i^{(t)} \propto \mathbf{y}_i^{(t)} + \sum_{j \in N_i^{(t)}} \boxed{w_{ij}} \tilde{\mathbf{y}}_j^{(t-1)}$$

$$w_{ij} = \exp\left(-\frac{d(\mathbf{f}_i, \mathbf{f}_j)}{\sigma^2}\right)$$

$$d(\mathbf{f}_i, \mathbf{f}_j) = (\mathbf{f}_i - \mathbf{f}_j)^T(\mathbf{f}_i - \mathbf{f}_j)$$

Metric Learning



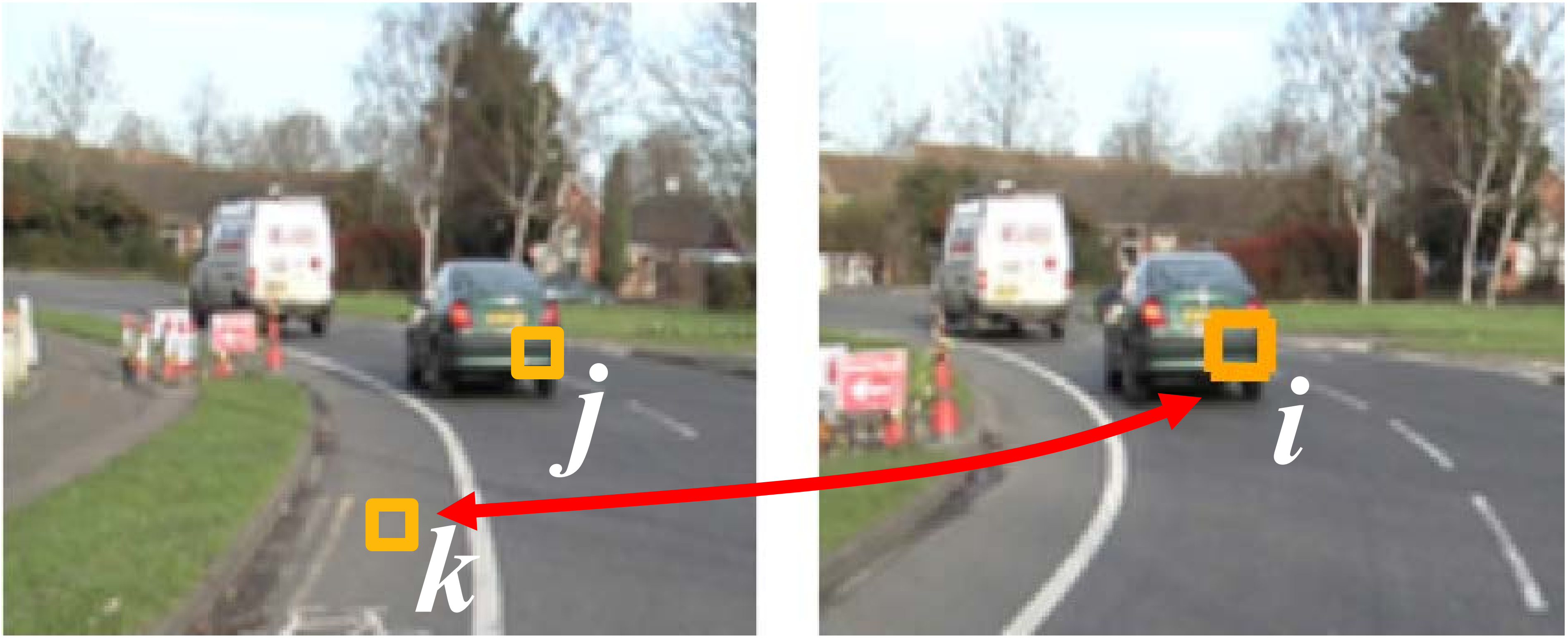
Xing 2003, Goldberger 2004, Hadsell 2006, Davis 2007, Chechik 2009, Weinberger 2009

Metric Learning

$$\boxed{d_{\mathbf{M}}(\mathbf{f}_i, \mathbf{f}_j)}$$

Xing 2003, Goldberger 2004, Hadsell 2006, Davis 2007, Chechik 2009, Weinberger 2009

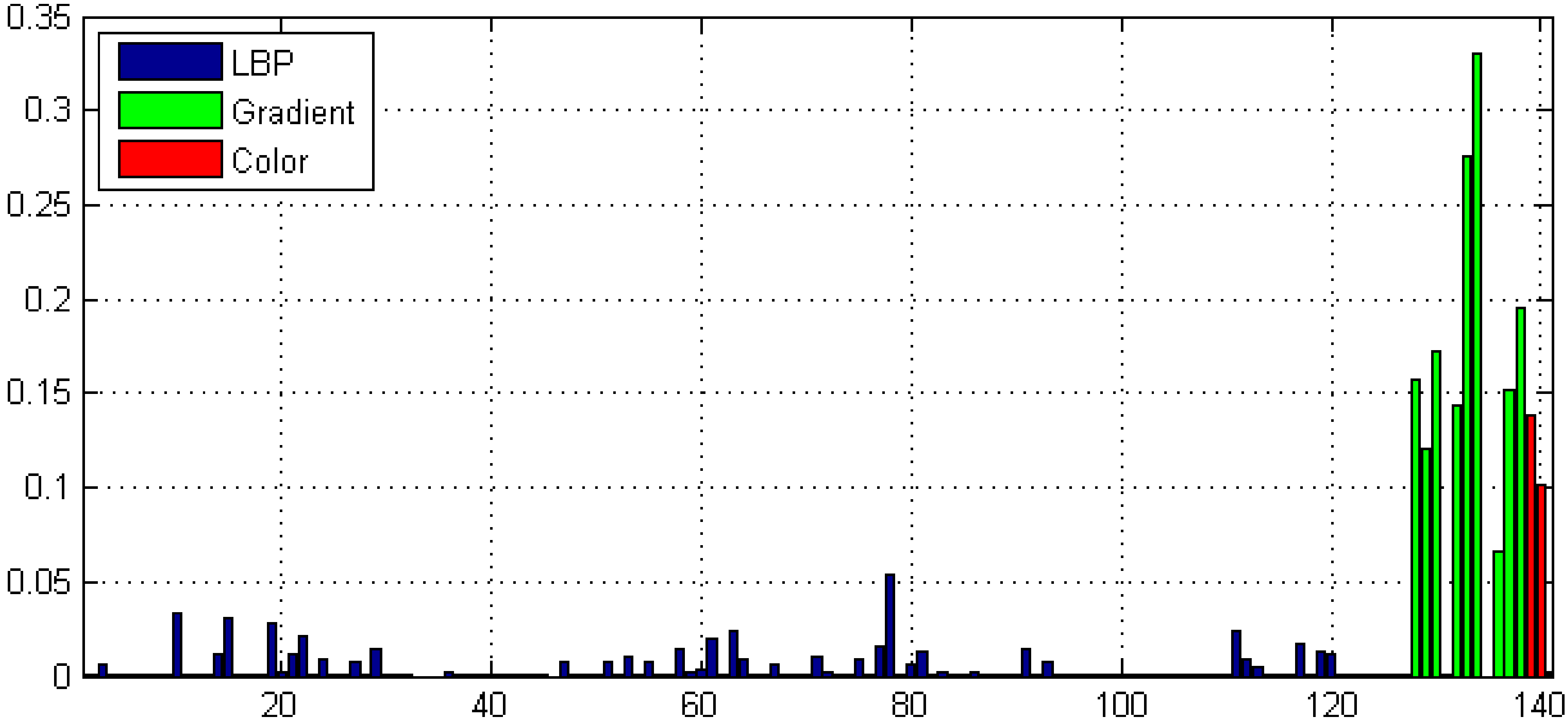
Metric Learning



$$d_{\mathbf{M}}(\mathbf{f}_i, \mathbf{f}_j) \leq \boxed{d_{\mathbf{M}}(\mathbf{f}_i, \mathbf{f}_k)} + 1$$

Xing 2003, Goldberger 2004, Hadsell 2006, Davis 2007, Chechik 2009, Weinberger 2009

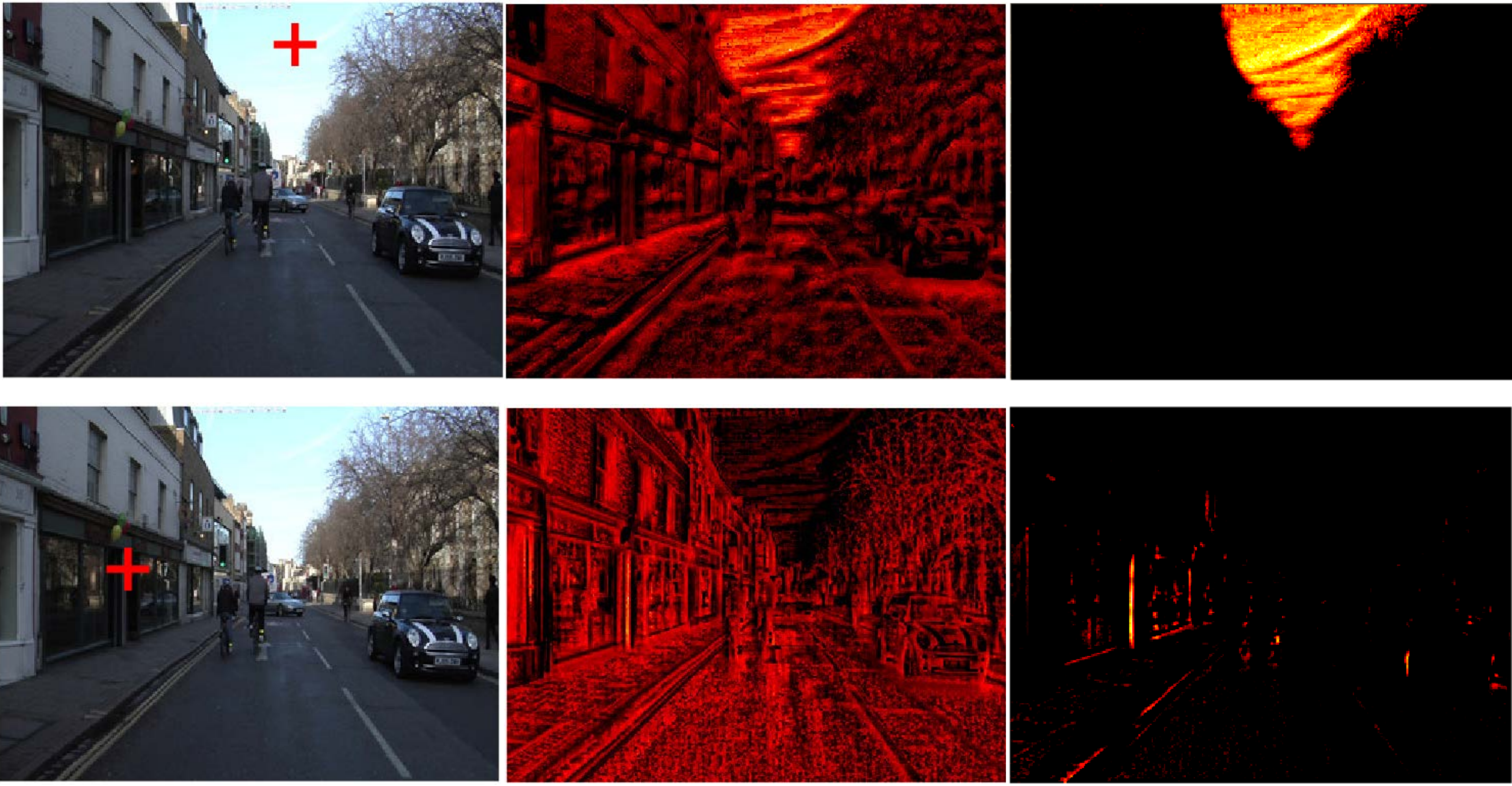
Learned Metric



$$d_{\mathbf{M}}(\mathbf{f}_i, \mathbf{f}_j) = (\mathbf{f}_i - \mathbf{f}_j)^T \mathbf{M} (\mathbf{f}_i - \mathbf{f}_j)$$

Euclidean vs. Learned

$$w_{ij} = \exp\left(-\frac{d(\mathbf{f}_i, \mathbf{f}_j)}{\sigma^2}\right)$$

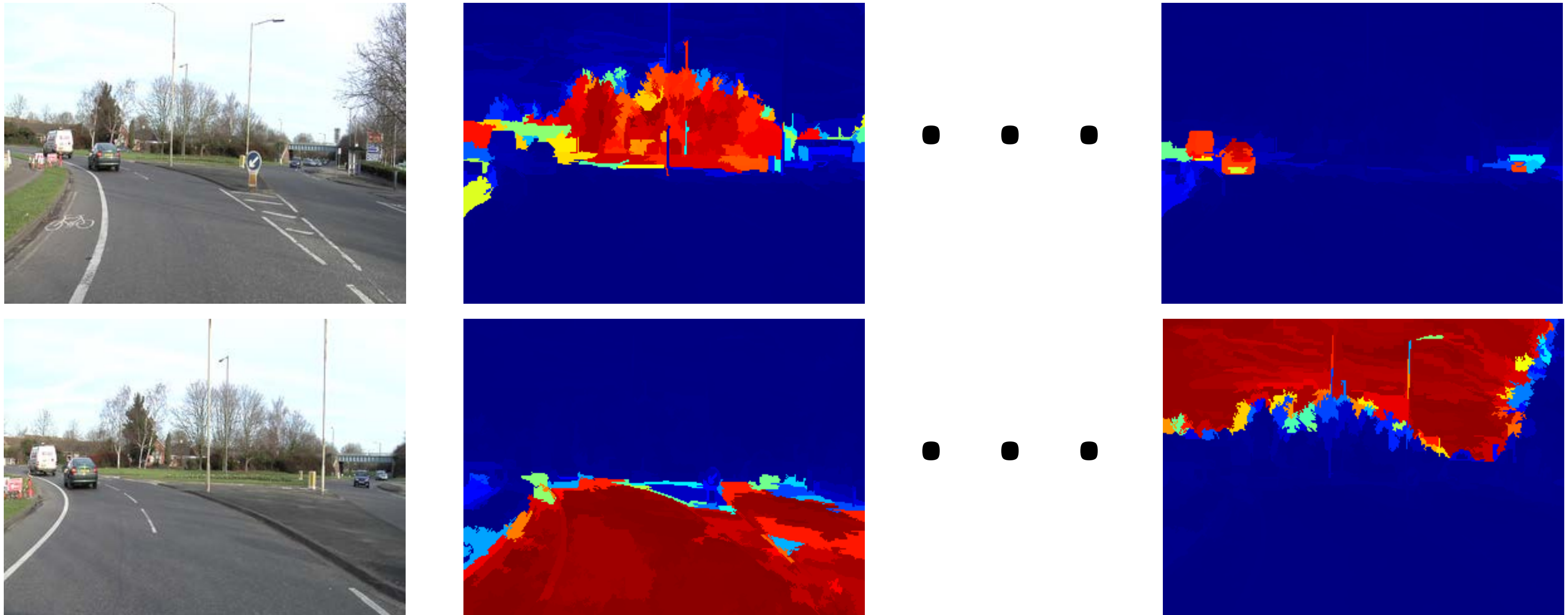


Summary

- Input: images and per-pixel probabilities

t-1

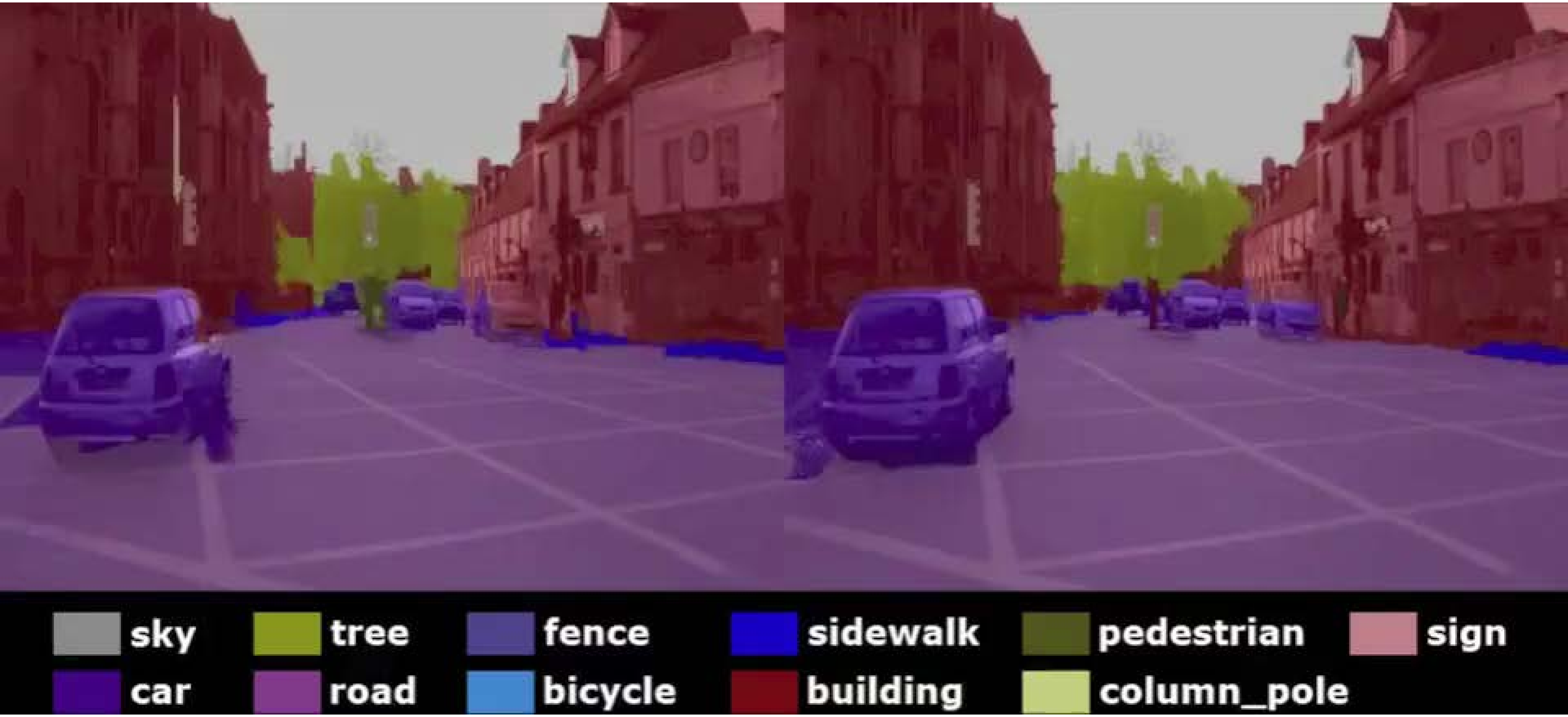
t



- Compute and apply optical flow (0.02 s)
- Compute pixel similarities and average (0.75 s)

$$\hat{\mathbf{y}}_i^{(t)} \propto \mathbf{y}_i^{(t)} + \sum_{j \in N_i^{(t)}} w_{ij} \tilde{\mathbf{y}}_j^{(t-1)}$$

CamVid Dataset (8 labels)



Per-frame
Munoz ECCV 2010

Temporal
This Work

NYU Scenes Dataset (11 labels)

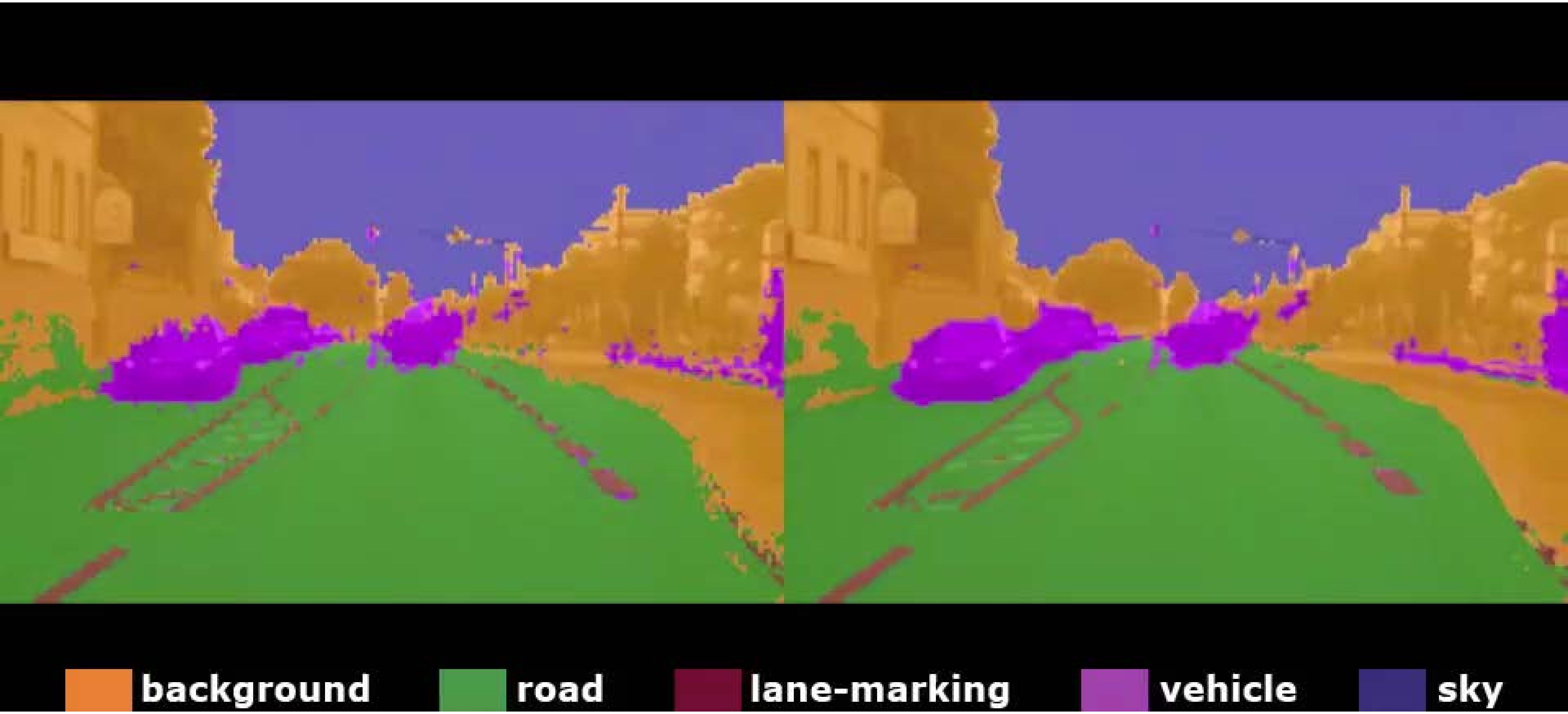
Per-frame

Farabet ICML 2012

Temporal

This Work

MPI-VehicleScenes Dataset (5 labels)



Per-frame
Wojek ECCV 2008*

Temporal
This Work

Overall Pixel Accuracy

Dataset	Per-frame	Temporal
CamVid-05VD	84.60	**86.85**
CamVid-16E5	87.37	**88.84**
NYUScenes	71.11	**75.31**
MPI-Vehicle	93.10	**93.55**

“Relative” Pixel Accuracy

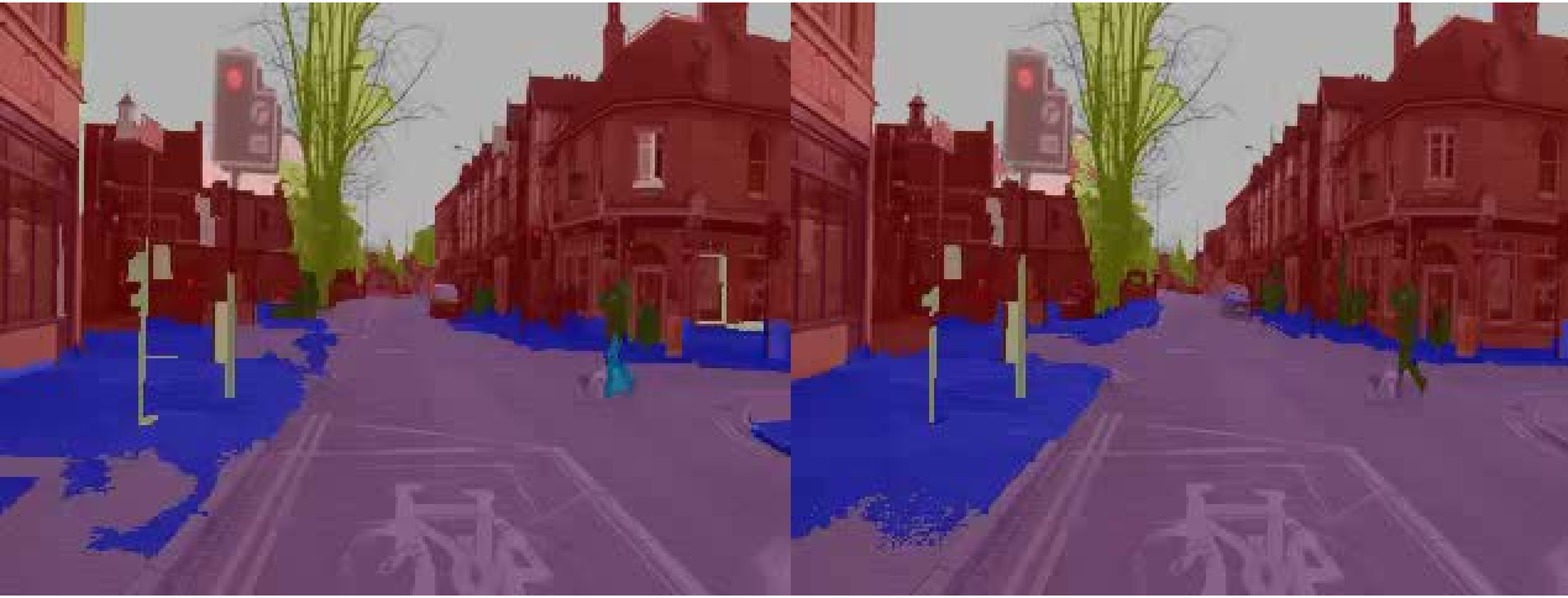
- Evaluate on pixels where **Per-frame** and **Temporal** predictions **disagree**

Dataset	Per-frame	Temporal
CamVid-05VD	26.1	**59.1**
CamVid-16E5	28.6	**53.0**
NYUScenes	20.9	**50.0**
MPI-Vehicle	40.7	**53.3**

Conclusion

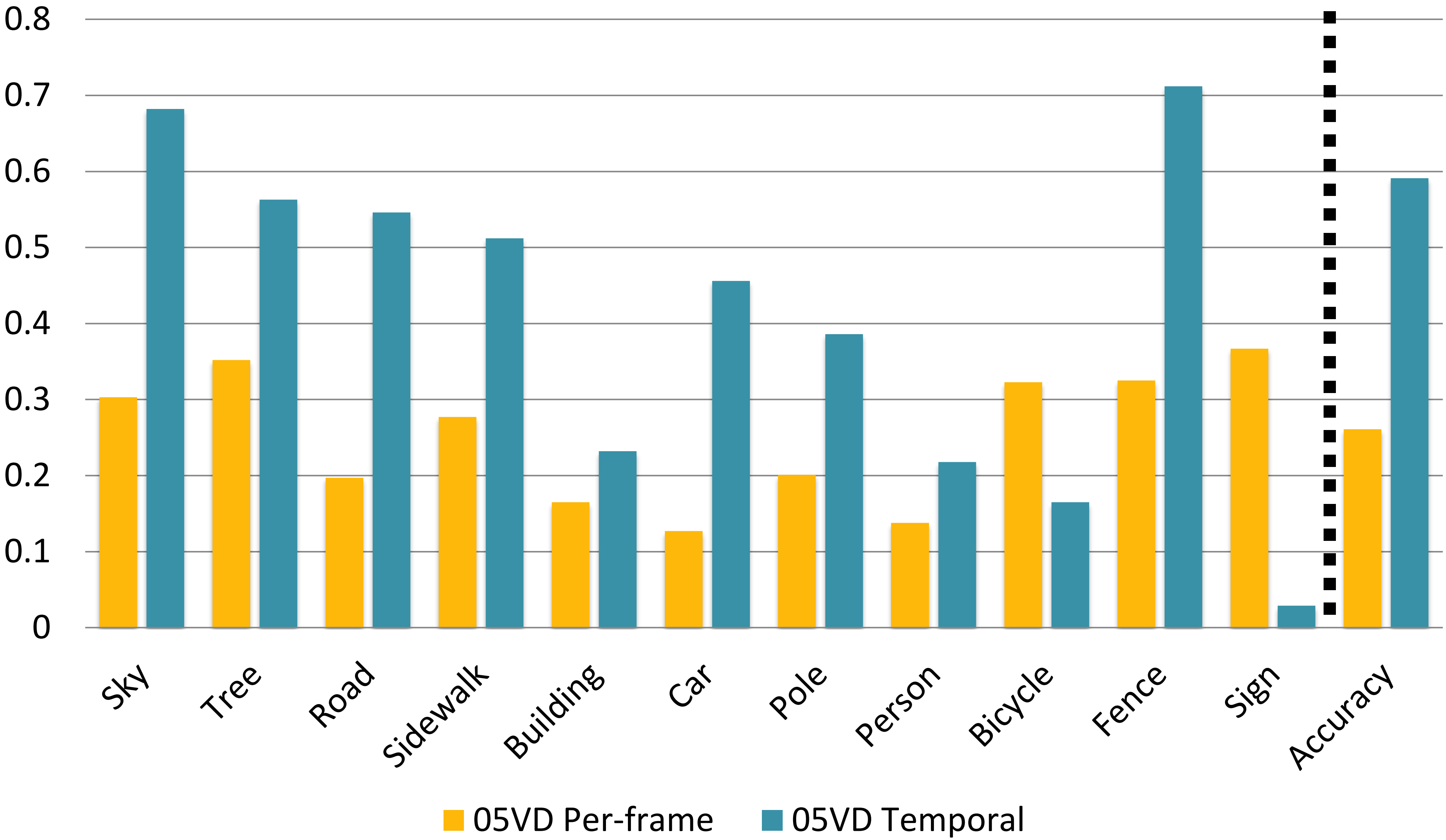
- Pros:
  - Efficient temporal consistency
  - Use your favorite scene parsing output
  - Easy to implement
- Cons:
  - Can't fix major errors from the scene parser
  - Can over-smooth small objects

Thanks!

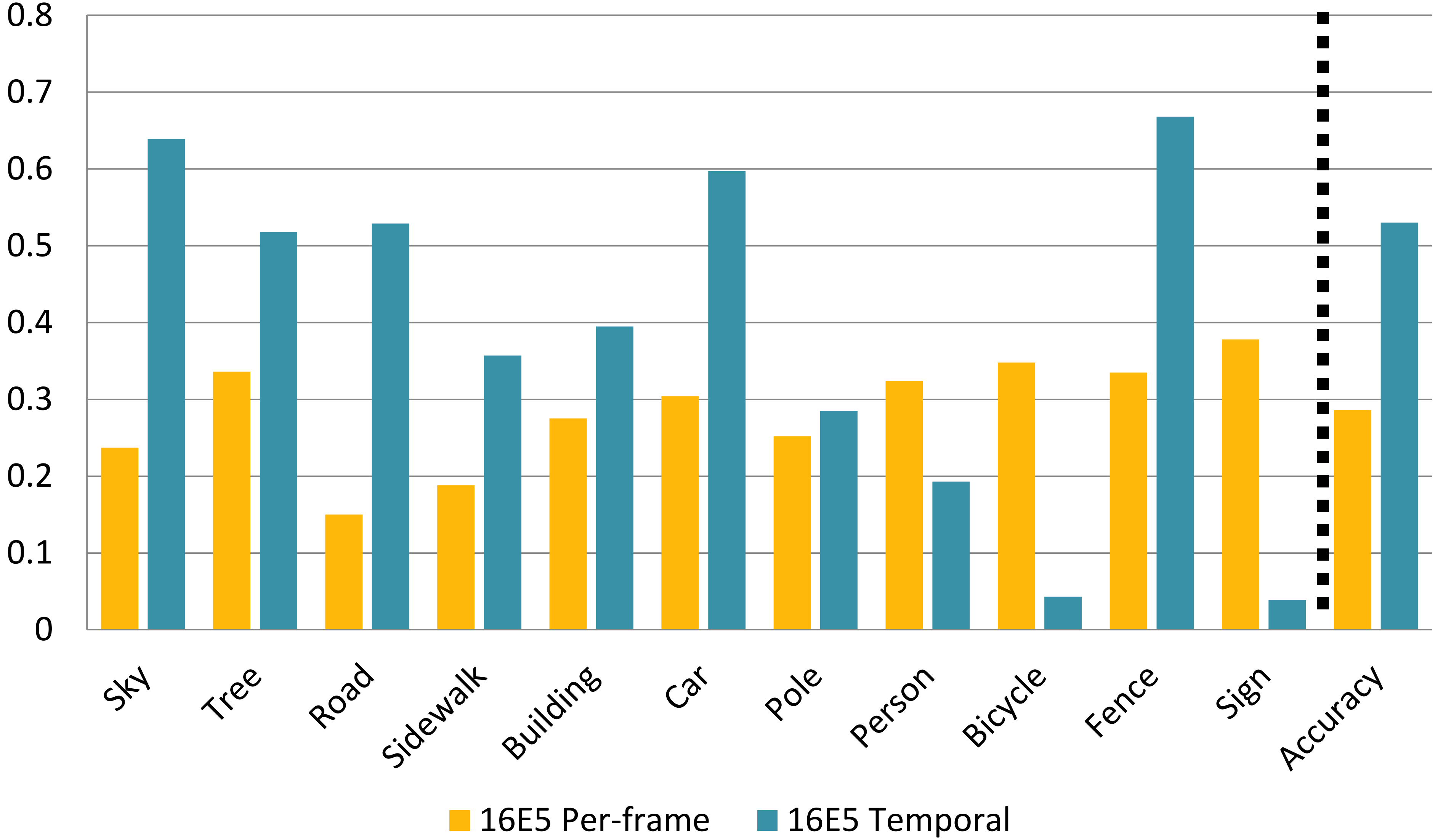


BACKUPS

CamVid (Sequence 05VD)



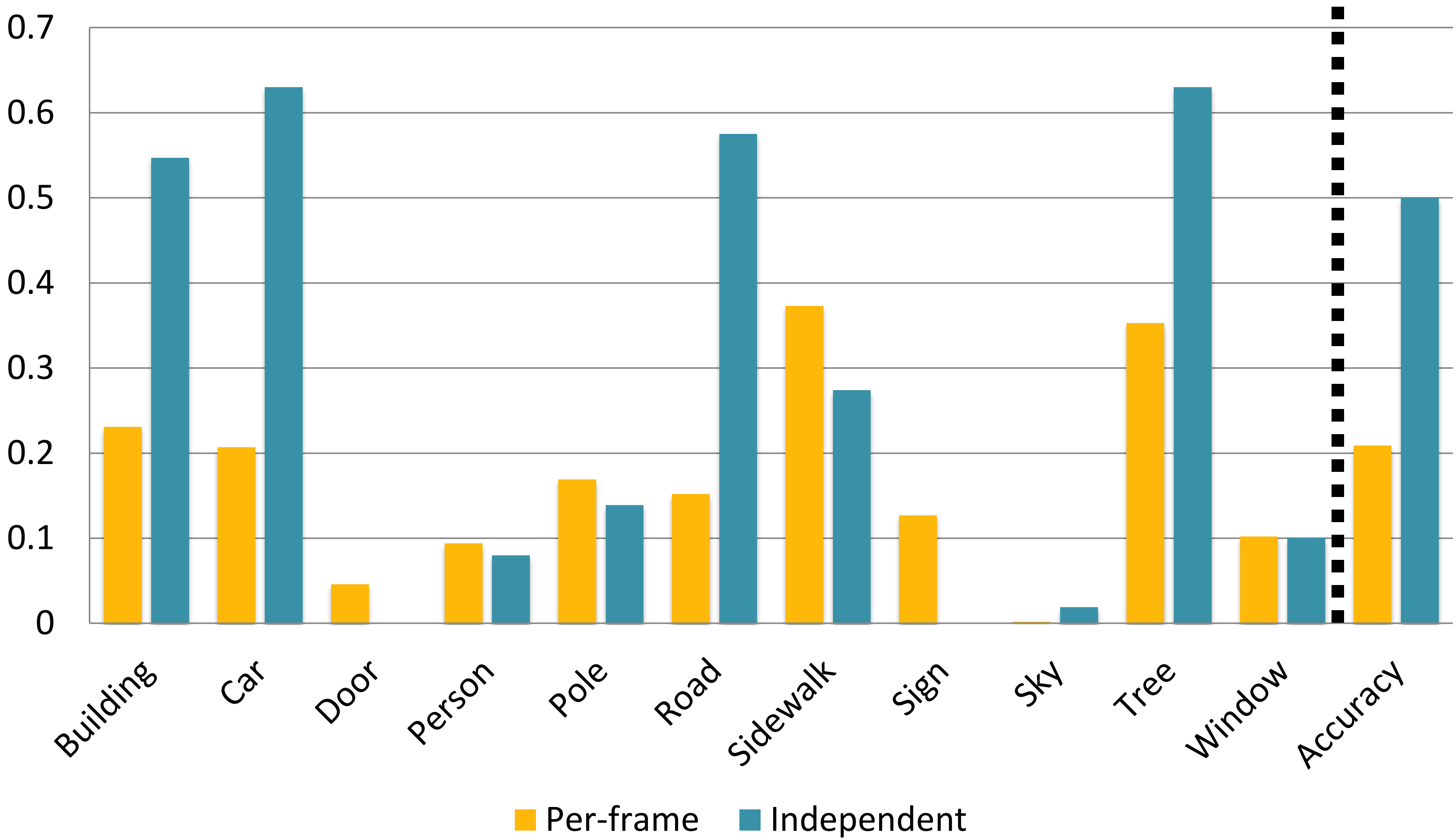
CamVid (Sequence 16E5)



CamVid

	05VD		16E5	
Class	**Per-frame**	**Temporal**	**Per-frame**	**Temporal**
Sky	.303	**.682**	.237	**.639**
Tree	.352	**.563**	.336	**.518**
Road	.197	**.546**	.150	**.529**
Sidewalk	.277	**.512**	.188	**.357**
Building	.165	**.232**	.275	**.395**
Car	.127	**.456**	.304	**.597**
Pole	.201	**.386**	.252	**.285**
Person	.138	**.218**	**.324**	.193
Bicycle	**.323**	.165	**.348**	.043
Fence	.325	**.712**	.335	**.668**
Sign	**.367**	.029	**.378**	.039
Accuracy	.261	**.591**	.286	**.530**

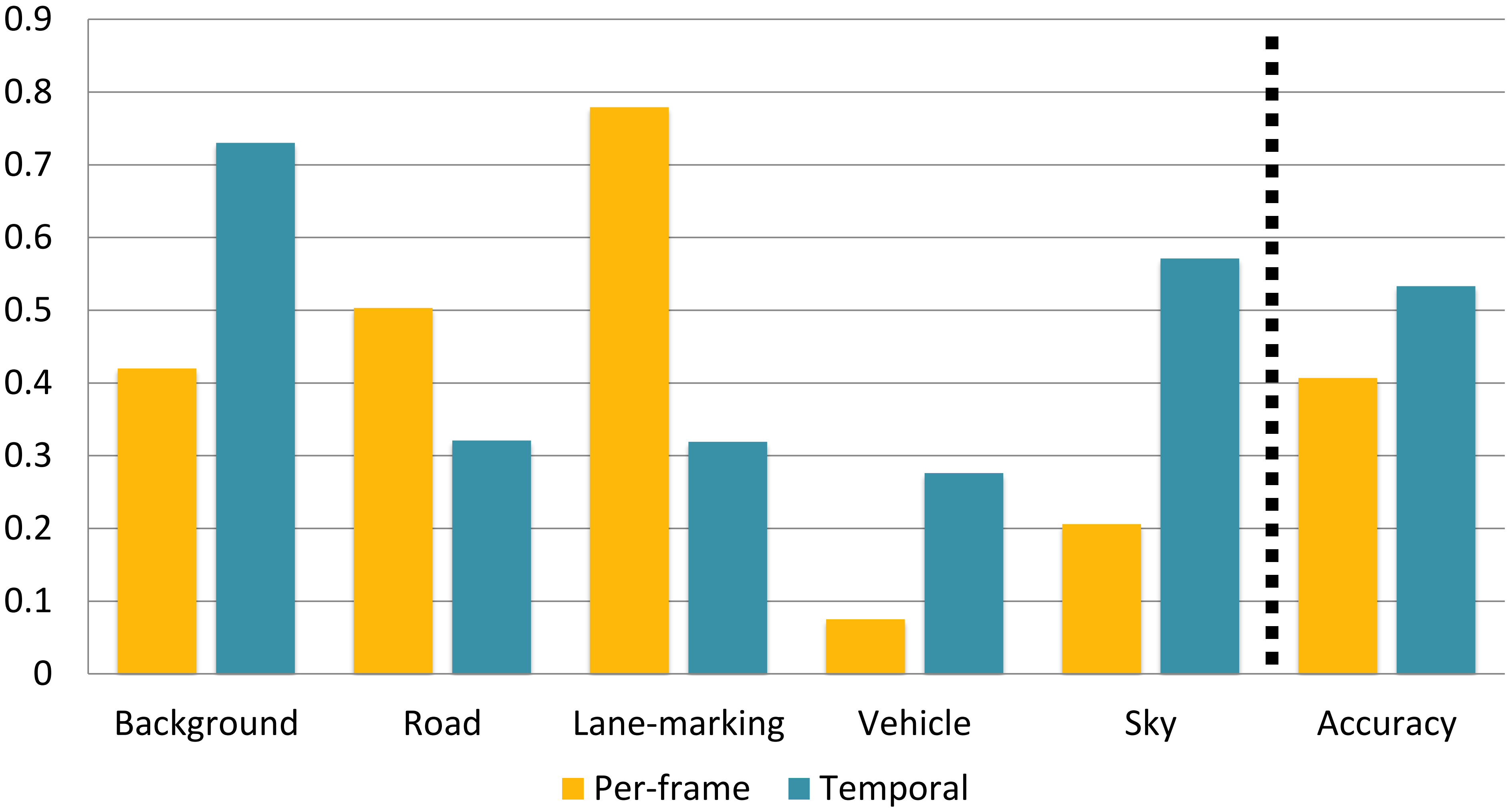
NYU Scenes Dataset (11 labels)



NYU Scenes Dataset (11 labels)

Class	Per-frame	Independent
Building	.231	**.547**
Car	.207	**.630**
Door	**.046**	.000
Person	.094	.080
Pole	.169	.139
Road	.152	**.575**
Sidewalk	**.373**	.274
Sign	**.127**	.000
Sky	.002	.019
Tree	.353	**.630**
Window	.102	.101
Accuracy	.209	**.500**

MPI-VehicleScenes Dataset (5 labels)



MPI-VehicleScenes Dataset (5 labels)

Class	Per-frame	Temporal
Background	.420	**.730**
Road	**.503**	.321
Lane-marking	**.779**	.319
Vehicle	.075	**.276**
Sky	.206	**.571**
Accuracy	.407	**.533**