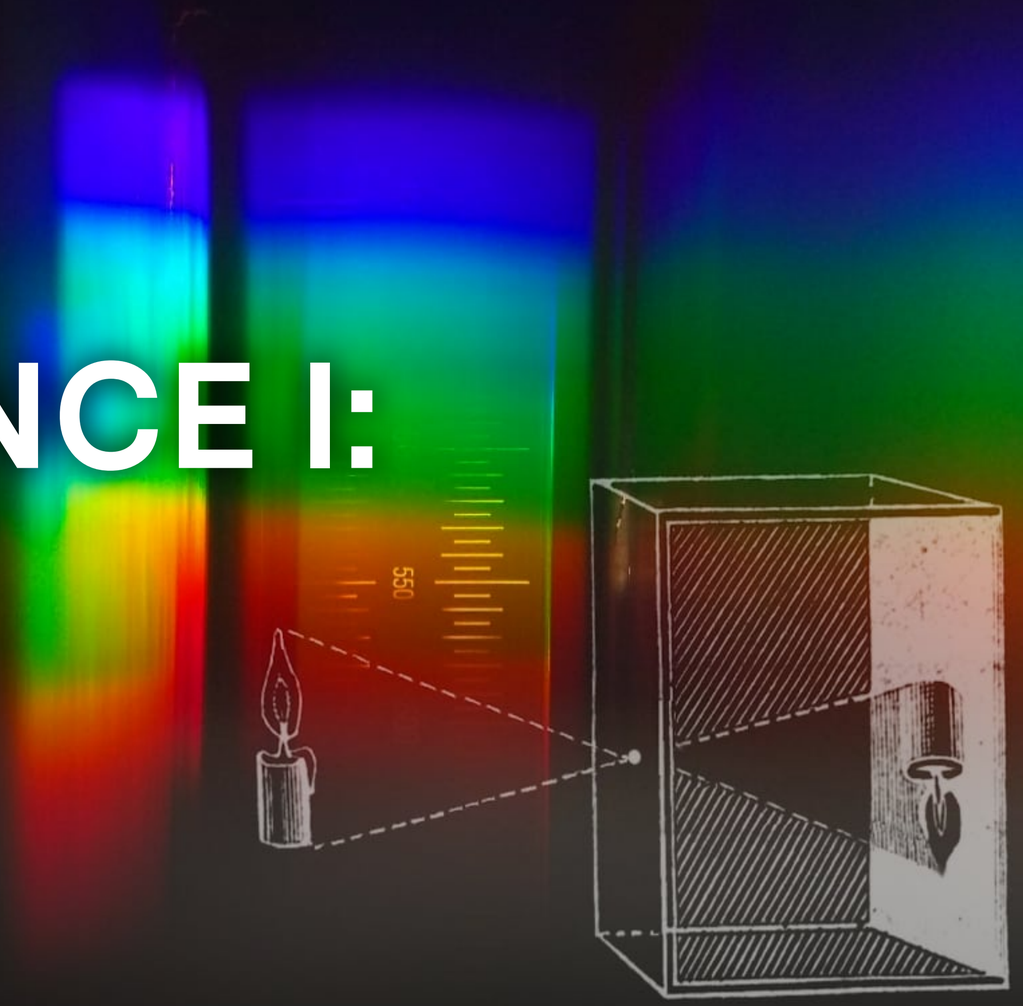
INVERSE GRAPHICS

MODULE 1:
Fundamentals of Image Formation

CORRESPONDENCE I: OPTICAL FLOW



6.S980

Credits

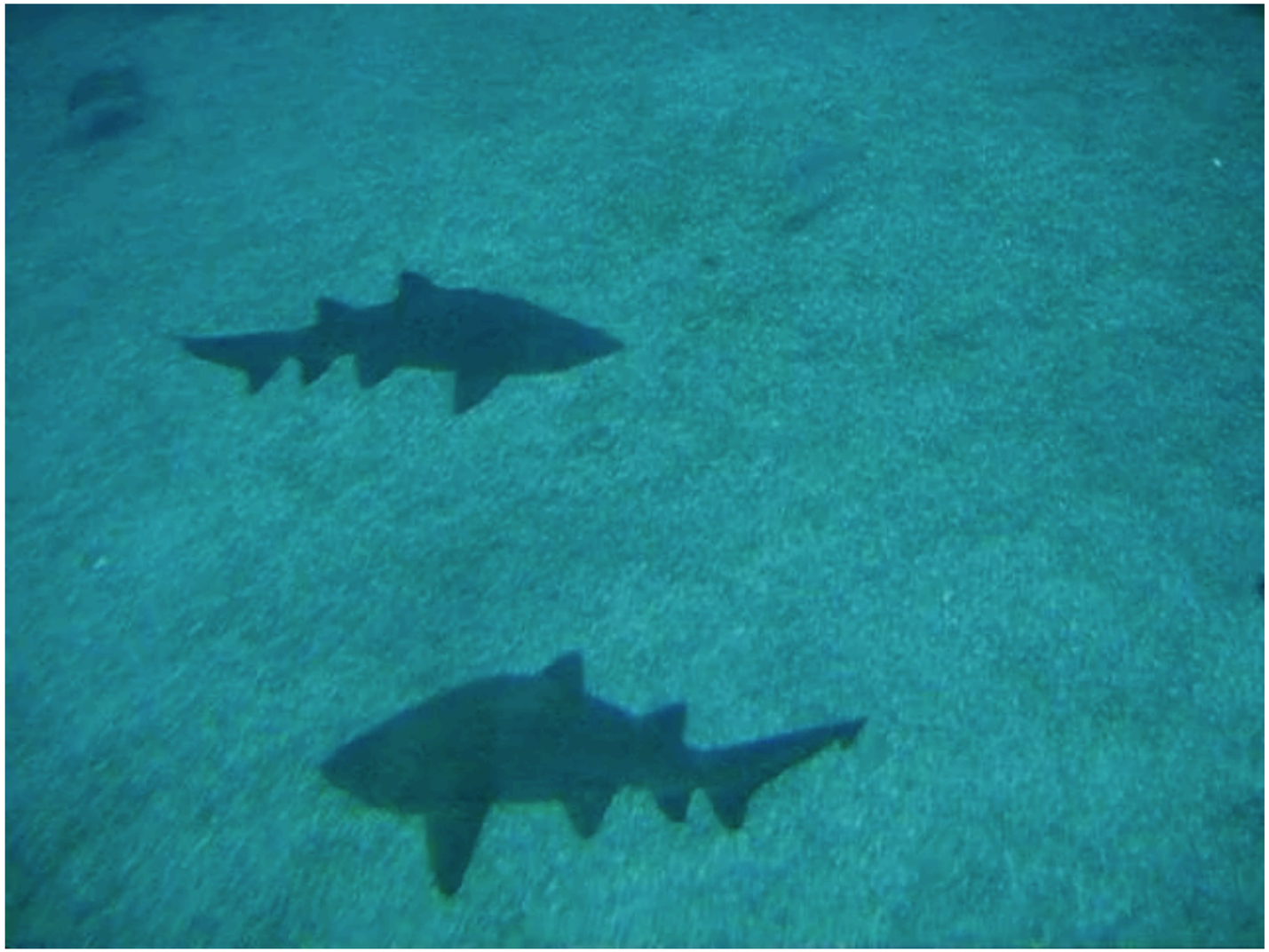
Base of this lecture: Deqing Sun & Charles Herrmann

Some slides, images, and videos from
- Dr. Ce Liu@Microsoft
- Dr. Huaizu Jiang@Northeastern
- Dr. David Fouhey@UMich
- Dr. Justin Johnson@UMich
- Dr. Svetlana Lazebnik@UIUC
- Dr. Shree K. Nayar@Columbia
- Dr. Jia Deng@Princeton
- Dr. Ming-Hsuan Yang@UC Merced
- Dr. Rick Szeliski's book
- Book by Antonio, Phillip, and Bill
- Dr. Christian Rupprecht@Oxford

Suggestions from Dr. Bill Freeman, Dr. Rick Szeliski, Dr. Noah Snavely and Dr. Junhwa Hur

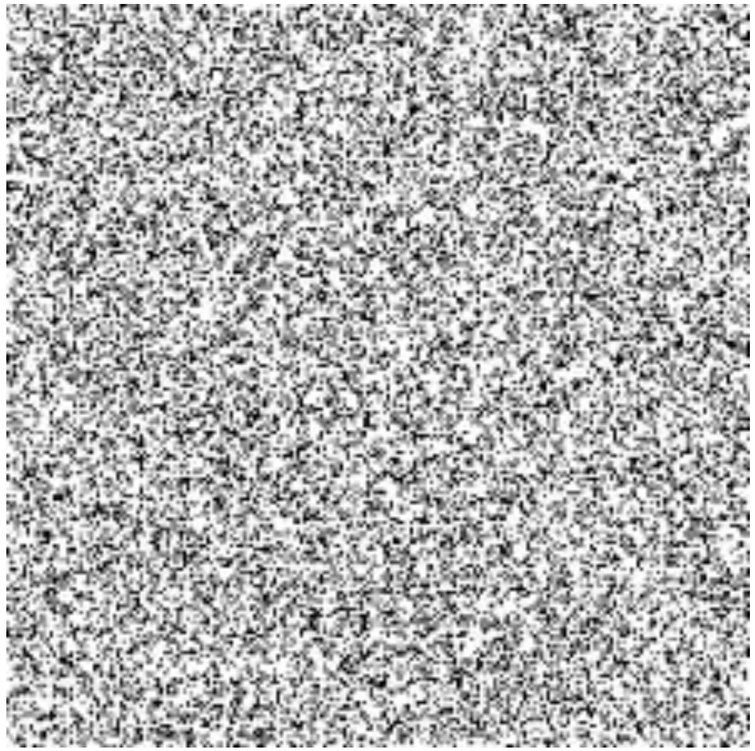
We live in a dynamic world

Perceiving, understanding and predicting motion is an important part of our daily lives



Motion and perceptual organization

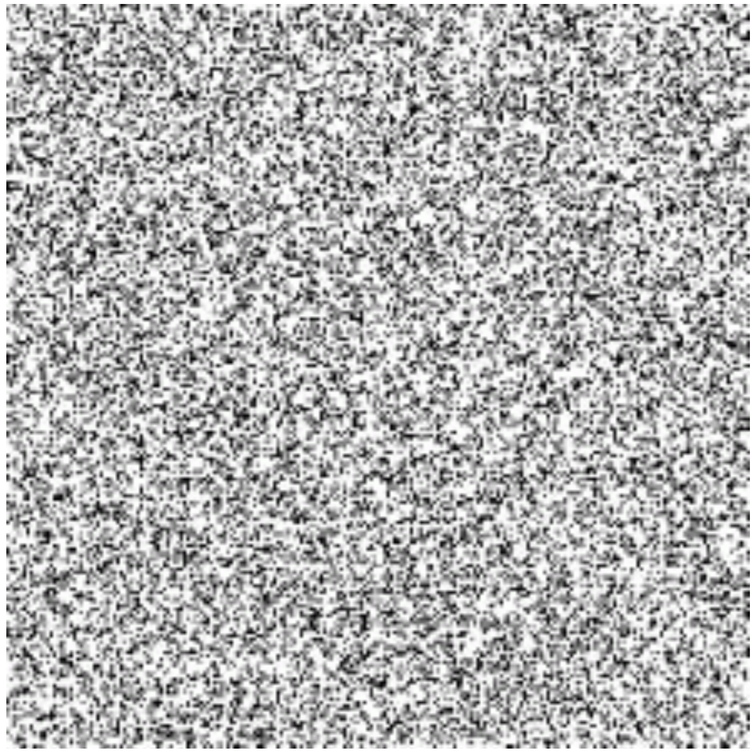
Sometimes motion is the only cue



Slide Credit: S. Lazebnik, but idea of random dot sterogram is due to B. Julesz

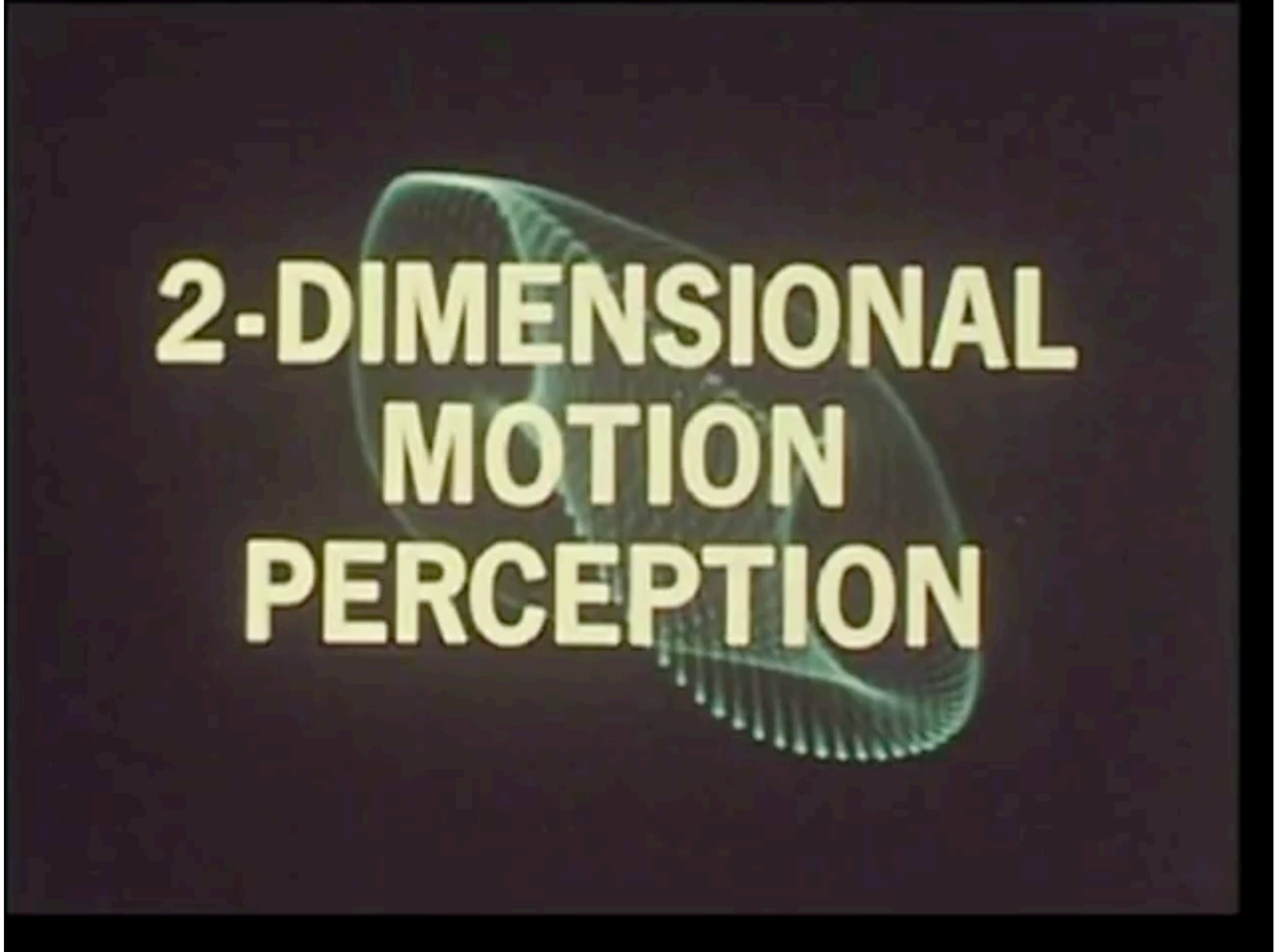
Motion and perceptual organization

Sometimes motion is the only cue!



Slide Credit: S. Lazebnik, but idea of random dot sterogram is due to B. Julesz

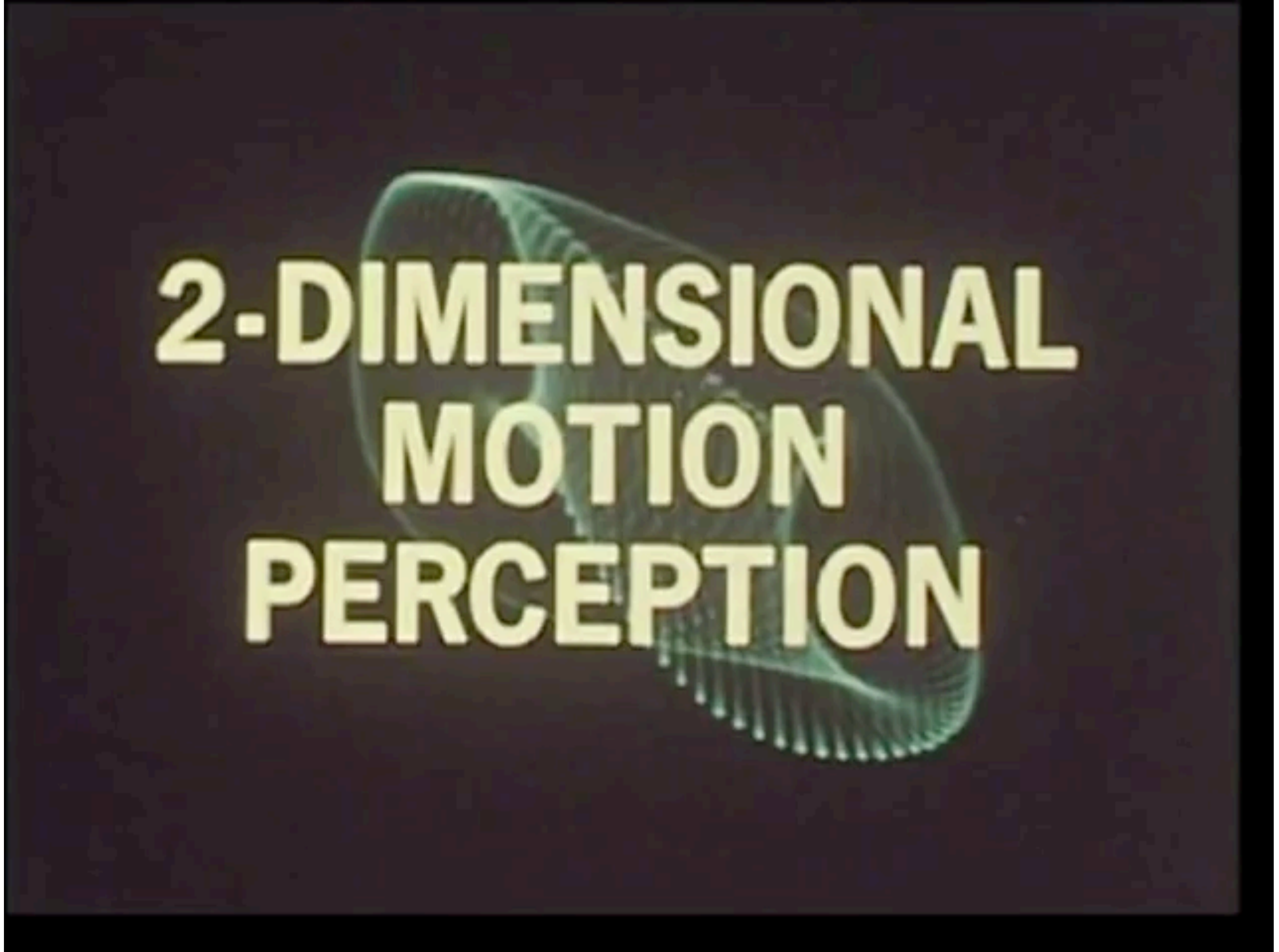
Slide Credit:
C. Rupprecht



Johansson, “Visual perception of biological motion and a model for its analysis.” 1973

source

Slide Credit:
C. Rupprecht

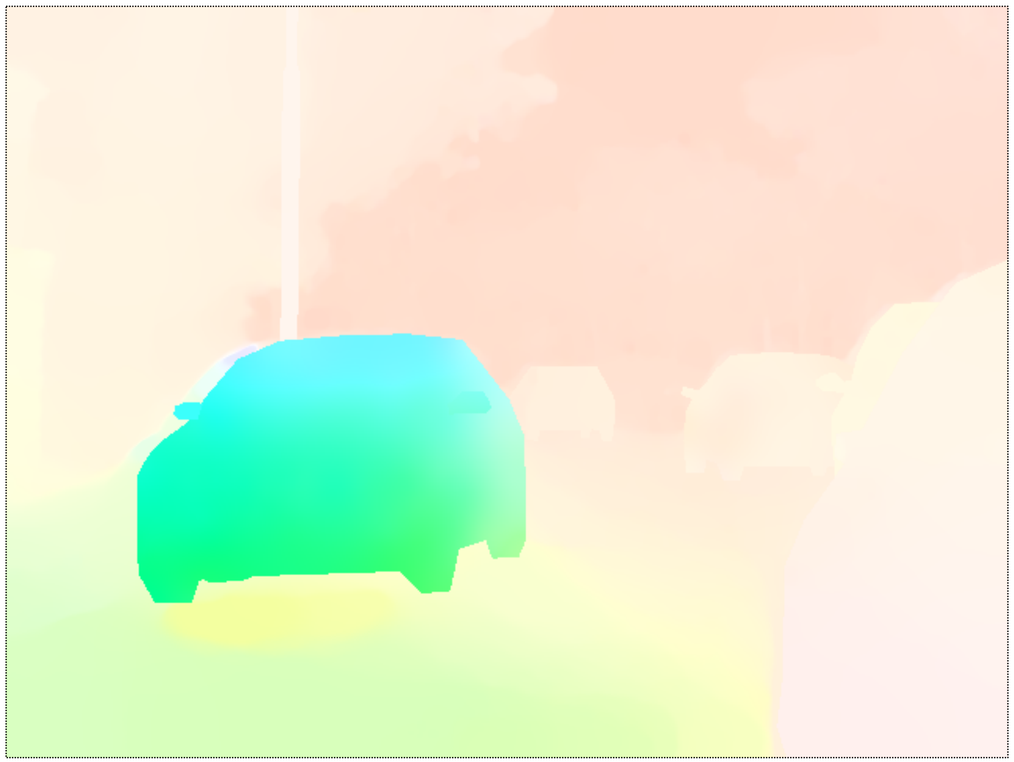


Johansson, “Visual perception of biological motion and a model for its analysis.” 1973

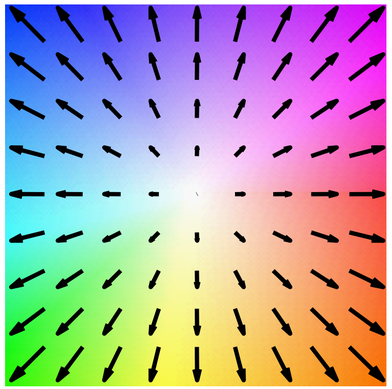
source

Optical flow: 2D motion of every pixel

Input [Liu *et al.* CVPR'08]



Optical flow (2D motion vector)



Color key
[Baker *et al.* IJCV'11]

Slide credit: Deqing Sun

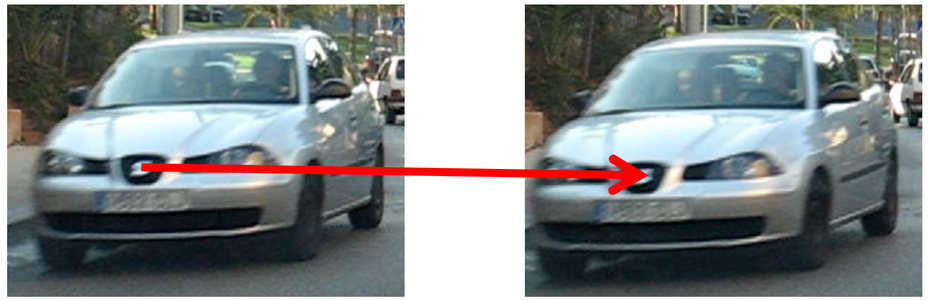
Fundamental assumption: Brightness constancy

[Horn & Schunck AI'81]

Assume a pixel at (x, y, t) with intensity $I(x, y, t)$ has moved by $(\Delta x, \Delta y)$ in space during a timestep Δt.

Color constancy assumption (CCA):

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$$



First image (t)

Second image (t+1)

Berthold Horn

Determining Optical Flow

Berthold K.P. Horn and Brian G. Schunck
Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

ABSTRACT
Optical flow cannot be computed locally, since only one independent measurement is available from the image sequence at a point, while the flow velocity has two components. A second constraint is needed. A method for finding the optical flow pattern is presented which assumes that the apparent velocity of the brightness pattern varies smoothly almost everywhere in the image. An iterative implementation is shown which successfully computes the optical flow for a number of synthetic image sequences. The algorithm is robust in that it can handle image sequences that are quantized rather coarsely in space and time. It is also insensitive to quantization of brightness levels and additive noise. Examples are included where the assumption of smoothness is violated at singular points or along lines in the image.

1. Introduction

Optical flow is the distribution of apparent velocities of movement of brightness patterns in an image. Optical flow can arise from relative motion of objects and the viewer [6, 7]. Consequently, optical flow can give important information about the spatial arrangement of the objects viewed and the rate of change of this arrangement [8]. Discontinuities in the optical flow can help in segmenting images into regions that correspond to different objects [27]. Attempts have been made to perform such segmentation using differences between successive image frames [15, 16, 17, 20, 25]. Several papers address the problem of recovering the motions of objects relative to the viewer from the optical flow [10, 18, 19, 21, 29]. Some recent papers provide a clear exposition of this enterprise [30, 31]. The mathematics can be made rather difficult, by the way, by choosing an inconvenient coordinate system. In some cases information about the shape of an object may also be recovered [3, 18, 19].

These papers begin by assuming that the optical flow has already been determined. Although some reference has been made to schemes for comput-

Artificial Intelligence **17** (1981) 185–203
0004–3702/81/0000–0000/$02.50

Horn & Schunck Optical Flow

Taylor expansion:

$$I(x+\Delta x, y+\Delta y, t+\Delta t) \approx I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t$$

With CCA:

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \quad \text{or} \quad \frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t} = 0$$

Using $\nabla I = \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right]^T$ and $\mu = \left[\frac{\Delta x}{\Delta t}, \frac{\Delta y}{\Delta t}\right]^T$

yields the **motion constraint equation:**

$$\nabla \mathbf{I}\mu = -\frac{\delta I}{\delta t}$$

Thinking About Horn & Schunck Optical Flow

$$\nabla \mathbf{I}\mu = -\frac{\delta I}{\delta t}$$

To use for optical flow estimation, put into a loss function:

$$E = \min_{\mu} \left[(\nabla^T \mathbf{I}_1 \mu - (\mathbf{I}_2 - \mathbf{I}_1))^2 \right]$$

It’s convex (can be solved via gradient descent or closed-form)!

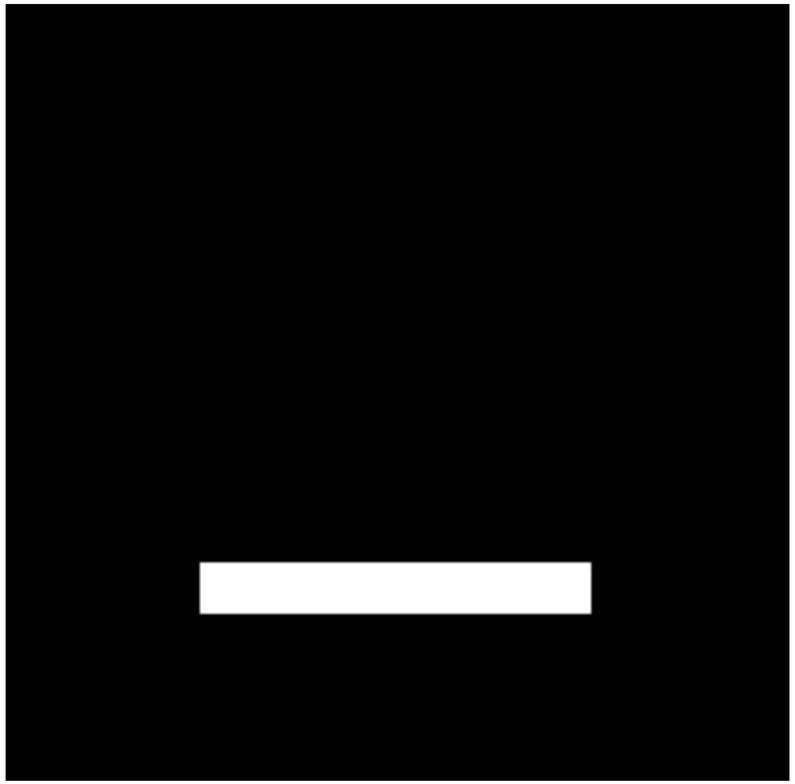
Thinking About Horn & Schunck Optical Flow

$$\nabla \mathbf{I}\mu = -\frac{\delta I}{\delta t}$$

Remarkable: Only the *image gradient* matters for time evolution!
Why? Think about it.

Thinking About Horn & Schunck Optical Flow

$$\nabla \mathbf{I}\mu = -\frac{\delta I}{\delta t}$$



Optical Flow – The Beginnings

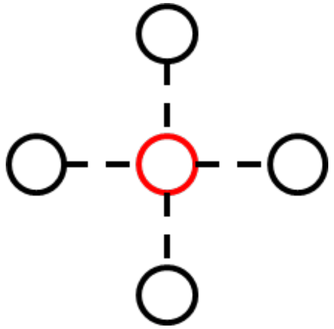
[Horn & Schunck AI'81]

Raw estimate

Optical Flow – The Beginnings

[Horn & Schunck AI'81]

Raw estimate



Prior term (Smoothness)

Optimization/energy minimization

[Horn & Schunck AI'81]

Data term (Constancy)

Prior term (Smoothness)

$$E = \min_{\mu} \left[(\nabla^T \mathbf{I}_1 \mu - (\mathbf{I}_2 - \mathbf{I}_1))^2 + \alpha(|\nabla \mu_x|^2 + |\nabla \mu_y|^2) \right]$$

Optical Flow – The Beginnings

[Horn & Schunck AI'81]

Raw estimate

[Horn & Schunck AI'81]

Smoothed estimate

Horn & Schunck on a more recent scene

Thinking About Horn & Schunck Optical Flow

$$\nabla \mathbf{I}\mu = -\frac{\delta I}{\delta t}$$

Some strong assumptions:
Was derived assuming *infinitesimal translation* $(\Delta x, \Delta y)$*!*
Get different equations for infinitesimal rotation, homography, scale…

From the gradient perspective, you can see that motion
is really assumed to be *small.*

Matching-based motion estimation

Generally: Energy to measure similarity between a pixel in image 1 with pixels in image 2:

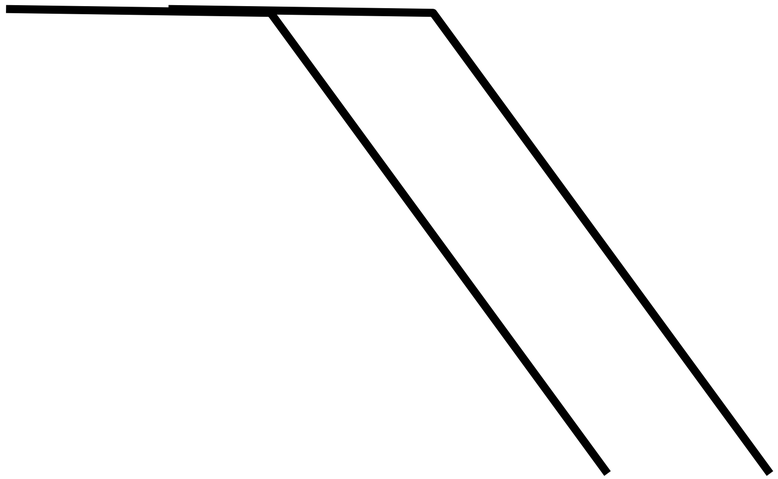
$$\min_{\mathbf{w_p}} \left(I_t(\mathbf{p}) - I_{t+1}\left(\mathbf{p} + \mathbf{w_p}\right)\right)^2$$

Image 1

Image 2

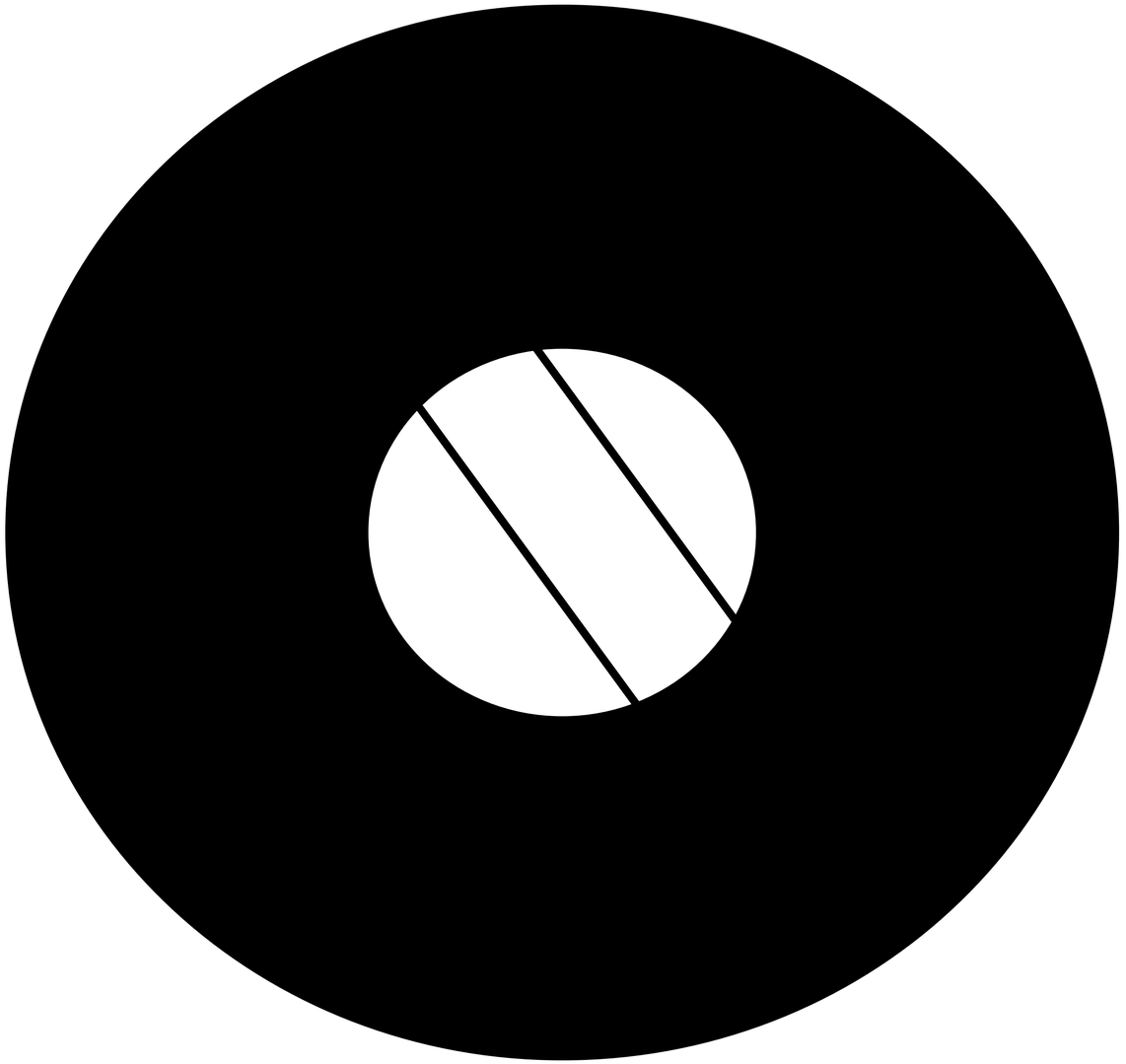
Comparing pixel colors directly

Estimating Motion in Direction of Constant Brightness: The Aperture Problem



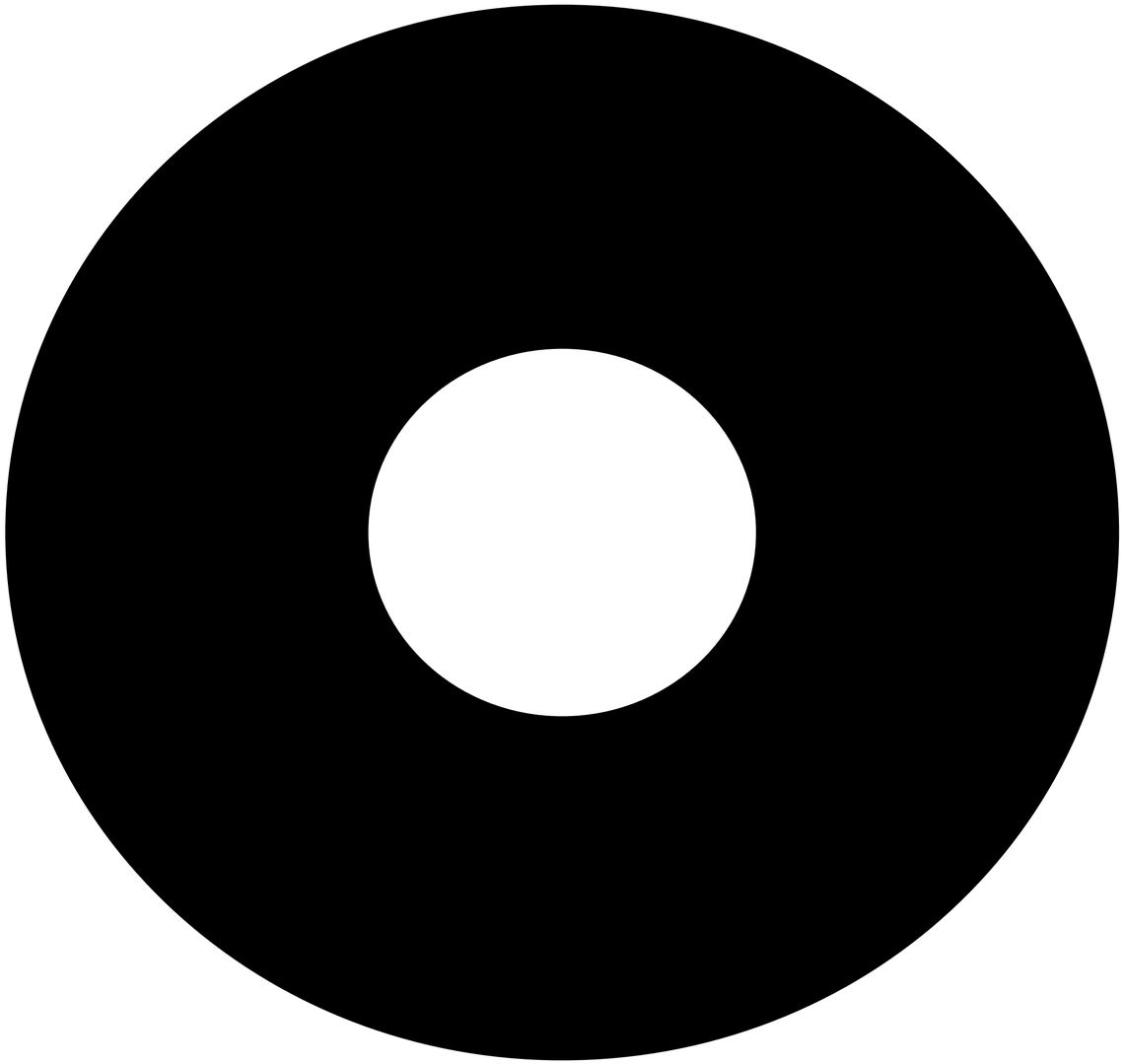
Slide Credit: S. Lazebnik

Estimating Motion in Direction of Constant Brightness: The Aperture Problem



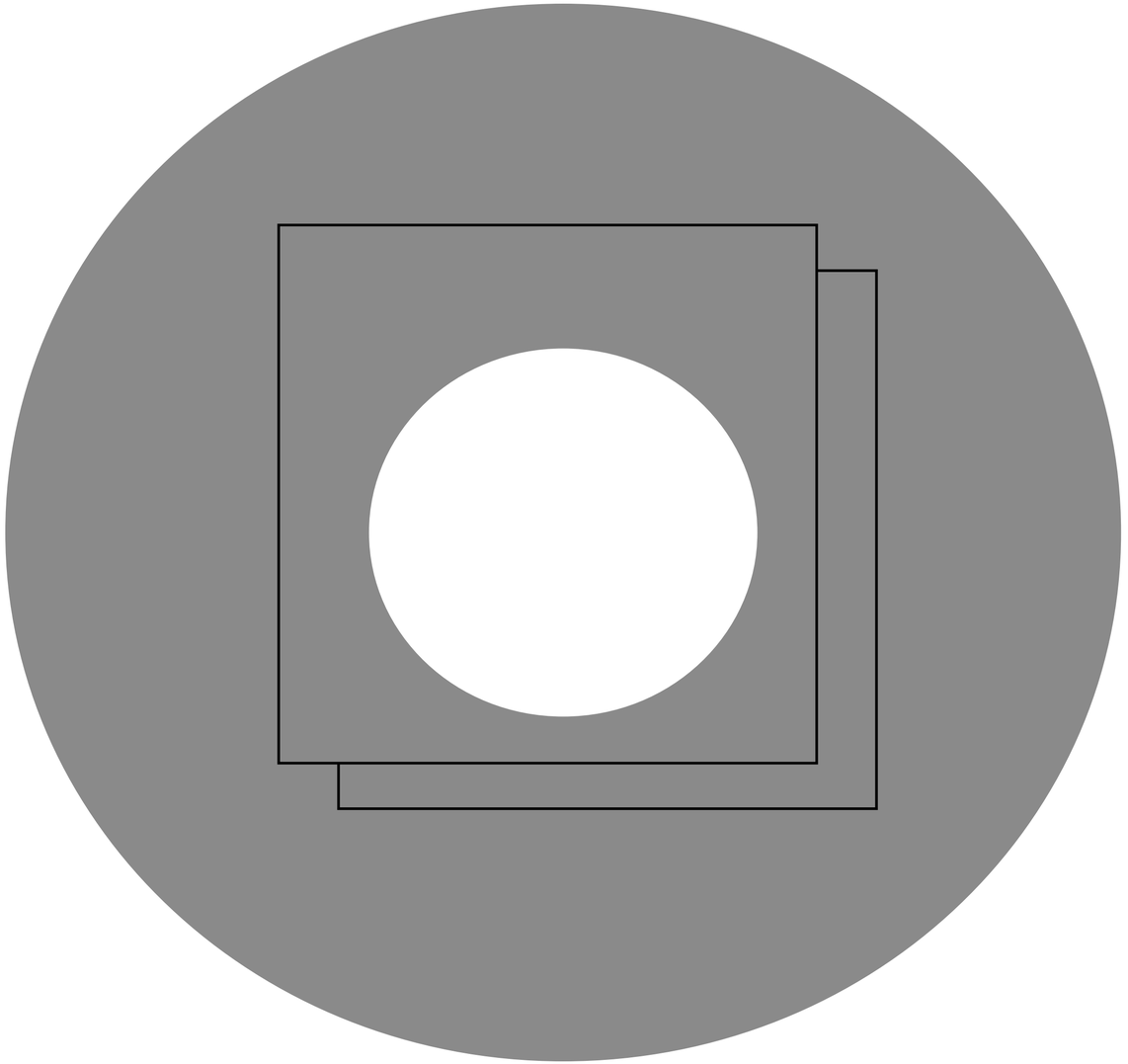
Slide Credit: S. Lazebnik

Other invisible flow



Slide credit: D Fouhey & J Johnson

Other invisible flow



Slide credit: D Fouhey & J Johnson

What do you perceive?

Matching-based motion estimation

Generally: Energy to measure similarity between a pixel in image 1 with pixels in image 2:

$$\min_{\mathbf{w_p}} \left(I_t(\mathbf{p}) - I_{t+1}\left(\mathbf{p} + \mathbf{w_p}\right)\right)^2$$

Image 1

Image 2

Resolving ambiguities: Lucas-Kanade

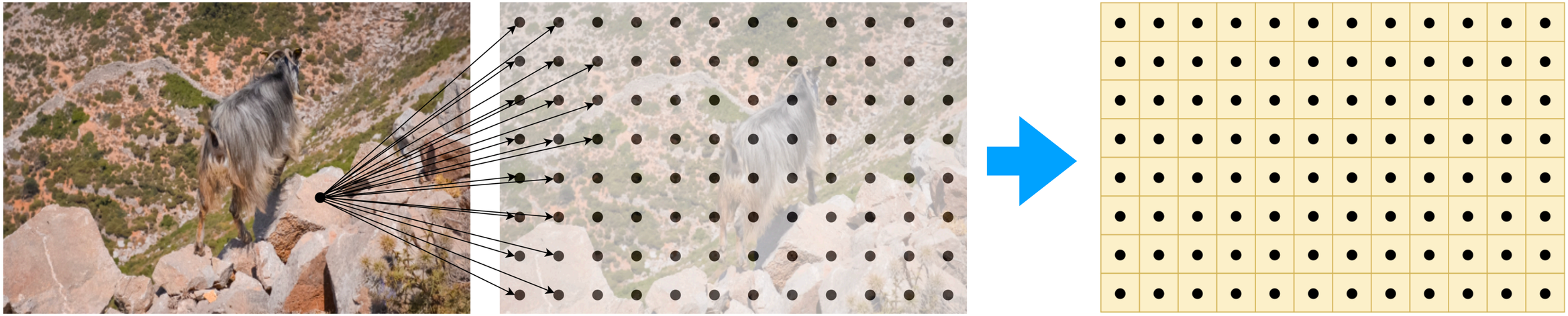
Insight: Pixels in the same patch generally share the same motion

$$\min_{\mathbf{w_p}} \sum_{\mathbf{q} \in N_p} \left(I_t(\mathbf{q}) - I_{t+1}\left(\mathbf{q} + \mathbf{w_p}\right) \right)^2$$

Image 1

Image 2

Cost / Correspondence Volumes



Idea: Correlate every pixel / patch in image 1 with every pixel in image 2, yielding a 4D **correlation volume.**

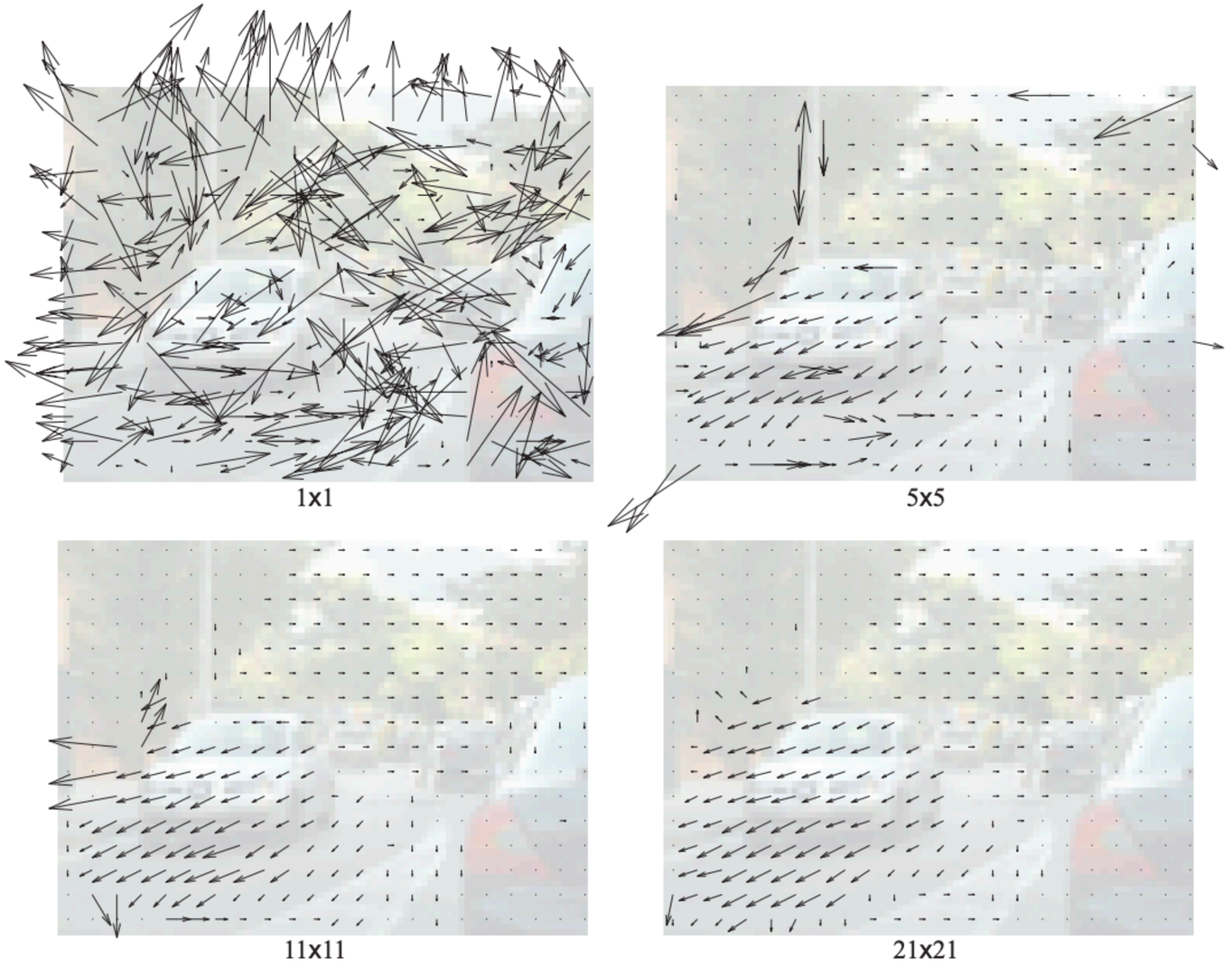
Image Courtesy: RAFT. *Teed et. al.*

Similarity between patches (cost volume)

Image 1

Cost volume (darker, more similar)

Effect of patch size



Issue: Brute force is too expensive

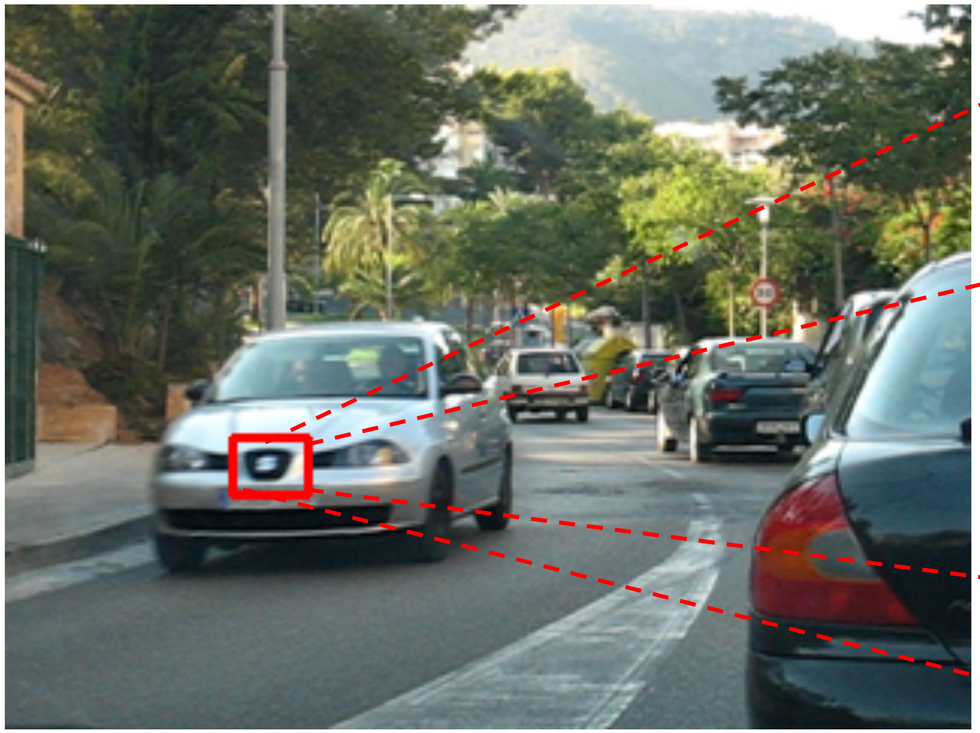


Image 1

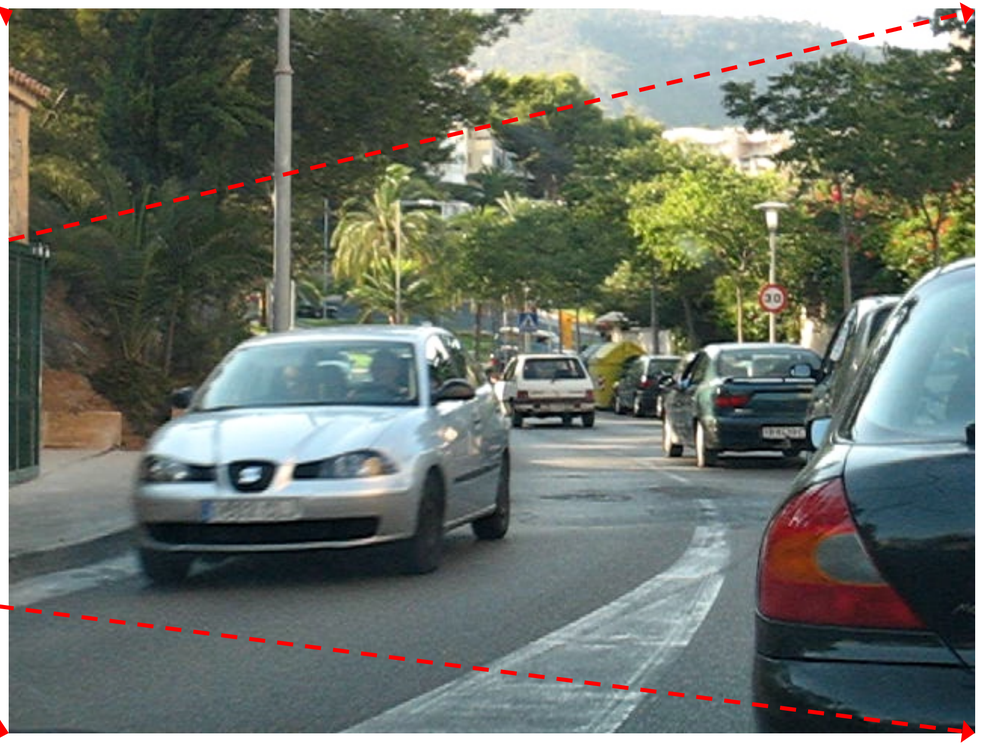
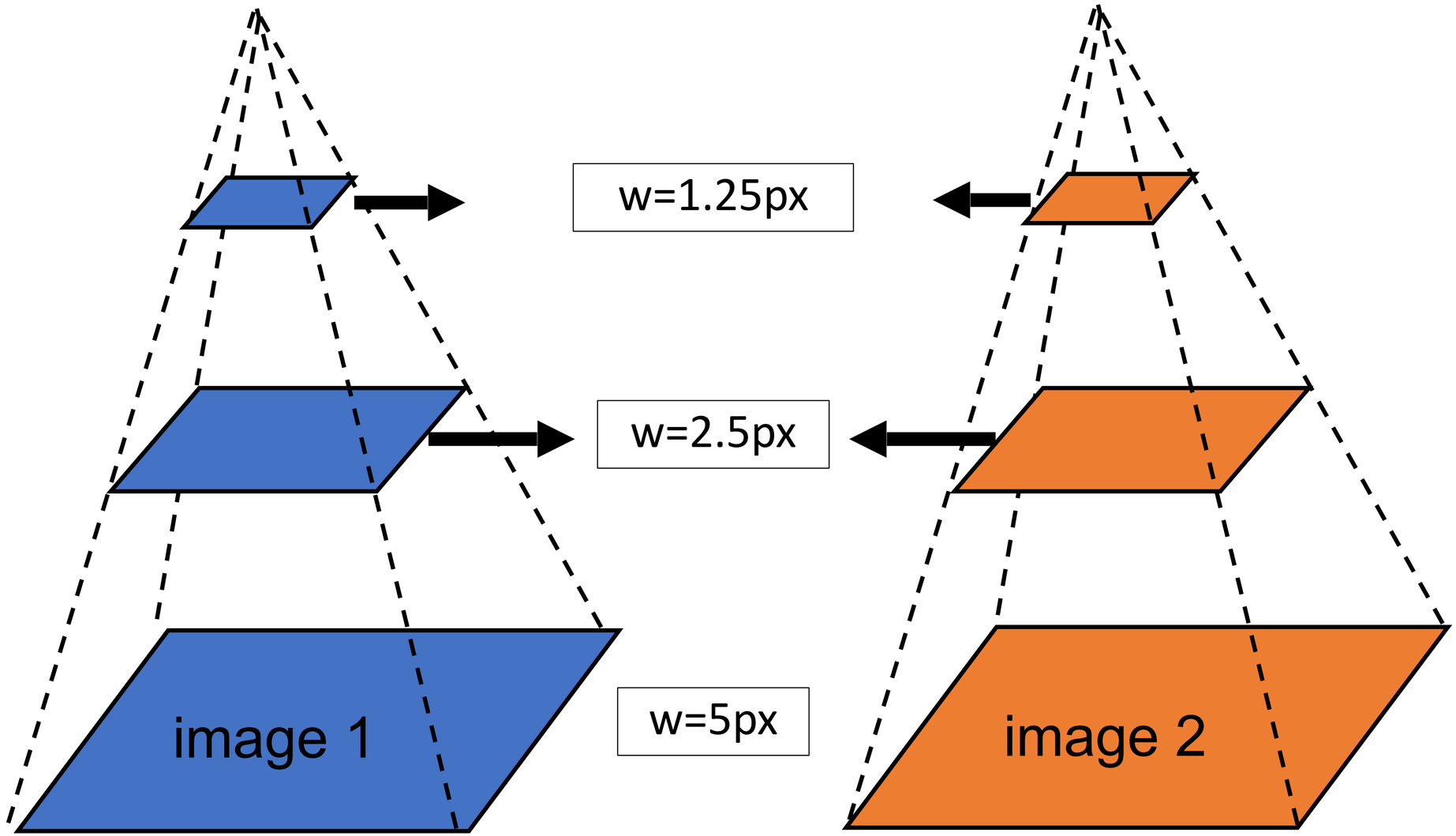


Image 2

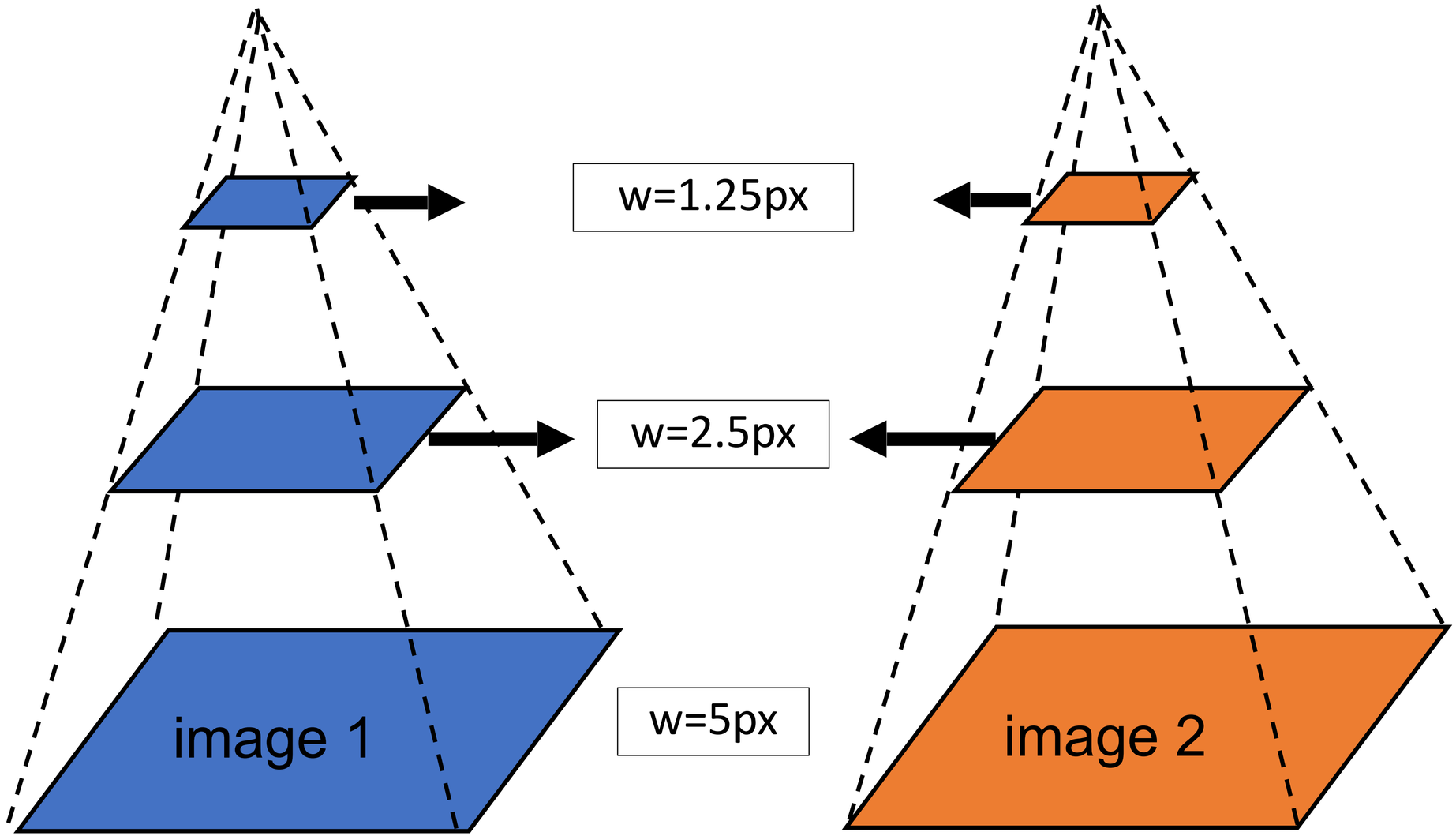
Coarse-to-fine iterative estimation



Start from top or bottom?

Slide credit: S. Lazebnik

Coarse-to-fine iterative estimation



How to use estimates from the upper level?

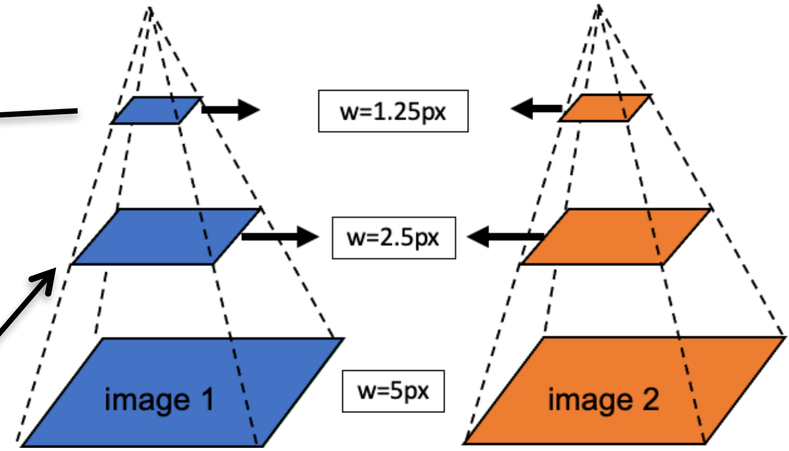
Slide credit: S. Lazebnik

Coarse-to-fine iterative estimation

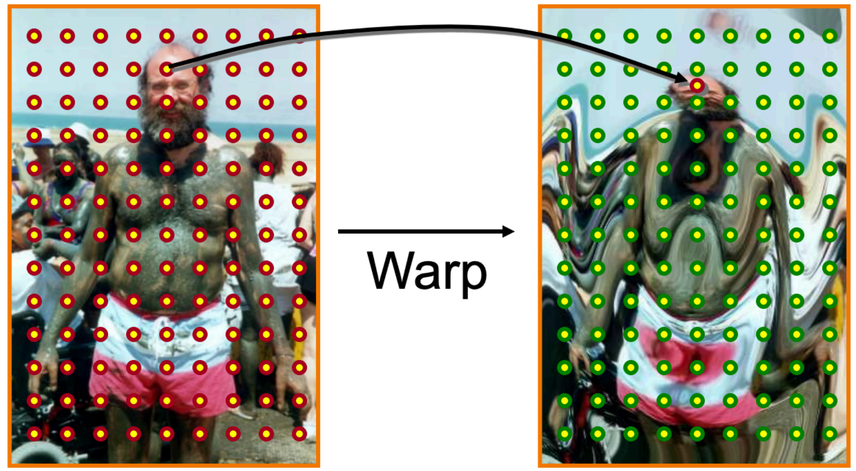
Current estimate

$$I_{t+1}\left(\mathbf{q}+\mathbf{w}_{\mathbf{p}}\right)=I_{t+1}\left(\mathbf{q}+\mathbf{w}_{\mathbf{p}}{}^{k}+\delta\mathbf{w}_{\mathbf{p}}\right)=\mathbf{I}_{\mathbf{w}}(\mathbf{q}+\delta\mathbf{w}_{\mathbf{p}})$$

Small increment



Warped image: $I_{\mathrm{w}}(\mathbf{q})=I_{t+1}\left(\mathbf{q}+\mathbf{w}_{\mathbf{p}}^{k}\right)$



Source image

Destination image

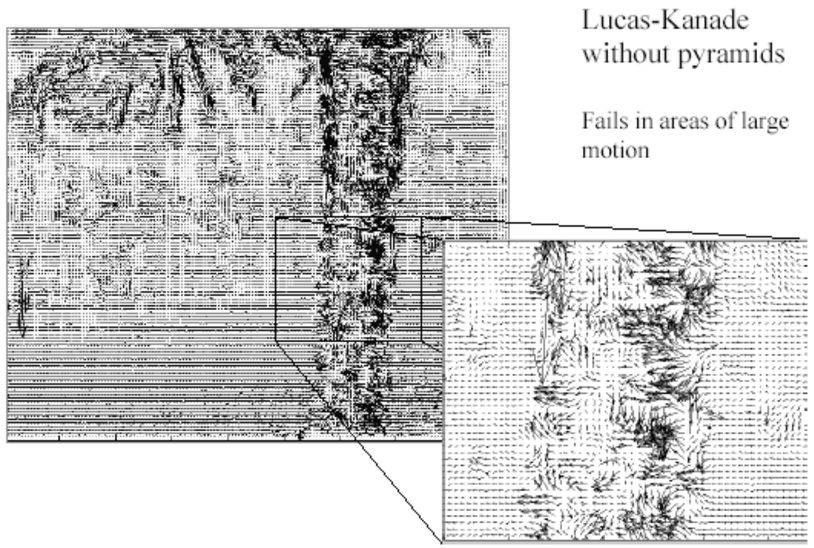
Warping operation

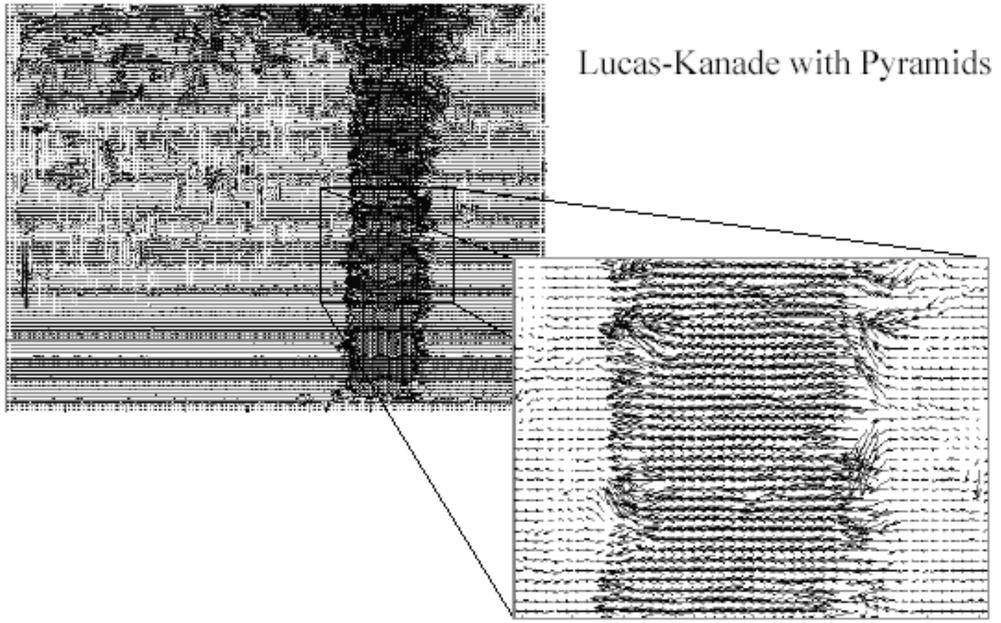
$$I_{\mathrm{w}}(\mathbf{q}) = I_{t+1}\left(\mathbf{q} + \mathbf{w}_{\mathbf{p}}^{k}\right)$$

Input images *t* and *t*+1

Image *t* and warped image

Optical Flow Results

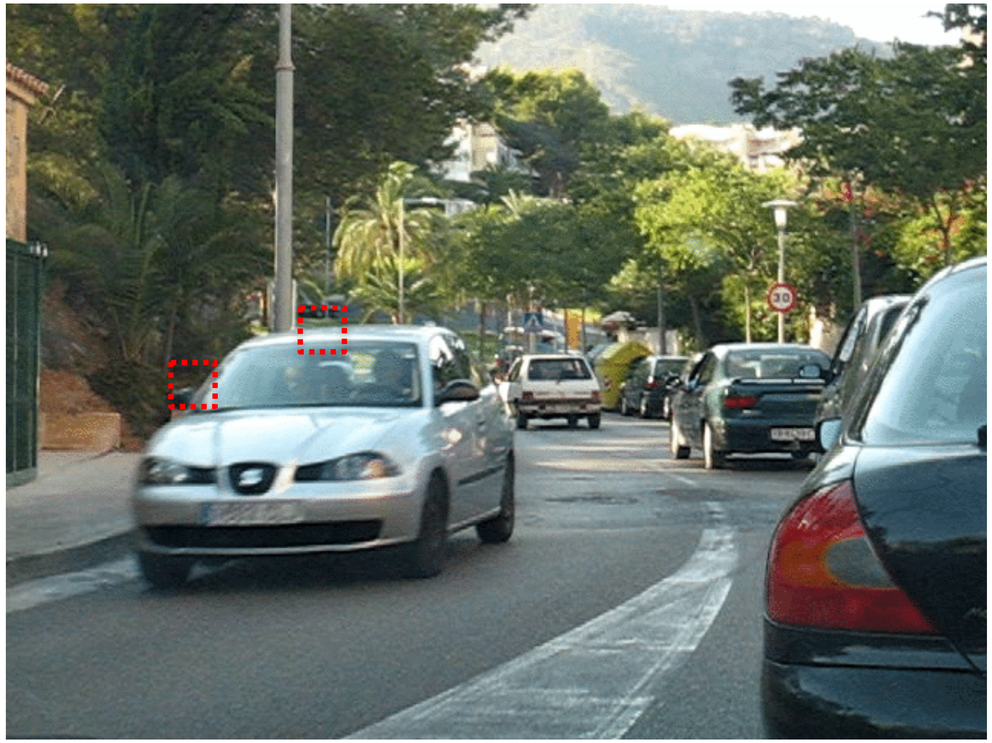


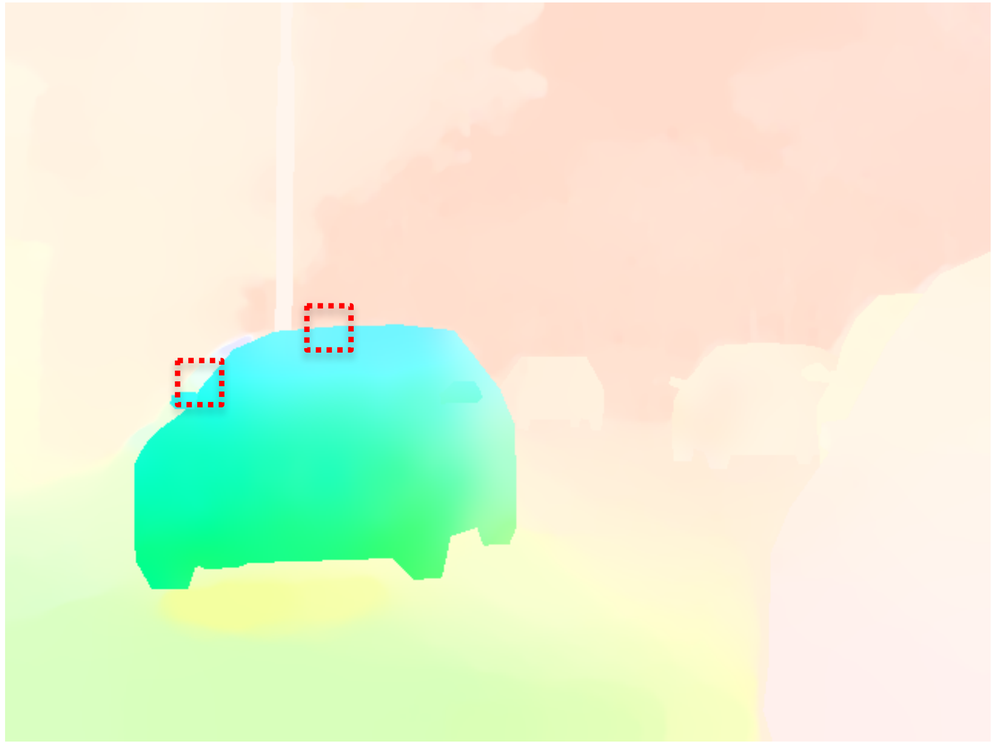


Slide credit: S. Lazebnik

Issue: Motion Boundaries

No matches can be found at motion boundaries…



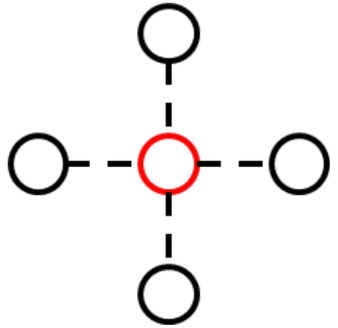


Issue: Ambiguous patch details

Flat regions of the image are underconstrained

Solving ambiguities: Horn & Schunck

Smoothness: neighboring pixels have similar motion

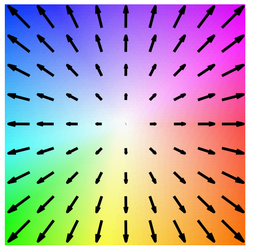


Horn & Schunck

Input

Horn & Schunck

Ground truth

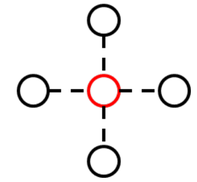


Color key
[Baker *et al.* IJCV'11]

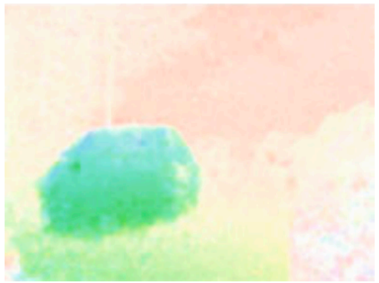
Improving Horn & Schunck

[Sun *et al.* CVPR'10, IJCV'14]

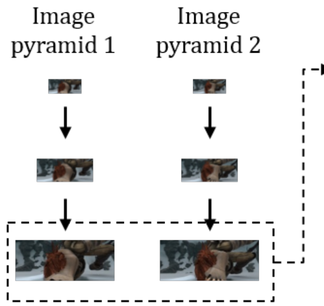
$$E(\mathbf{w}) = \sum_{\mathbf{p}} \left| I_t(\mathbf{p}) - I_{t+1}\left(\mathbf{p} + \mathbf{w}_\mathbf{p}\right) \right|^2 + \lambda \sum_{\mathbf{q} \in N_\mathbf{p}} \left| \mathbf{w}_\mathbf{p} - \mathbf{w}_\mathbf{q} \right|^2$$

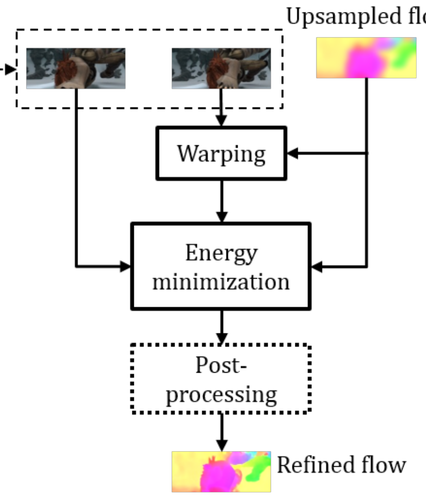


"Old"

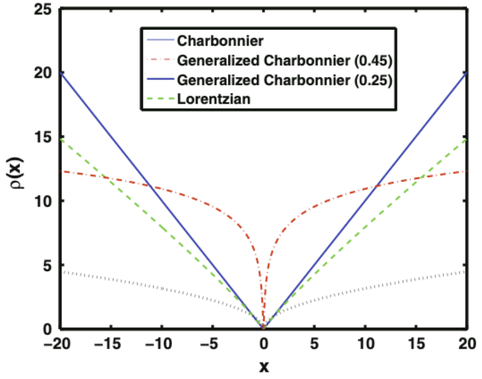


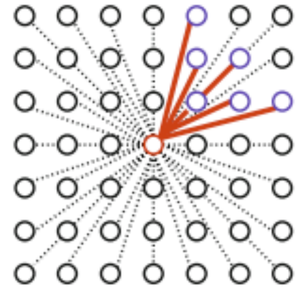
Multi-Scale Pyramid





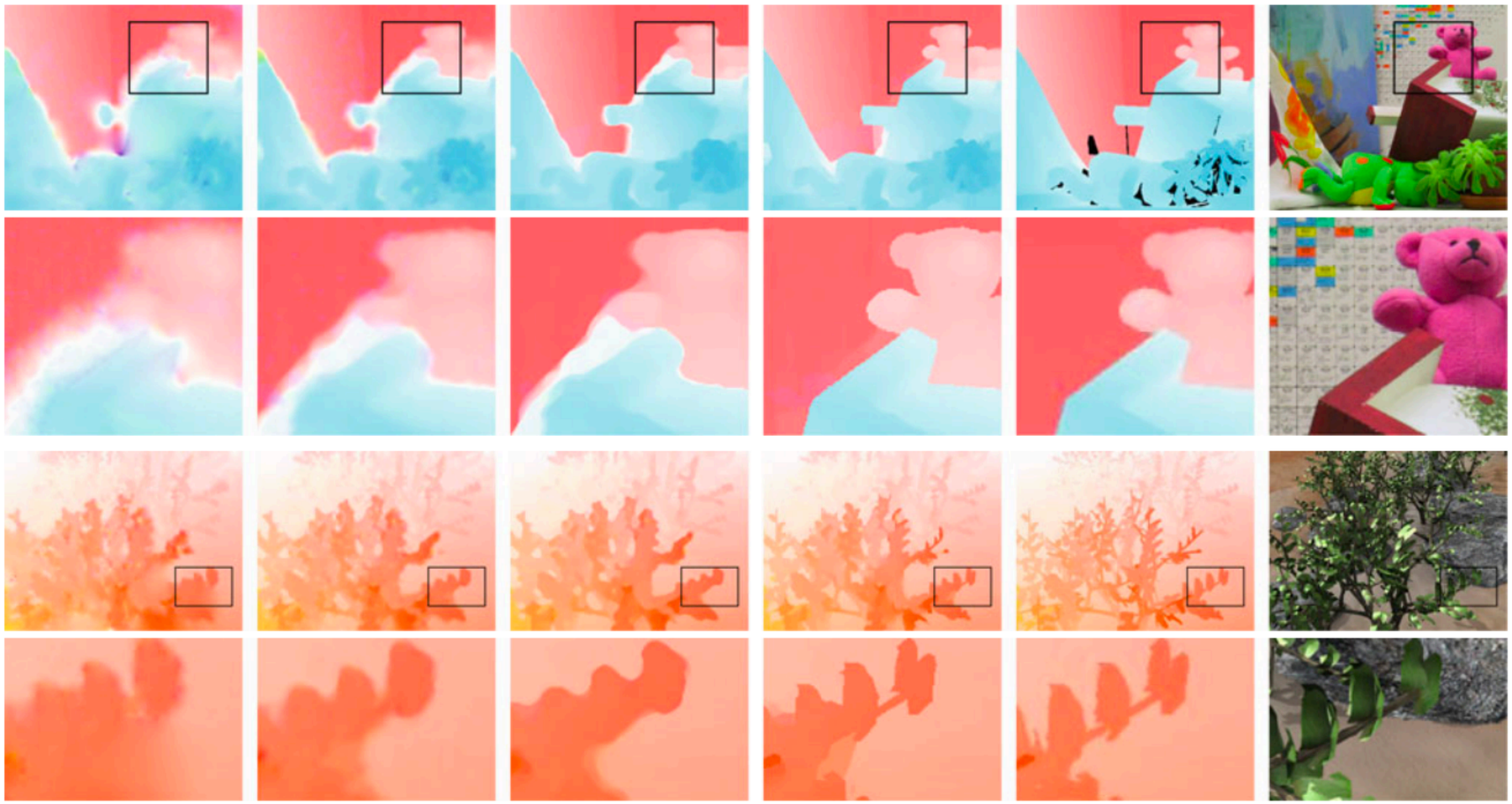
Ground truth





Improving Horn & Schunck

[Sun *et al.* CVPR'10, IJCV'14]



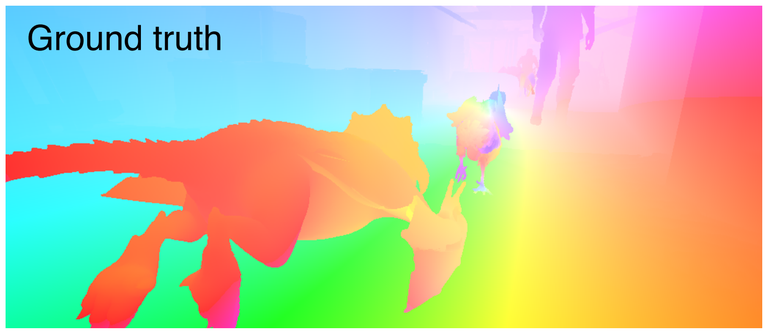
(a) "Old" **HS** [58] **(b)** "New" **HS** **(c)** **Classic++** **(d)** **Classic+NL** **(e)** Ground truth **(f)** First frame

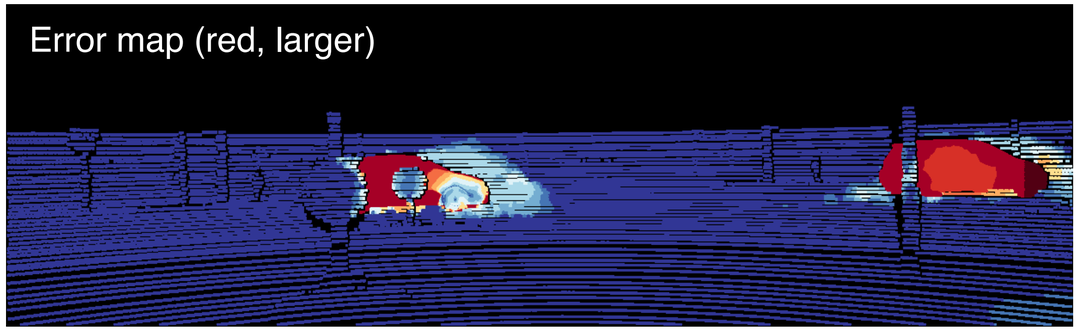
Implementation Robust function Non-local smoothness term

Challenges for classical methods

Large motion
Motion blur
Occlusions
Lighting changes
Noise…

Hard to modify objective function and even harder to optimize it





Summary

- Classical approach
  - Constancy assumption -> matching by comparison (cost volume)

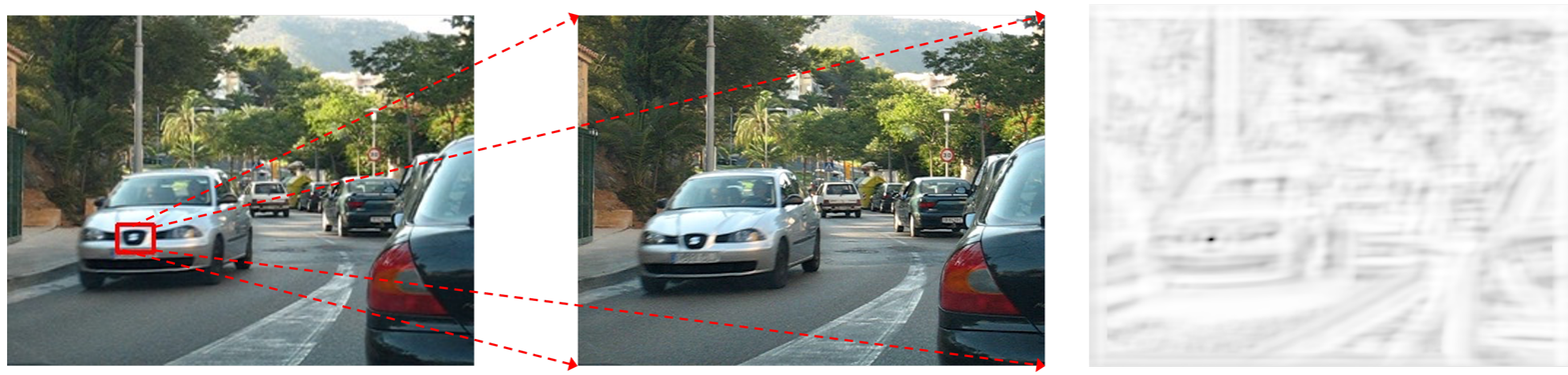
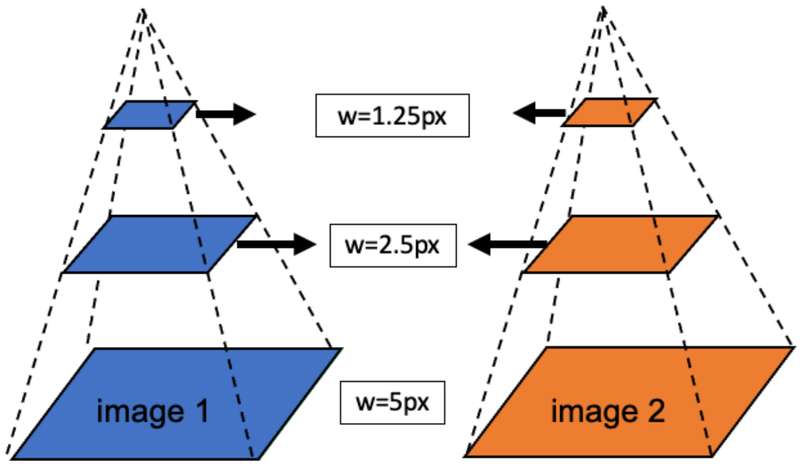


Image 1 Image 2 Cost volume

Summary

- Classical approach
  - Constancy assumption -> matching by comparison (cost volume)
  - Coarse-to-fine, warping-based iterative estimation

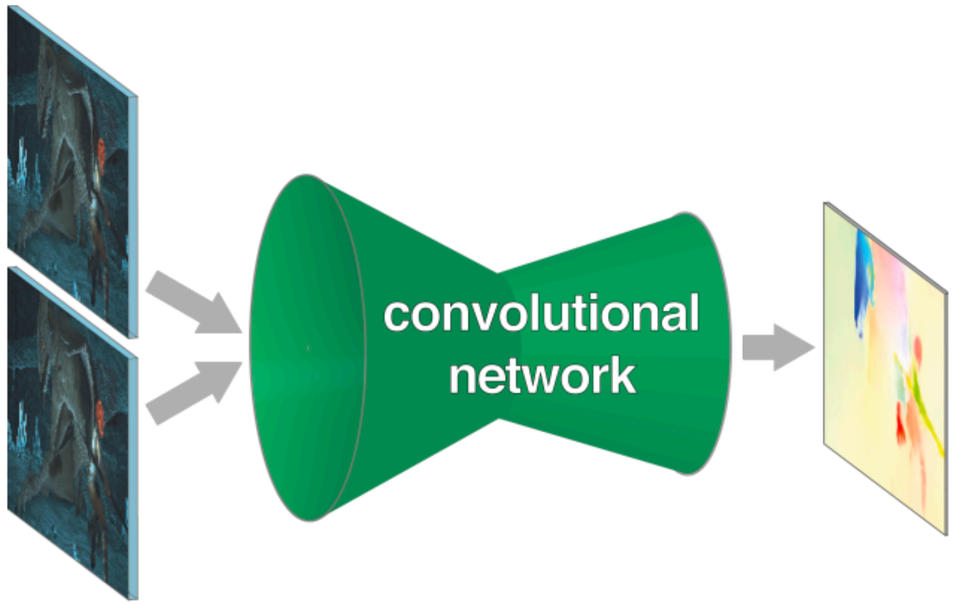


Input images *t* and *t*+1

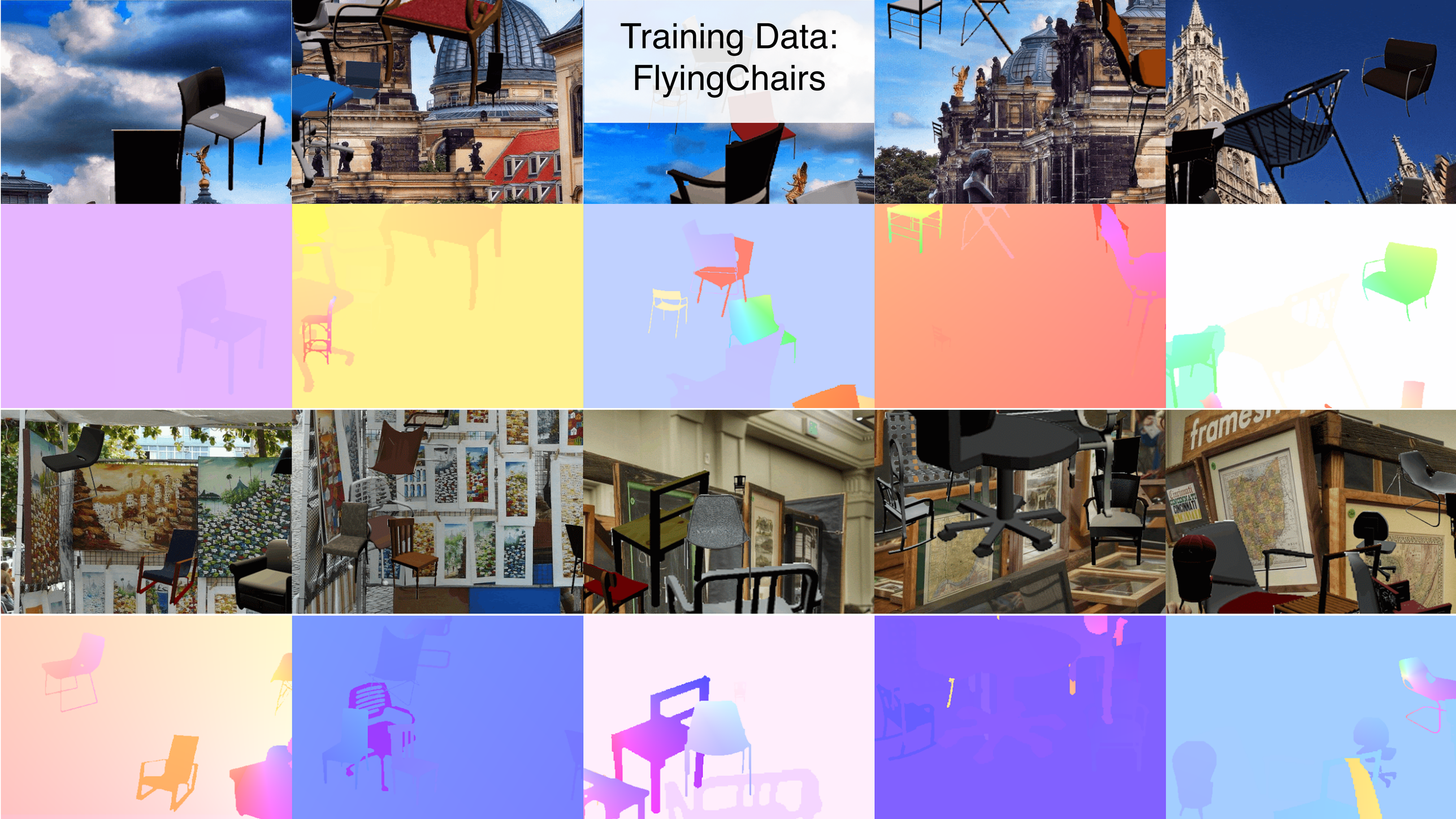
Image *t* and warped image

Supervised optical flow

[Dosovitskiy *et al.* ICCV'15]



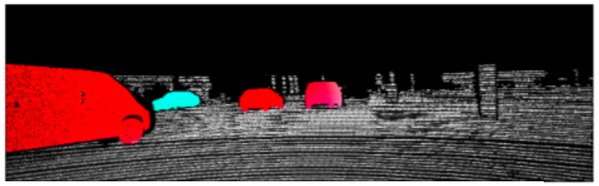
What is the training/test data?

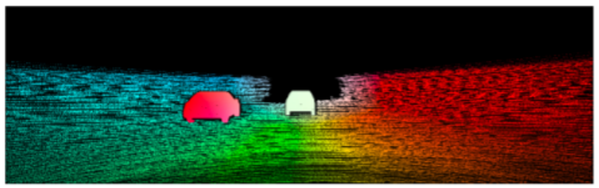

Training Data:
FlyingChairs

Two widely-used benchmarks for optical flow

Sintel (Blender movie)

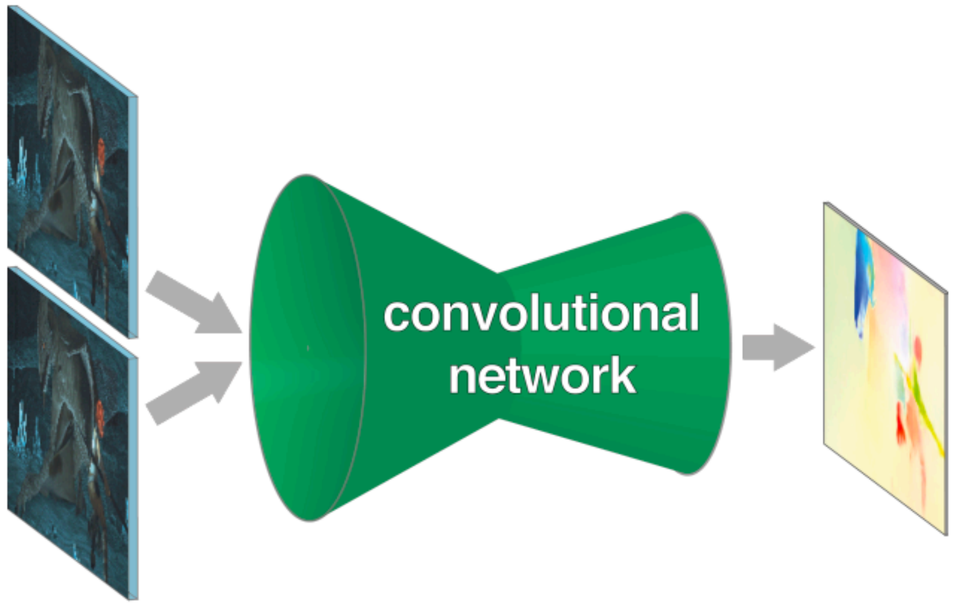
KITTI (driving)





Supervised optical flow

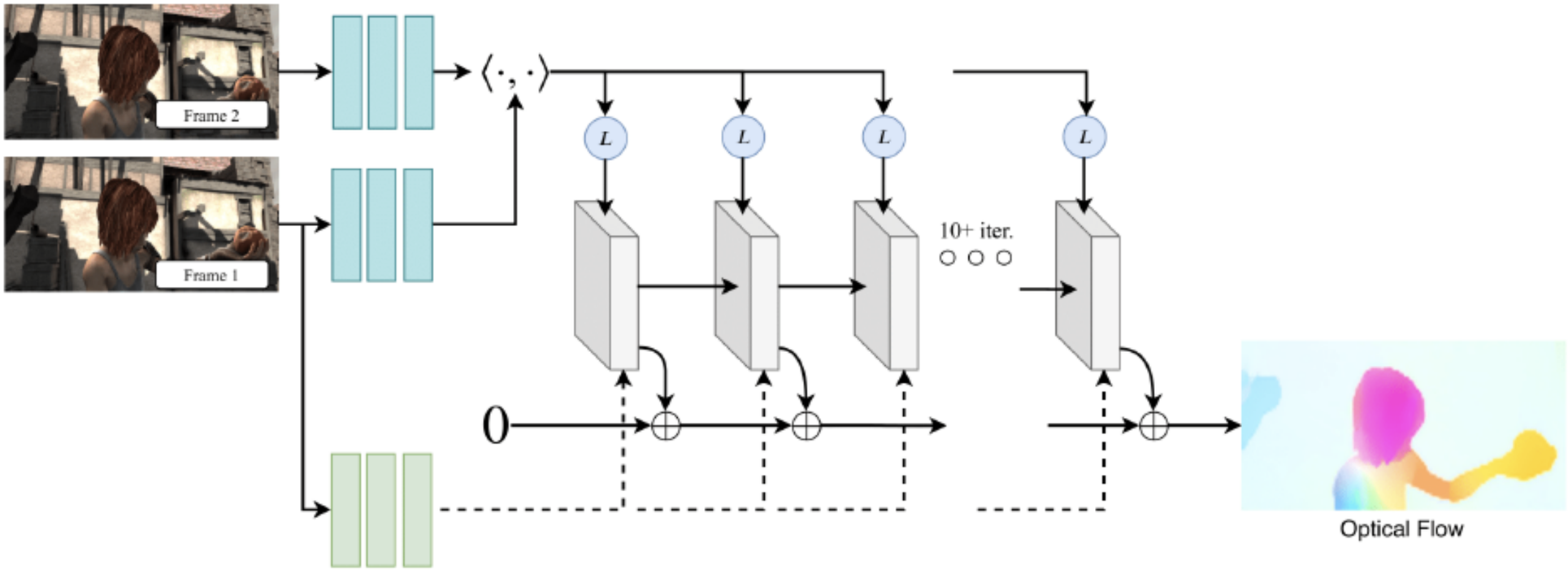
[Dosovitskiy *et al.* ICCV'15]



What is the network/architecture?

RAFT: Recurrent All-pairs Field Transforms

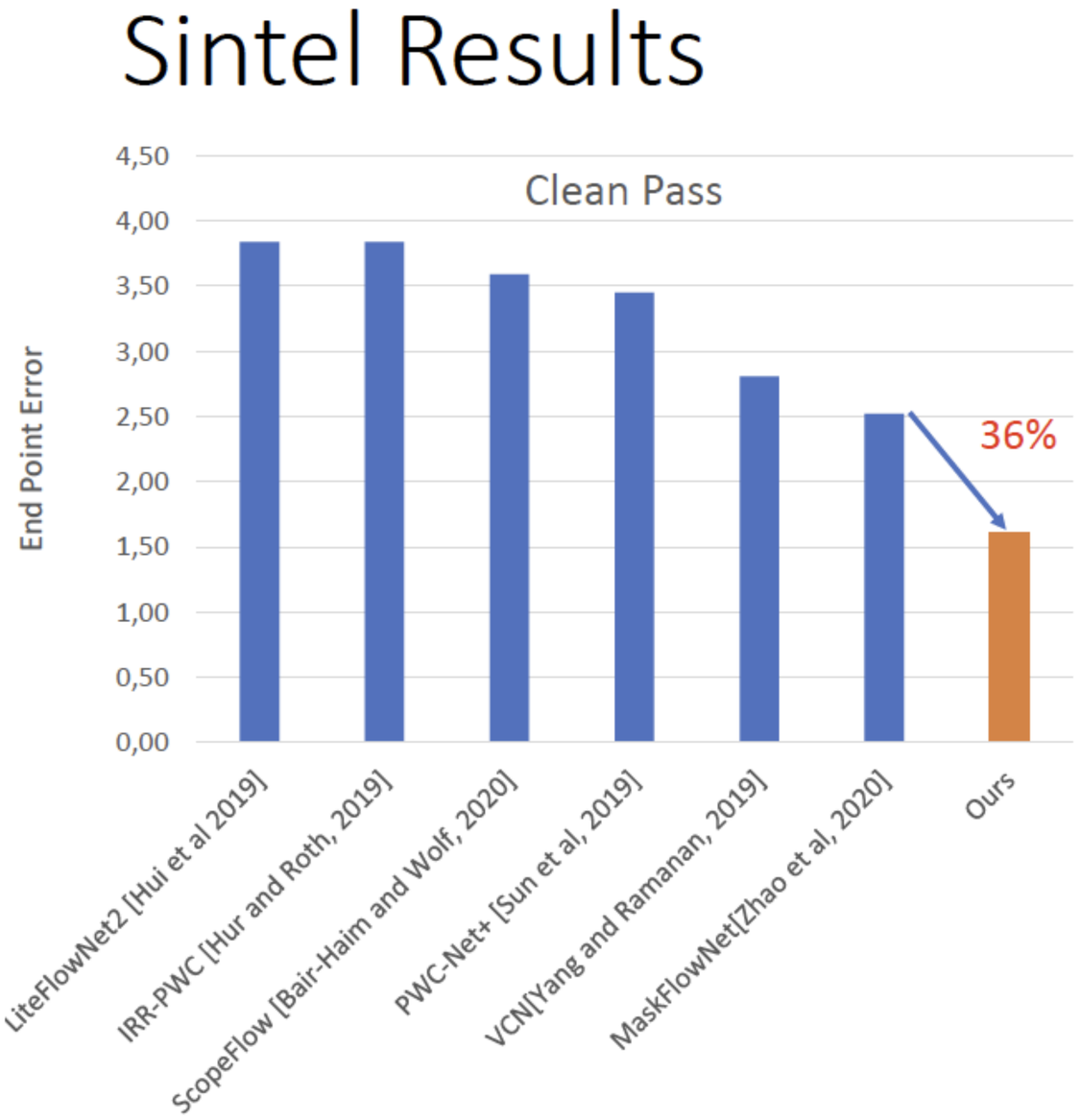
[Teed and Deng ECCV 2020 Best paper]



Slide credit: Teed and Deng

Significant improvement over prior art

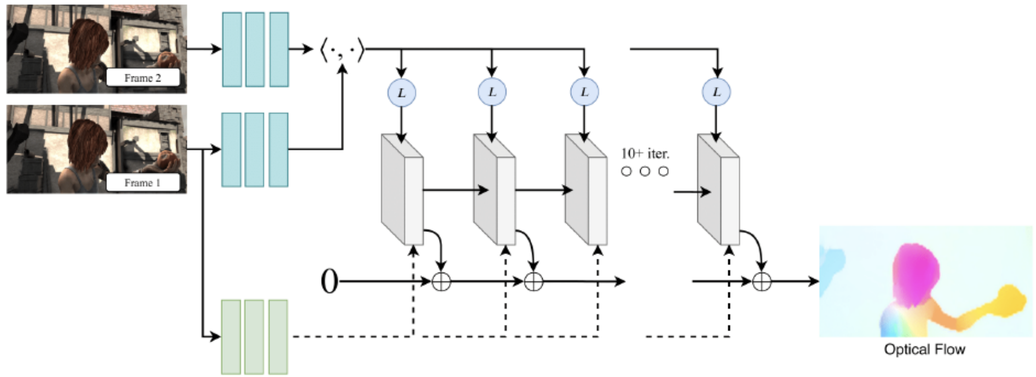
[Teed and Deng ECCV 2020 Best paper]



Slide credit: Teed and Deng

RAFT: Recurrent All-pairs Field Transforms

[Teed and Deng ECCV 2020 Best paper]



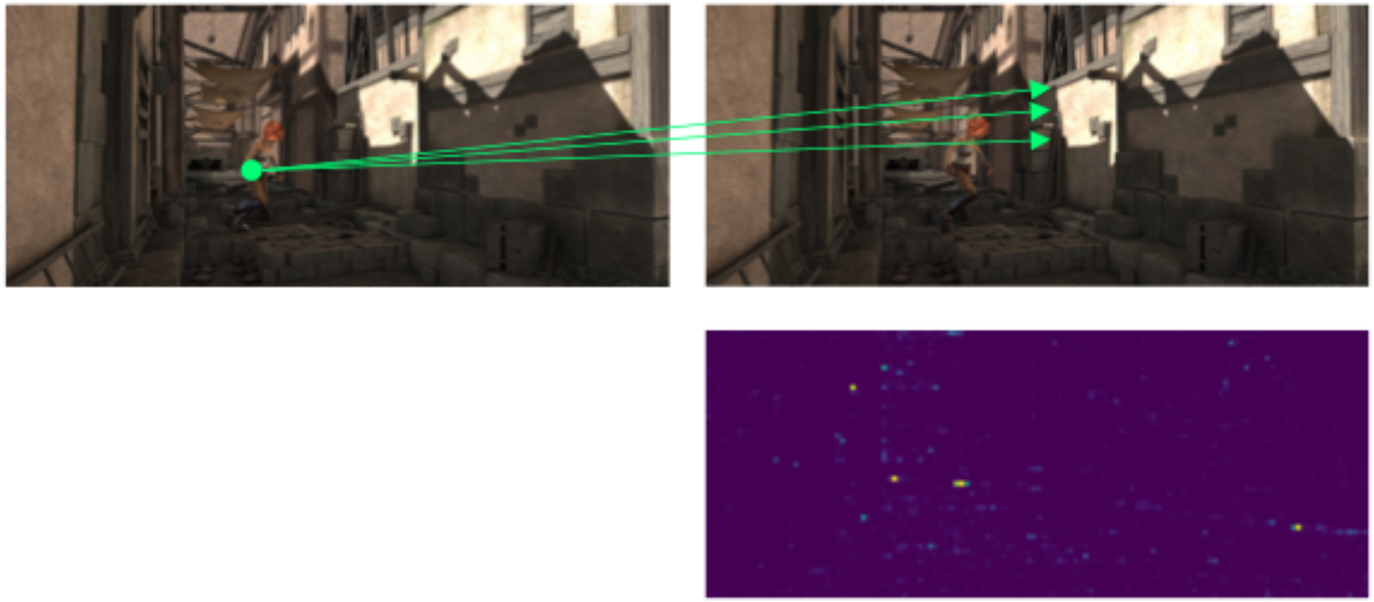
Key ideas

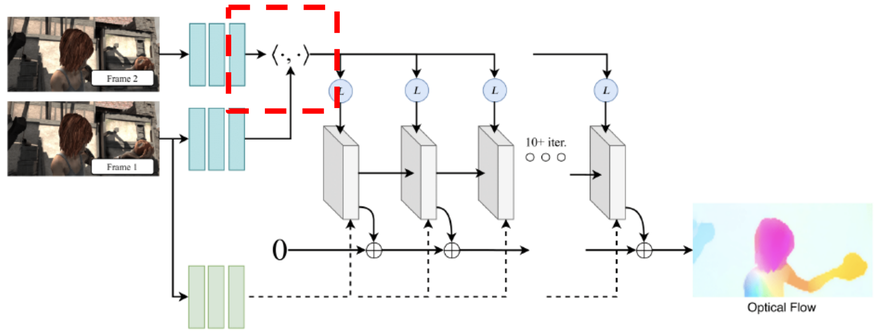
- All-pairs cost volume consistent across scales
- Recurrent module for iterative refinement of flow
- (Secret) Better training schemes

Slide credit: Teed and Deng

All-pairs visual similarity (cost volume)

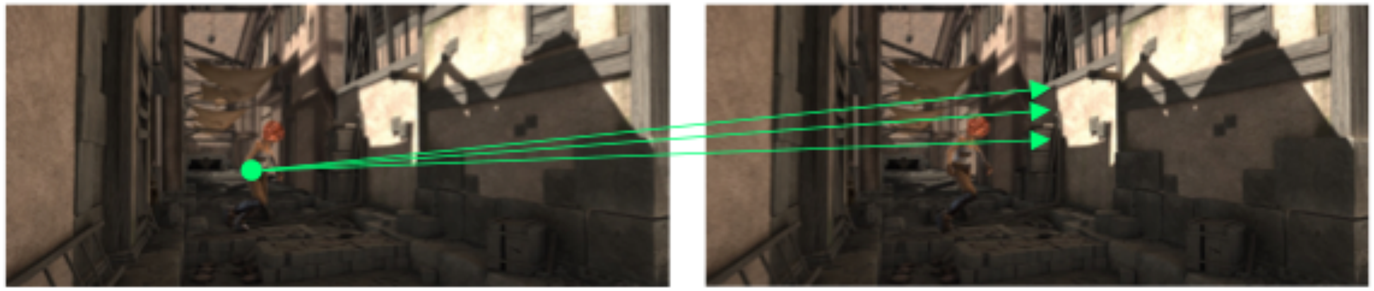
Inner product/correlation between features

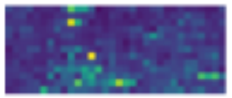


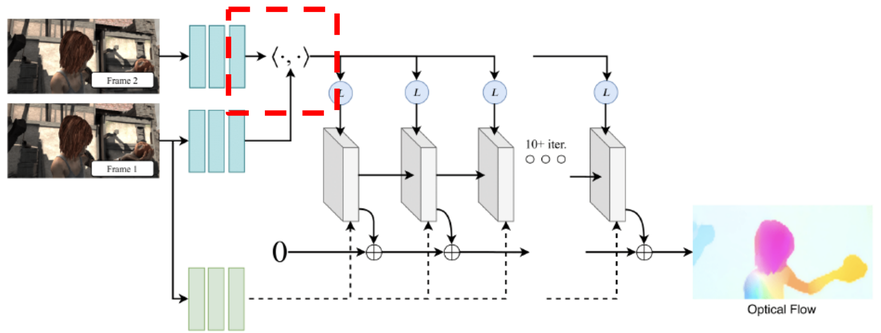


Cost volume pyramid

Spatial pooling

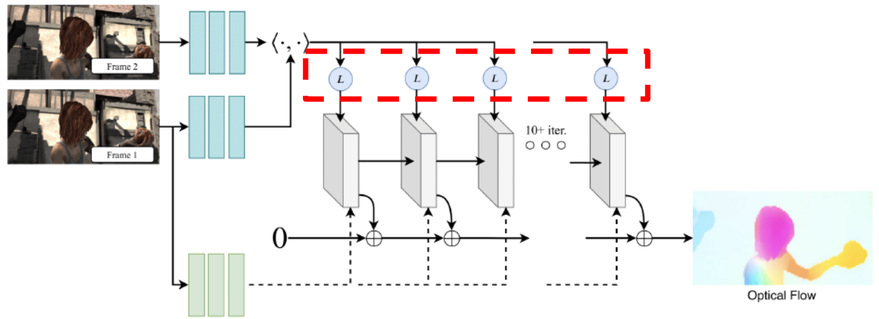


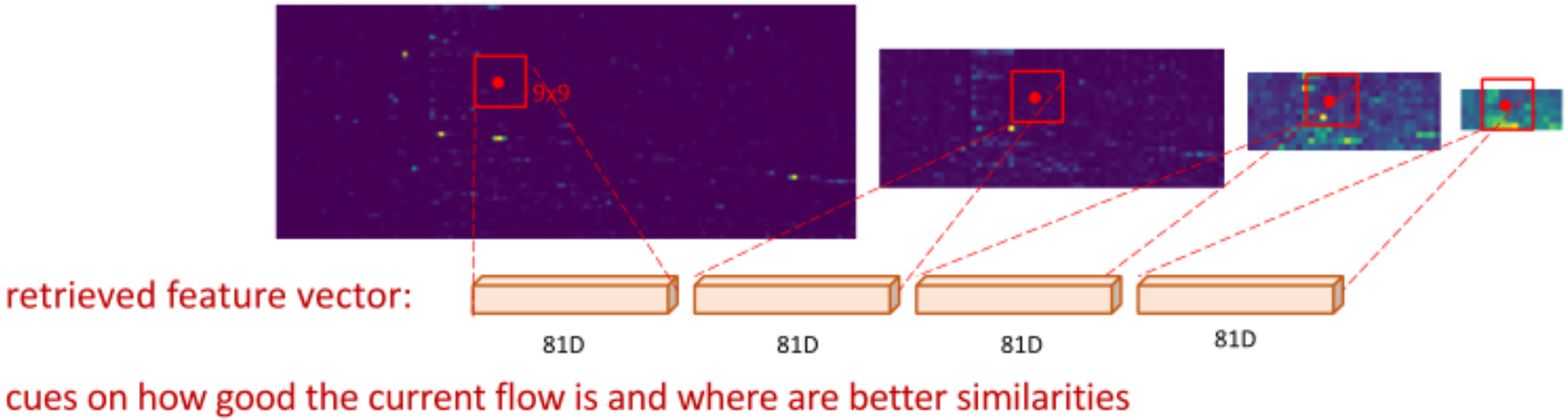




Look up cost volume

Retrieve using current motion estimates

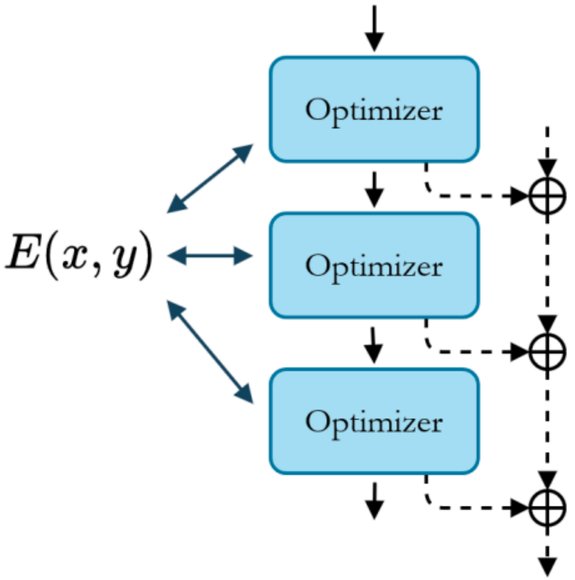


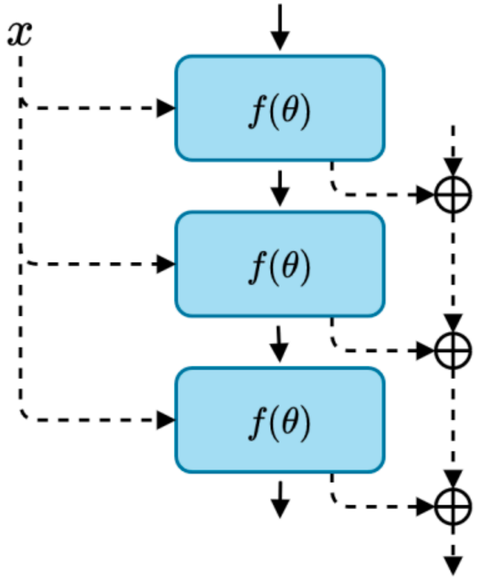


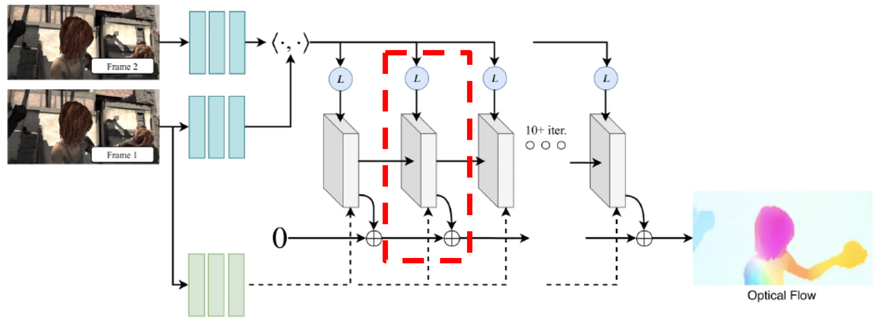
cues on how good the current flow is and where are better similarities

Recurrent update

Like classical optimization algorithms

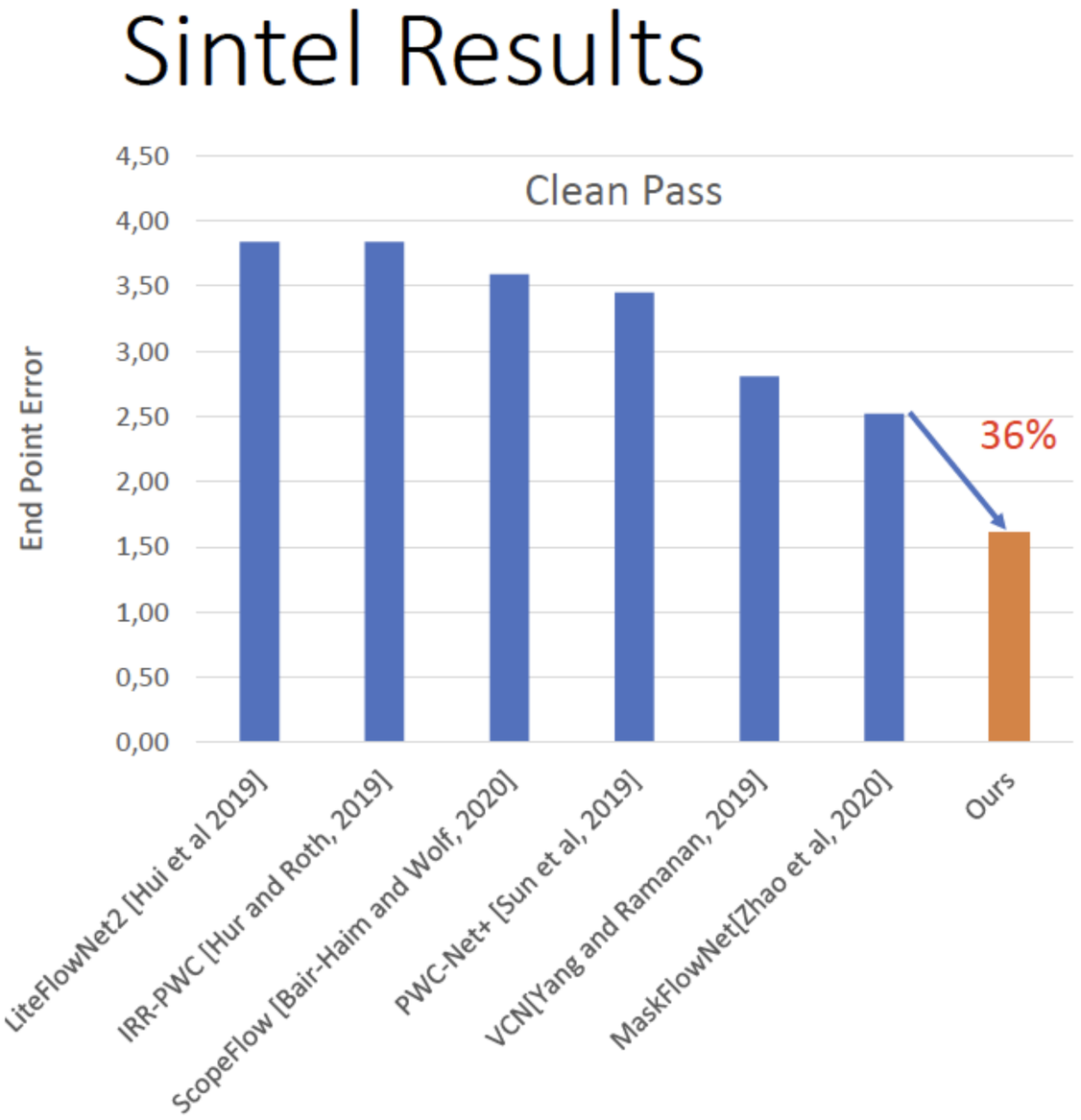






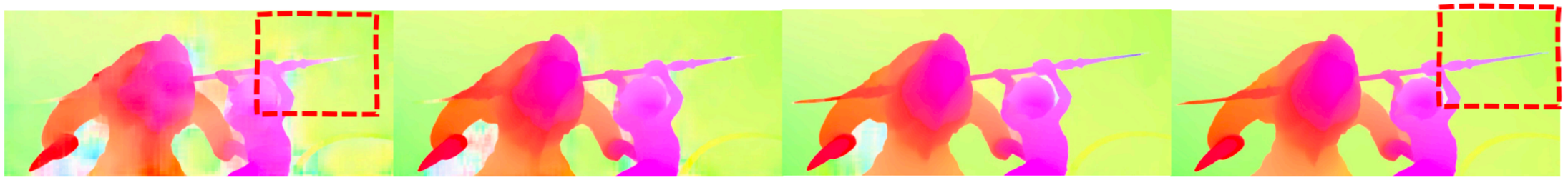
Significant improvement over prior art

[Teed and Deng ECCV 2020 Best paper]



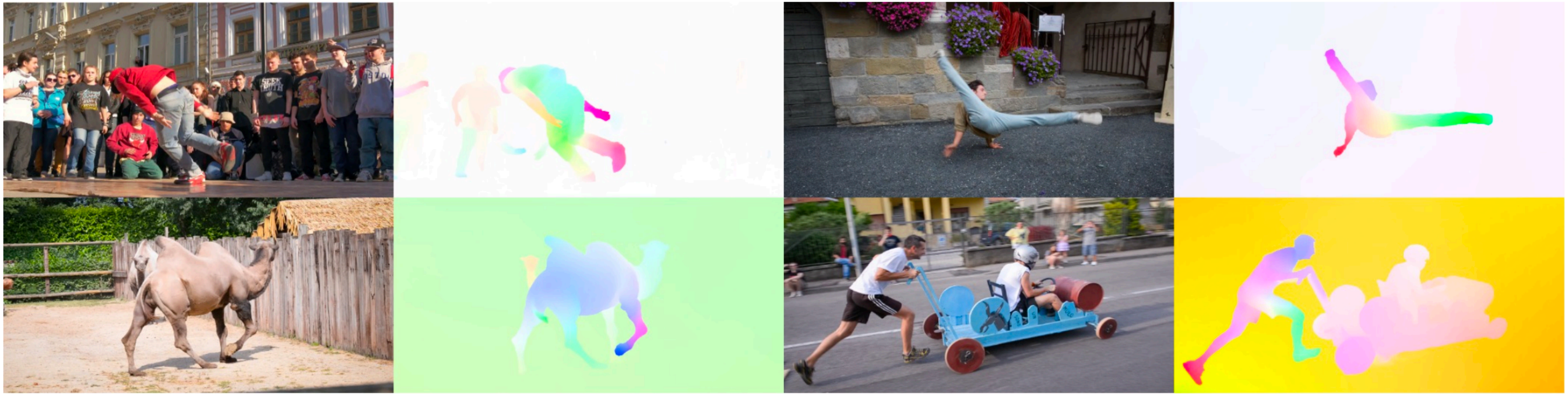
Slide credit: Teed and Deng

Recurrent update



1 Iteration | 2 Iterations | 5 Iterations | 32 Iterations

Visual results on Davis (real-world)



Is the hand-crafted structure (cost-volume, unrolled optimization on cost volume) really necessary?

The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, David J. Fleet*

{srbs,irwinherrmann,junhwahur,abhiskar,deqingsun,davidfleet}@google.com

Google DeepMind and Google Research

Abstract

Denoising diffusion probabilistic models have transformed image generation with their impressive fidelity and diversity. We show that they also excel in estimating optical flow and monocular depth, surprisingly, without task-specific architectures and loss functions that are predominant for these tasks. Compared to the point estimates of conventional regression-based methods, diffusion models also enable Monte Carlo inference, e.g., capturing uncertainty and ambiguity in flow and depth. With self-supervised pre-training, the combined use of synthetic and real data for supervised training, and technical innovations (infilling and step-unrolled denoising diffusion training) to handle noisy-incomplete training data, and a simple form of coarse-to-fine refinement, one can train state-of-the-art diffusion models for depth and optical flow estimation. Extensive experiments focus on quantitative performance against benchmarks, ablations, and the model's ability to capture uncertainty and multimodality, and impute missing values. Our model, DDVM (Denoising Diffusion Vision Model), obtains a state-of-the-art relative depth error of 0.074 on the indoor NYU benchmark and an Fl-all outlier rate of 3.26% on the KITTI optical flow benchmark, about 25% better than the best published method. For an overview see diffusion-vision.github.io

1 Introduction

Diffusion models have emerged as powerful generative models for high fidelity image synthesis, capturing rich knowledge about the visual world [21, 48, 55, 62]. However, at first glance, it is unclear whether these models can be as effective on many classical computer vision tasks. For example, consider two dense vision estimation tasks, namely, optical flow, which estimates frame-to-frame correspondences, and monocular depth perception, which makes depth predictions based on a single image. Both tasks are usually treated as regression problems and addressed with specialized architectures and task-specific loss functions, *e.g.*, cost volumes, feature warps, or suitable losses for depth. Without these specialized components or the regression framework, general generative techniques may be ill-equipped and vulnerable to both generalization and performance issues.

In this paper, we show that these concerns, while valid, can be addressed and that, surprisingly, a generic, conventional diffusion model for image to image translation works impressively well on both tasks, often outperforming the state of the art. In addition, diffusion models provide valuable benefits over networks trained with regression; in particular, diffusion allows for approximate inference with multi-modal distributions, capturing uncertainty and ambiguity (*e.g.* see Figure 1).

One key barrier to training useful diffusion models for monocular depth and optical flow inference concerns the amount and quality of available training data. Given the limited availability of labelled

*DF is also affiliated with the University of Toronto and the Vector Institute.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Is the hand-crafted structure (cost-volume, unrolled optimization on cost volume) really necessary?

The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, David J. Fleet*

{srbs,irwinherrmann,junhwahur,abhiskar,deqingsun,davidfleet}@google.com

Google DeepMind and Google Research

Abstract

Denoising diffusion probabilistic models have transformed image generation with their impressive fidelity and diversity. We show that they also excel in estimating optical flow and monocular depth, surprisingly, without task-specific architectures and loss functions that are predominant for these tasks. Compared to the point estimates of conventional regression-based methods, diffusion models also enable Monte Carlo inference, e.g., capturing uncertainty and ambiguity in flow and depth. With self-supervised pre-training, the combined use of synthetic and real data for supervised training, and technical innovations (infilling and step-unrolled denoising diffusion training) to handle noisy-incomplete training data, and a simple form of coarse-to-fine refinement, one can train state-of-the-art diffusion models for depth and optical flow estimation. Extensive experiments focus on quantitative performance against benchmarks, ablations, and the model's ability to capture uncertainty and multimodality, and impute missing values. Our model, DDVM (Denoising Diffusion Vision Model), obtains a state-of-the-art relative depth error of 0.074 on the indoor NYU benchmark and an Fl-all outlier rate of 3.26% on the KITTI optical flow benchmark, about 25% better than the best published method. For an overview see diffusion-vision.github.io

1 Introduction

Diffusion models have emerged as powerful generative models for high fidelity image synthesis, capturing rich knowledge about the visual world [21, 48, 55, 62]. However, at first glance, it is unclear whether these models can be as effective on many classical computer vision tasks. For example, consider two dense vision estimation tasks, namely, optical flow, which estimates frame-to-frame correspondences, and monocular depth perception, which makes depth predictions based on a single image. Both tasks are usually treated as regression problems and addressed with specialized architectures and task-specific loss functions, *e.g.*, cost volumes, feature warps, or suitable losses for depth. Without these specialized components or the regression framework, general generative techniques may be ill-equipped and vulnerable to both generalization and performance issues.

In this paper, we show that these concerns, while valid, can be addressed and that, surprisingly, a generic, conventional diffusion model for image to image translation works impressively well on both tasks, often outperforming the state of the art. In addition, diffusion models provide valuable benefits over networks trained with regression; in particular, diffusion allows for approximate inference with multi-modal distributions, capturing uncertainty and ambiguity (*e.g.* see Figure 1).

One key barrier to training useful diffusion models for monocular depth and optical flow inference concerns the amount and quality of available training data. Given the limited availability of labelled

*DF is also affiliated with the University of Toronto and the Vector Institute.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Key contributions:

- Merge four large synthetic flow dataset for pre-training: FlyingThings3D, AutoFlow, Kubric, TartanAir

The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, David J. Fleet*

{srbs,irwinherrmann,junhwahur,abhiskar,deqingsun,davidfleet}@google.com

Google DeepMind and Google Research

Abstract

Denoising diffusion probabilistic models have transformed image generation with their impressive fidelity and diversity. We show that they also excel in estimating optical flow and monocular depth, surprisingly, without task-specific architectures and loss functions that are predominant for these tasks. Compared to the point estimates of conventional regression-based methods, diffusion models also enable Monte Carlo inference, e.g., capturing uncertainty and ambiguity in flow and depth. With self-supervised pre-training, the combined use of synthetic and real data for supervised training, and technical innovations (infilling and step-unrolled denoising diffusion training) to handle noisy-incomplete training data, and a simple form of coarse-to-fine refinement, one can train state-of-the-art diffusion models for depth and optical flow estimation. Extensive experiments focus on quantitative performance against benchmarks, ablations, and the model's ability to capture uncertainty and multimodality, and impute missing values. Our model, DDVM (Denoising Diffusion Vision Model), obtains a state-of-the-art relative depth error of 0.074 on the indoor NYU benchmark and an Fl-all outlier rate of 3.26% on the KITTI optical flow benchmark, about 25% better than the best published method. For an overview see diffusion-vision.github.io

1 Introduction

Diffusion models have emerged as powerful generative models for high fidelity image synthesis, capturing rich knowledge about the visual world [21, 48, 55, 62]. However, at first glance, it is unclear whether these models can be as effective on many classical computer vision tasks. For example, consider two dense vision estimation tasks, namely, optical flow, which estimates frame-to-frame correspondences, and monocular depth perception, which makes depth predictions based on a single image. Both tasks are usually treated as regression problems and addressed with specialized architectures and task-specific loss functions, *e.g.*, cost volumes, feature warps, or suitable losses for depth. Without these specialized components or the regression framework, general generative techniques may be ill-equipped and vulnerable to both generalization and performance issues.

In this paper, we show that these concerns, while valid, can be addressed and that, surprisingly, a generic, conventional diffusion model for image to image translation works impressively well on both tasks, often outperforming the state of the art. In addition, diffusion models provide valuable benefits over networks trained with regression; in particular, diffusion allows for approximate inference with multi-modal distributions, capturing uncertainty and ambiguity (*e.g.* see Figure 1).

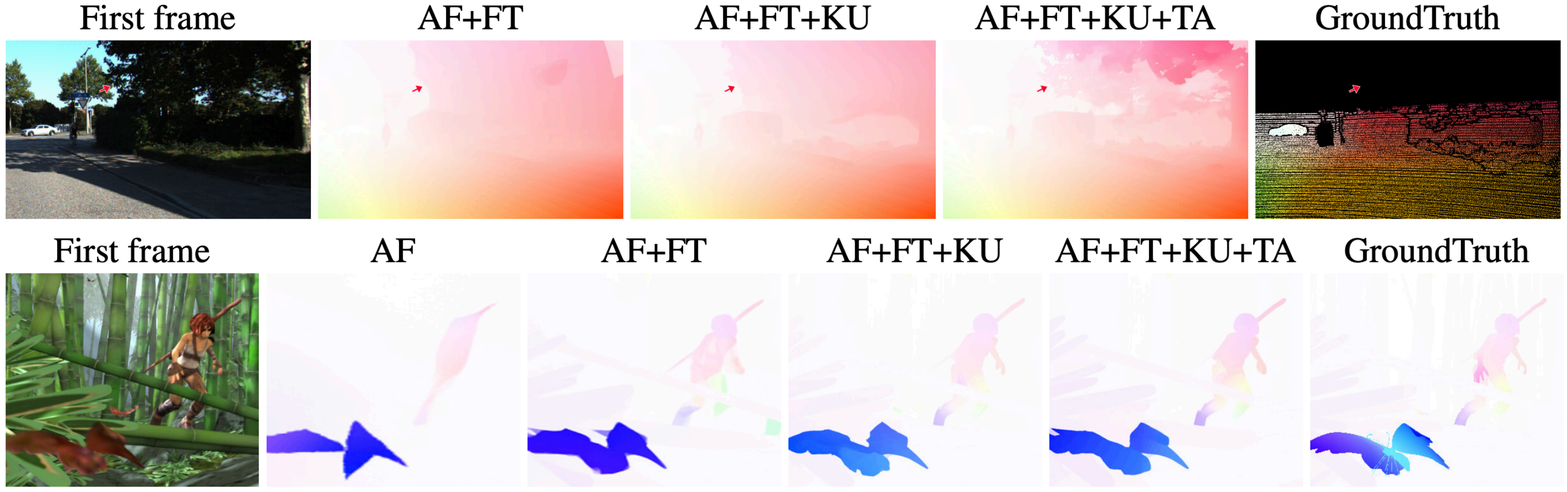
One key barrier to training useful diffusion models for monocular depth and optical flow inference concerns the amount and quality of available training data. Given the limited availability of labelled

*DF is also affiliated with the University of Toronto and the Vector Institute.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Key contributions:

- Merge four large synthetic flow dataset for pre-training: FlyingThings3D, AutoFlow, Kubric, TartanAir



The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, David J. Fleet*

{srbs,irwinherrmann,junhwahur,abhiskar,deqingsun,davidfleet}@google.com

Google DeepMind and Google Research

Abstract

Denoising diffusion probabilistic models have transformed image generation with their impressive fidelity and diversity. We show that they also excel in estimating optical flow and monocular depth, surprisingly, without task-specific architectures and loss functions that are predominant for these tasks. Compared to the point estimates of conventional regression-based methods, diffusion models also enable Monte Carlo inference, e.g., capturing uncertainty and ambiguity in flow and depth. With self-supervised pre-training, the combined use of synthetic and real data for supervised training, and technical innovations (infilling and step-unrolled denoising diffusion training) to handle noisy-incomplete training data, and a simple form of coarse-to-fine refinement, one can train state-of-the-art diffusion models for depth and optical flow estimation. Extensive experiments focus on quantitative performance against benchmarks, ablations, and the model's ability to capture uncertainty and multimodality, and impute missing values. Our model, DDVM (Denoising Diffusion Vision Model), obtains a state-of-the-art relative depth error of 0.074 on the indoor NYU benchmark and an Fl-all outlier rate of 3.26% on the KITTI optical flow benchmark, about 25% better than the best published method. For an overview see diffusion-vision.github.io

1 Introduction

Diffusion models have emerged as powerful generative models for high fidelity image synthesis, capturing rich knowledge about the visual world [21, 48, 55, 62]. However, at first glance, it is unclear whether these models can be as effective on many classical computer vision tasks. For example, consider two dense vision estimation tasks, namely, optical flow, which estimates frame-to-frame correspondences, and monocular depth perception, which makes depth predictions based on a single image. Both tasks are usually treated as regression problems and addressed with specialized architectures and task-specific loss functions, *e.g.*, cost volumes, feature warps, or suitable losses for depth. Without these specialized components or the regression framework, general generative techniques may be ill-equipped and vulnerable to both generalization and performance issues.

In this paper, we show that these concerns, while valid, can be addressed and that, surprisingly, a generic, conventional diffusion model for image to image translation works impressively well on both tasks, often outperforming the state of the art. In addition, diffusion models provide valuable benefits over networks trained with regression; in particular, diffusion allows for approximate inference with multi-modal distributions, capturing uncertainty and ambiguity (*e.g.* see Figure 1).

One key barrier to training useful diffusion models for monocular depth and optical flow inference concerns the amount and quality of available training data. Given the limited availability of labelled

*DF is also affiliated with the University of Toronto and the Vector Institute.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Key contributions:

- Merge four large synthetic flow dataset: FlyingThings3D, AutoFlow, Kubric, TartanAir
- Training with real data is tricky: Has holes! Diffusion model would generate optical flow maps with holes as well…

The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, David J. Fleet*

{srbs,irwinherrmann,junhwahur,abhiskar,deqingsun,davidfleet}@google.com

Google DeepMind and Google Research

Abstract

Denoising diffusion probabilistic models have transformed image generation with their impressive fidelity and diversity. We show that they also excel in estimating optical flow and monocular depth, surprisingly, without task-specific architectures and loss functions that are predominant for these tasks. Compared to the point estimates of conventional regression-based methods, diffusion models also enable Monte Carlo inference, e.g., capturing uncertainty and ambiguity in flow and depth. With self-supervised pre-training, the combined use of synthetic and real data for supervised training, and technical innovations (infilling and step-unrolled denoising diffusion training) to handle noisy-incomplete training data, and a simple form of coarse-to-fine refinement, one can train state-of-the-art diffusion models for depth and optical flow estimation. Extensive experiments focus on quantitative performance against benchmarks, ablations, and the model's ability to capture uncertainty and multimodality, and impute missing values. Our model, DDVM (Denoising Diffusion Vision Model), obtains a state-of-the-art relative depth error of 0.074 on the indoor NYU benchmark and an Fl-all outlier rate of 3.26% on the KITTI optical flow benchmark, about 25% better than the best published method. For an overview see diffusion-vision.github.io

1 Introduction

Diffusion models have emerged as powerful generative models for high fidelity image synthesis, capturing rich knowledge about the visual world [21, 48, 55, 62]. However, at first glance, it is unclear whether these models can be as effective on many classical computer vision tasks. For example, consider two dense vision estimation tasks, namely, optical flow, which estimates frame-to-frame correspondences, and monocular depth perception, which makes depth predictions based on a single image. Both tasks are usually treated as regression problems and addressed with specialized architectures and task-specific loss functions, *e.g.*, cost volumes, feature warps, or suitable losses for depth. Without these specialized components or the regression framework, general generative techniques may be ill-equipped and vulnerable to both generalization and performance issues.

In this paper, we show that these concerns, while valid, can be addressed and that, surprisingly, a generic, conventional diffusion model for image to image translation works impressively well on both tasks, often outperforming the state of the art. In addition, diffusion models provide valuable benefits over networks trained with regression; in particular, diffusion allows for approximate inference with multi-modal distributions, capturing uncertainty and ambiguity (*e.g.* see Figure 1).

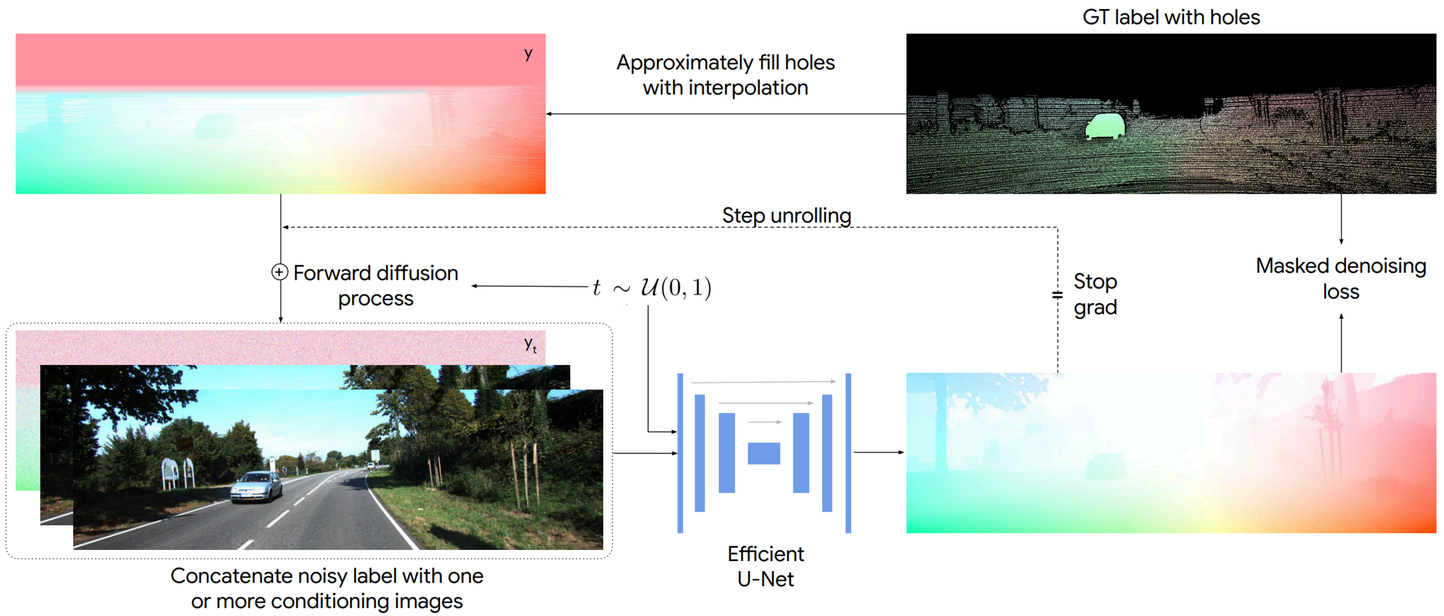
One key barrier to training useful diffusion models for monocular depth and optical flow inference concerns the amount and quality of available training data. Given the limited availability of labelled

*DF is also affiliated with the University of Toronto and the Vector Institute.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Key contributions:

- Merge four large synthetic flow dataset: FlyingThings3D, AutoFlow, Kubric, TartanAir
- Training with real data is tricky: Has holes! Diffusion model would generate optical flow maps with holes as well…



The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, David J. Fleet*

{srbs,irwinherrmann,junhwahur,abhiskar,deqingsun,davidfleet}@google.com

Google DeepMind and Google Research

Abstract

Denoising diffusion probabilistic models have transformed image generation with their impressive fidelity and diversity. We show that they also excel in estimating optical flow and monocular depth, surprisingly, without task-specific architectures and loss functions that are predominant for these tasks. Compared to the point estimates of conventional regression-based methods, diffusion models also enable Monte Carlo inference, e.g., capturing uncertainty and ambiguity in flow and depth. With self-supervised pre-training, the combined use of synthetic and real data for supervised training, and technical innovations (infilling and step-unrolled denoising diffusion training) to handle noisy-incomplete training data, and a simple form of coarse-to-fine refinement, one can train state-of-the-art diffusion models for depth and optical flow estimation. Extensive experiments focus on quantitative performance against benchmarks, ablations, and the model's ability to capture uncertainty and multimodality, and impute missing values. Our model, DDVM (Denoising Diffusion Vision Model), obtains a state-of-the-art relative depth error of 0.074 on the indoor NYU benchmark and an Fl-all outlier rate of 3.26% on the KITTI optical flow benchmark, about 25% better than the best published method. For an overview see diffusion-vision.github.io

1 Introduction

Diffusion models have emerged as powerful generative models for high fidelity image synthesis, capturing rich knowledge about the visual world [21, 48, 55, 62]. However, at first glance, it is unclear whether these models can be as effective on many classical computer vision tasks. For example, consider two dense vision estimation tasks, namely, optical flow, which estimates frame-to-frame correspondences, and monocular depth perception, which makes depth predictions based on a single image. Both tasks are usually treated as regression problems and addressed with specialized architectures and task-specific loss functions, *e.g.*, cost volumes, feature warps, or suitable losses for depth. Without these specialized components or the regression framework, general generative techniques may be ill-equipped and vulnerable to both generalization and performance issues.

In this paper, we show that these concerns, while valid, can be addressed and that, surprisingly, a generic, conventional diffusion model for image to image translation works impressively well on both tasks, often outperforming the state of the art. In addition, diffusion models provide valuable benefits over networks trained with regression; in particular, diffusion allows for approximate inference with multi-modal distributions, capturing uncertainty and ambiguity (*e.g.* see Figure 1).

One key barrier to training useful diffusion models for monocular depth and optical flow inference concerns the amount and quality of available training data. Given the limited availability of labelled

*DF is also affiliated with the University of Toronto and the Vector Institute.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Key contributions:

- Merge four large synthetic flow dataset: FlyingThings3D, AutoFlow, Kubric, TartanAir
- Training with real data is tricky: Has holes! Diffusion model would generate optical flow maps with holes as well…
- Coarse-to-fine training strategy

The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation

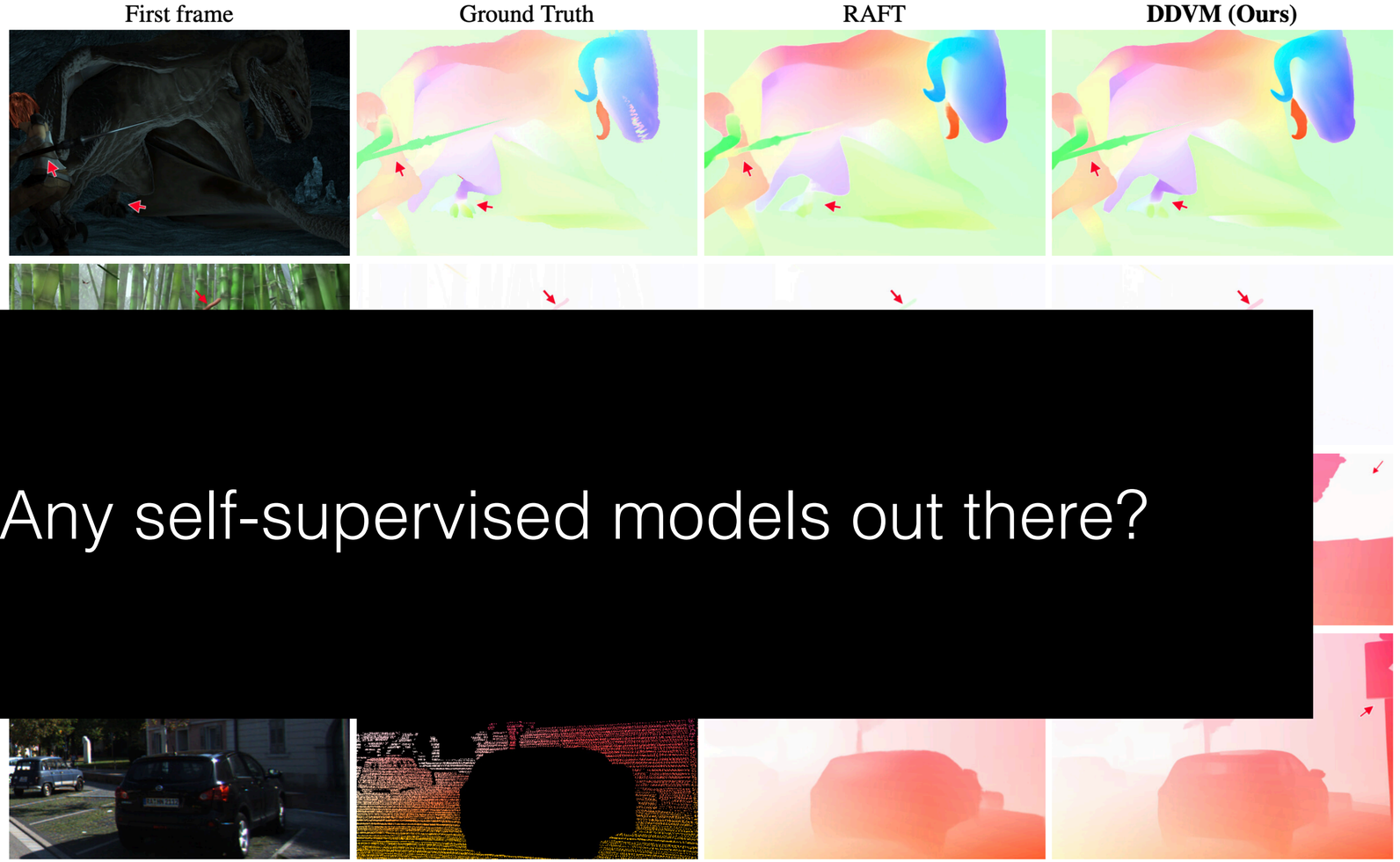
Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, David J. Fleet*

{srbs,irwinherrmann,junhwahur,abhiskar,deqingsun,davidfleet}@google.com

Google DeepMind and Google Research

Abstract

Denoising diffusion probabilistic models have transformed image generation with their impressive fidelity and diversity. We show that they also excel in estimating optical flow and monocular depth, surprisingly, without task-specific architectures and loss functions that are predominant for these tasks. Compared to the point estimates of conventional regression-based methods, diffusion models also enable Monte Carlo inference, e.g., capturing uncertainty and ambiguity in flow and depth.



Fully supervised… Any self-supervised models out there?

for depth. Without these specialized components or the regression framework, general generative techniques may be ill-equipped and vulnerable to both generalization and performance issues.

In this paper, we show that these concerns, while valid, can be addressed and that, surprisingly, a generic, conventional diffusion model for image to image translation works impressively well on both tasks, often outperforming the state of the art. In addition, diffusion models provide valuable benefits over networks trained with regression; in particular, diffusion allows for approximate inference with multi-modal distributions, capturing uncertainty and ambiguity (*e.g.* see Figure 1).

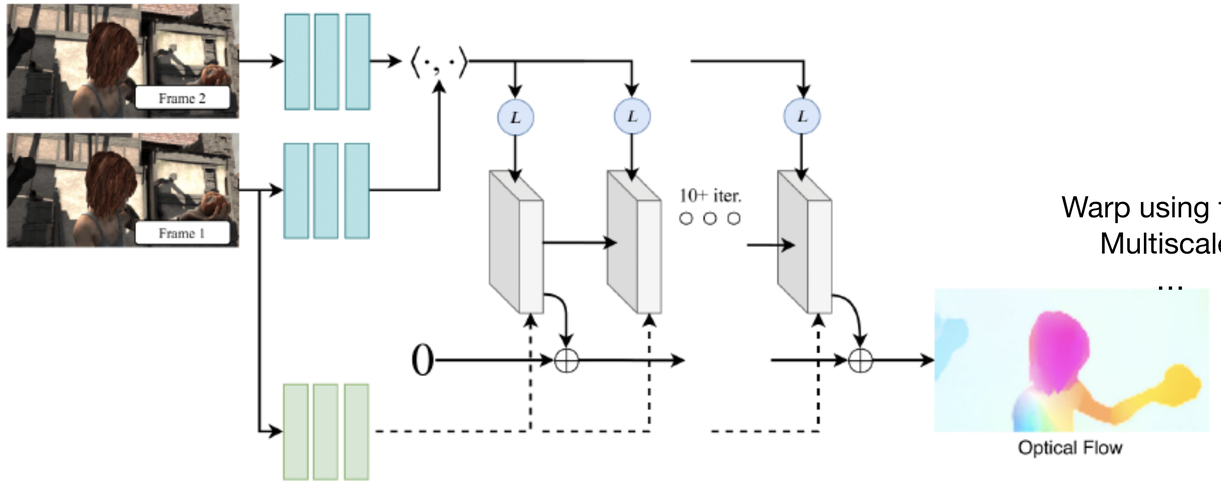
One key barrier to training useful diffusion models for monocular depth and optical flow inference concerns the amount and quality of available training data. Given the limited availability of labelled

*DF is also affiliated with the University of Toronto and the Vector Institute.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Concepts of Self-Supervised Optical Flow

[Stone, Maurer et al. 2021, Jonschkowski et al. 2020, Meister et al. 2018, …]
(see SMURF for good discussion!)



Warp using flow,
Multiscale,
…

$$\sum_{\mathbf{p}} \left| I_t(\mathbf{p}) - I_{t+1}\left(\mathbf{p} + \mathbf{w}_\mathbf{p}\right) \right|^2$$
$$+\lambda \sum_{\mathbf{q} \in N_p} \left| \mathbf{w}_\mathbf{p} - \mathbf{w}_\mathbf{q} \right|^2$$
$$+ \ldots$$

Estimate flow with Neural Network…

…but *supervise* flow with warping-based losses & regularizers!

Teacher-Student for Optical Flow

SMURF: Self-Teaching Multi-Frame Unsupervised RAFT with Full-Image Warping

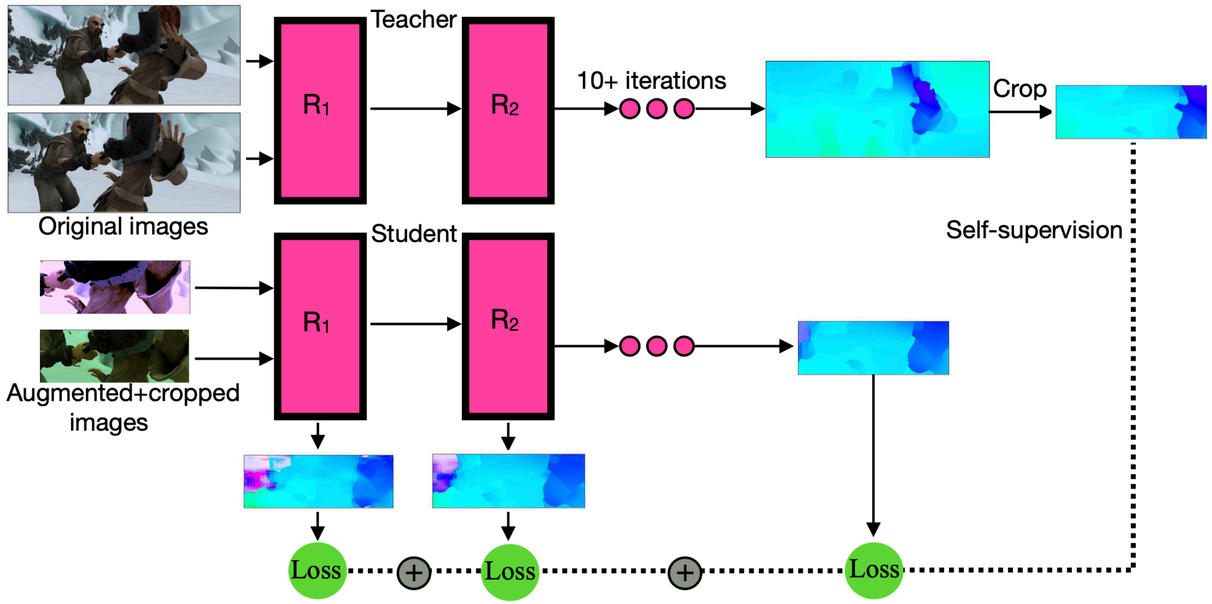
Austin Stone*[1] Daniel Maurer*[2] Alper Ayvaci[2] Anelia Angelova[1] Rico Jonschkowski[1]
[1]Robotics at Google [2]Waymo
{austinstone, anelia, rjon}@google.com {maurerd, ayvaci}@waymo.com
https://github.com/google-research/google-research/tree/master/smurf

Abstract

We present SMURF, a method for unsupervised learning of optical flow that improves state of the art on all benchmarks by 36% to 40% (over the prior best method UFlow) and even outperforms several supervised approaches such as PWC-Net and FlowNet2. Our method integrates architecture improvements from supervised optical flow, i.e. the RAFT model, with new ideas for unsupervised learning that include a sequence-aware self-supervision loss, a technique for handling out-of-frame motion, and an approach for learning effectively from multi-frame video data while still only requiring two frames for inference.

Unsupervised learning is a promising direction to address this issue as it allows training optical flow models from unlabeled videos of any domain. The unsupervised approach works by combining ideas from classical methods and supervised-learning – training the same neural networks as in supervised approaches but optimizing them with objectives such as smoothness and photometric similarity from classical methods. Unlike those classical methods, unsupervised approaches perform optimization not per image pair but jointly for the entire training set.

Since unsupervised optical flow takes inspiration from classical and supervised learning methods, we can make substantial progress by properly combining novel ideas with insights from these two directions. In this paper, we do exactly that and make the following three contributions:



Motivation:

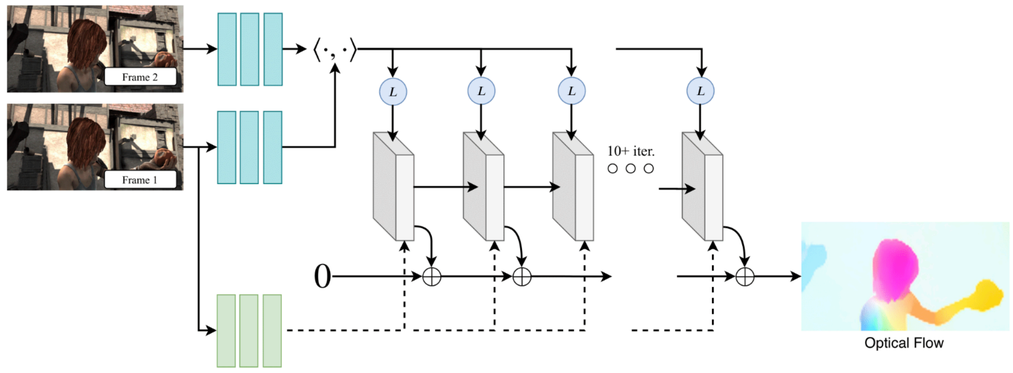
1. The model learns to ignore photometric augmentations
2. The model learns to make better predictions at the borders and in occluded areas of the image (with other tricks, see paper)
3. Early iterations of the recurrent model learn from the output at the final iteration

Two principles of Motion Estimation

Optical Flow

Dense correspondences between a pair of frames

RAFT



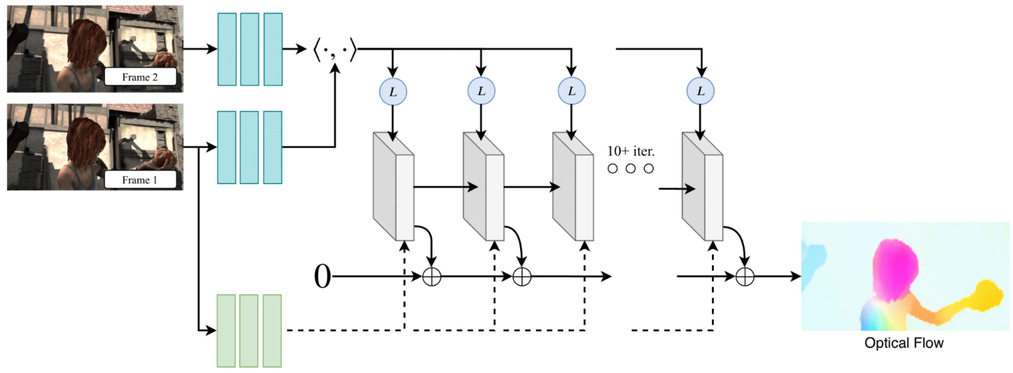
Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022
TAP-Vid: A Benchmark for Tracking Any Point in a Video. Doersch et al., NeurIPS D&B 2022
RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Teed et al. , ECCV 2020

Two principles of Motion Estimation

Optical Flow

Dense correspondences between a pair of frames

RAFT

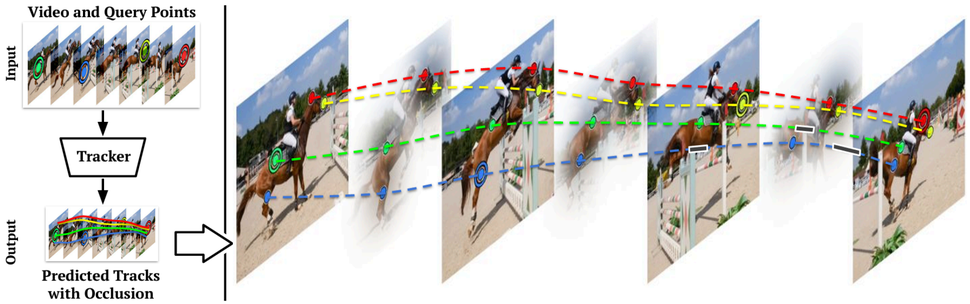


Point Tracking

Long-term tracking of individual points

PIPs

TAP-Net



Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022
TAP-Vid: A Benchmark for Tracking Any Point in a Video. Doersch et al., NeurIPS D&B 2022
RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Teed et al. , ECCV 2020

Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories

Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki

Carnegie Mellon University
{aharley, zhaoyuaf, katef}@cs.cmu.edu
Project page: https://particle-video-revisited.github.io

Abstract. Tracking pixels in videos is typically studied as an optical flow estimation problem, where every pixel is described with a displacement vector that locates it in the next frame. Even though wider temporal context is freely available, prior efforts to take this into account have yielded only small gains over 2-frame methods. In this paper, we revisit Sand and Teller's "particle video" approach, and study pixel tracking as a long-range motion estimation problem, where every pixel is described with a trajectory that locates it in multiple future frames. We re-build this classic approach using components that drive the current state-of-the-art in flow and object tracking, such as dense cost maps, iterative optimization, and learned appearance updates. We train our models using long-range amodal point trajectories mined from existing optical flow data that we synthetically augment with multi-frame occlusions. We test our approach in trajectory estimation benchmarks and in keypoint label propagation tasks, and compare favorably against state-of-the-art optical flow and feature tracking methods.

Particle Video Revisited

Initialize positions and features

(x_1, y_1)

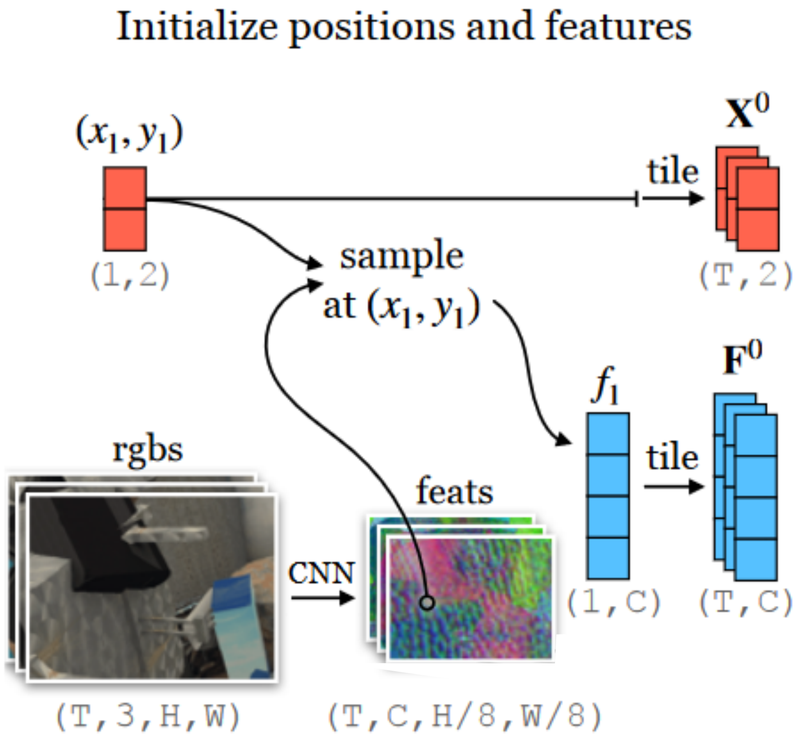
(1,2)

rgbs

(T,3,H,W)

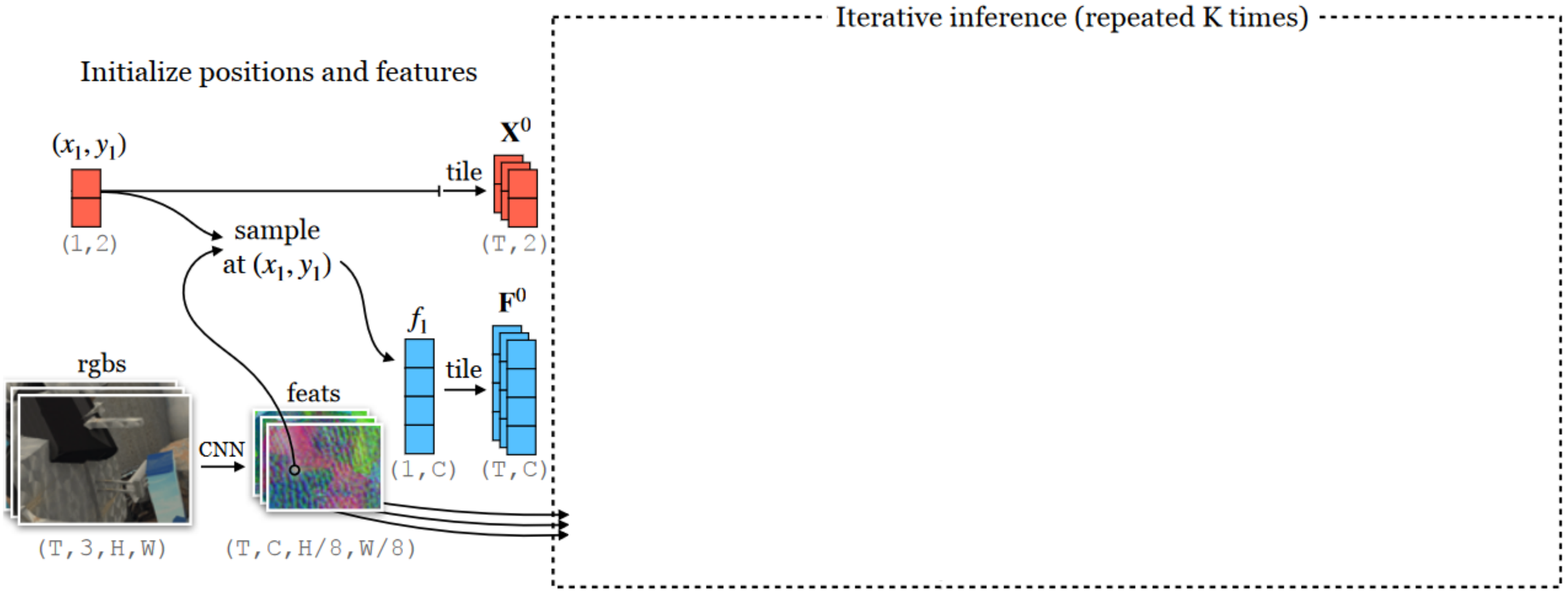
Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022

Particle Video Revisited



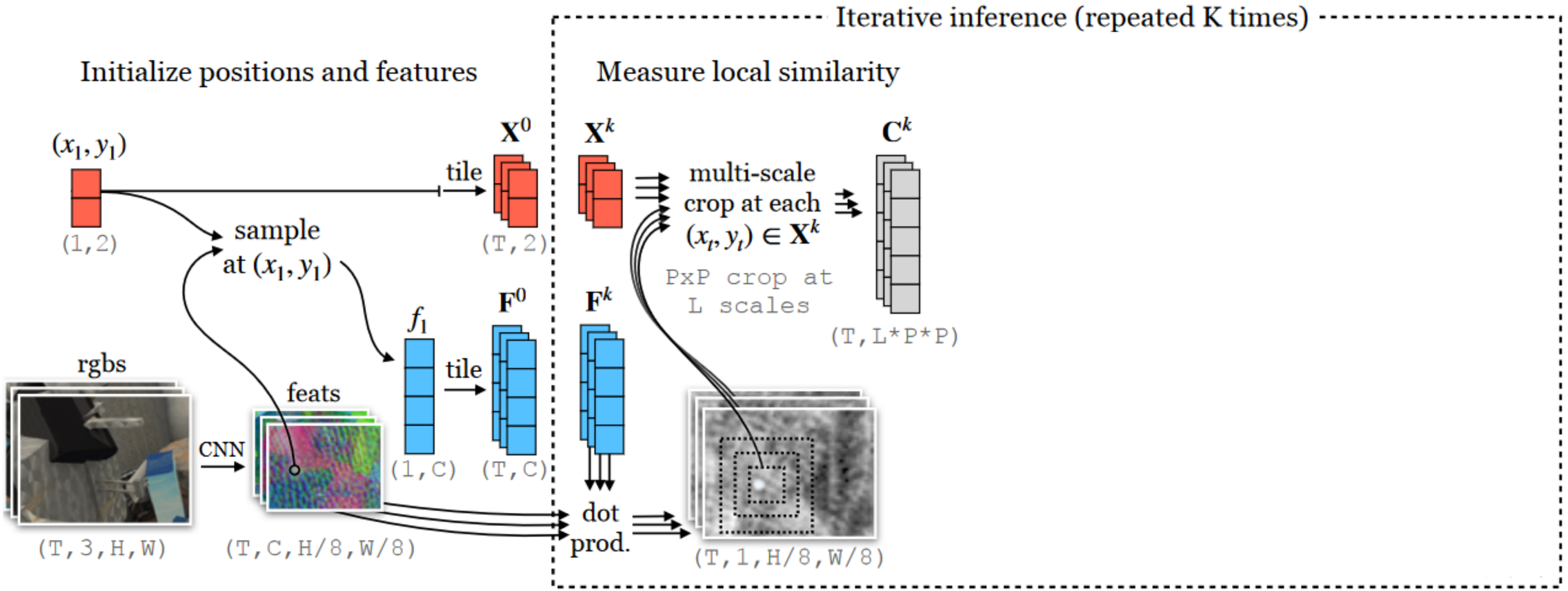
Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022

Particle Video Revisited



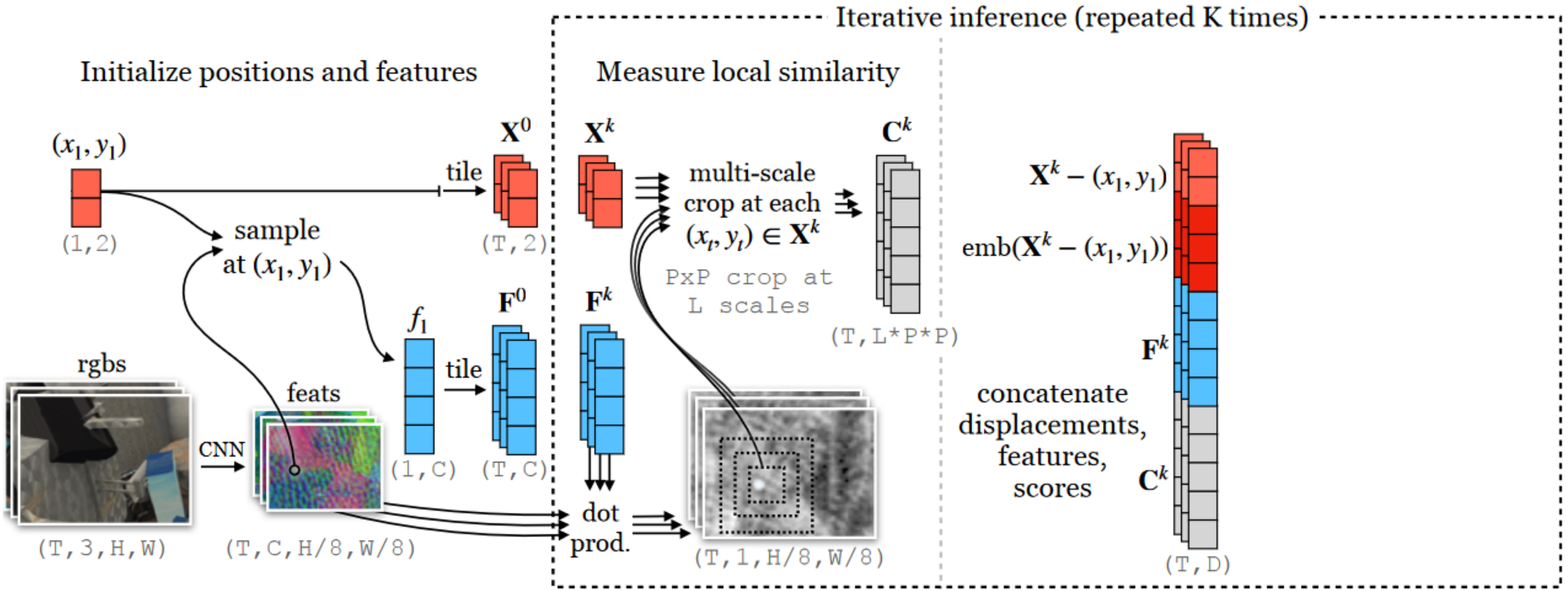
Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022

Particle Video Revisited



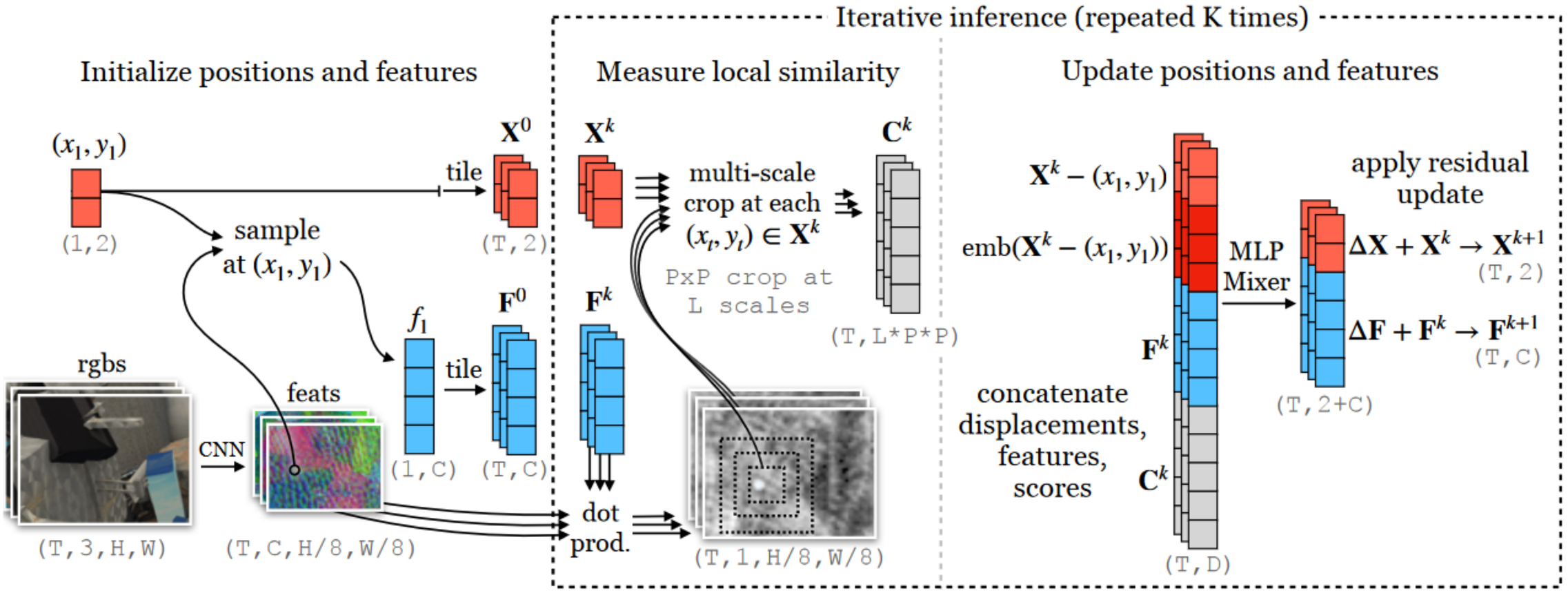
Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022

Particle Video Revisited



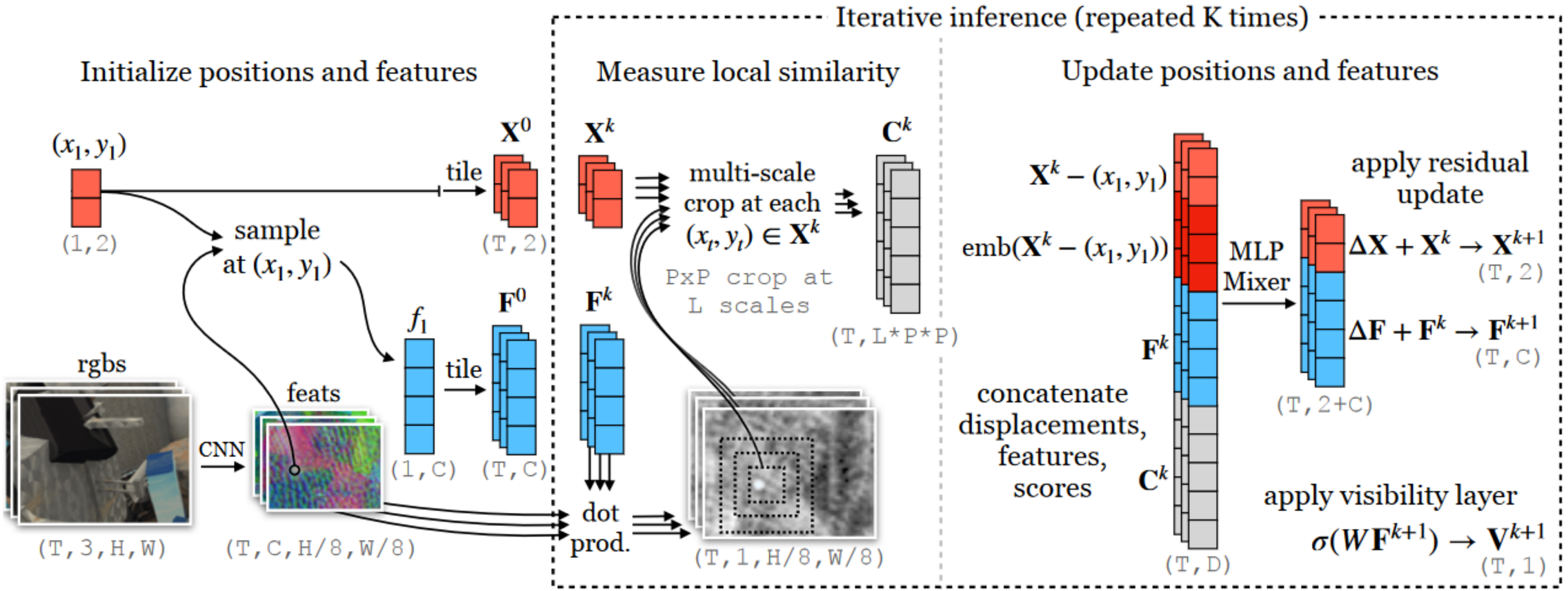
Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022

Particle Video Revisited



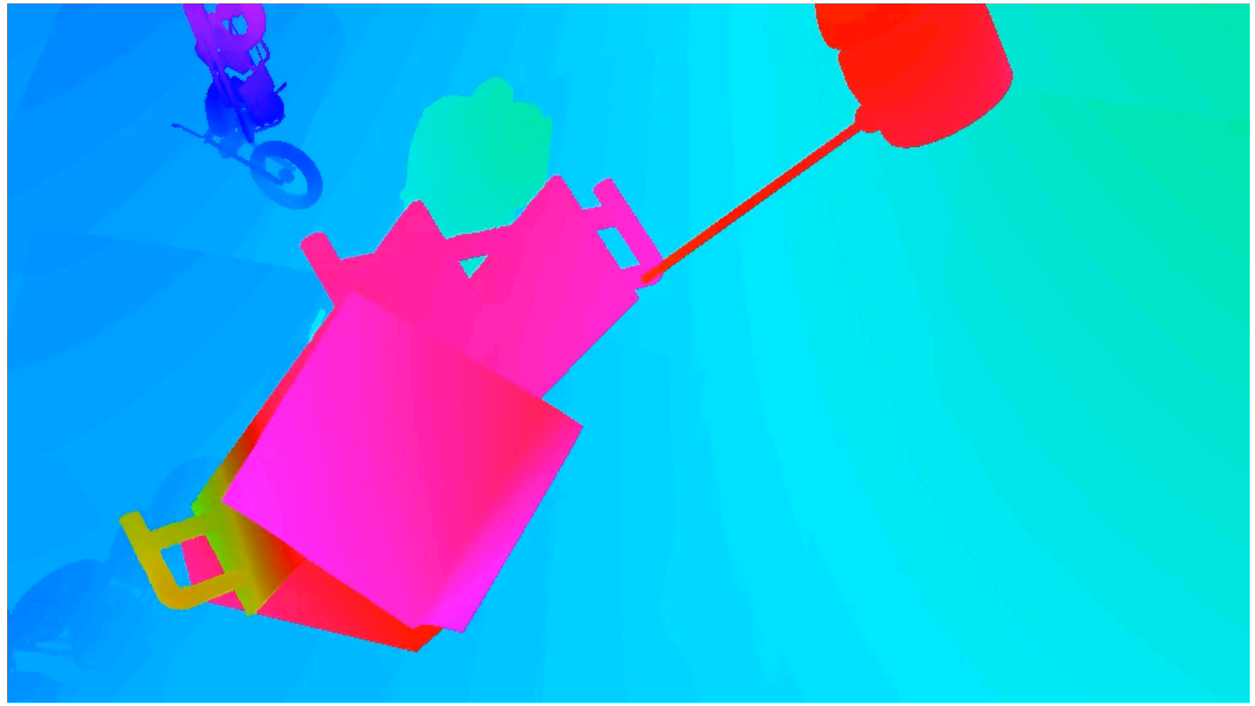
Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022

Particle Video Revisited

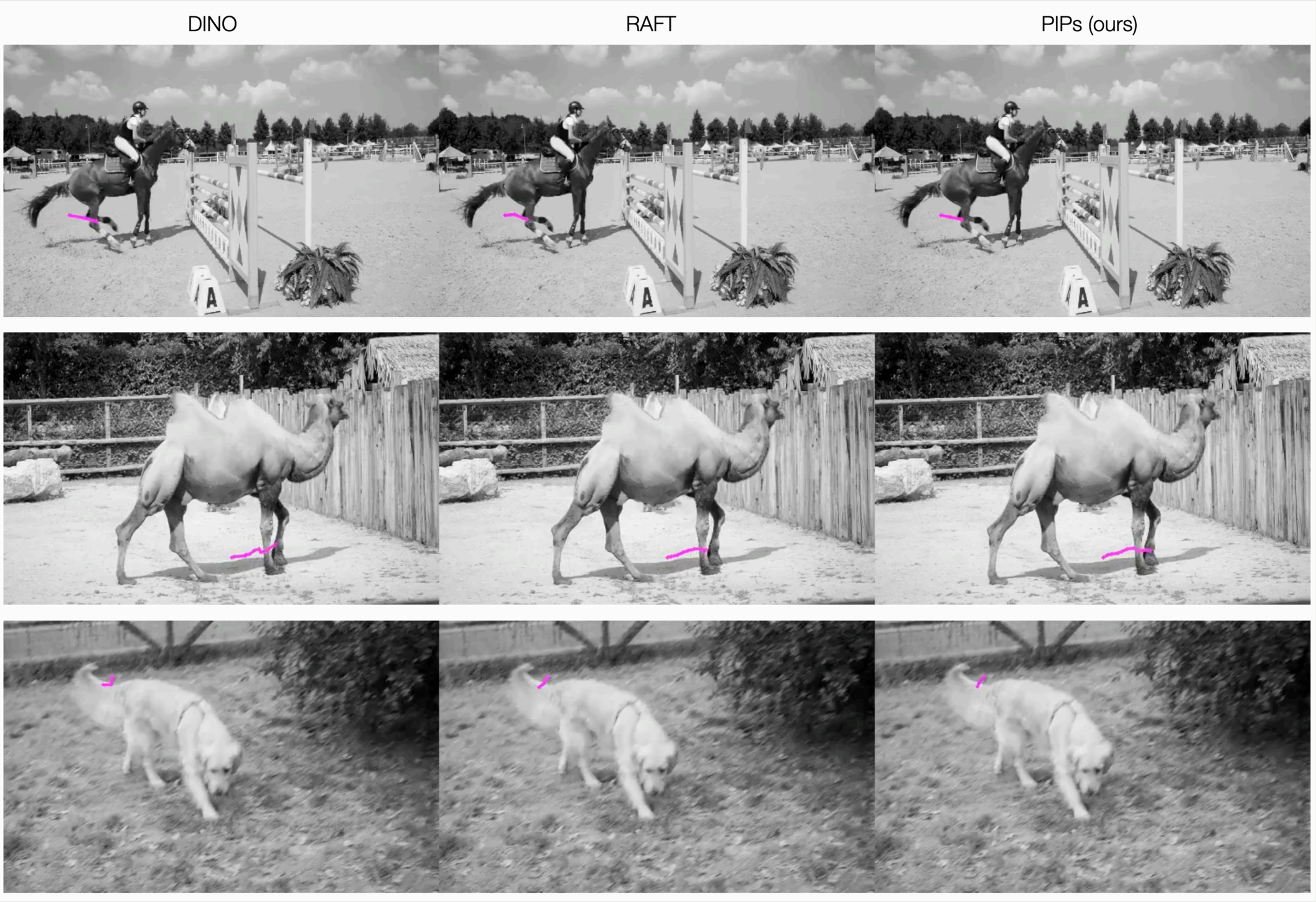


Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. Harley et. al. ECCV 2022

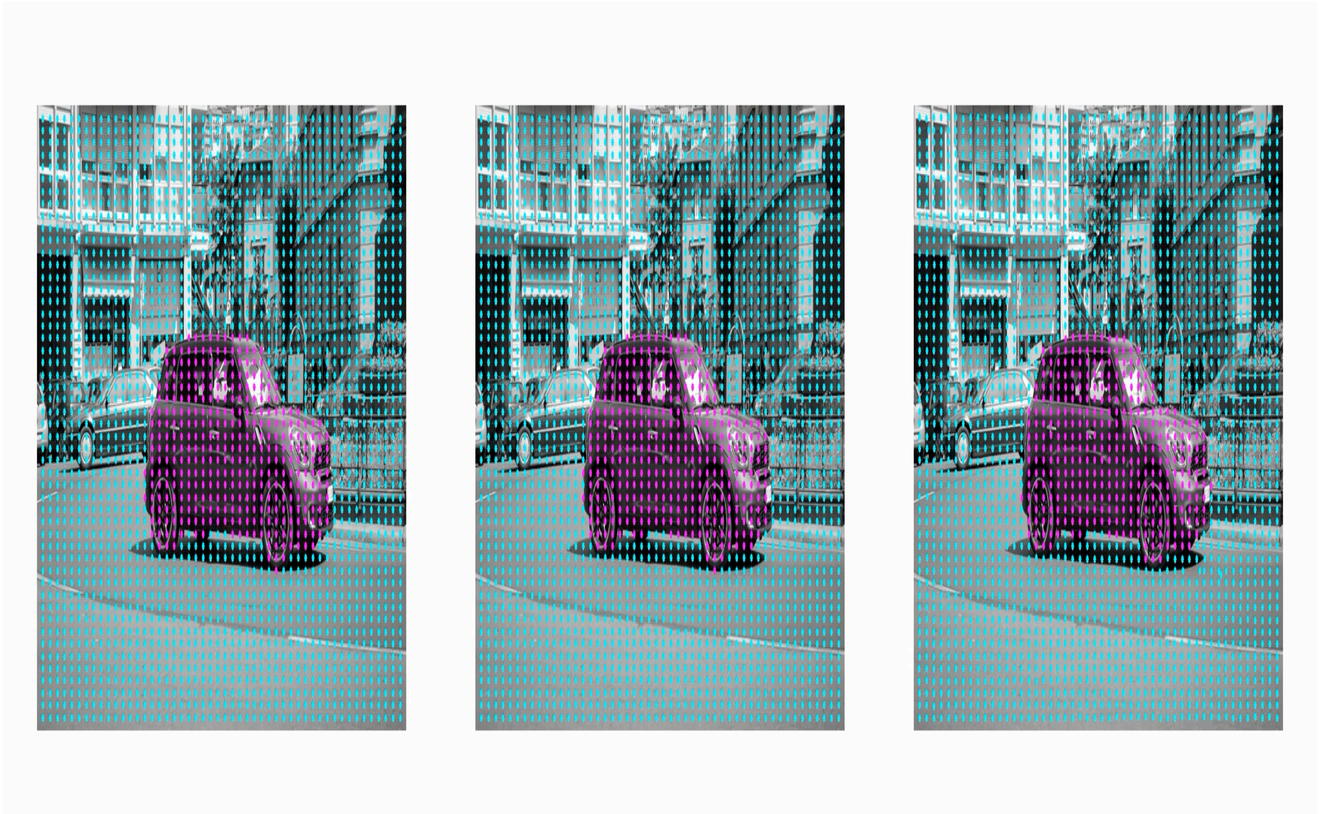
Trained Fully Supervised on Synthetic Data



A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation, Mayer et al., 2016


DINO
RAFT
PIPs (ours)

Single Point Tracking

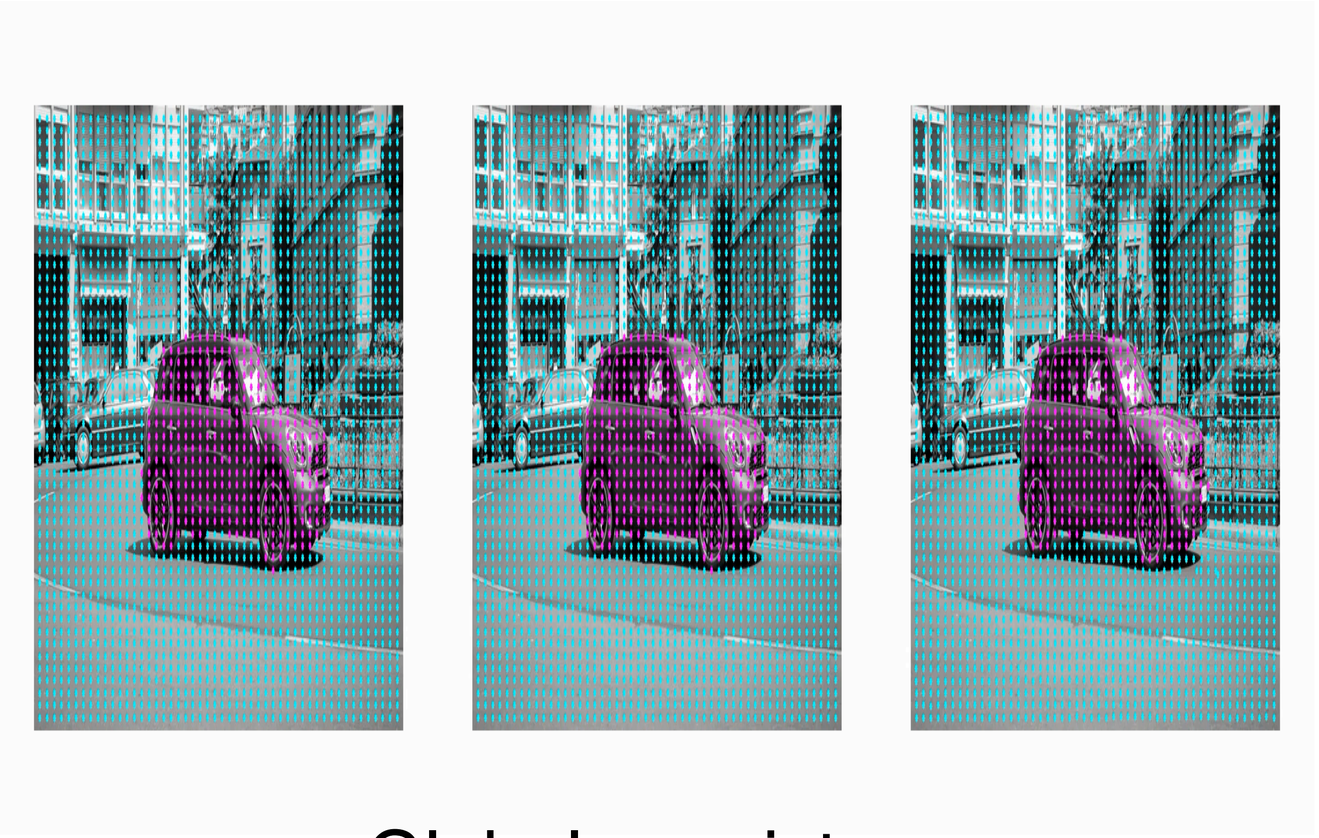


- Background points
- Object points

Tracking single points lacks global consistency

Points drift relative to each other

Tracking with Optical Flow

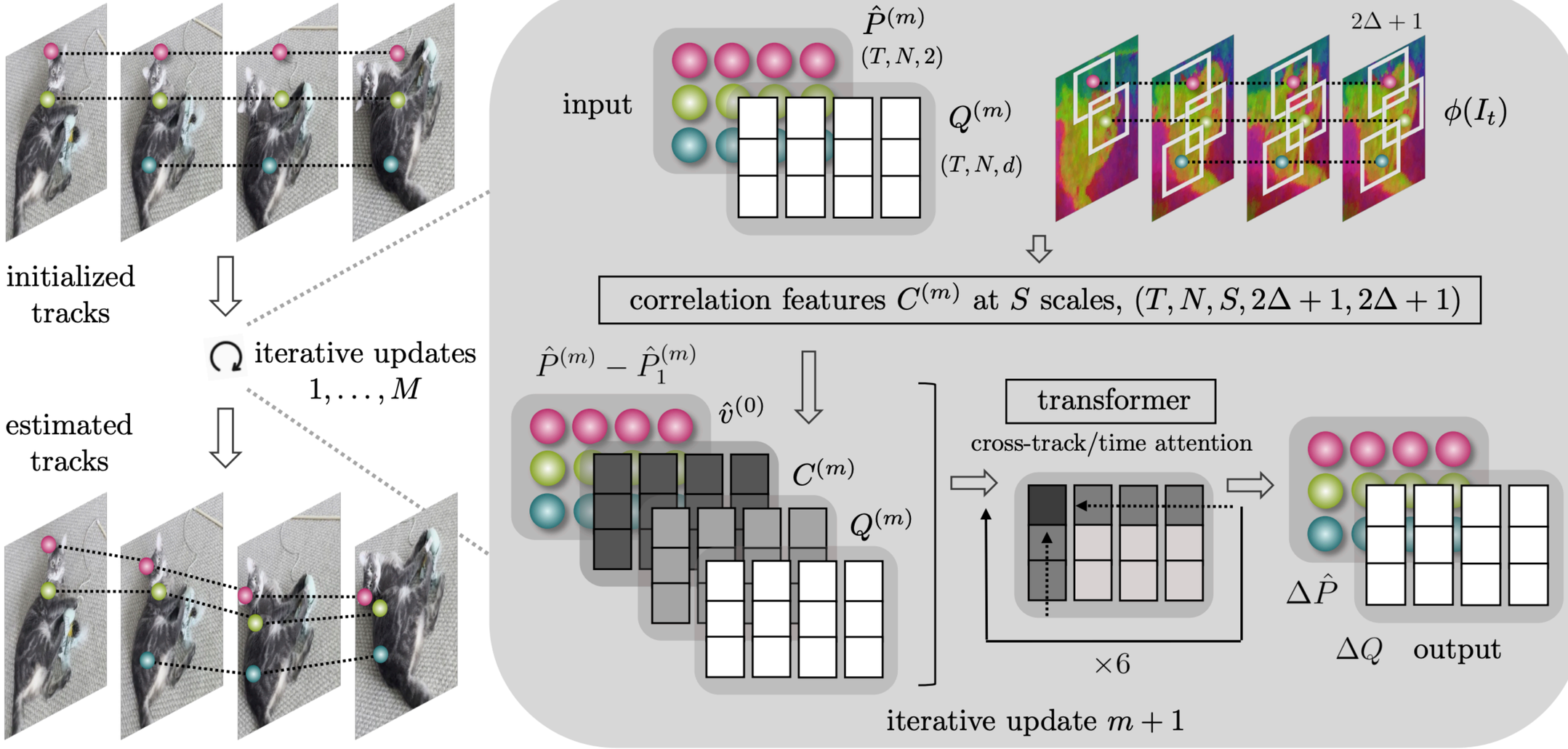


- Background points
- Object points

Global consistency

Occlusion problem: background accumulates on the object

CoTracker: It is Better to Track Together

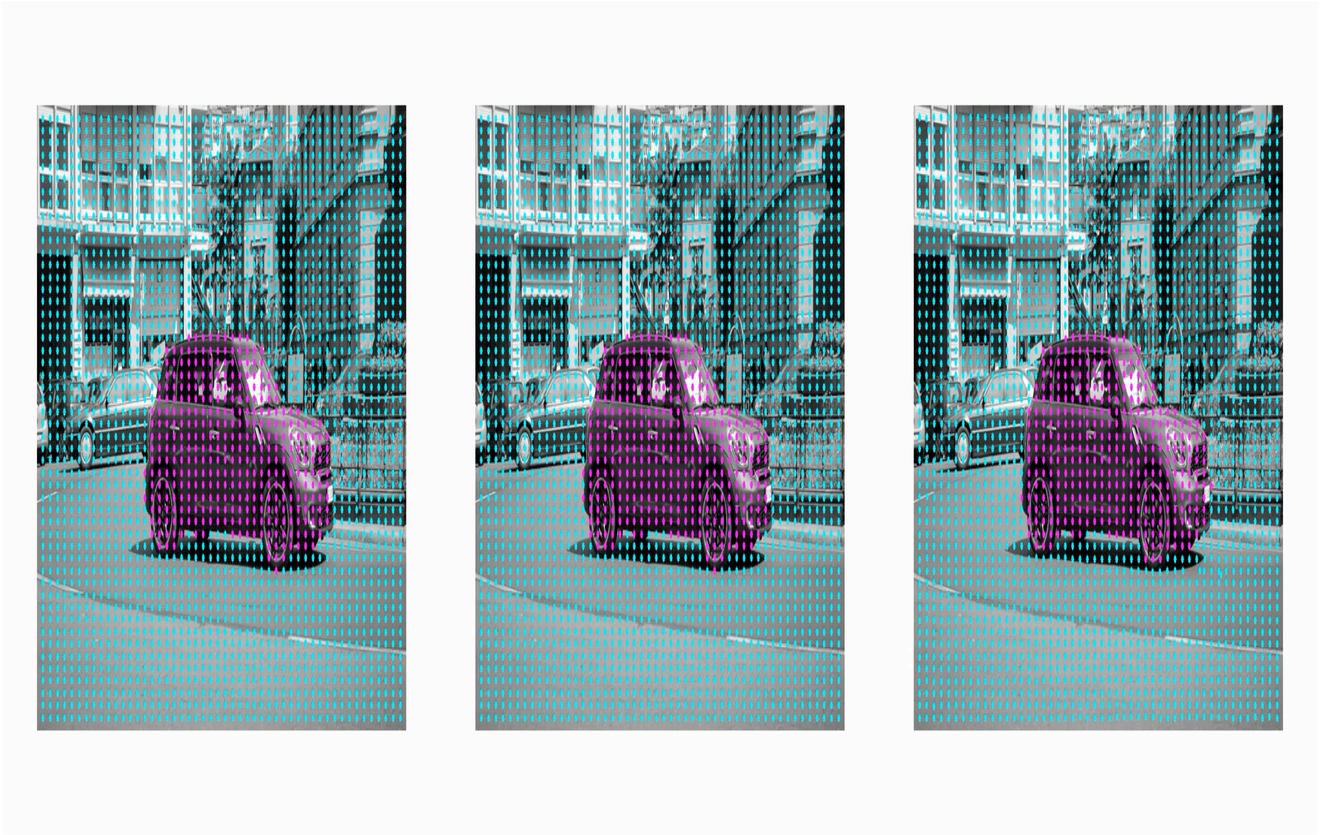


Tracking Points Together

- Global consistency from tracking many points
- Occlusion handling from long-term tracking
- Train on synthetic data only
- Main mechanism:
  - Transformer architecture
  - Factorised attention: space, time, and group (across tracks)

CoTracker: It is Better to Track Together, N Karaev et al., 2023

Tracking Points Together



- Background points
- Object points

- Better global consistency
- Better occlusion handling

Tracking Points Together

Tracking Points Together

Summary

- Discussed two key approaches to motion tracking: Optical flow and point tracking
- Went in-depth on how to think about the motion estimation problem: infinitesimal perspective, finite perspective, cost volumes, smoothness, etc.
- Discussed SOTA supervised optical flow estimators.
- Discussed options for self-supervised methods.
- Discussed point tracking paradigm.