Optical Flow and Motion-Based SSL

2025-02-20

Chris Hoang

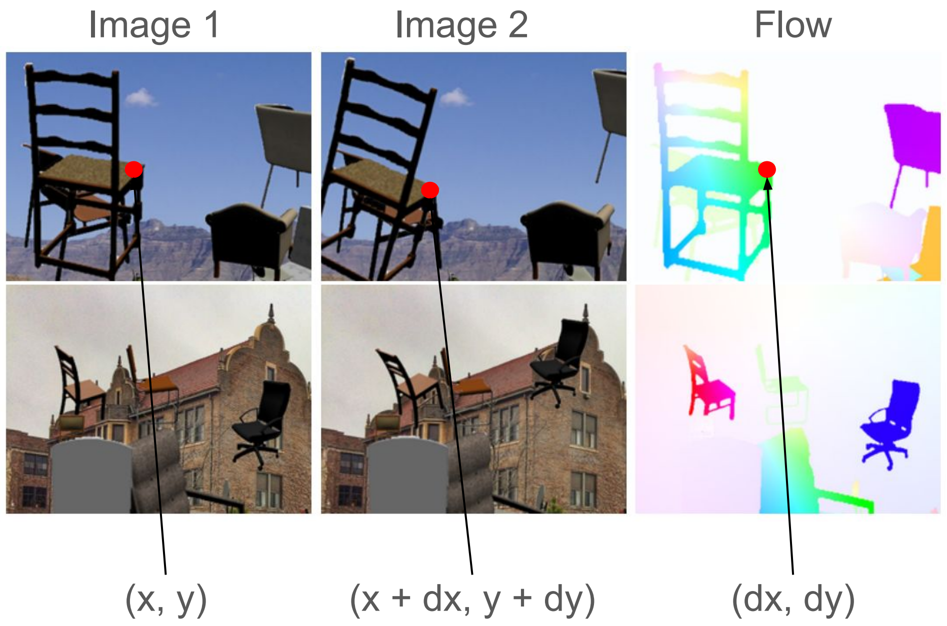
Optical flow example

Optical flow problem

Task: estimate motion of pixels between video frames

Estimate the flow field that contains the motion of each pixel coordinate (x, y) from image1 → image2

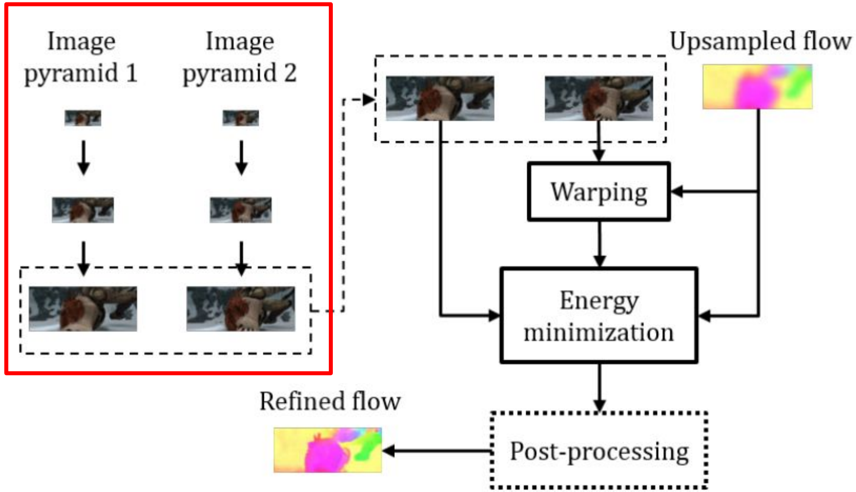
- flow[x, y] = dx, dy
- image1(x, y) ↔ image2(x + dx, y + dy)



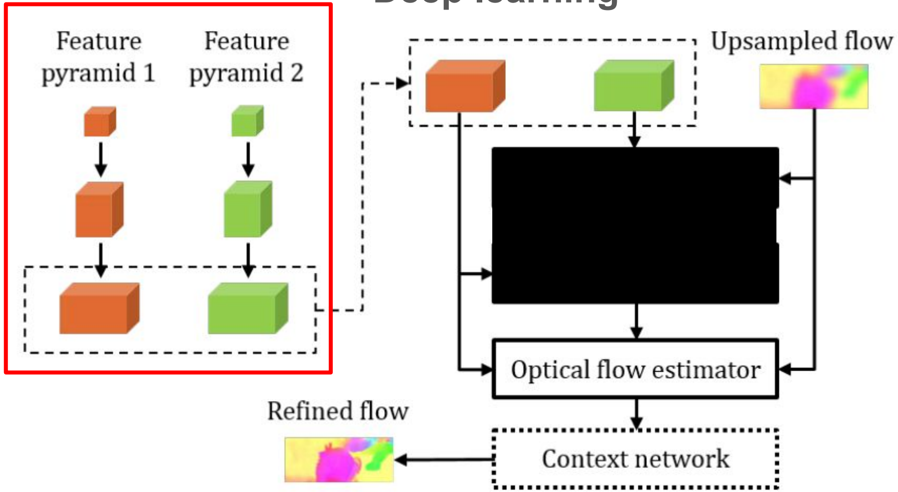
Deep learning for optical flow

Replace image pyramids and hand-crafted features with end-to-end neural networks that produce **feature pyramids**

Classical



Deep learning



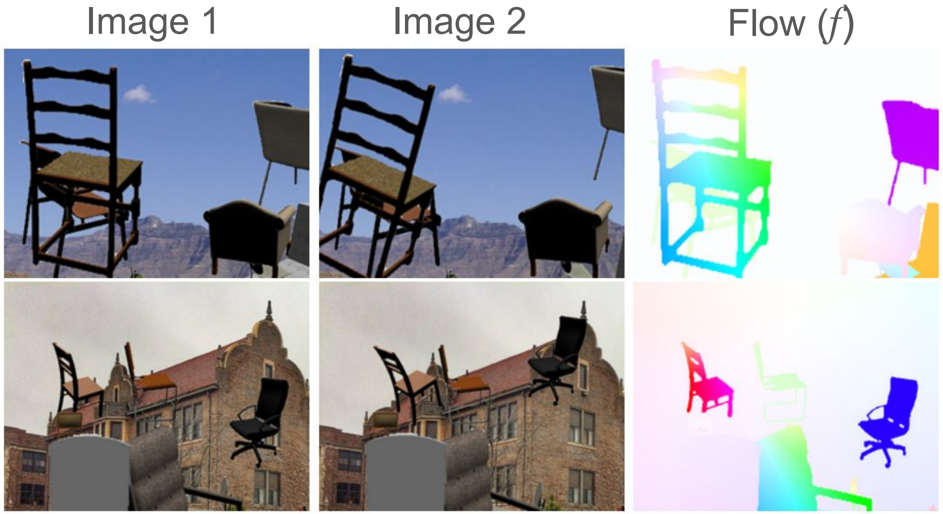
PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. Sun et al. CVPR 2018

Model architectures for optical flow

Suppose we have the features for each image

We are trying to learn how to match the two feature maps

We use correlations between the two features as useful information for flow



Correlation layer

Inputs: two tensors ***u, v*** that are each of dimension H x W x D. For example, features from two images

Output: one tensor ***z*** of dimension H x W x H’ x W’

Correlation layer

Inputs: two tensors ***u, v*** that are each of dimension H x W x D. For example, features from two images

Output: one tensor ***z*** of dimension H x W x H’ x W’

- **H’ < H / W’ < W if we want to only search a local neighborhood for each point**

Intermediate flow predictions

For flow prediction, we can start with predictions to match coarse, high-level features

Refine these predictions to match more fine-grained features

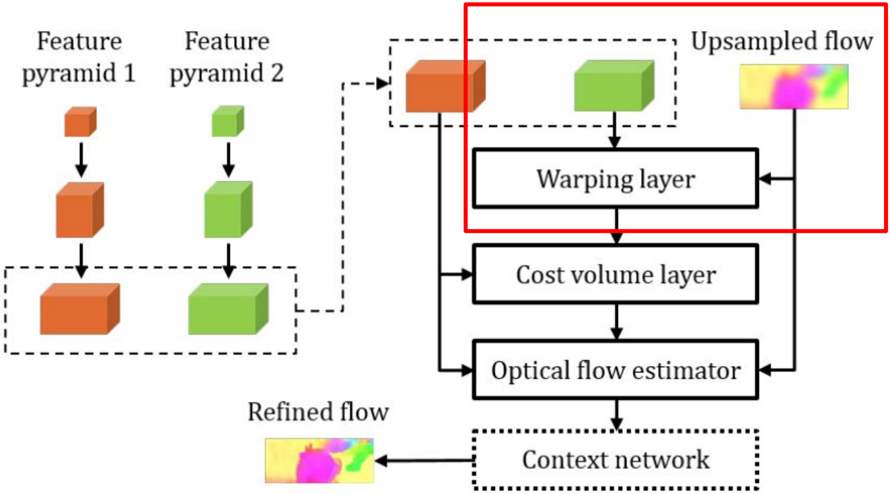
Warp (align) features using flow predictions before computing correlations

Intermediate flow predictions

For flow prediction, we can start with predictions to match coarse, high-level features

Refine these predictions to match more fine-grained features

Warp (align) features using flow predictions before computing correlations



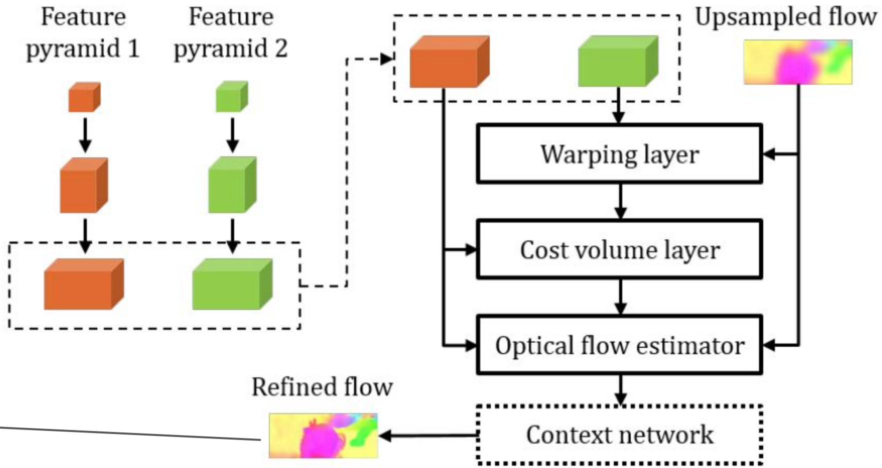
Learning objectives for optical flow models

To train the model to predict flow, we will minimize the error between the model's predicted flow and the ground-truth flow

Error is averaged over entire flow map

Can compute loss over intermediate flow predictions

$$\mathcal{L} = \sum_{i=1}^{N} \gamma^{N-i} ||\mathbf{f}_{gt} - \mathbf{f}_i||_1$$



Datasets for optical flow models

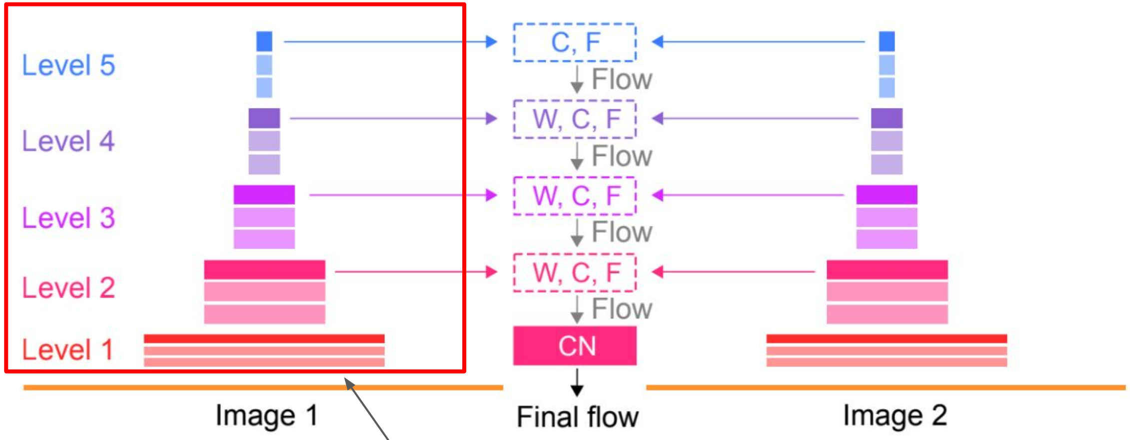
Synthetically rendered datasets

1. FlyingChairs
2. FlyingThings
3. **MPI Sintel**

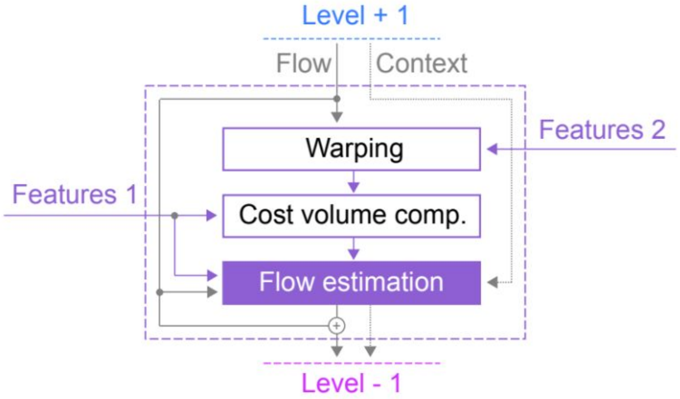
Real datasets

1. KITTI

Optical flow architectures: PWC-Net

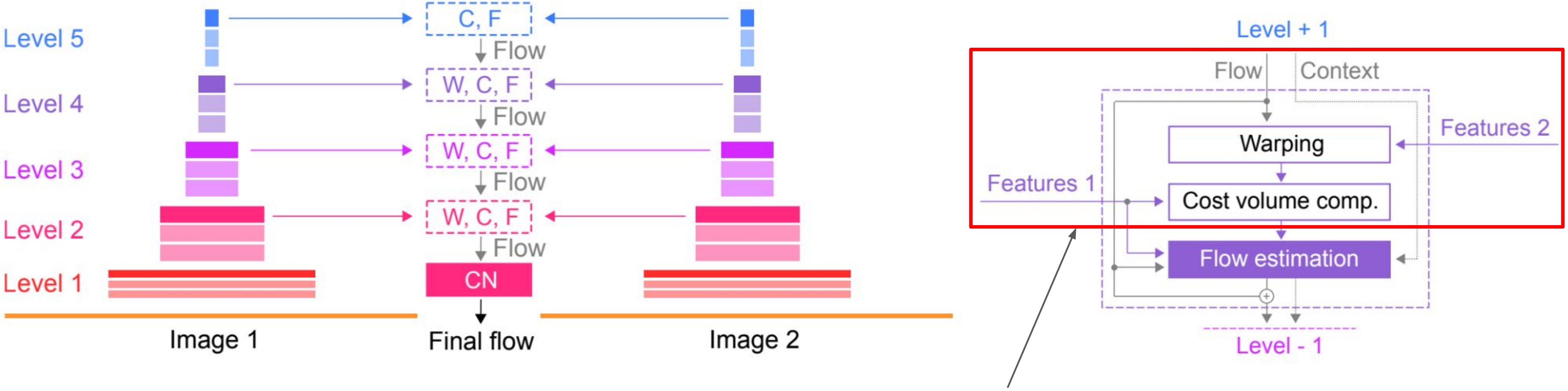


Feature encoder pyramid



PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. Sun et al. CVPR 2018

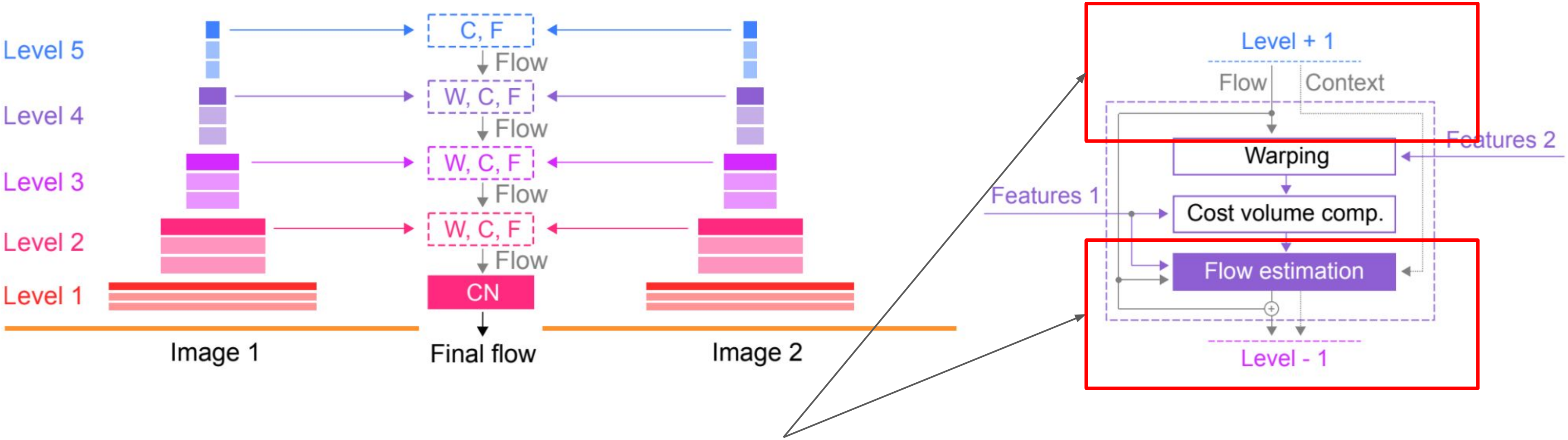
Optical flow architectures: PWC-Net



Cost volume between features

PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. Sun et al. CVPR 2018

Optical flow architectures: PWC-Net



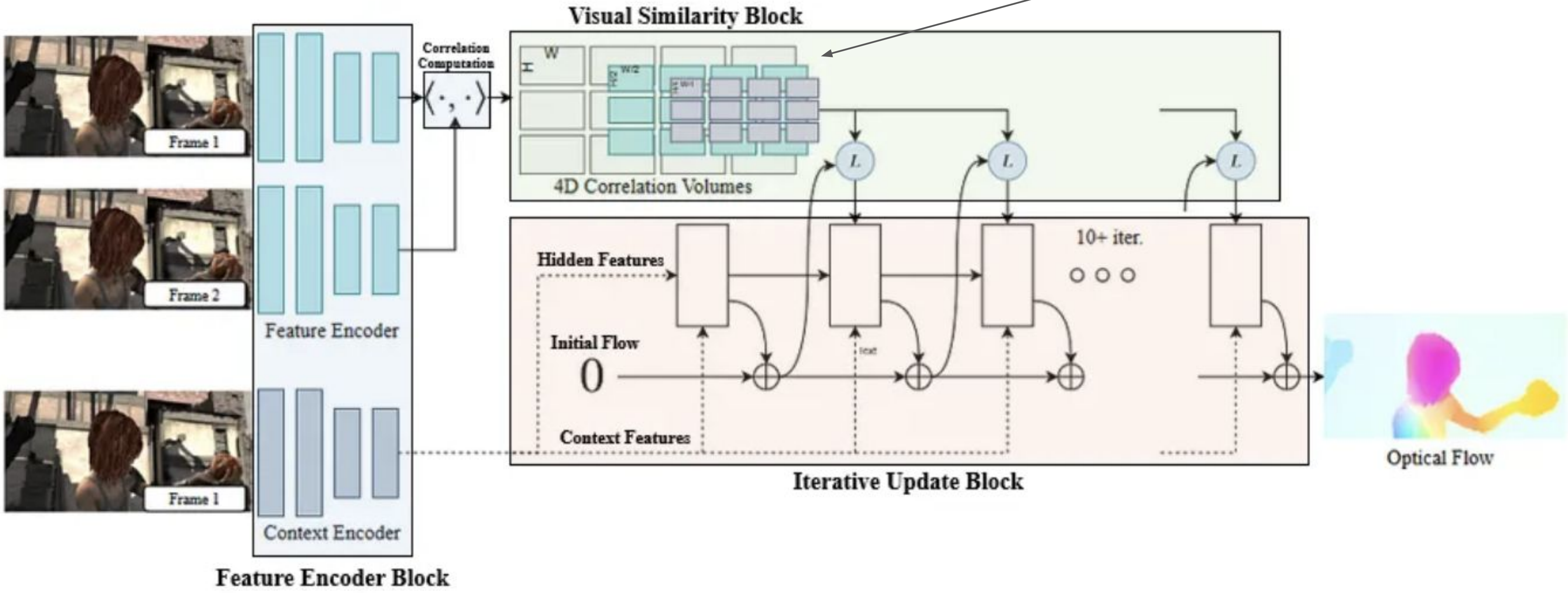
Top-down flow refinement
through feature layers

PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. Sun et al. CVPR 2018

PWC-Net: Jupyter Notebook

Optical flow architectures: RAFT

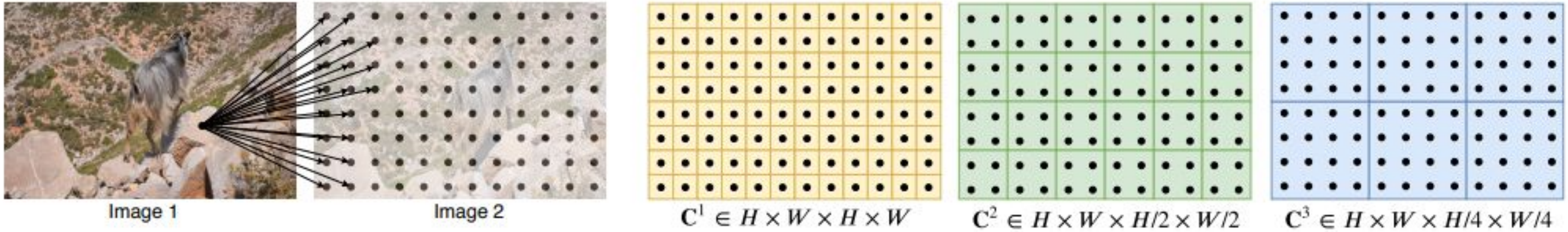
Correlation volumes pooled at different resolutions



RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Teed and Deng. ECCV 2020

Optical flow architectures: RAFT

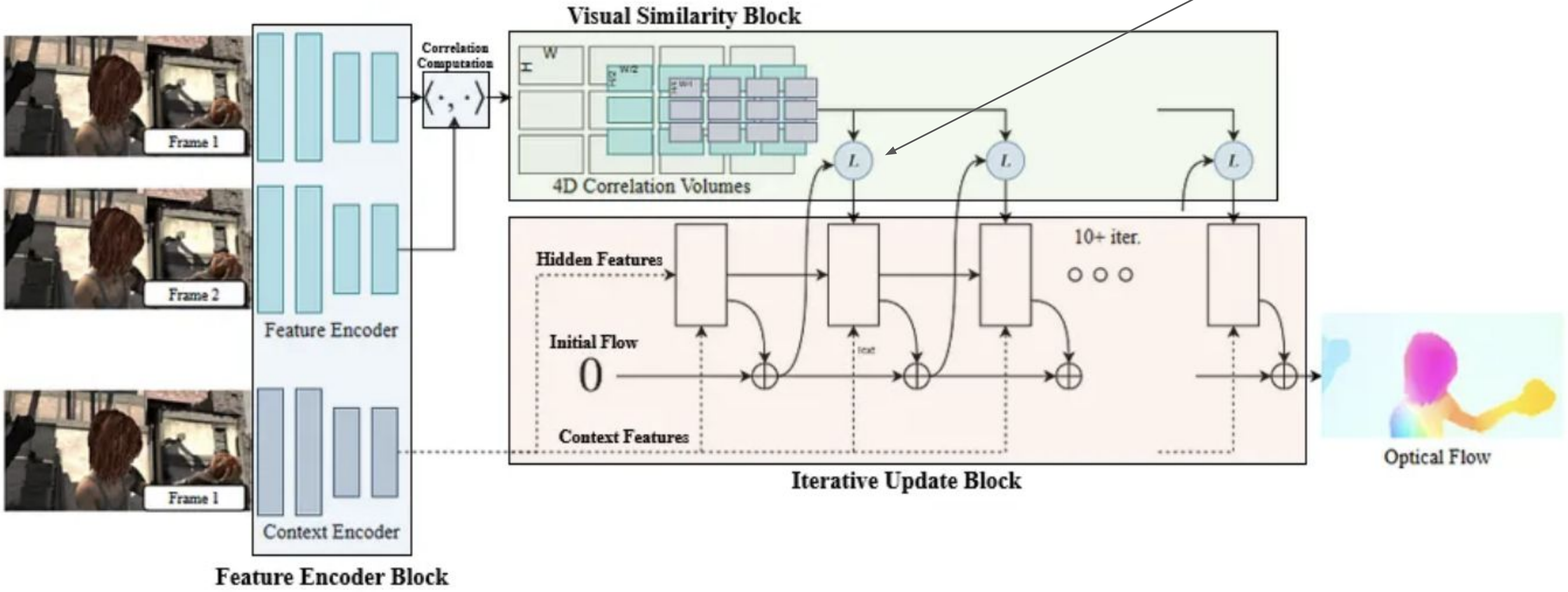
Correlation volumes pooled at different resolutions



RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Teed and Deng. ECCV 2020

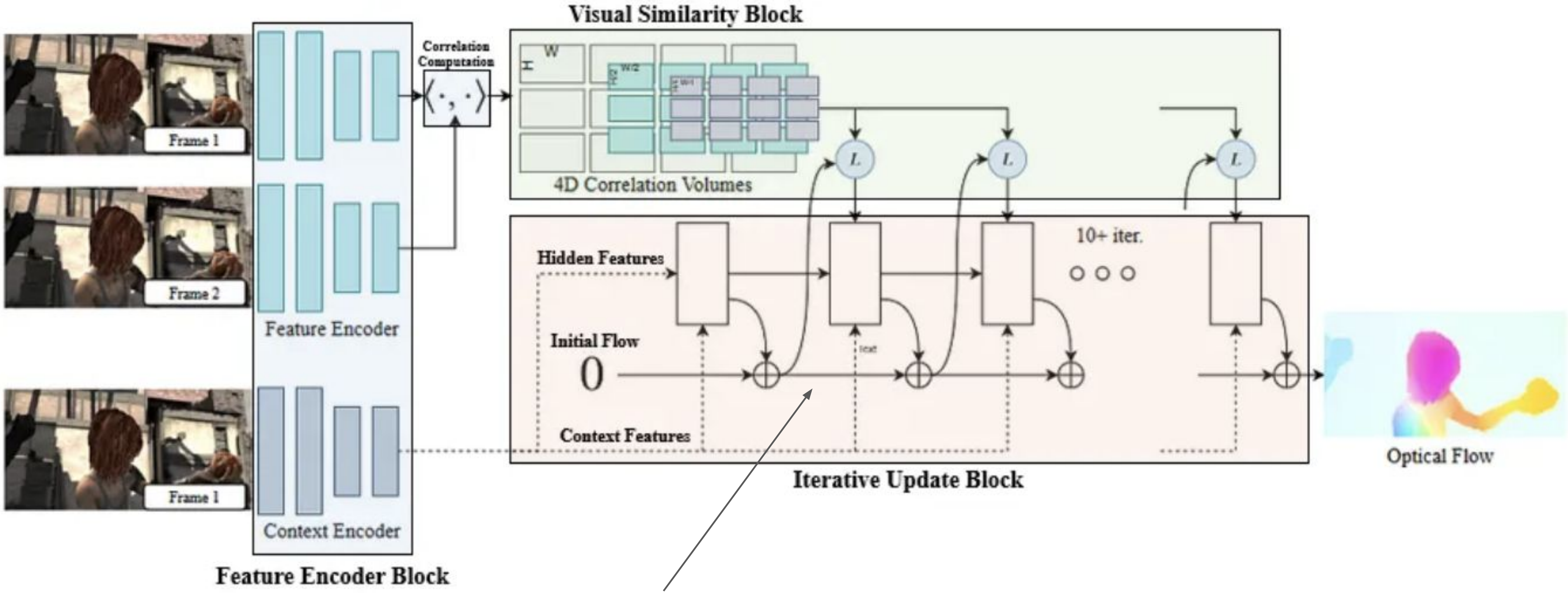
Optical flow architectures: RAFT

Use current flow prediction index into correlation volume



RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Teed and Deng. ECCV 2020

Optical flow architectures: RAFT



Iteratively update using conv + GRU block

RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. Teed and Deng. ECCV 2020

Recent advances: long-range flow estimation

Predict dense + long-range motion

Model is optimized at inference time for an entire video

Quasi-3D volume representation + bijections to map volume to local frames

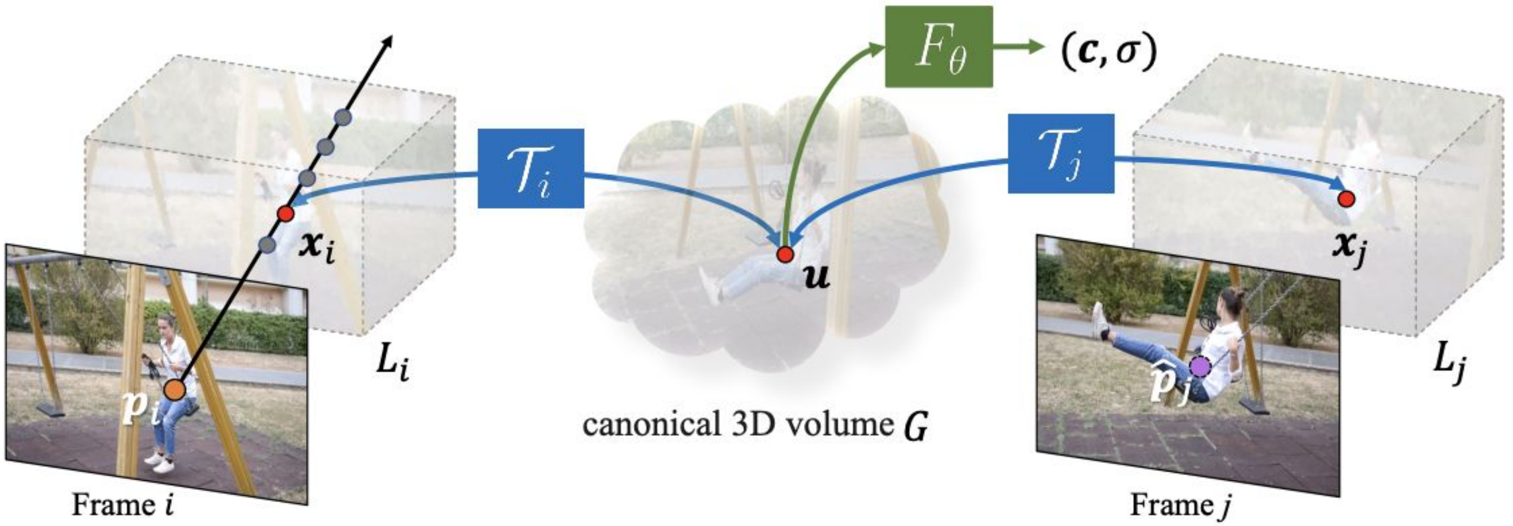
Supervised learning on flow estimates from RAFT

Tracking Everything Everywhere All at Once. Wang et al. ICCV 2023

OmniMotion visualizations

https://omnimotion.github.io/

OmniMotion: Method



(a) OmniMotion representation

Tracking Everything Everywhere All at Once. Wang et al. ICCV 2023

Recent advances: object segmentation with flow

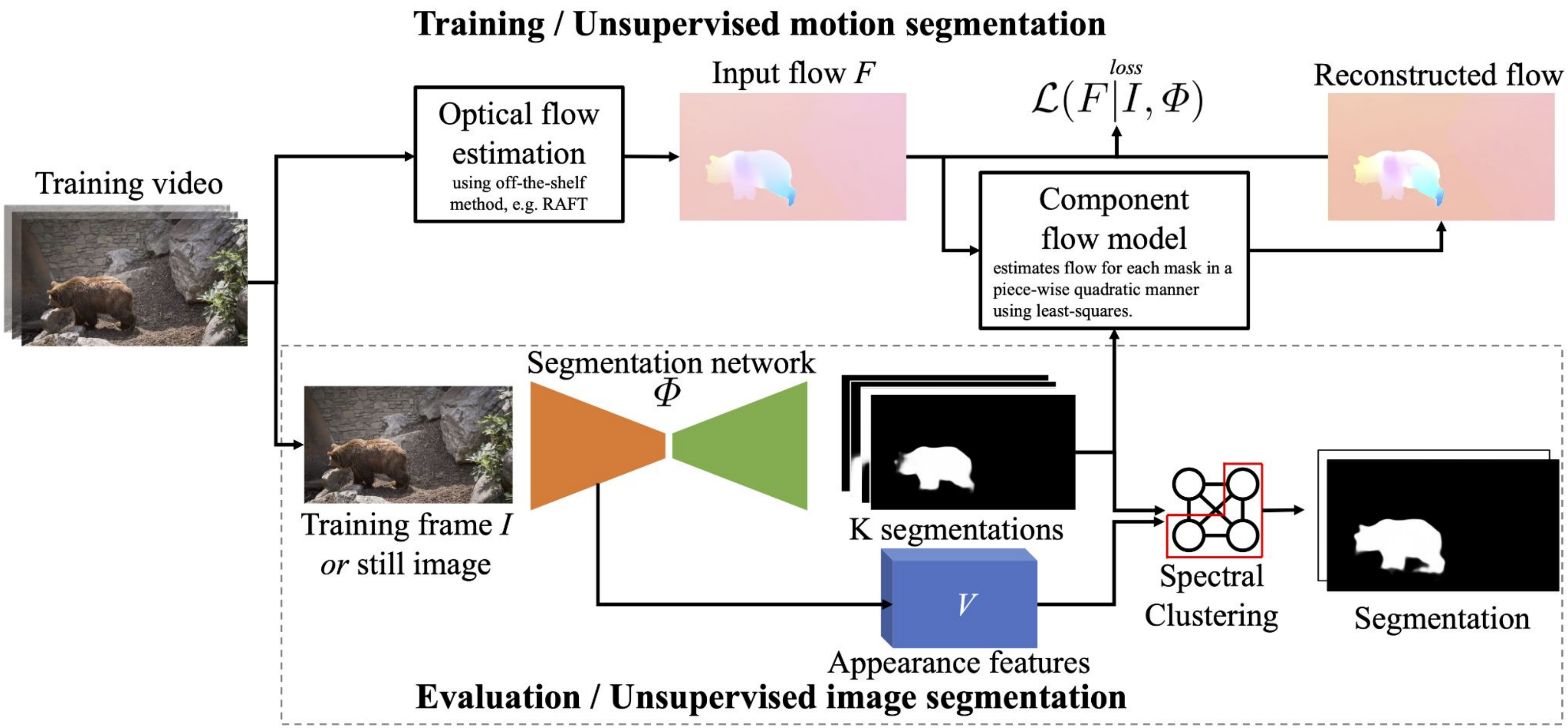
Use optical flow as a cue for objects in images and videos

Have prior notion that we can model the motion of objects with simple parametrics

Learn segmentation network using constrained flow reconstruction as the supervising signal

Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating Motion. Choudhury et al. BMVC 2022

Recent advances: object segmentation with flow



Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating Motion. Choudhury et al. BMVC 2022

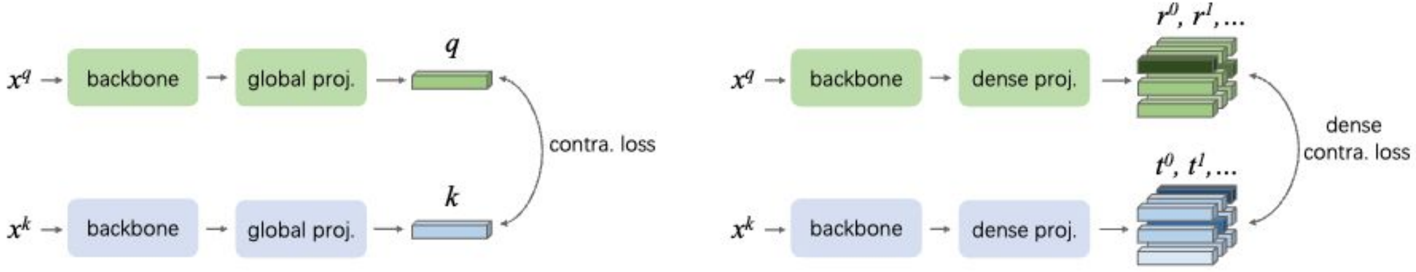
Global versus dense self-supervised learning

Global versus dense self-supervised learning

Joint-embedding SSL: push together embeddings of “positive pairs”

Global SSL: embeddings are single global vectors of images

Dense SSL: embeddings are dense (h x w) local vectors of images



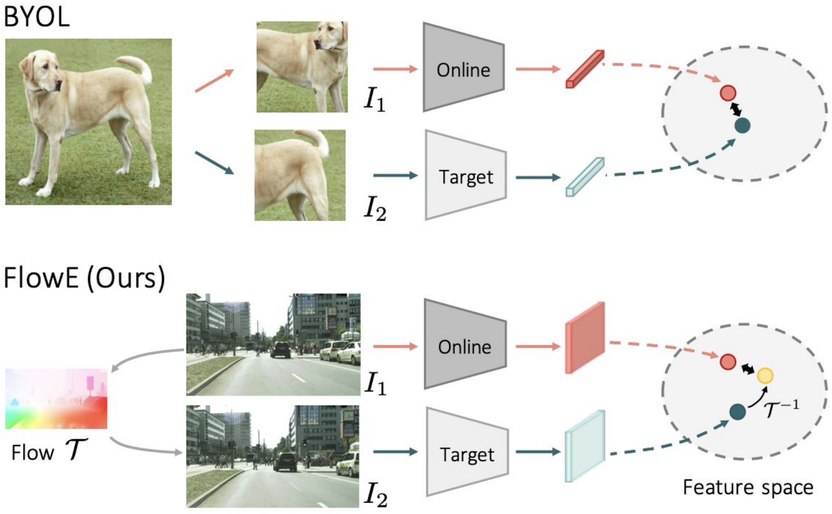
Dense Contrastive Learning for Self-Supervised Visual Pre-Training. Wang et al., 2021.

Flow for dense self-supervised learning

SSL (e.g., BYOL): minimize ***global*** features between different views of an image

1. ***In-the-wild*** data may contain cluttered scenes with ***many objects***

Dense SSL: compare ***dense*** feature maps between different views of an image



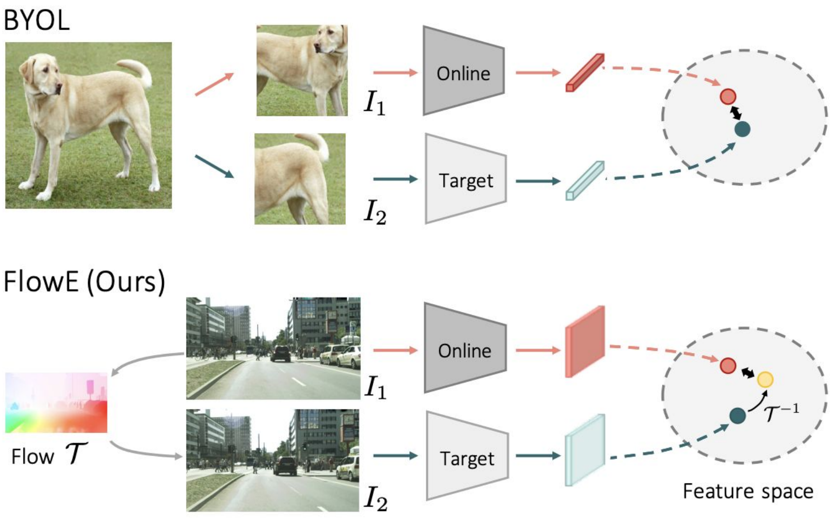
Self-Supervised Representation Learning from Flow Equivariance.
Xiong, Ren et al. ICCV 2021

Flow Equivariance (FlowE)

SSL (e.g., BYOL): minimize ***global*** features between different views of an image

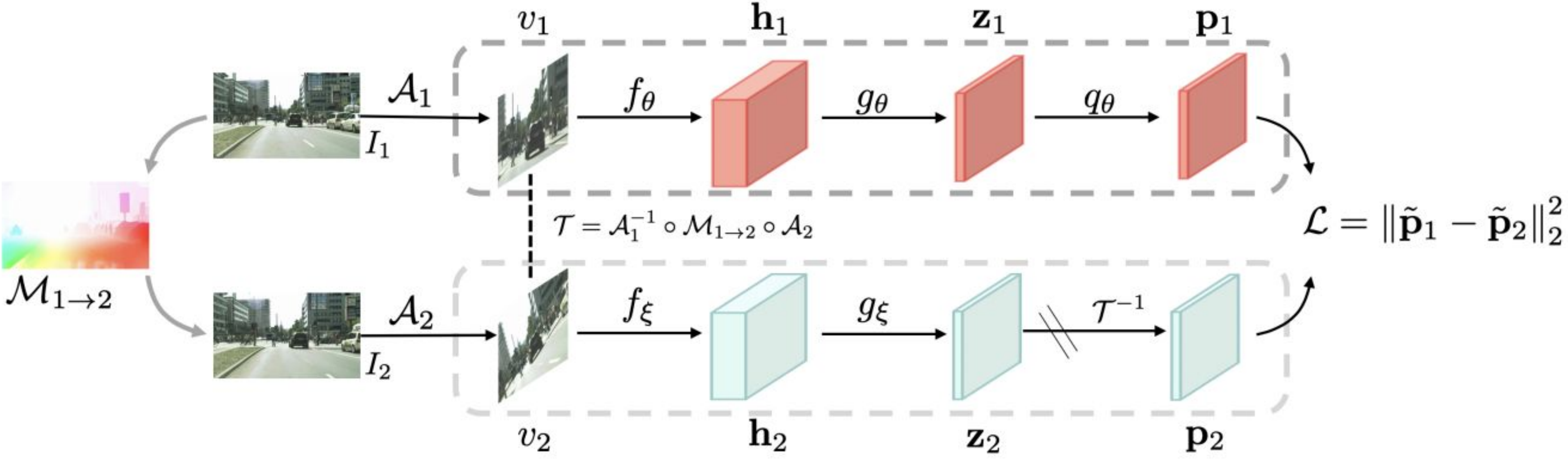
Dense SSL: compare ***dense*** feature maps between different views of an image

FlowE: use ***flow*** to align dense feature maps between frames of a video



Self-Supervised Representation Learning from Flow Equivariance.
Xiong, Ren et al. ICCV 2021

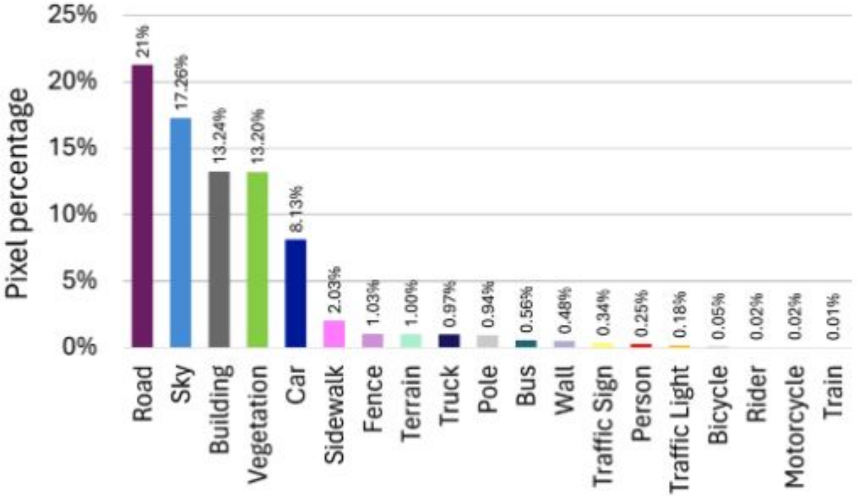
FlowE Method



Self-Supervised Representation Learning from Flow Equivariance.
Xiong, Ren et al. ICCV 2021

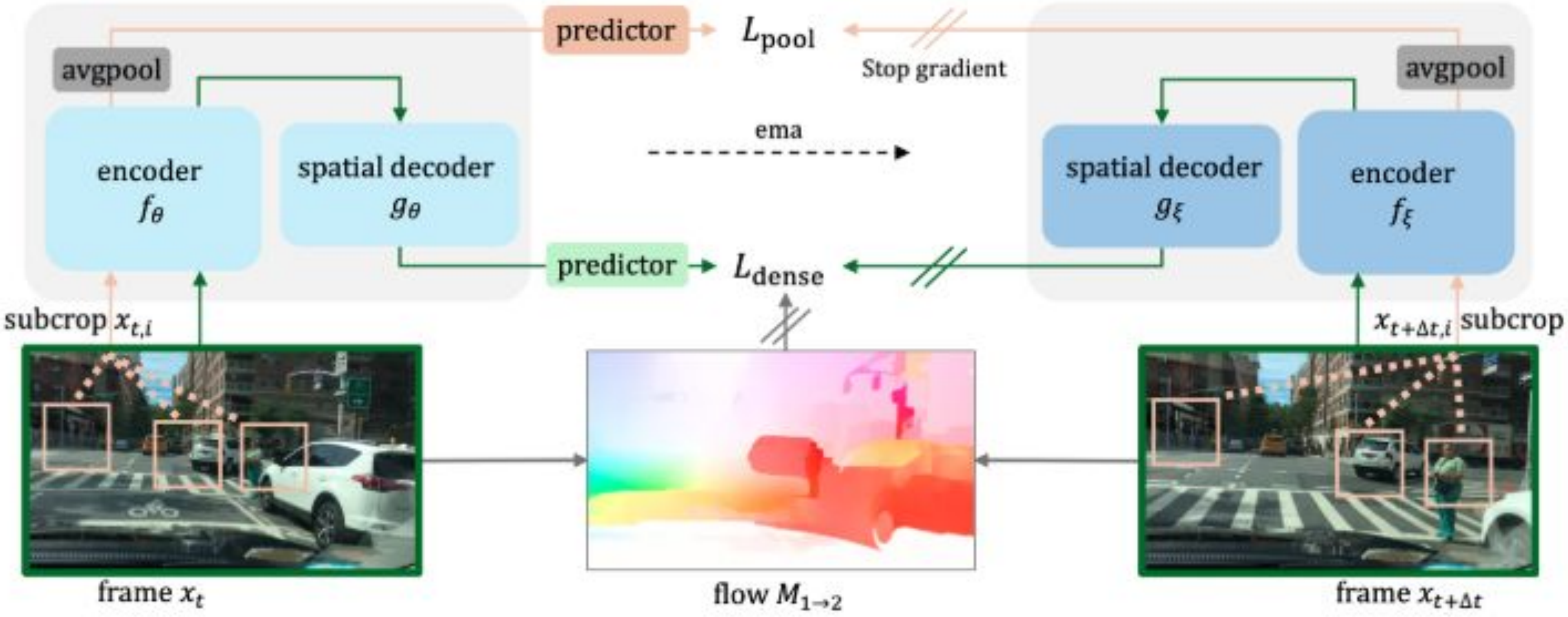
Dense SSL's spatial region imbalance problem

Semantic segmentation map for multi-object scene from BDD100K



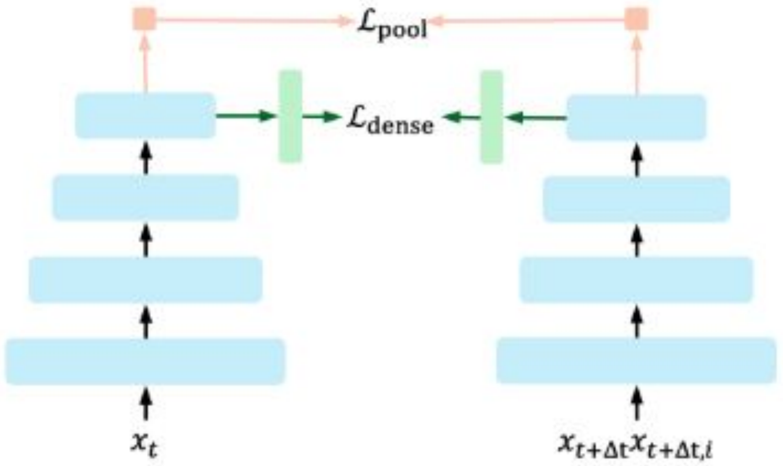
Crucial foreground objects only represent a small proportion of pixels

Pooled and Dense Learning (PooDLe)

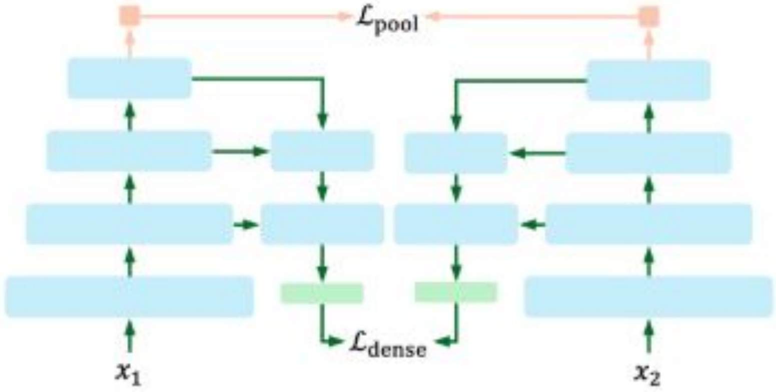


PooDLe: Pooled and dense self-supervised learning from naturalistic videos.
Wang*, Hoang* et al., ICLR 2025

PooDLe: encoder-decoder design



Naive: place both objectives at last encoder layer



PooDLe: pooled objective on last encoder layer; dense objective on high-resolution output from spatial decoder

Using subcrops as pseudo-iconic views of objects

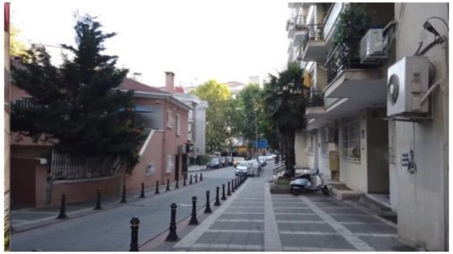
100% 75% 50% 25% 5%

Semantic segmentation results

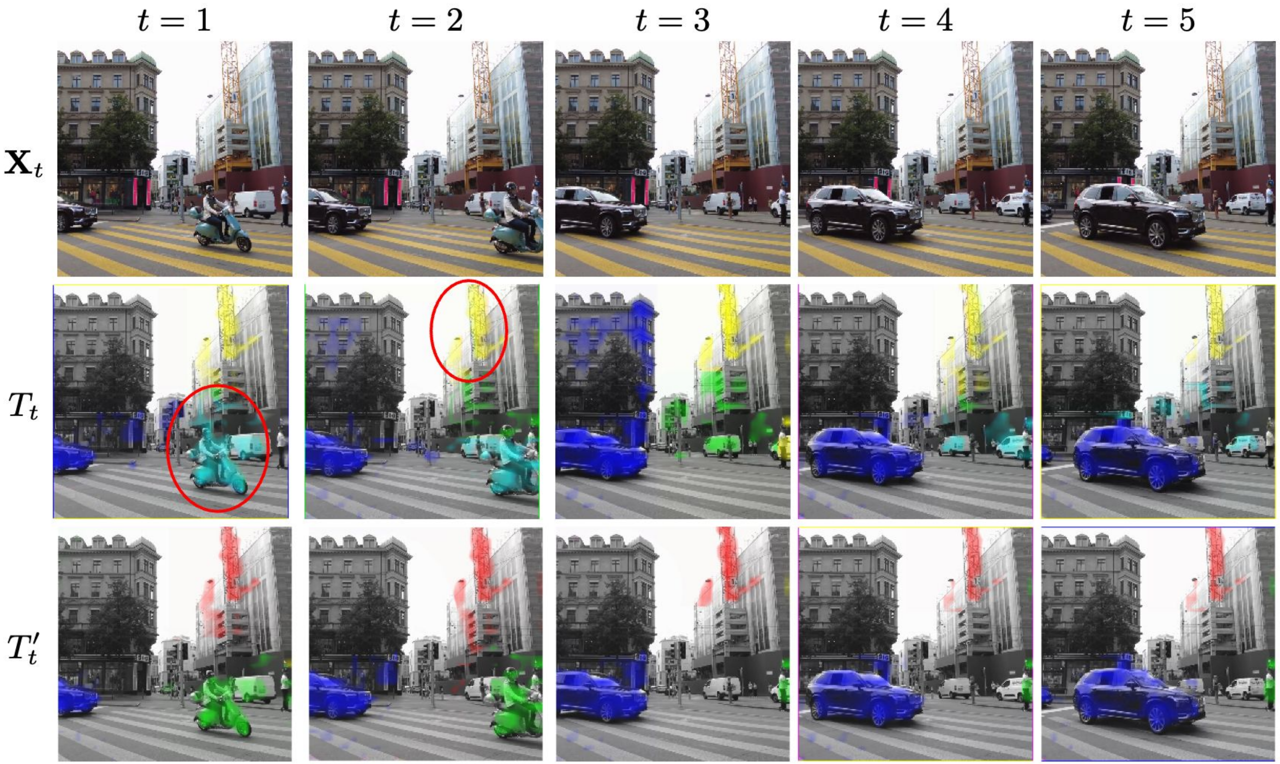
https://agenticlearning.ai/poodle

Discover and tRack Objects (DoRA)






Discover and tRack Objects (DoRA)



Discover and tRack Objects (DoRA)