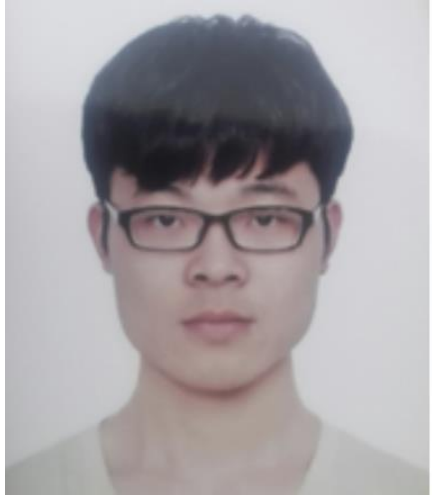
Microsoft Research
Berkeley University of California
Johns Hopkins University

The Emergence of Objectness: Learning Zero-shot Segmentation from Videos




Runtao Liu   Zhirong Wu   Stella X. Yu   Stephen Lin

Current Usage of Self-supervised Learning

Pre-training

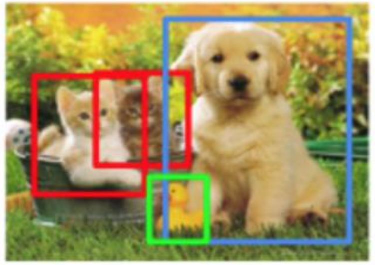
Self-supervised pretrained model

Fine-tuning Applications

Classification

Segmentation

Detection



Current Usage of Self-supervised Learning

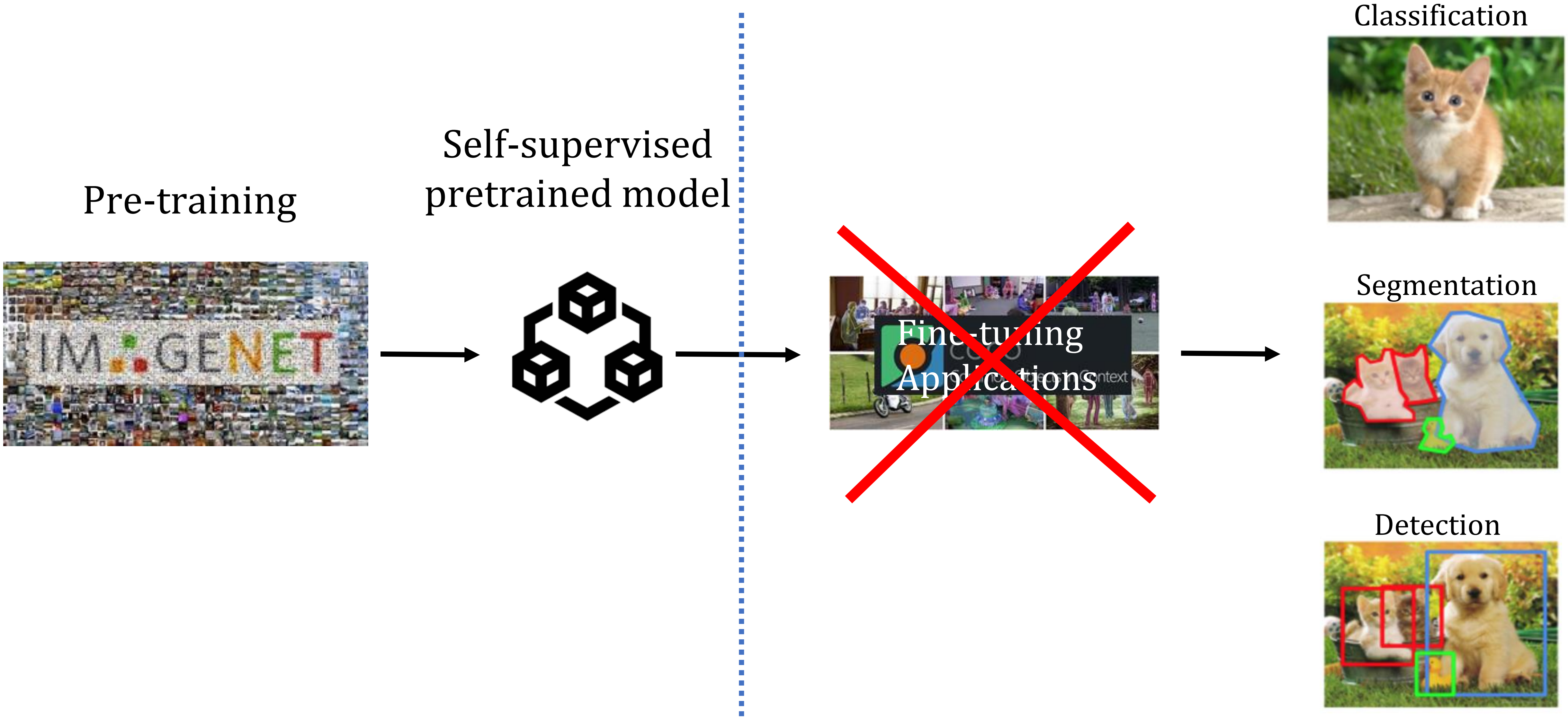
Pre-training

Self-supervised
pretrained model

A representation model

Not directly useful

Our Goal of Self-supervised Learning



Our Goal of Zero-shot Learning

Pre-training

Self-supervised pretrained model

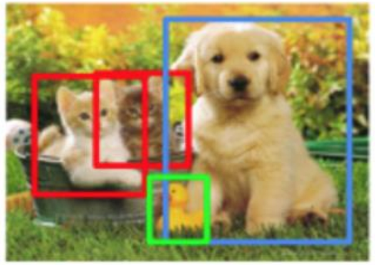
Direct Support for Applications

No labels
No finetuning

Classification

Segmentation

Detection

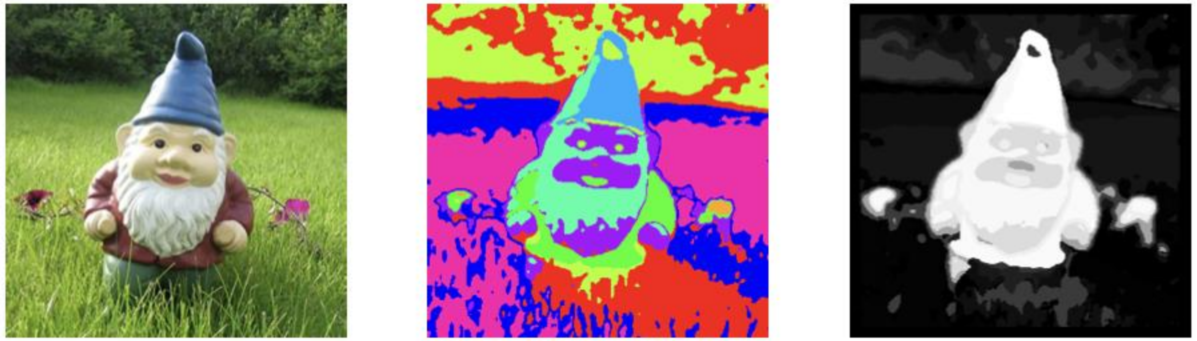


The Problem:
Segment the Primary Object From an Image

input

output

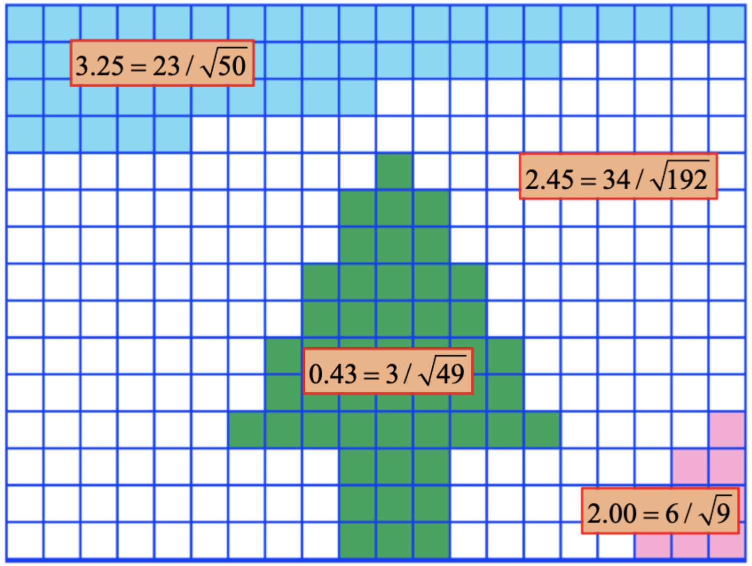
Existing Bottom-up Cues for Objectness Detection



Scharfenberger et al. 2014

the texture prior

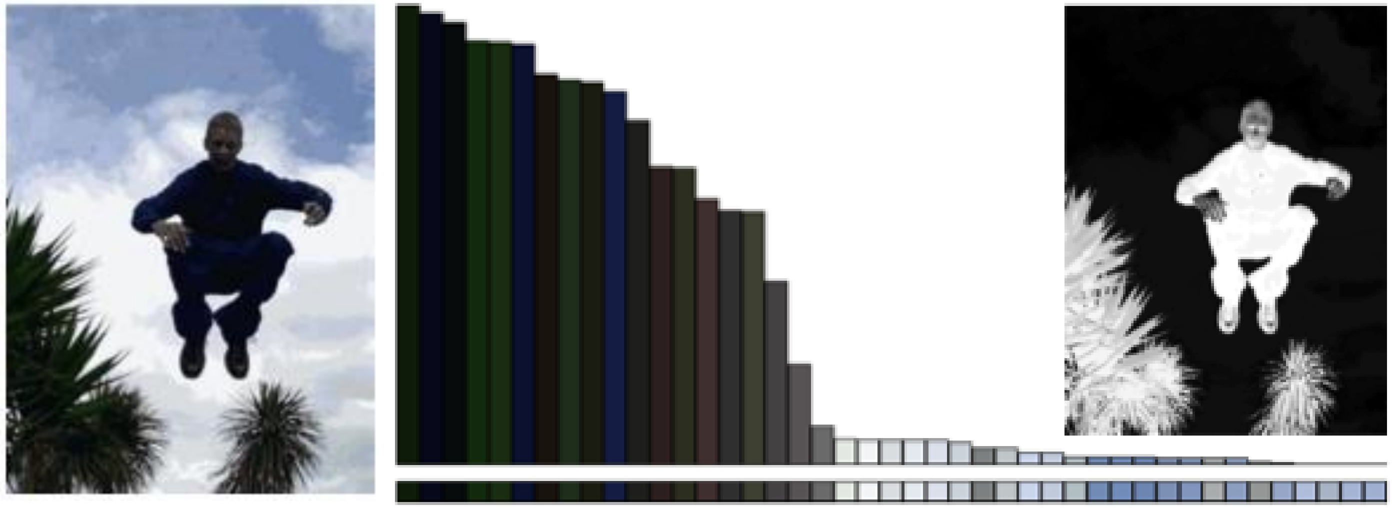
Distinctive region of textures different from the rest of the scene



Zhu et al. 2014

the center prior

regions that has least connectivity to the image bounder tend to be foreground



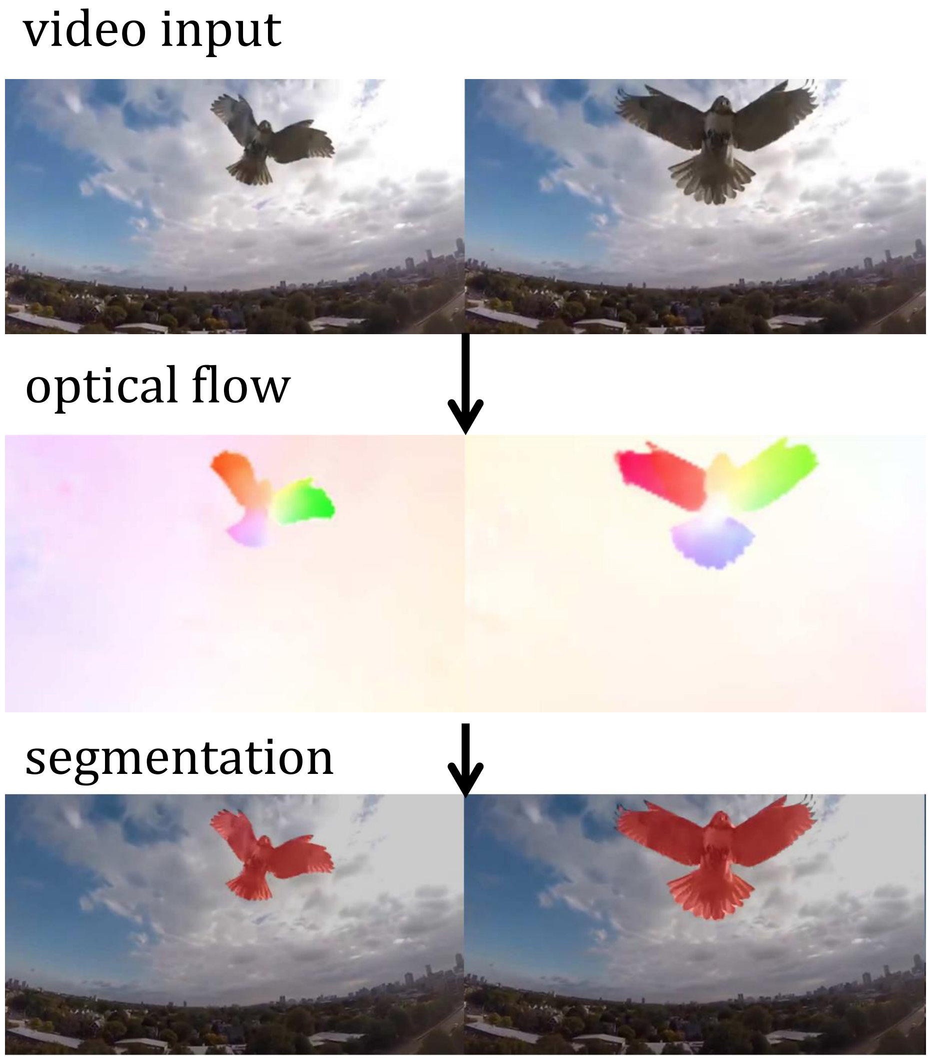
Cheng et al. 2016

the color contrast prior

high color contrast pixels tend to be the foreground

Bottom-up Motion Cues – Motion Segmentation

Group pixels having similar motions into a single region, following the common fate principle.



A series of work differs in:

[Sun 2012, Kumar 2008, Shi 1998, Yang 2019]

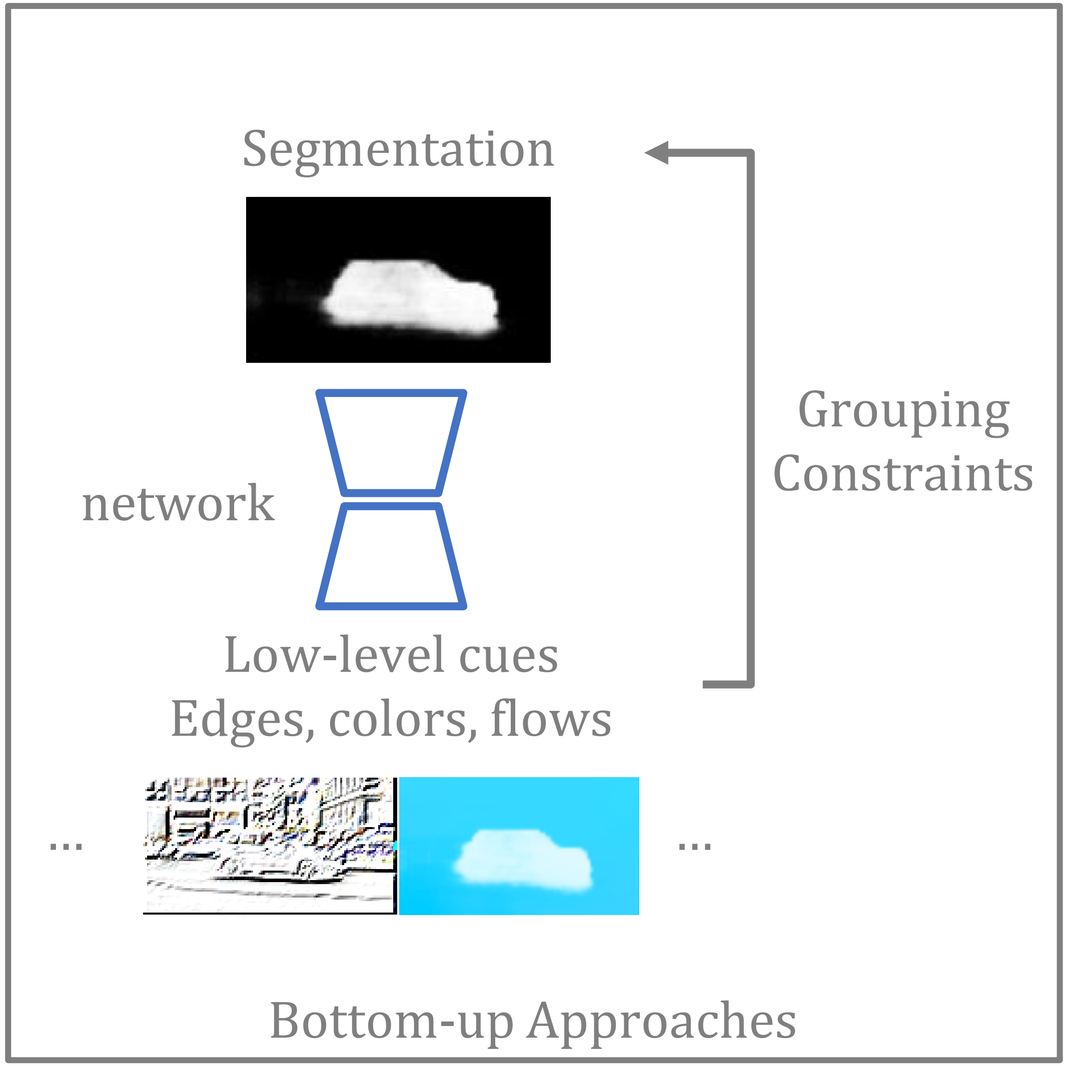
- Short-term or long-term analysis
- Whether to model occlusion
- Energy function to cluster the pixels

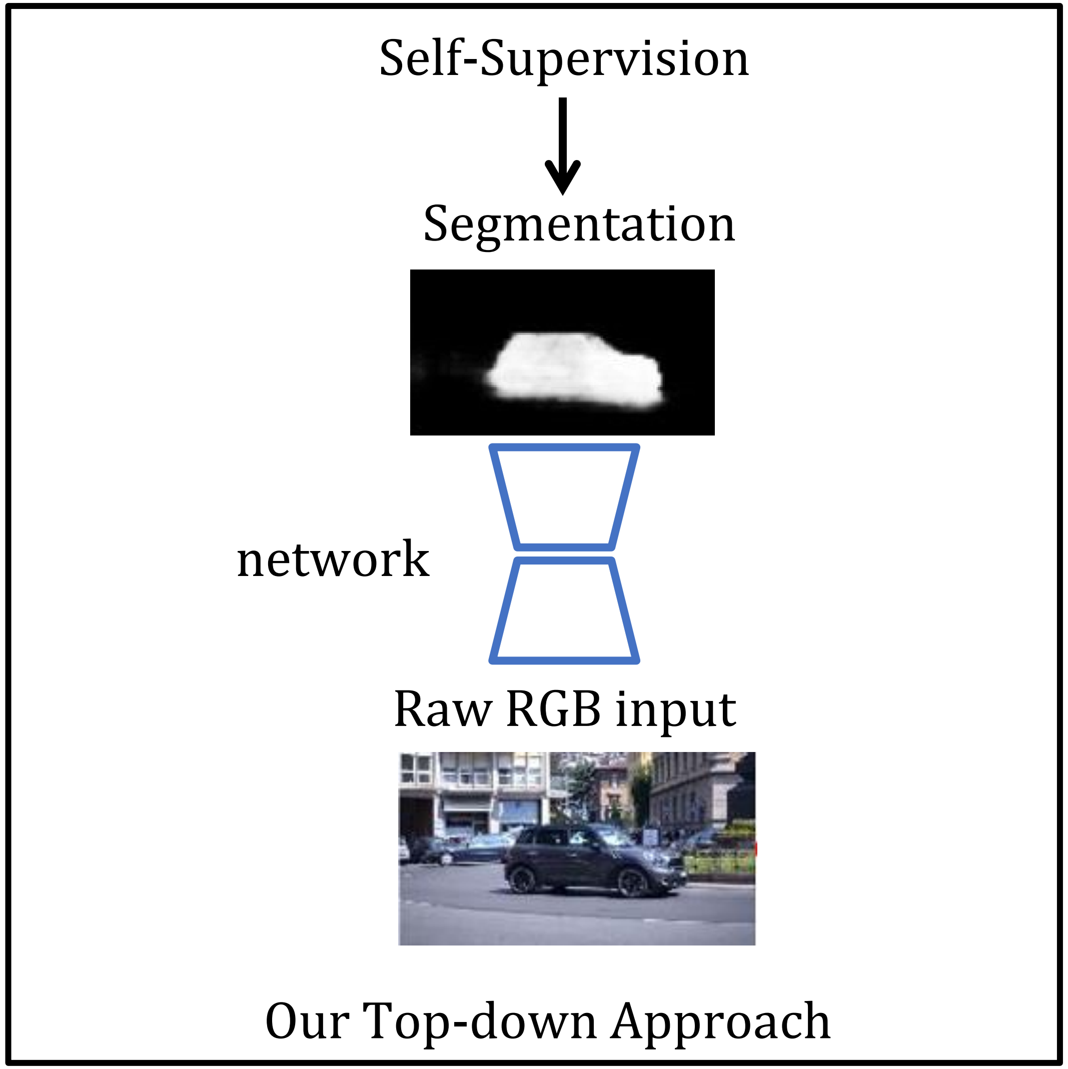
All requires a low-level input:

- Dense optical flow

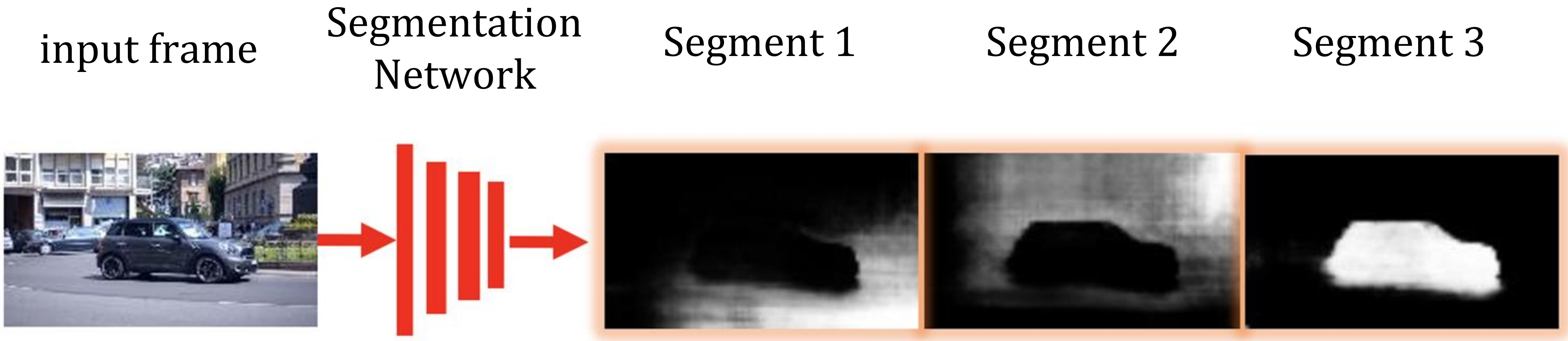
Our Top-down Approach:

Appearance and Motion Decomposition for Videos

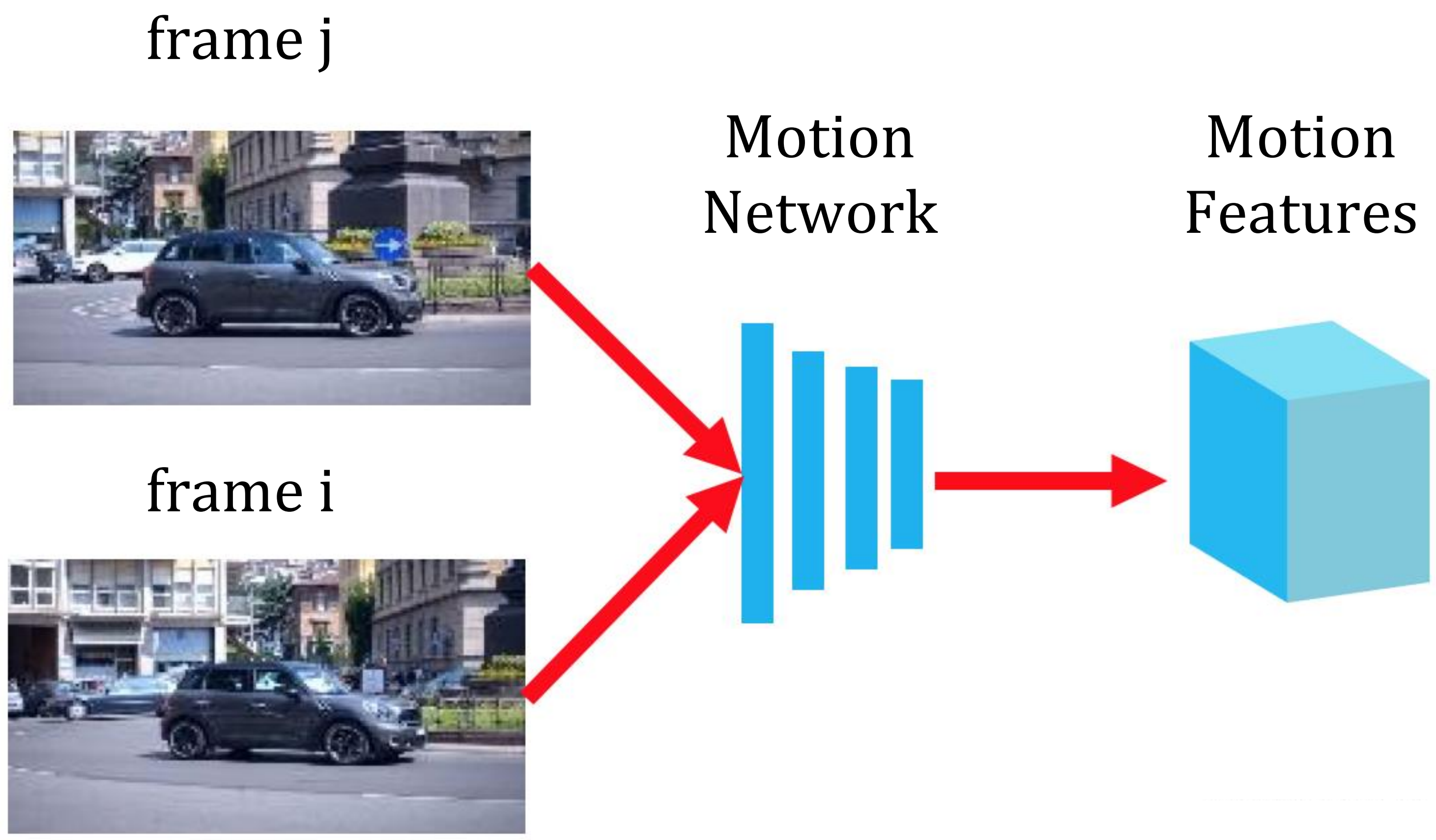




Appearance Pathway to Segment Object

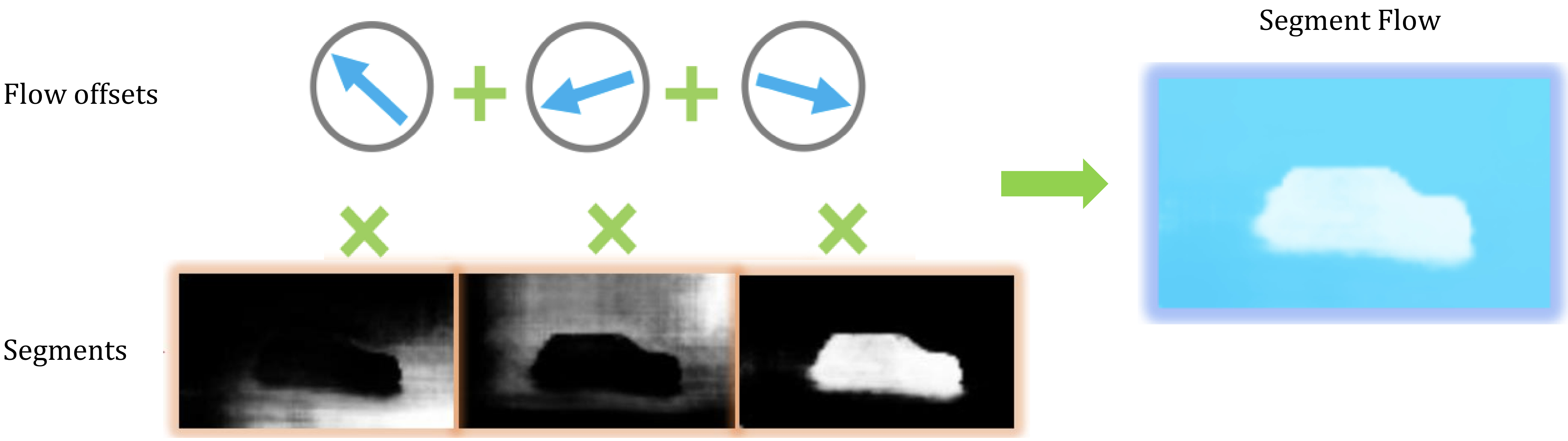


Motion Pathway to Extract Correspondence Features



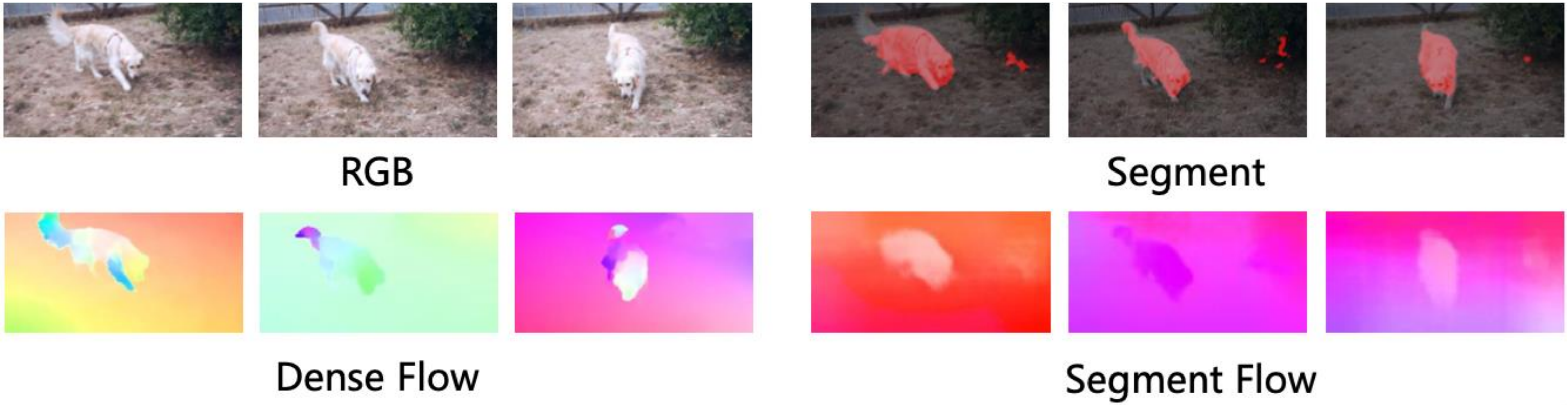
Segment Flow Representation

- Predict a flow vector for each segment produced by the appearance pathway.
- Broardcast each flow vector to pixels within each segment.

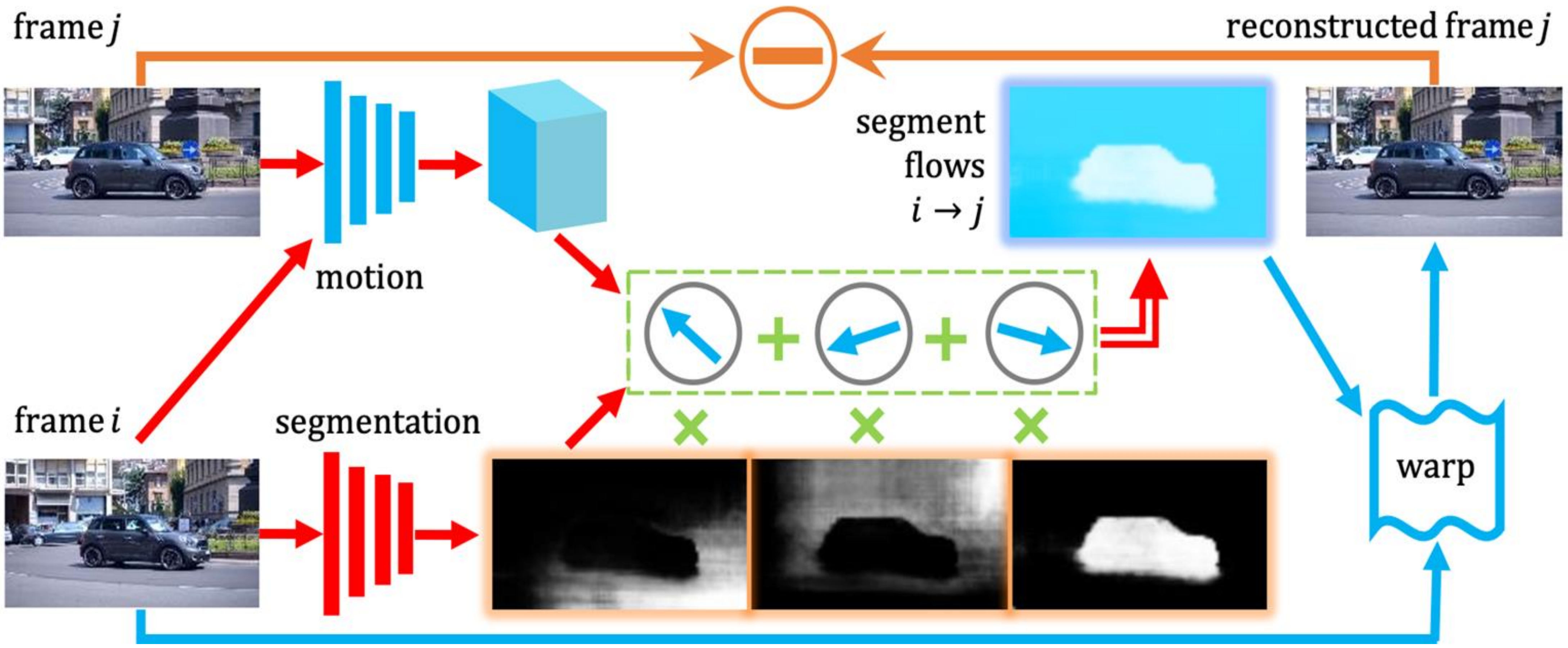


Segment Flow Representation

- Segment flow gives inaccurate estimation of per-pixel movement.
- Segment flow preserves the structural organization of the scene.



View Synthesis as the Self-supervision Signal



Applications of Our Model

Self-supervised pretraining on the Youtube-VOS dataset with 4000 videos. Then transfer to:

1. Zero-shot object segmentation from images.
2. Zero-shot moving object segmentation from videos.
3. Fine-tuning on labeled data for semantic segmentation.

Zero-shot Object Segmentations from Images

Take the appearance pathway for single image inference.

Evaluation dataset:

the testing split of the DUTS saliency detection benchmark.

	Model	F_β	MAE
Non-learning approaches using priors such as color, edge contrast, and image borders.	RBD[55]	51.0	0.20
	HS[65]	52.1	0.23
	MC[56]	52.9	0.19
	DSR[66]	55.8	0.14
	DRFI[57]	55.2	0.15
Ours	**AMD**	**60.2**	**0.13**

Zero-shot Object Segmentations from Images

Zero-shot Object Segmentations from Videos

Per-img: take the appearance pathway for individual images in a video.

Per-vid: unsupervised test-time adaptation for a video with both pathways.

Test-time adaptation for 100 iterations

Testing
Video

Pretrained
AMD model

Unsupervised
View synthesis
Objective

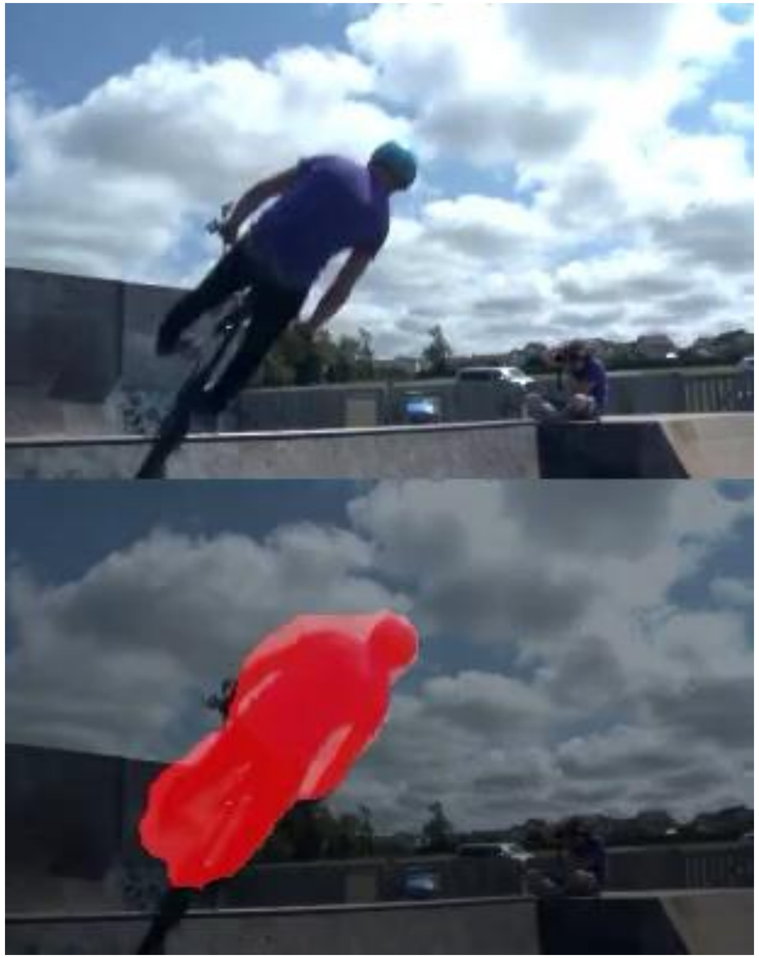
Zero-shot Object Segmentations from Videos

Per-img: take the appearance pathway for individual images in a video.

Per-vid: unsupervised test-time adaptation for a video with both pathways.

	Model	e2e	Sup.	Flow	DAVIS 2016	SegTrackv2	FBMS59
traditional	SAGE[65]	✗	✗	LDOF[66]	42.6	57.6	61.2
	NLC[14]	✗	edge	SIFTFlow[67]	55.1	67.2	51.5
	CUT[28]	✗	✗	LDOF[66]	55.2	54.3	57.2
	FTS[16]	✗	✗	LDOF[68]	55.8	47.8	47.7
	ARP[15]	✗	saliency	CPMFlow[69]	76.2	57.2	59.8
learning	CIS[18]	✗	✗	PWC[20]	59.2	45.6	36.8
	MG[19]	✗	✗	ARFlow[6]	53.2	37.8*	50.4*
	AMD (per-img)	✓	✗	✗	45.7	28.7	42.9
	AMD (per-vid)	✓	✗	✗	57.8	57.0	47.5

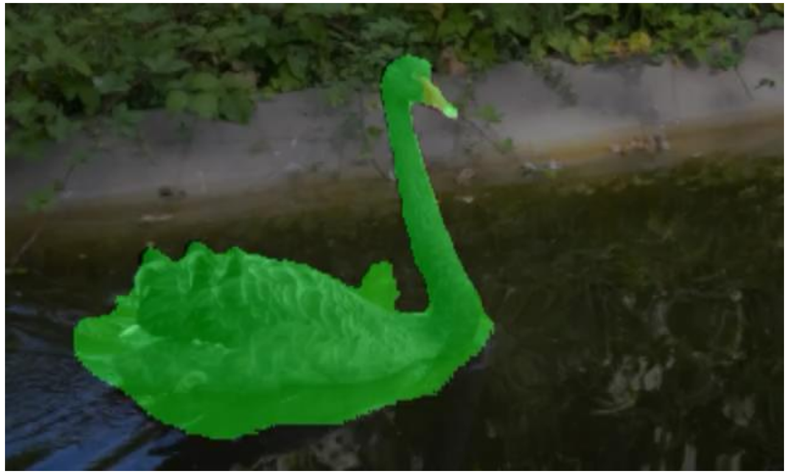
Zero-shot Object Segmentations from Videos



Comparison with Prior Approach CIS

Input

Ours



CIS

Fine-tuning for Semantic Segmentation:

Evaluation dataset: Pascal VOC 2012

Our model does not rely on heavy augmentations.

Pretraining with light image augmentation
Resize(384), Crop(384)

Model	Data	mIoU
Scratch	–	48.0
TimeCyle[62]	VLOG	52.8
MoCo-v2[2]	YTB	61.5
AMD	YTB	**62.0**

Pretraining with heavy augmentation
ResizedCrop(384), ColorJitter, GrayScaling

Model	Data	mIoU
MoCo-v2[2]	YTB	**62.8**
AMD	YTB	62.1

Summary

The first end-to-end self-supervised approach for zero-shot segmentations.

- Learning from raw videos without precomputed flows.
- Works with minimal image augmentations.
- Applicable to image object segmentation under zero-shot.
- Applicable to video moving object segmentation under zero-shot.
- Fine-tuning for semantic segmentation.