Self-supervised Learning for Video Correspondence Flow

British Machine Vision Conference (BMVC), 2019

Zihang Lai, Weidi Xie
University of Oxford, Oxford, United Kingdom

Presented by: Ricardo B. Sousa ([up201503004](up201503004))

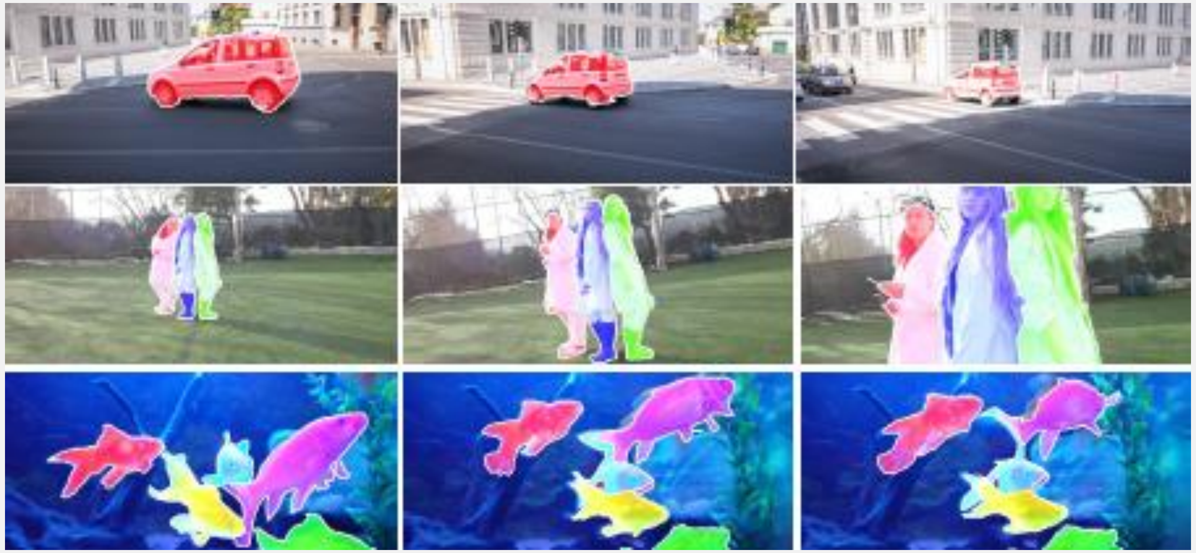
FEUP, PDEEC, Computer Vision, 2020/2021

OUTLINE

- Introduction
  - Context
  - Motivation
  - Goals
- Method
  - Full vs Restricted Attention
  - Long-term Correspondence Flow
- Experiments and Analysis
  - DAVIS-2017
  - JHMDB
- Conclusions
- Open Questions

Introduction: Context

- **Correspondence matching:** discerning which parts of images correspondent between each other
    - Depth estimation
    - Optical flow
    - Segmentation and tracking
    - 3D reconstruction



Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

INTRODUCTION: MOTIVATION

- Supervised learning for correspondence matching can be prohibitively expensive
- Self-supervised learning does not require labelled data
- Videos have an almost infinite supply (e.g., YouTube)
- Spatio-temporal coherence is inherent to videos

Introduction: Goals

- Train a CNN-based model on videos assuming the spatio-temporal coherence
- Channel-wise (RGB) dropout and colour jittering added intentionally on input frames

- Benchmark the model on video segmentation and keypoint tracking

METHOD

- Train an embedding network w/ self-supervised learning for pixelwise correspondence matching
  - Exploit spatial-temporal coherence in videos

- **Requirements for training:**
  - Ground-truth annotation for the first frame
  - Model should be trained on full-colour and high-resolution (the model resizes all frames to 256 x 256 x 3) over long video sequences

Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

METHOD

- **Overview:**

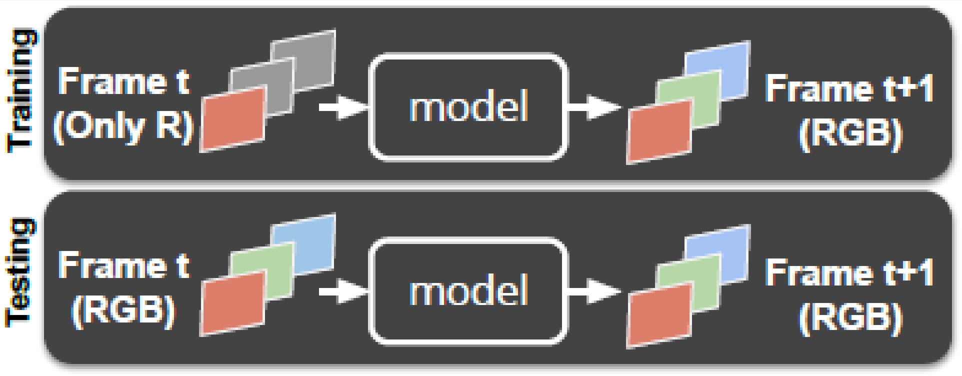
Method

- Given a collection of frames $\{I_1, I_2, \ldots, I_N\}$ from a video:

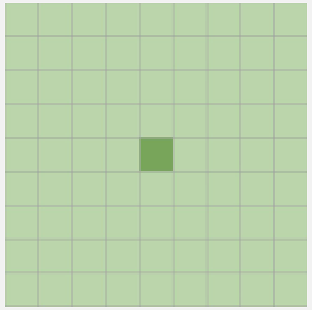
$$f_i = \Phi(g(I_i); \theta)$$

  - Φ: ResNet-18
  - $g(\cdot)$: information bottleneck

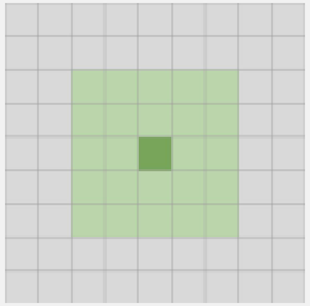
- **Information bottleneck:**
  - Randomly zero out 0, 1, or 2 channels in each input frame
  - Randomly perturb the brightness, contrast and saturation of an image by up to 10%
  - 2 benefits:
    - Input jitterings and stochastic dropout prevents the model from co-adaptation of low-level colours
    - Data augmentation



METHOD: FULL VS RESTRICTED ATTENTION



- **Full attention**
  - All pairs of pixels in target and reference frames are correlated
  - Memory and computational consumption grow $O(n^2)$ with the spatial footprint of the feature maps
    - dimension { Feature maps } = H x W → 4D tensor of dimension of H x W x H x W

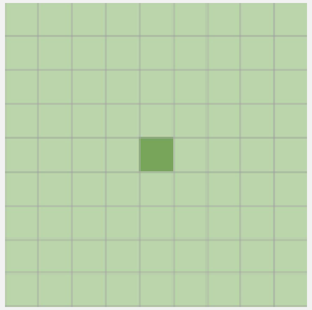


- **Restricted attention:**
  - Pixels in the reference frame are searched for locally in a square patch of size (2M+1) x (2M+1)
  - Maximum disparity of M defines the size of the square
    - dimension { Feature maps } = H x W → 4D tensor of dimension of H x W x (2M+1) x (2M+1)

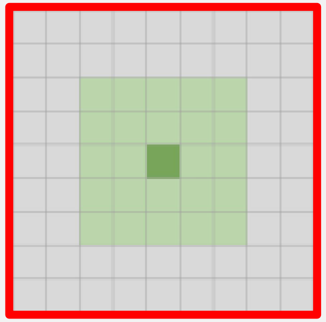
Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

METHOD: FULL VS RESTRICTED ATTENTION

- **Full attention**
  - All pairs of pixels in target and reference frames are correlated
  - Memory and computational consumption grow quadratically with the spatial footprint of the feature maps
    - dimension { Feature maps } = H x W → 4D tensor of dimension of H x W x H x W



- **Restricted attention:**
  - Pixels in the reference frame are searched for locally in a square patch of size (2M+1) x (2M+1)
  - Maximum disparity of M defines the size of the square
    - dimension { Feature maps } = H x W → 4D tensor of dimension of H x W x (2M+1) x (2M+1)



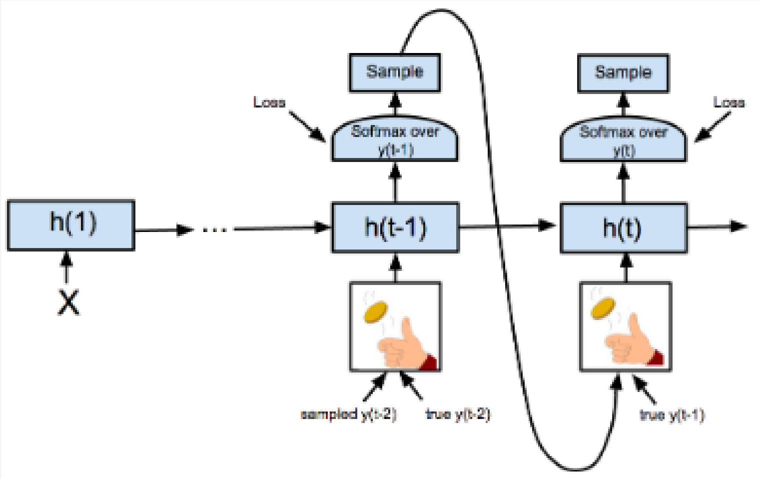
Method: Long-term Correspondence Flow

- IF 2 frames are sampled closely in time, THEN objects remain unchanged
- IF frames are sampled with a large temporal stride, THEN the assumption of using reconstruction as supervision may fail

Method: Long-term Correspondence Flow

- IF 2 frames are sampled closely in time, THEN objects remain unchanged
- IF frames are sampled with a large temporal stride, THEN the assumption of using reconstruction as supervision may fail

- **Scheduled sampling:**
  - Initial probability of using ground-truth frames = 0.9
  - Uniformly annealed to 0.6



METHOD: LONG-TERM CORRESPONDENCE FLOW

- IF 2 frames are sampled closely in time, THEN objects remain unchanged
- IF frames are sampled with a large temporal stride, THEN the assumption of using reconstruction as supervision may fail

- **Scheduled sampling**
- **Cycle Consistency:**
  - Tracking is performed forward and backward in time
  - Inconsistency between start and end points are the loss function

Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

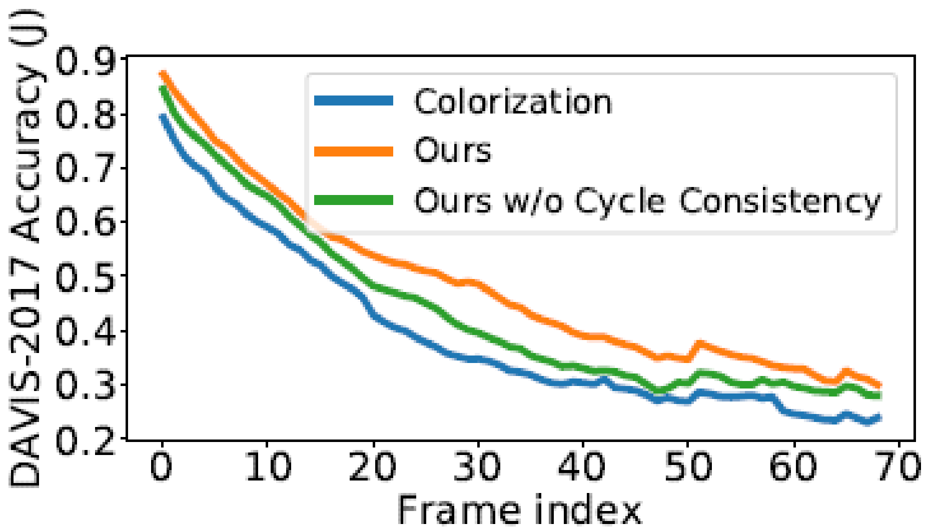
Experiments and Analysis

- **Training details:**
  - Only used video sequences from Kinetics dataset not finetuned for any target task
  - Frame rate of 6fps and resize all frames to 256 x 256 x 3
  - ResNet-18

- **Evaluation metrics:**
  - Video segmentation:
    - J: intersection over union
    - $\mathcal{F}$: contour accuracy
  - Keypoint tracking:
    - $PCK_{instance}$: IF normalized Euclidean distance error < α, THEN keypoint is considered correct
    - PCK_{max}: IF keypoint located within $\alpha \cdot \max(w, h)$ pixels of the ground-truth, THEN keypoint is accepted

Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

Experiments and Analysis: DAVIS-2017



Method	$\mathcal{J}$(Mean)	$\mathcal{F}$(Mean)
Ours (Full Model)	47.7	51.3
Ours w/o Colour Dropout	40.5	39.5
Ours w/o Restricted Attention	40.8	39.7
Ours w/o Scheduled Sampling	40.2	39.2
Ours w/o Cycle Consistency	41.0	40.4

Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

EXPERIMENTS AND ANALYSIS: DAVIS-2017

Method	Supervised	Dataset	$\mathcal{J}\&\mathcal{F}$(Mean)	$\mathcal{J}$(Mean)	$\mathcal{J}$(Recall)	$\mathcal{F}$(Mean)	$\mathcal{F}$(Recall)
Identity	✗	-	22.9	22.1	15.9	23.6	11.7
Optical Flow (FlowNet2) [15]	✗	-	26.0	26.7	-	25.2	-
SIFT Flow [28]	✗	-	34.0	33.0	-	35.0	-
Transitive Inv. [42]	✗	-	29.4	32.0	-	26.8	-
DeepCluster [50]	✗	YFCC100M	35.4	37.5	-	33.2	-
Video Colorization [40]	✗	Kinetics	34.0	34.6	34.1	32.7	26.8
CycleTime (ResNet-50) [43]	✗	VLOG	40.7	41.9	40.9	39.4	33.6
Ours (Full Model ResNet-18)	✗	Kinetics [23]	**49.5**	**47.7**	**53.2**	**51.3**	**56.5**
Ours (Full Model ResNet-18)	✗	OxUvA [39]	**50.3**	**48.4**	**53.2**	**52.2**	**56.0**
ImageNet (ResNet-50) [13]	✓	ImageNet	49.7	50.3	-	49.0	-
SiamMask [41]	✓	YouTube-VOS	53.1	51.1	60.5	55.0	64.3
OSVOS[6]	✓	DAVIS	60.3	56.6	63.8	63.9	73.8

Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

EXPERIMENTS AND ANALYSIS: JHMDB

Method	Supervised	Dataset	$PCK_{instance}$		PCK_{max}	
			@.1	@.2	@.1	@.2
SIFT Flow[28]	✗	-	49.0	68.6	-	-
Video Colorization [40]	✗	Kinetics	45.2	69.6	-	-
CycleTime (ResNet-50) [43]	✗	VLOG	57.7	78.5	-	-
Ours (Full Model ResNet-18)	✗	Kinetics	**58.5**	**78.8**	**71.9**	**88.3**
ImageNet (ResNet-50) [13]	✓	ImageNet	58.4	78.4	-	-
Fully Supervised [38]	✓	JHMDB	-	-	68.7	81.6

Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

CONCLUSIONS

- Potential of self-supervised learning
- State-of-the-art performance on both video segmentation and keypoint tracking
- Key improvements in terms of reducing the drift over time

Open Questions

- Why ResNet-18?
- Could the consideration of middle losses led to improvements?
- What about the drift over time from other methods in the literature (only 1 was used for comparison)?

Self-supervised Learning for Video Correspondence Flow

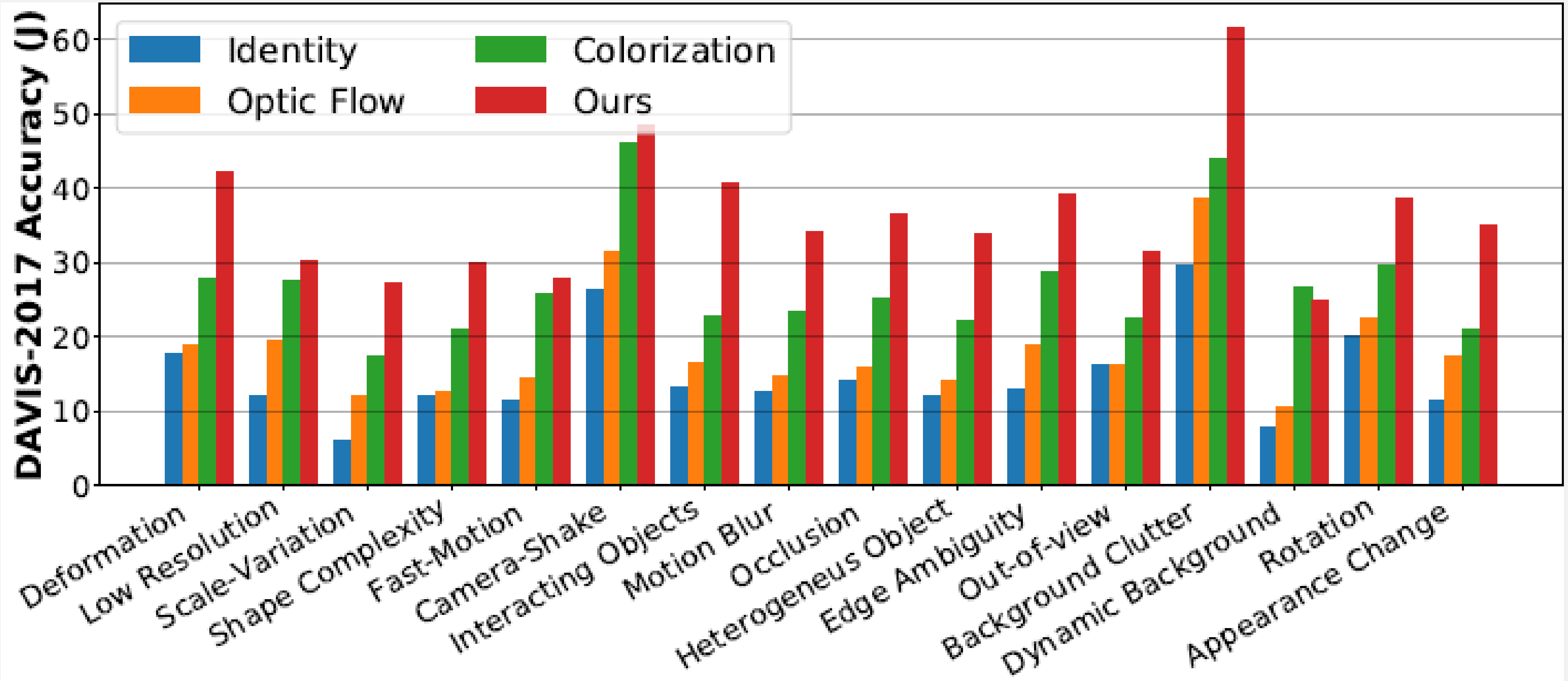
British Machine Vision Conference (BMVC), 2019

Zihang Lai, Weidi Xie
University of Oxford, Oxford, United Kingdom

Presented by: Ricardo B. Sousa (up201503004)

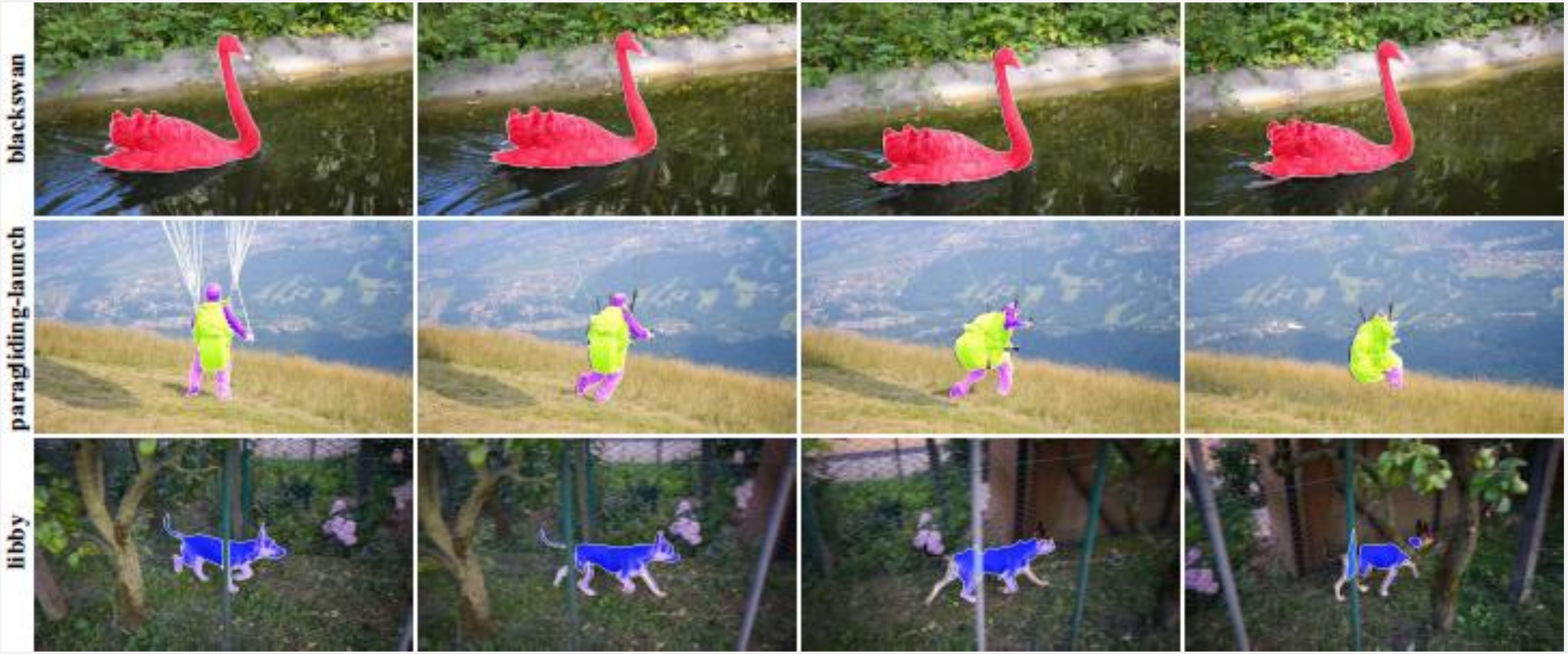
FEUP, PDEEC, Computer Vision, 2020/2021

Experiments and Analysis: DAVIS-2017



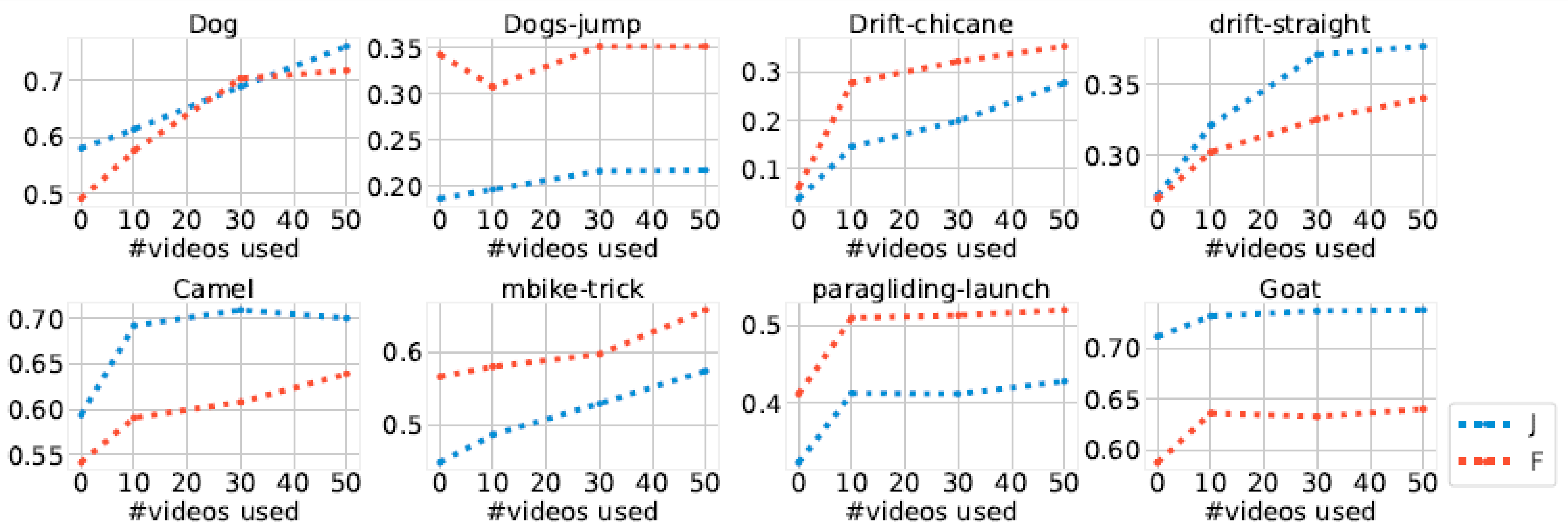
Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

EXPERIMENTS AND ANALYSIS: DAVIS-2017



Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

EXPERIMENTS AND ANALYSIS: DAVIS-2017 + YOUTUBE



Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)

Experiments and Analysis

Method	Dataset	Dog	Dog-j	Drift-c	Drift-s	Camel	Mbike	Paragliding	Goat
Self-supervised ($\mathcal{J}$)	Kinetics	58.0	18.6	3.9	27.1	59.4	44.8	32.4	71.1
Self-supervised ($\mathcal{J}$)	Additional	76.1	21.7	27.8	37.7	70.0	57.4	42.8	73.7
Supervised ($\mathcal{J}$) [51]	COCO+DAVIS	87.7	38.8	4.9	66.4	88.4	72.5	38.0	80.4
Self-supervised ($\mathcal{F}$)	Kinetics	49.1	34.3	6.3	26.9	54.2	56.6	41.2	58.8
Self-supervised ($\mathcal{F}$)	Additional	71.8	35.2	35.3	34.0	63.9	65.8	52.0	64.0
Supervised ($\mathcal{F}$) [51]	COCO+DAVIS	84.6	45.2	8.4	57.7	92.2	76.7	58.1	74.7

Lai, Z., Xie, W.: Self-supervised Learning for Video Correspondence Flow (2019)