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

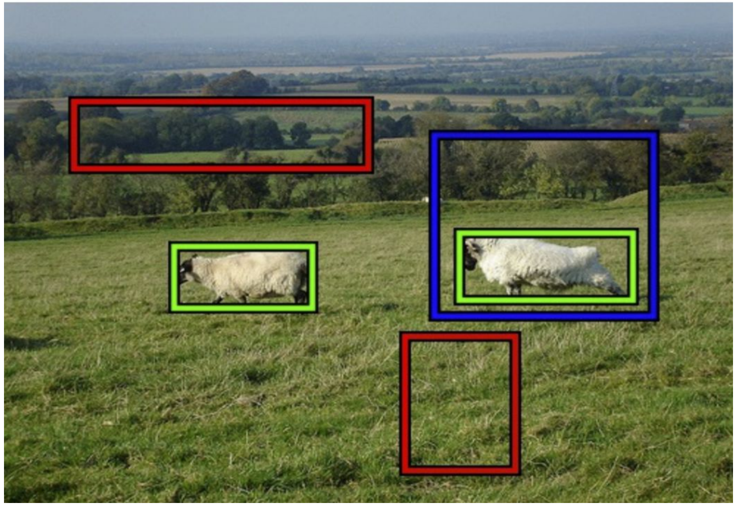
•••

Runtao Liu, Zhirong Wu, Stella X. Yu, Stephen Lin

Presenter: Katsumi Ibaraki

Introduction

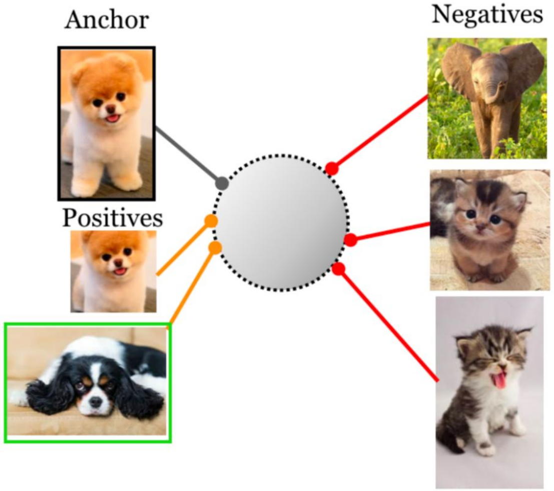
- Objectness
  - Measure of how likely an object exists
  - High objectness
    - Likely contains object
    - Uniqueness
    - Tight boundaries



https://medium.com/@zhao.nathan/understanding-objectness-in-object-detection-models-5d8c9d032488

Introduction

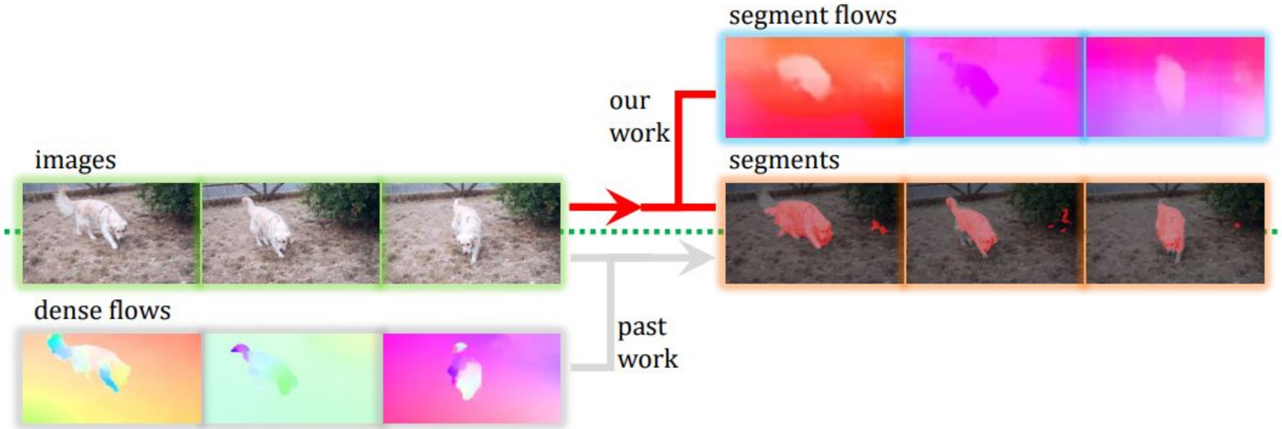
- Contrastive learning
  - High performance
  - Reliance on hand-crafted image augmentations
  - Additional labeled data and fine-tuning for downstream applications



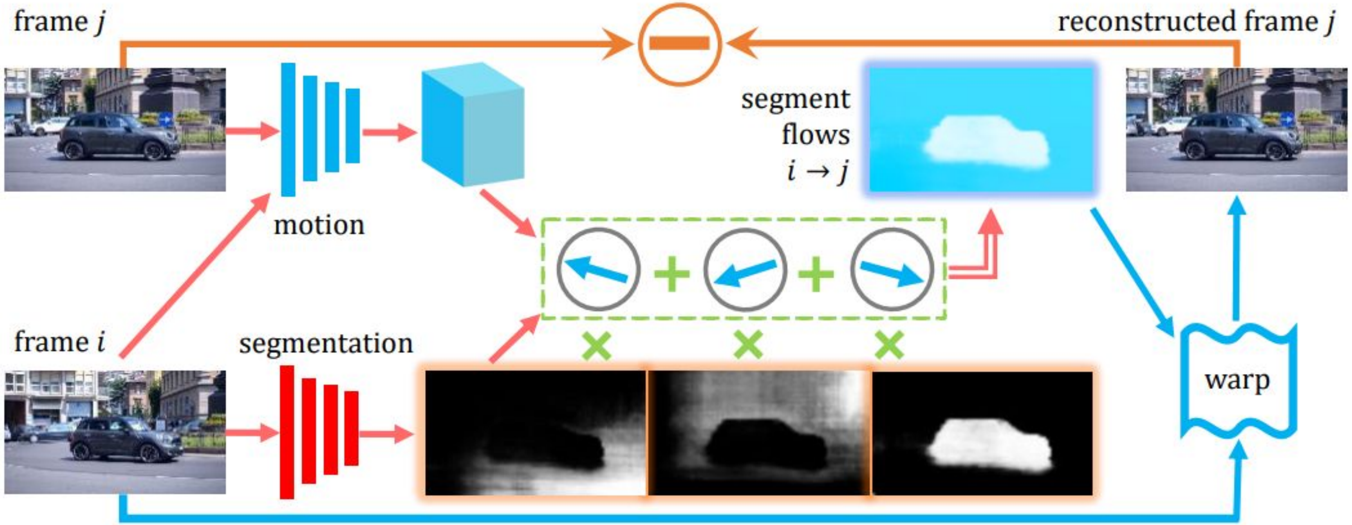
https://ai.googleblog.com/2021/06/extending-contrastive-learning-to.html

Introduction

- Goal: zero-shot image model to detect and segment objects
  - Learn from unlabeled videos
  - Dynamic sequence of observations
    - Reveals boundary and hierarchical information



Model


frame j
reconstructed frame j
segment flows $i \to j$
motion
frame i
segmentation
warp

Model

- Appearance Pathway for Segmentation
  - Convolutional Neural Network
  - Segment static RGB image into regions
- Given image $X_i \in \mathbb{R}^{3 \times h \times w}$, segment into *c* regions by

$$S = f_A(X_i) \in \mathbb{R}^{c \times h \times w}$$

Model

- Motion Pathway for Correspondence
  - Extract pixel-wise motion features
  - PWC-Net
- Given inputs X_i and X_j, extract features V by

$$V = f_M(X_i, X_j) \in \mathbb{R}^{d_v \times h \times w}$$

where d_v is the dimension of motion features

Model

- Segment Flow Representation
- Obtain mask motion feature vector from

$$V_m = \frac{\sum (V \odot S_m)}{\sum S_m} \in \mathbb{R}^{d_v}, \quad m = 1, \ldots, c$$

Model

- Segment Flow Representation
- Obtain optical flow vector for each mask by

$$F_m = g(V_m) \in \mathbb{R}^2, \quad m = 1, ..., c$$

Model

- Segment Flow Representation
- Obtain optical flow vector for each mask by

$$F_m = g(V_m) \in \mathbb{R}^2, \quad m = 1, ..., c$$

- Obtain novel flow representation for full image by

$$F = \sum_m F_m \odot S_m, \quad m = 1, ..., c,$$

Model

- Reconstruction
  - Warp frame X_i to X_j by

$$\hat{X}_j(p) = X_i(p + F(p))$$

- Supervision for reconstructed frame provided by

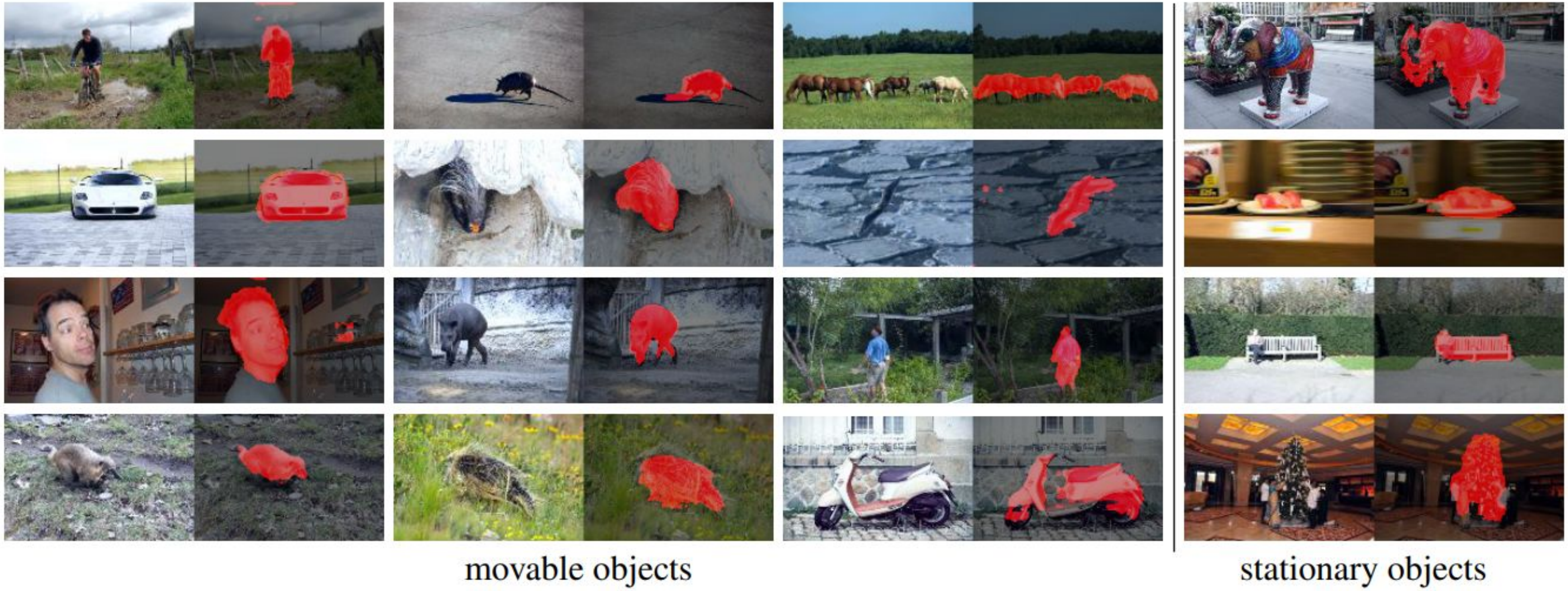
$$\mathcal{L} = D(X_j, \hat{X}_j)$$

Experiments

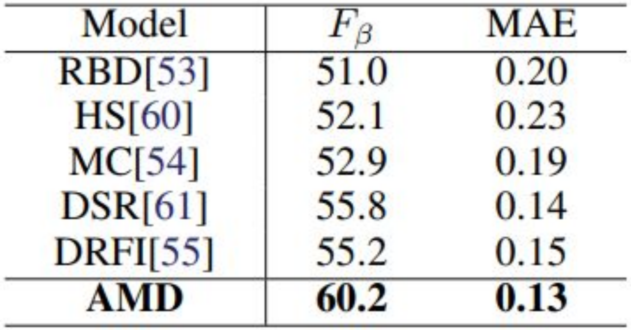
- Train all model parameters from scratch
    - No external pretraining
    - ResNet50 for segmentation network with convolutional head
    - PWC-Net for motion network
- Some augmentations
    - Resize input image
    - Random crop
    - Random horizontal flipping

Experiments

- Pretrained segmentation network tested on DUTS dataset



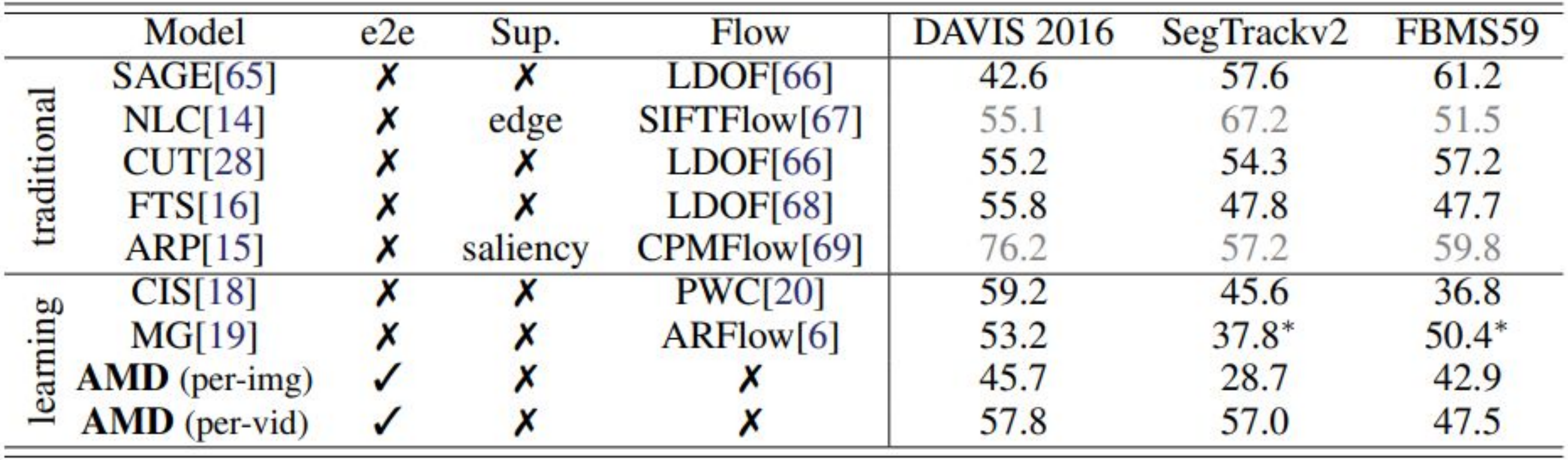
Experiments



Model	F_β	MAE
RBD[53]	51.0	0.20
HS[60]	52.1	0.23
MC[54]	52.9	0.19
DSR[61]	55.8	0.14
DRFI[55]	55.2	0.15
AMD	**60.2**	**0.13**

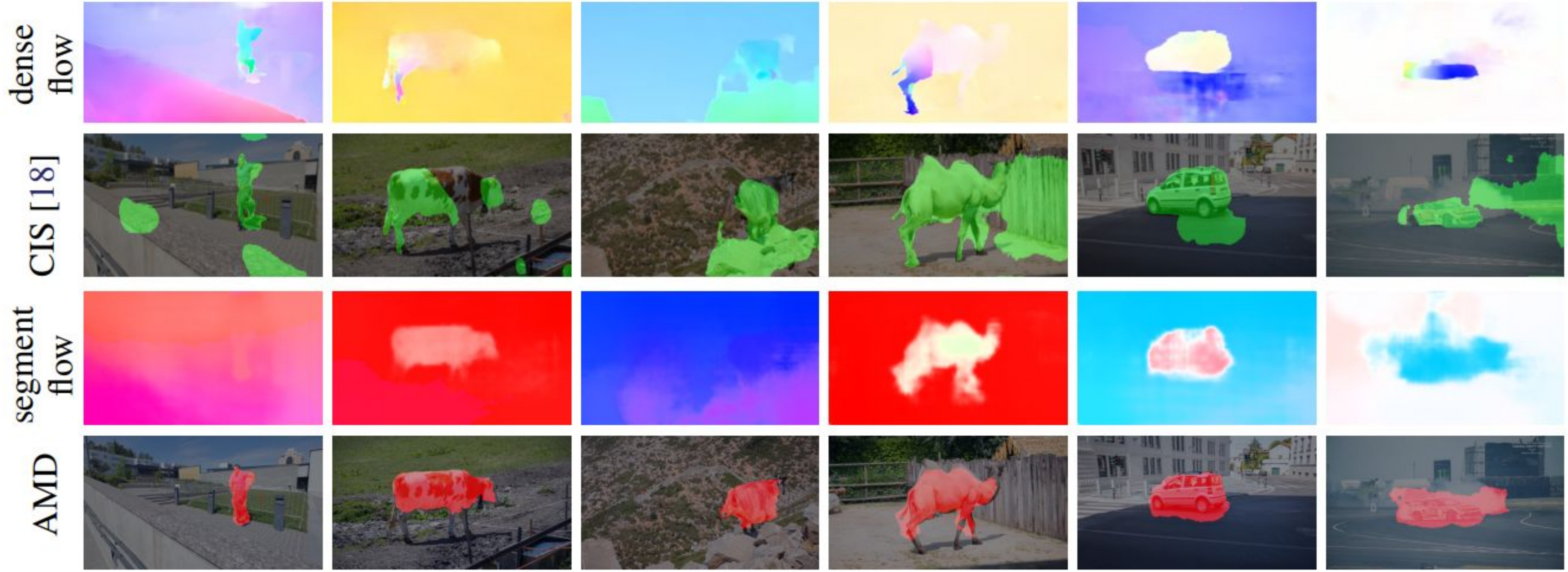
Saliency Detection

Experiments (Zero-Shot Object Segmentation)

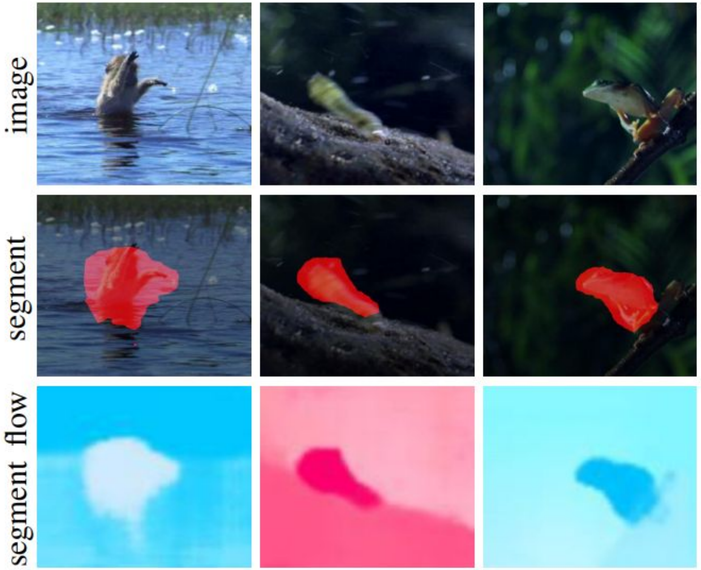


	Model	e2e	Sup.	Flow	DAVIS 2016	SegTrackv2	FBMS59
traditional	SAGE[65]	✗	✗	LDOF[66]	42.6	57.6	61.2
	NLC[14]	✗	edge	SIFTFlow[67]	55.1	67.2	51.5
	CUT[28]	✗	✗	LDOF[66]	55.2	54.3	57.2
	FTS[16]	✗	✗	LDOF[68]	55.8	47.8	47.7
	ARP[15]	✗	saliency	CPMFlow[69]	76.2	57.2	59.8
learning	CIS[18]	✗	✗	PWC[20]	59.2	45.6	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

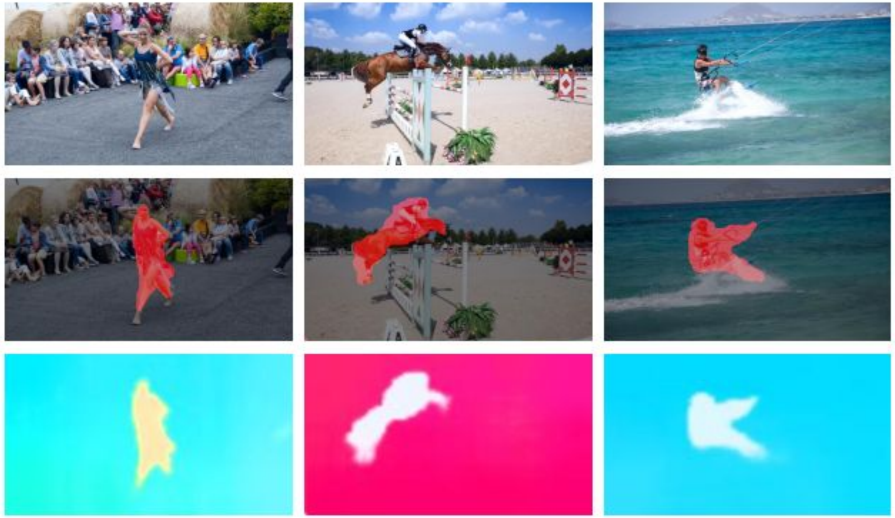
Experiments (Zero-Shot Object Segmentation)



Experiments

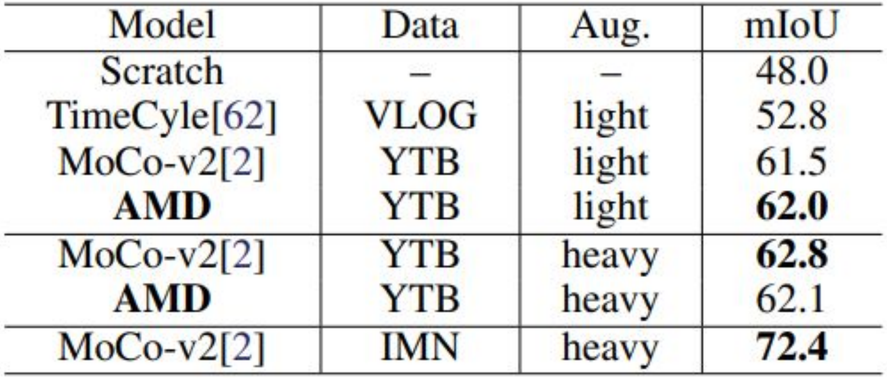


Results of SegTrackv2



Results of DAVIS 2016

Experiments



Model	Data	Aug.	mIoU
Scratch	–	–	48.0
TimeCyle[62]	VLOG	light	52.8
MoCo-v2[2]	YTB	light	61.5
AMD	YTB	light	**62.0**
MoCo-v2[2]	YTB	heavy	**62.8**
AMD	YTB	heavy	62.1
MoCo-v2[2]	IMN	heavy	**72.4**

Semantic Segmentation

Discussion Questions

@55_f3

This is kind of a small detail, *but I was surprised/found it interesting that the primary salient object appears in a particular mask channel across the training videos. Does anyone have a clearer idea on why this might be?* I wonder if the other mask channels consistently encoded the same/similar information across training videos.

From class:

NOT a small detail. It is the crux of the paper. If the salient object did not appear in the same mask channel across the different videos, we would not be able to analyze the way we did, as we would not know where to look for the objects. Also, this feature was not intended, hence, the paper is called *The **Emergence** of Objectness*.

Discussion Questions

@55_f2

I also wonder if the method in this paper depends strongly on photographer bias?

Is photographer bias a potential limitation of the proposed model? If so, are there other ways to train the model?

- Perhaps explore alternative training procedures or data sources that may be less susceptible to bias.
- More diverse data sets?

Discussion

@55_f1

- One significant idea behind the AMD model is to decouple the object's appearance and motion
- Segment flows are sometimes more robust than computing dense pixel-wise flows for the whole image
- Optical flow prediction can be affected by changes of appearance

Discussion

@55_f5

- The concept of AMD is rather simple but intuitive
- Intuitive since human infants learn object segmentation and detection in an unsupervised manner
- Single image classification may be the opposite of what infants are doing with their “unlabeled video stream” input

Thank you!