Towards Weakly Supervised Object Segmentation & Scene Parsing

Yunchao Wei

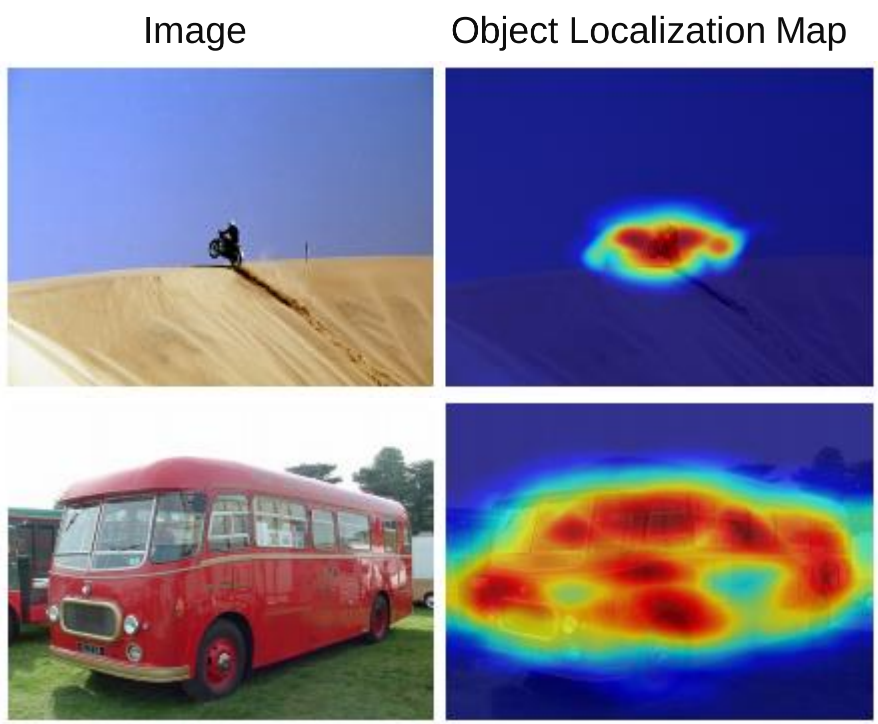
IFP, Beckman Institute, University of Illinois at Urbana-Champaign, IL, USA

Self-Erasing Network for Integral Object Attention

Qibin Hou[1], *Peng-Tao Jiang*[1], *Yunchao Wei* [2], *Ming-Ming Cheng*[1]

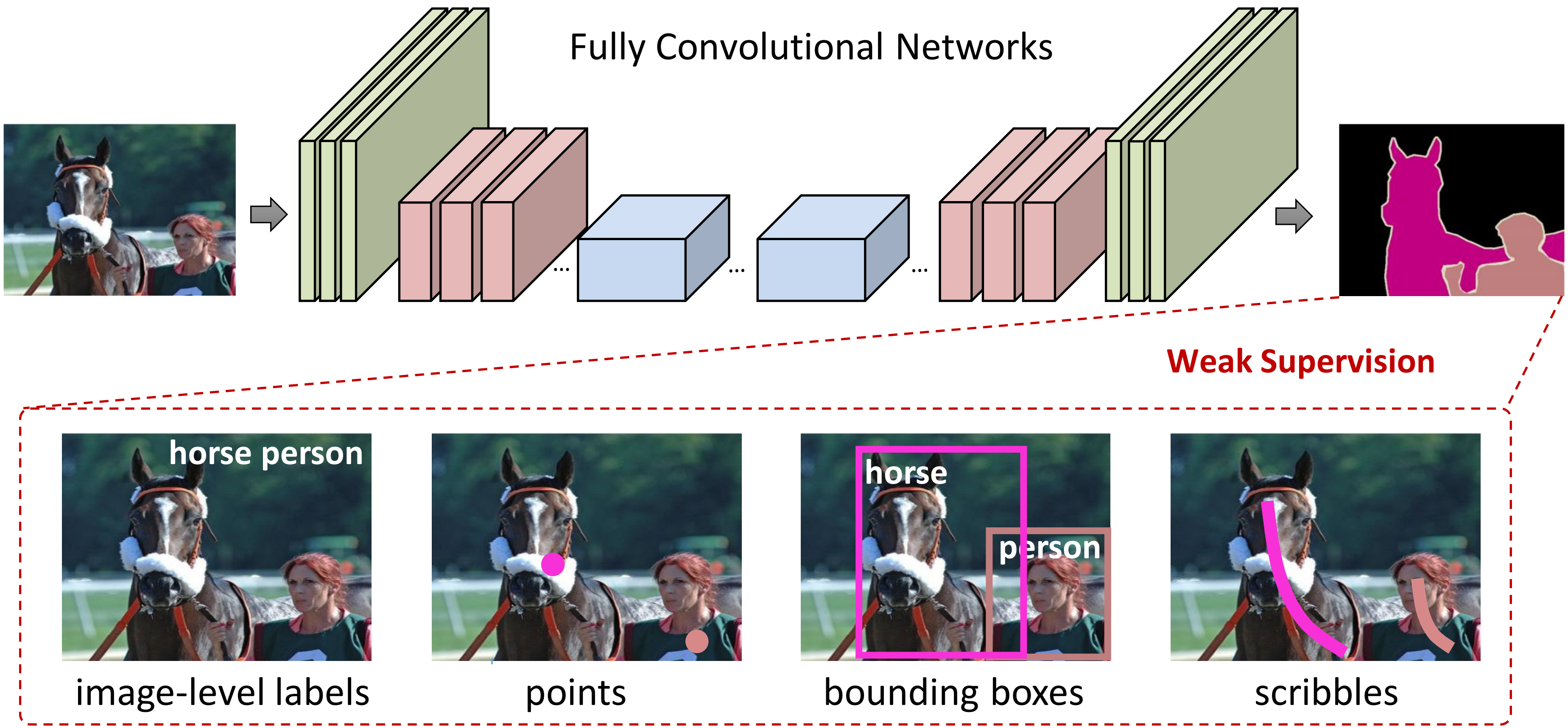
[1]College of Computer Science, Nankai University, Beijing, China

[2]IFP, Beckman Institute, University of Illinois at Urbana-Champaign, IL, USA



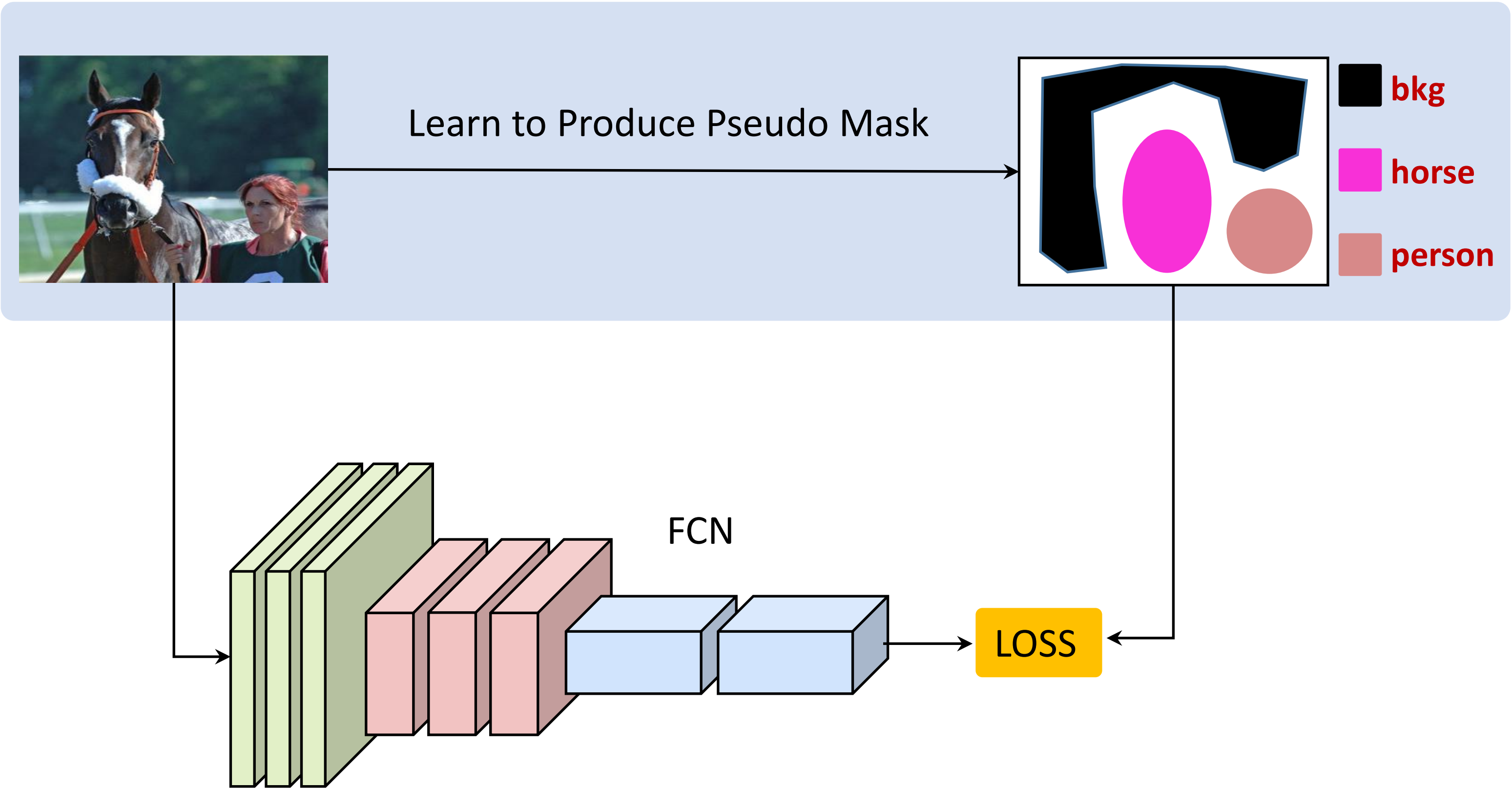
Background

Weak Supervision: Lower degree (or **cheaper**, **simper**) annotations at **training stage** than the required outputs at the **testing stage**.



Background

The Popular Pipeline



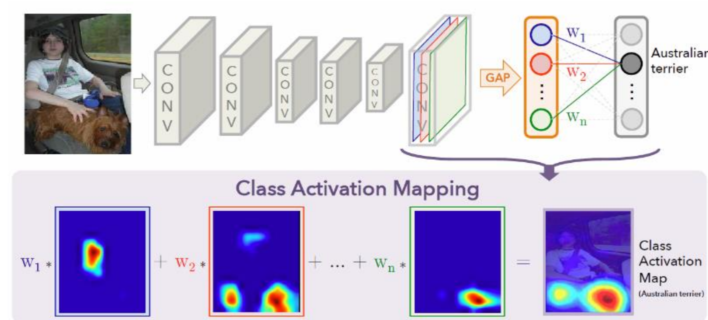
Background

Our Target & Current Issue

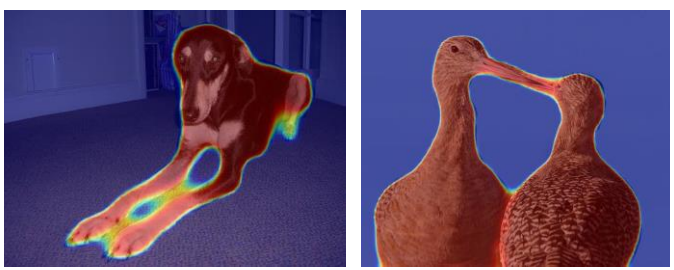
dog

bird

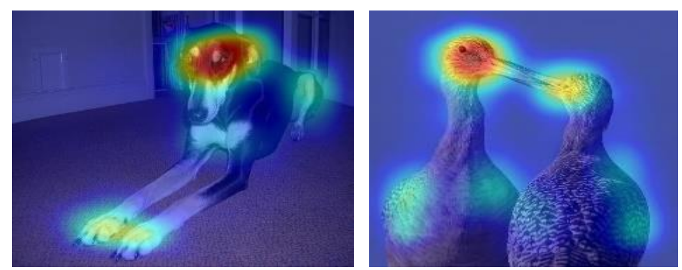
Class Activation Mapping [Zhou CVPR16]



Our target



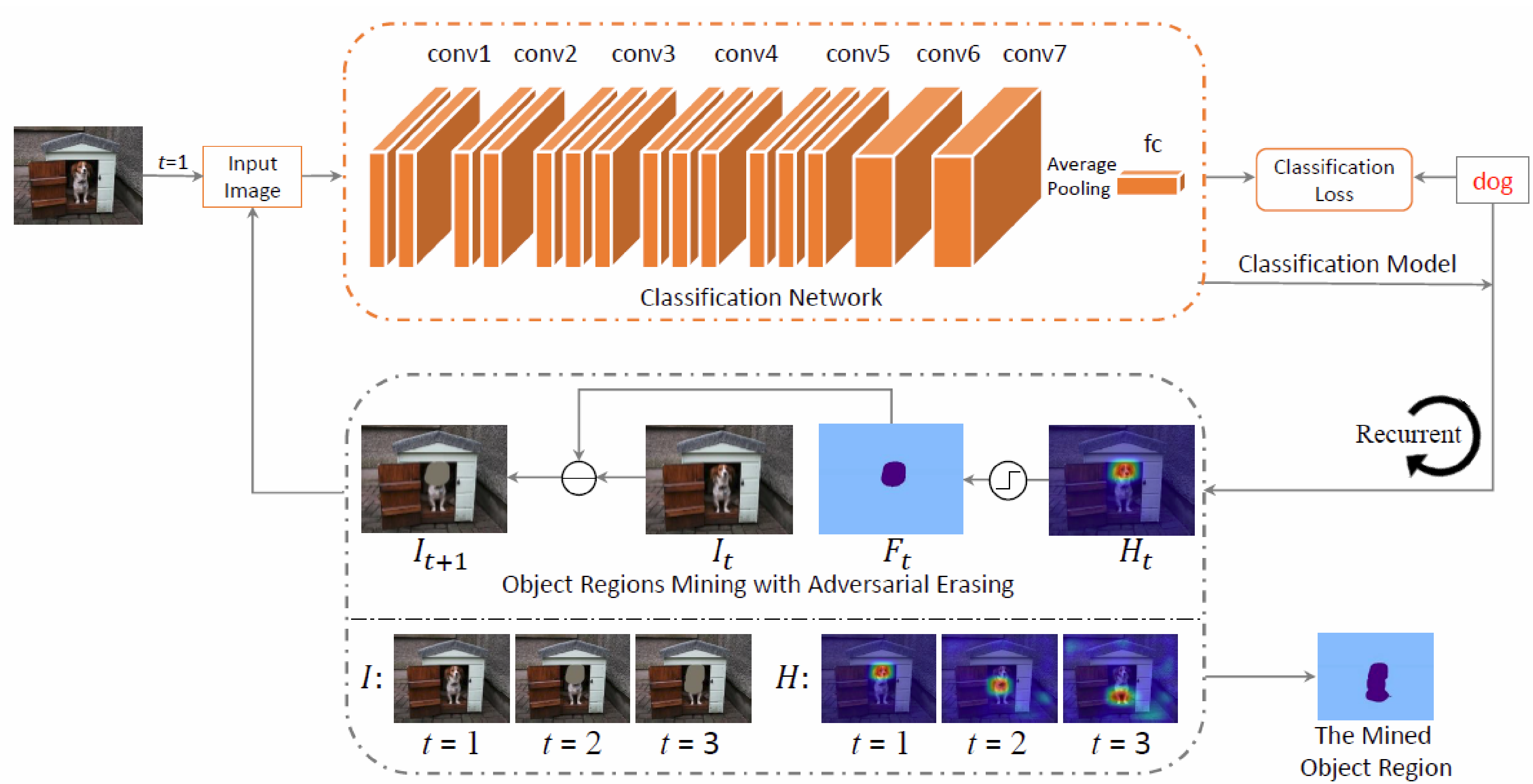
Dense and integral object localization maps



Small and sparse object localization maps

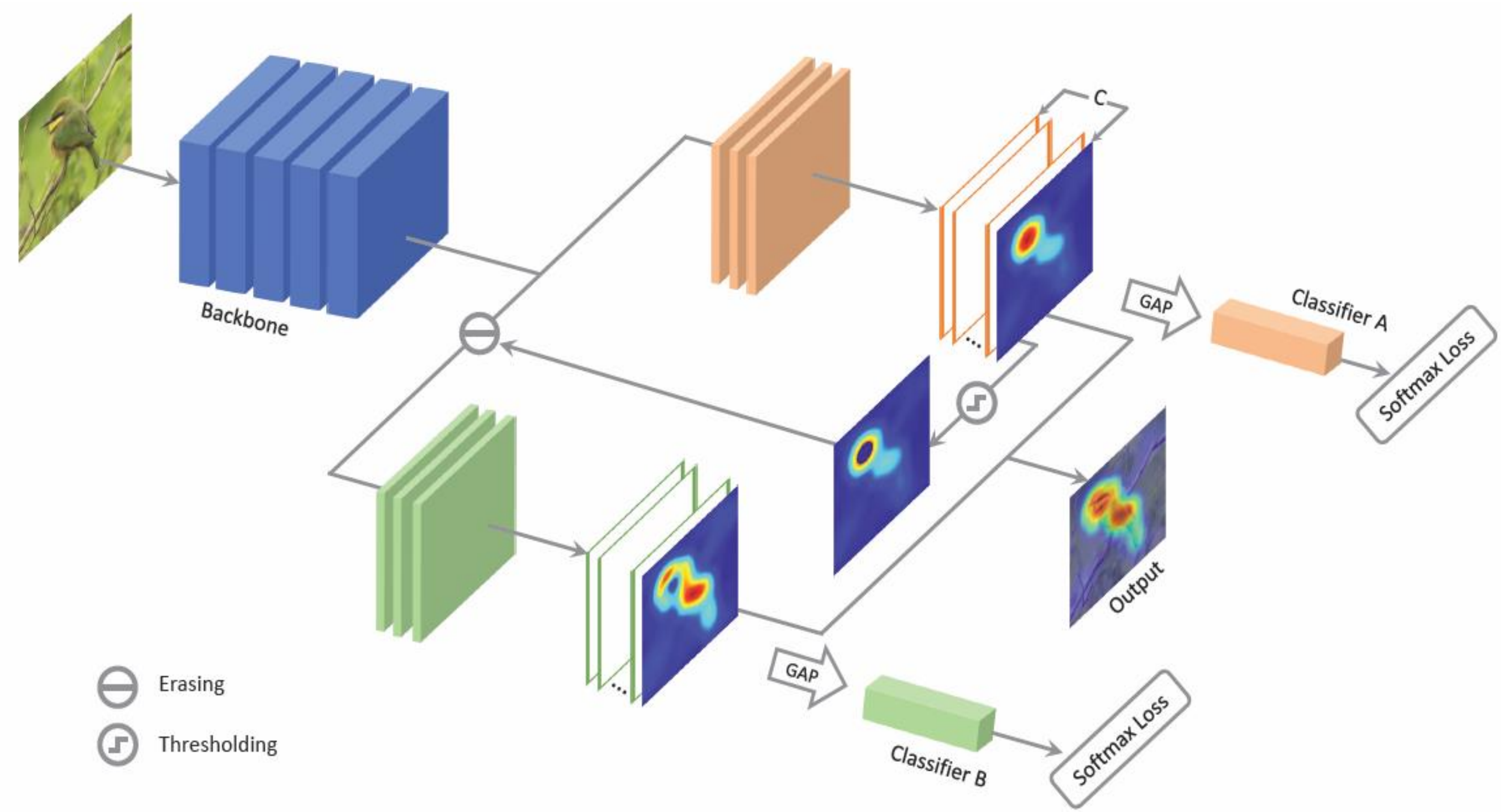
Revisit Adversarial Erasing

Object Region Mining with Adversarial Erasing [Wei CVPR17]



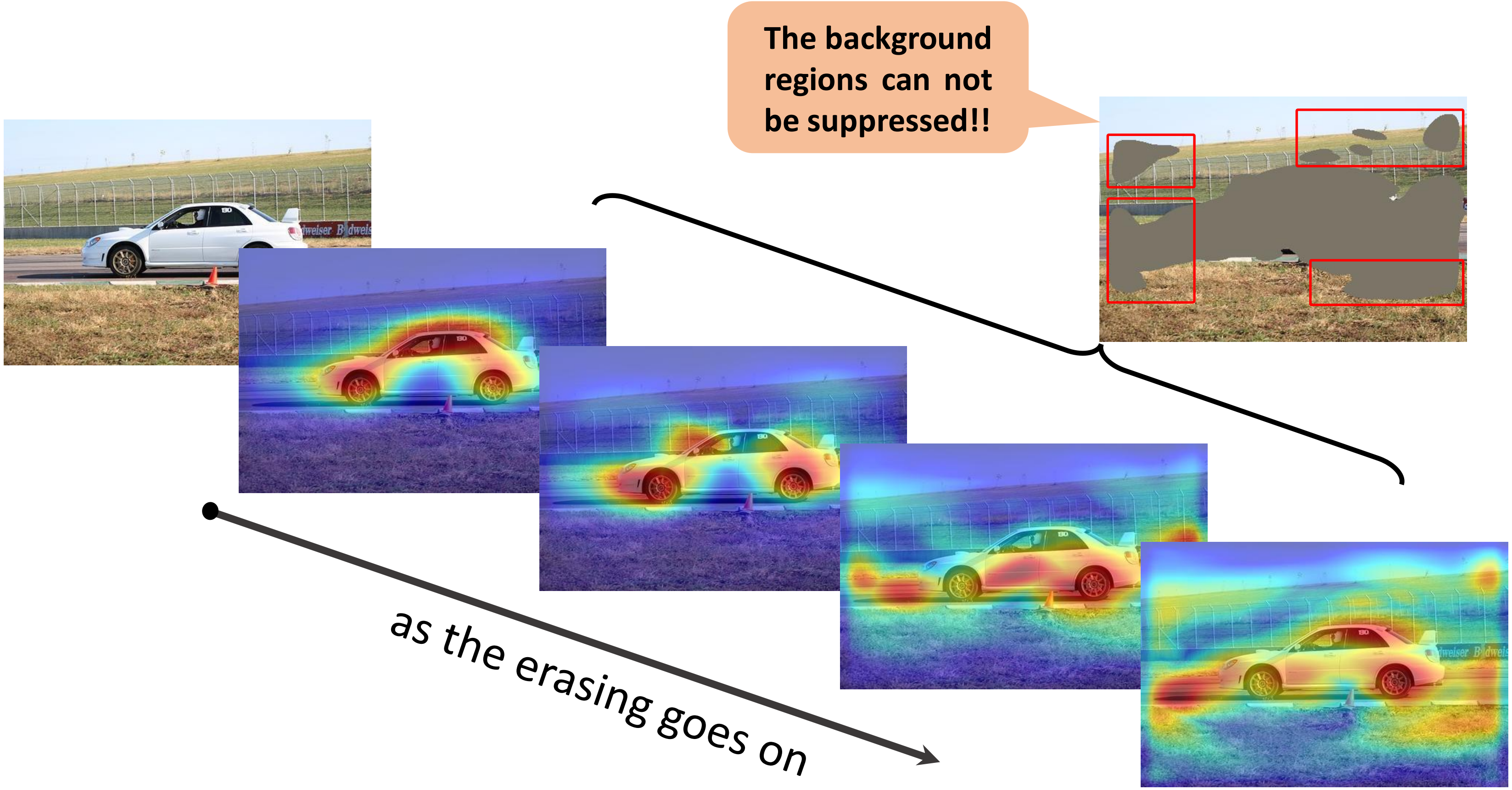
Revisit Adversarial Erasing

Adversarial Complementary Learning [Zhang CVPR18]



Revisit Adversarial Erasing

Over Erasing: The Failure Case of Adversarial Erasing



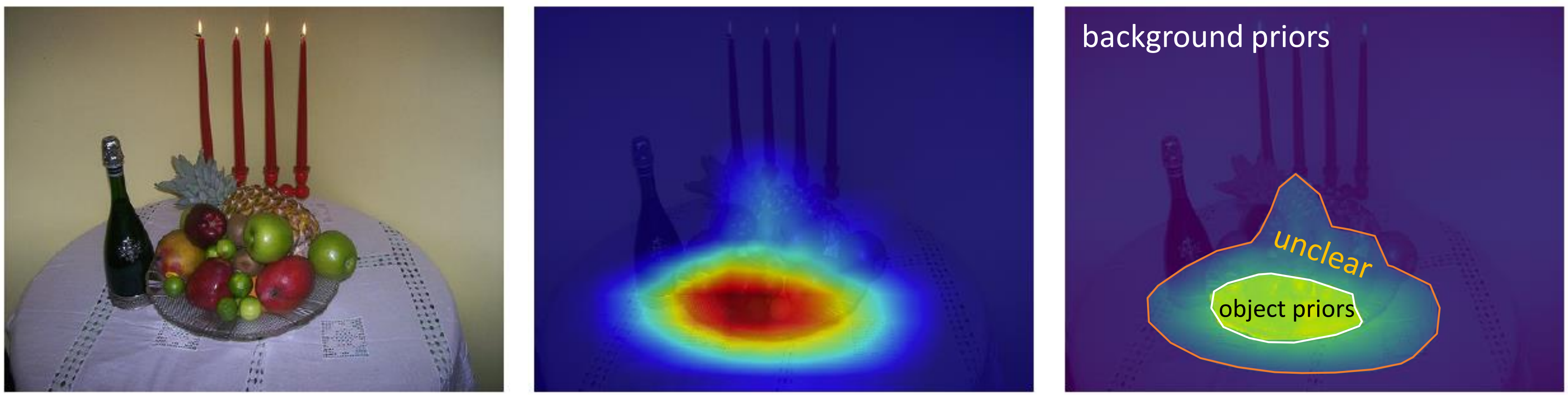
Our Solution: Self-Erasing Network

Motivation

Image

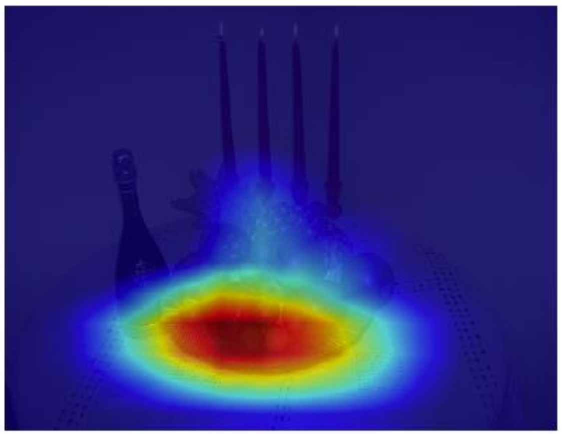
Attention Map

Ternary Mask



Our Solution: Self-Erasing Network

Attention Map



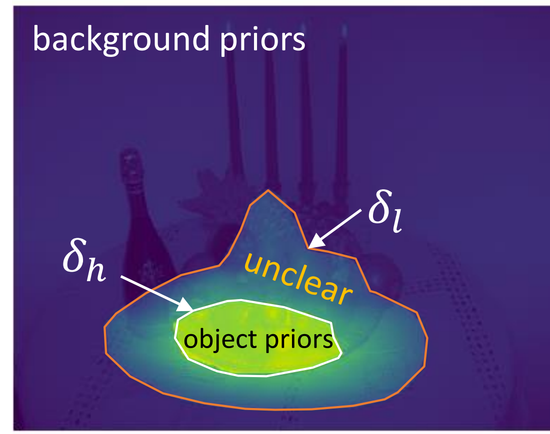
M_A

$$T_{A,(i,j)} = 0 \text{ if } M_{A,(i,j)} \geq \delta_h$$

$$T_{A,(i,j)} = -1 \text{ if } M_{A,(i,j)} < \delta_l$$

$$T_{A,(i,j)} = 1 \text{ otherwise}$$

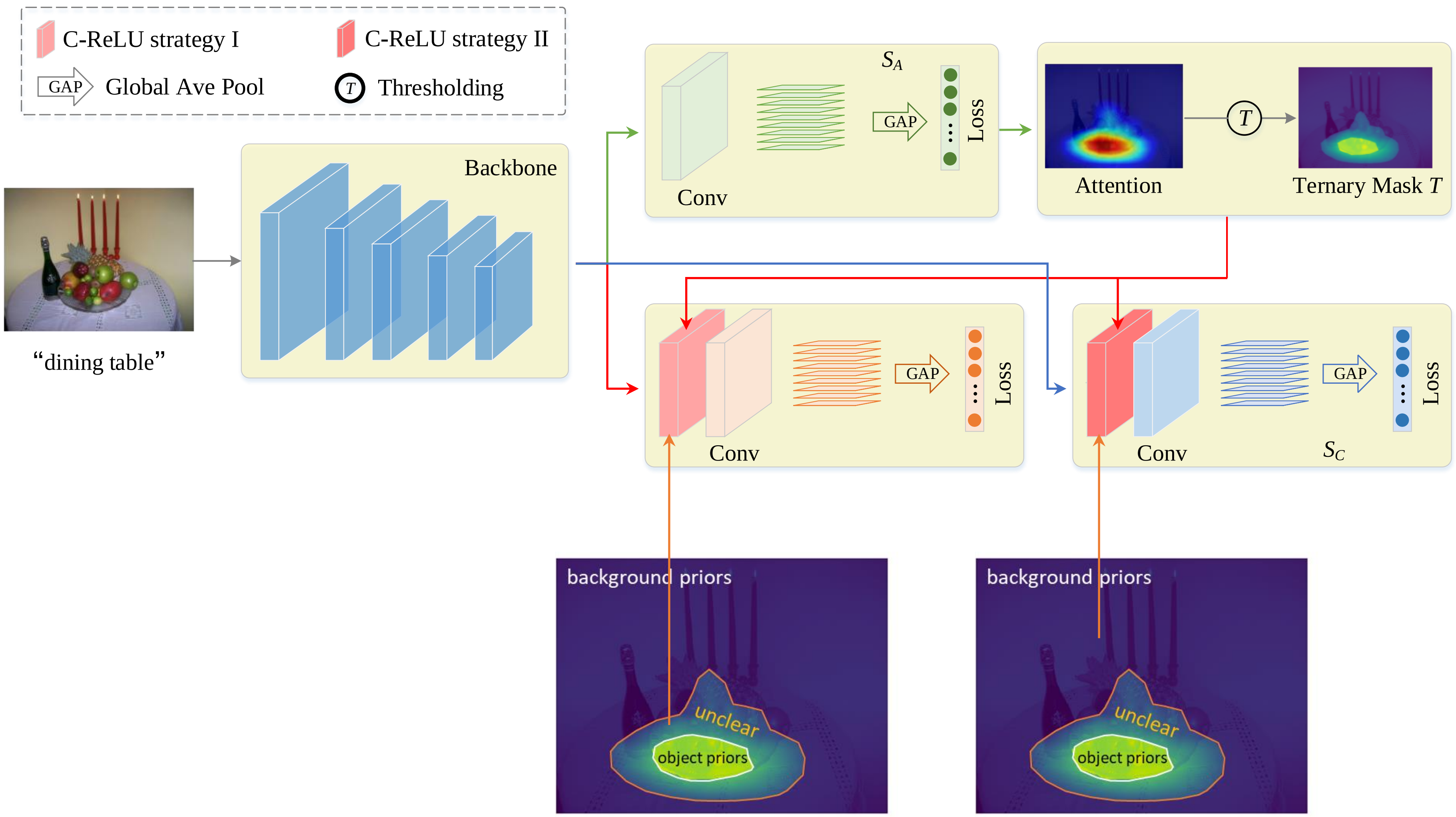
Ternary Mask



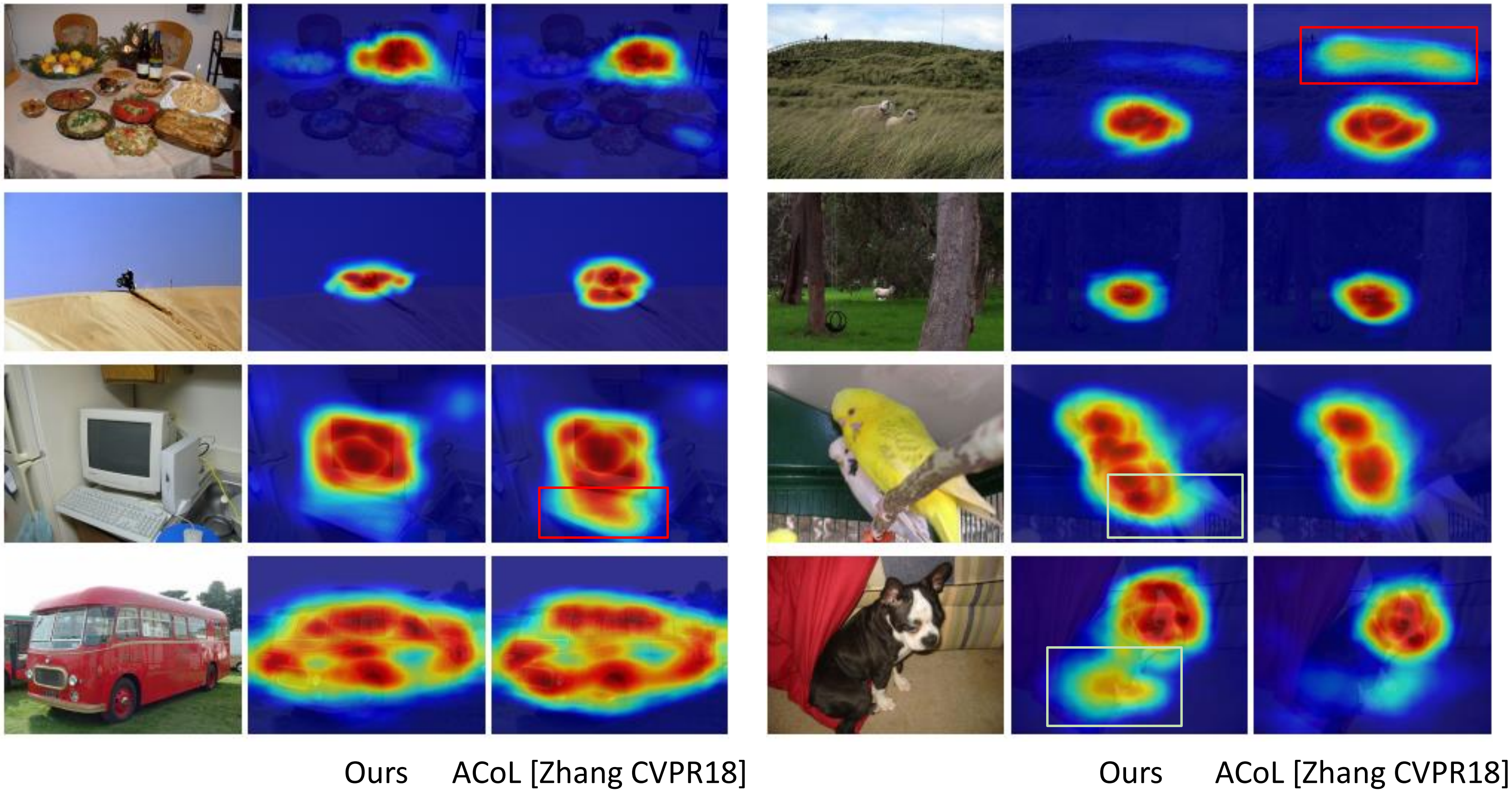
T_A

Our Solution: Self-Erasing Network

Framework



Experimental Results

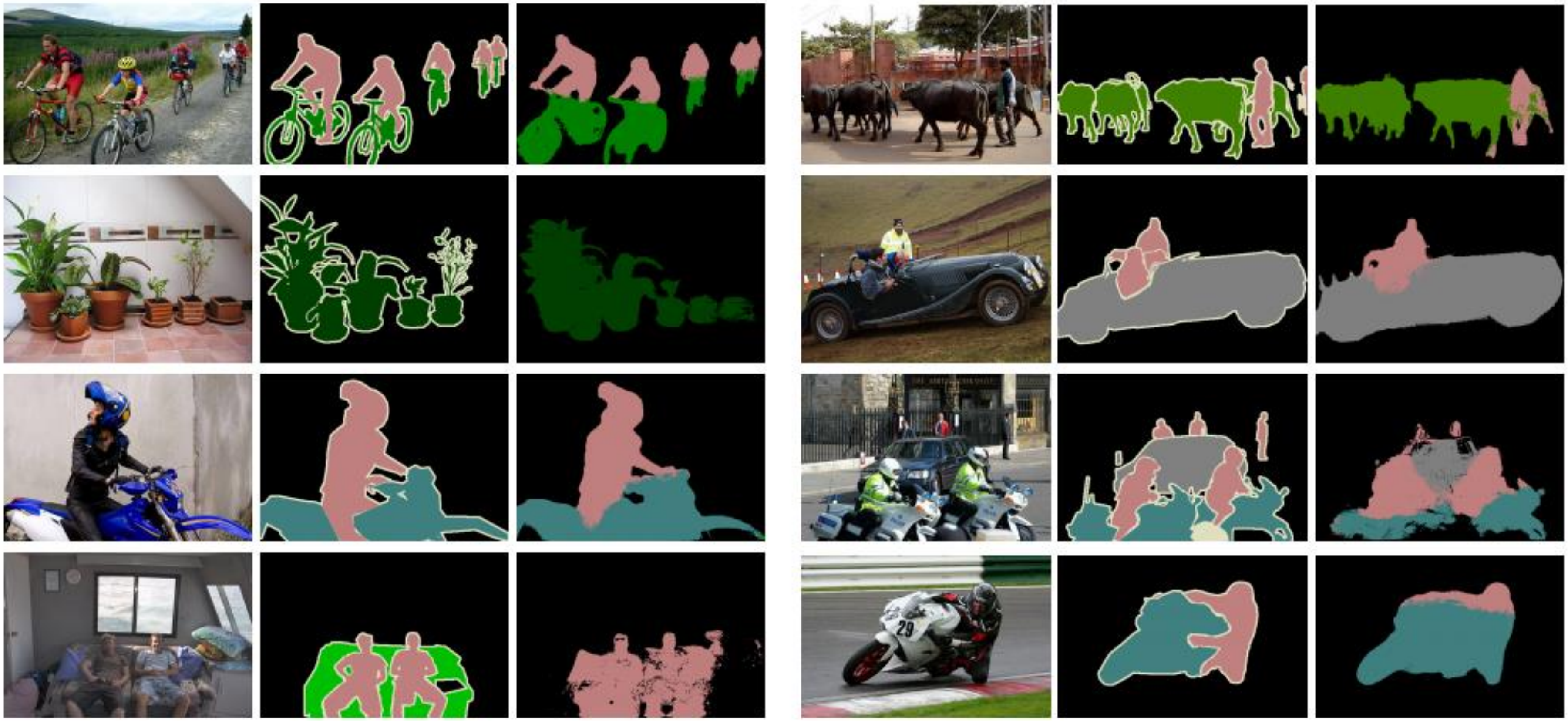


Experimental Results

Pascal VOC 2012

Methods	Publication	Supervision	mIoU (val)		mIoU (test)
			w/o CRF	w/ CRF	w/ CRF
CCNN [25]	ICCV'15	10K weak	33.3%	35.3%	-
EM-Adapt [24]	ICCV'15	10K weak	-	38.2%	39.6%
MIL [26]	CVPR'15	700K weak	42.0%	-	-
DCSM [30]	ECCV'16	10K weak	-	44.1%	45.1%
SEC [16]	ECCV'16	10K weak	44.3%	50.7%	51.7%
AugFeed [27]	ECCV'16	10K weak + bbox	50.4%	54.3%	55.5%
STC [35]	PAMI'16	10K weak + sal	-	49.8%	51.2%
Roy et al. [28]	CVPR'17	10K weak	-	52.8%	53.7%
Oh et al. [23]	CVPR'17	10K weak + sal	51.2%	55.7%	56.7%
AE-PSL [34]	CVPR'17	10K weak + sal	-	55.0%	55.7%
Hong et al. [9]	CVPR'17	10K + video weak	-	58.1%	58.7%
WebS-i2 [14]	CVPR'17	19K weak	-	53.4%	55.3%
DCSP-VGG16 [3]	BMVC'17	10K weak + sal	56.5%	58.6%	59.2%
DCSP-ResNet101 [3]	BMVC'17	10K weak + sal	59.5%	60.8%	61.9%
TPL [15]	ICCV'17	10K weak		53.1%	53.8%
GAIN [39]	CVPR'18	10K weak + sal	-	55.3%	56.8%
SeeNet (Ours, VGG16)	-	10K weak + sal	59.9%	61.1%	60.7%
SeeNet (Ours, ResNet101)	-	10K weak + sal	62.6%	63.1%	62.8%

Experimental Results



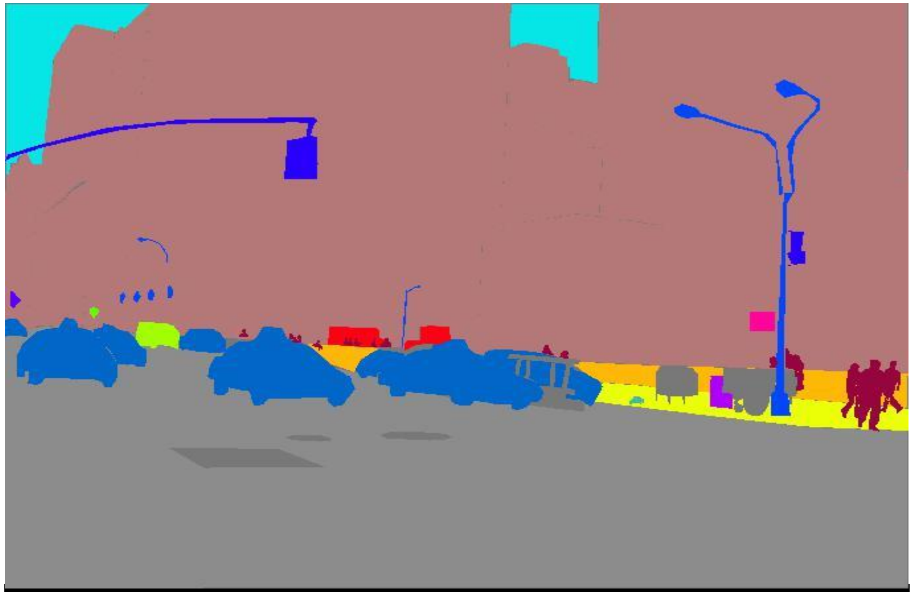
Weakly Supervised Scene Parsing with Point-based Distance Metric Learning

Rui Qian[1,3], Yunchao Wei [3], Honghui Shi [2,3], Jiachen Li [3], Jiaying Liu[1] and Thomas Huang[3]

[1]Institute of Computer Science and Technology, Peking University, Beijing, China

[2]IBM T.J. Waston Research Center, [3]IFP, Beckman, UIUC

Point Annotation



Full Annotation

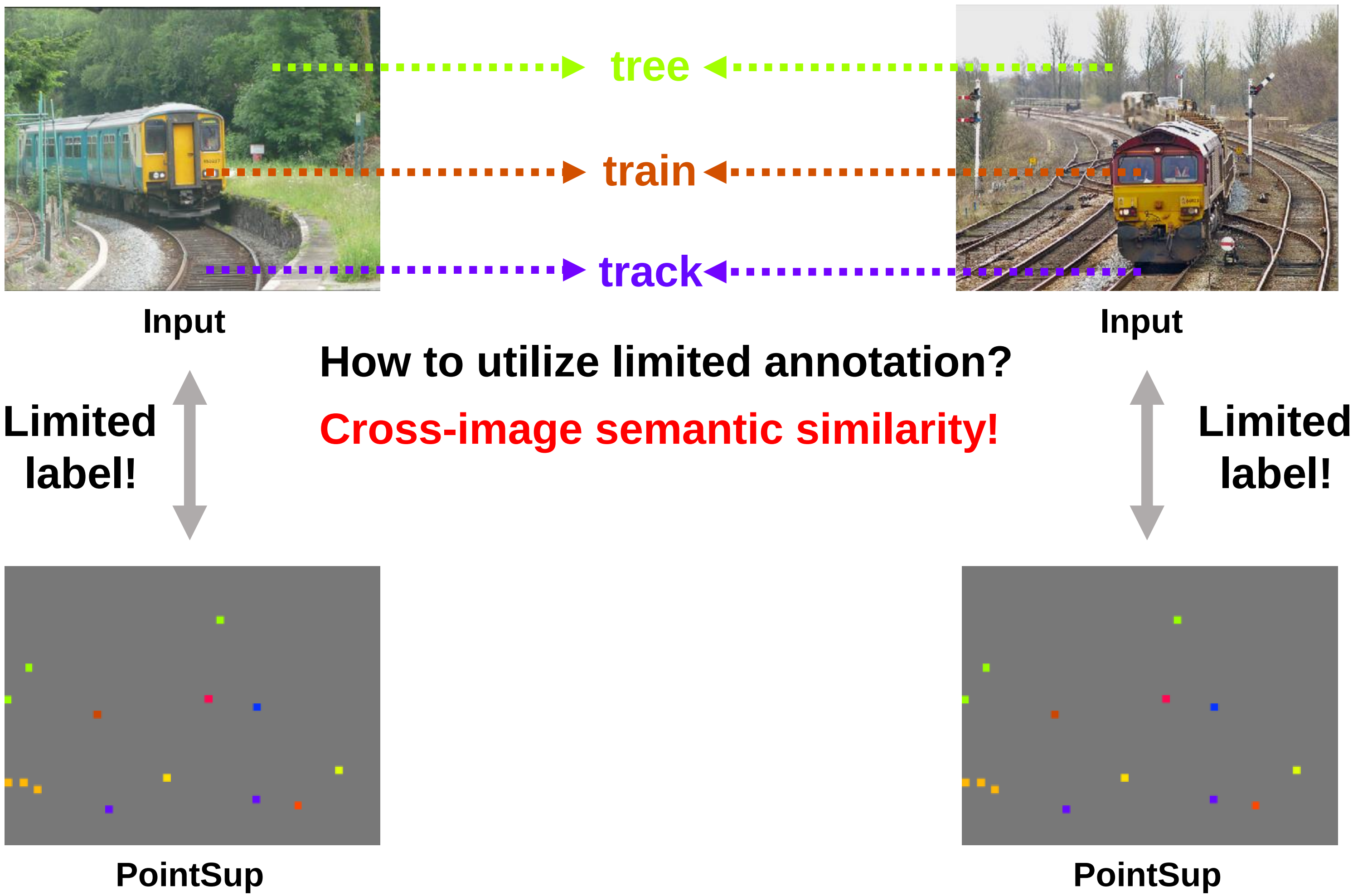
Review on Weakly Supervised Methods

- **Weakly supervised methods for scene parsing**
  - Image-level
  - Box supervision
  - Scribble supervision
  - Point supervision

Annotation burden comparison

Method	Full	Scribble	Point
Average Anno.pixel/Image	170K	1817.48	**12.26**

Motivation



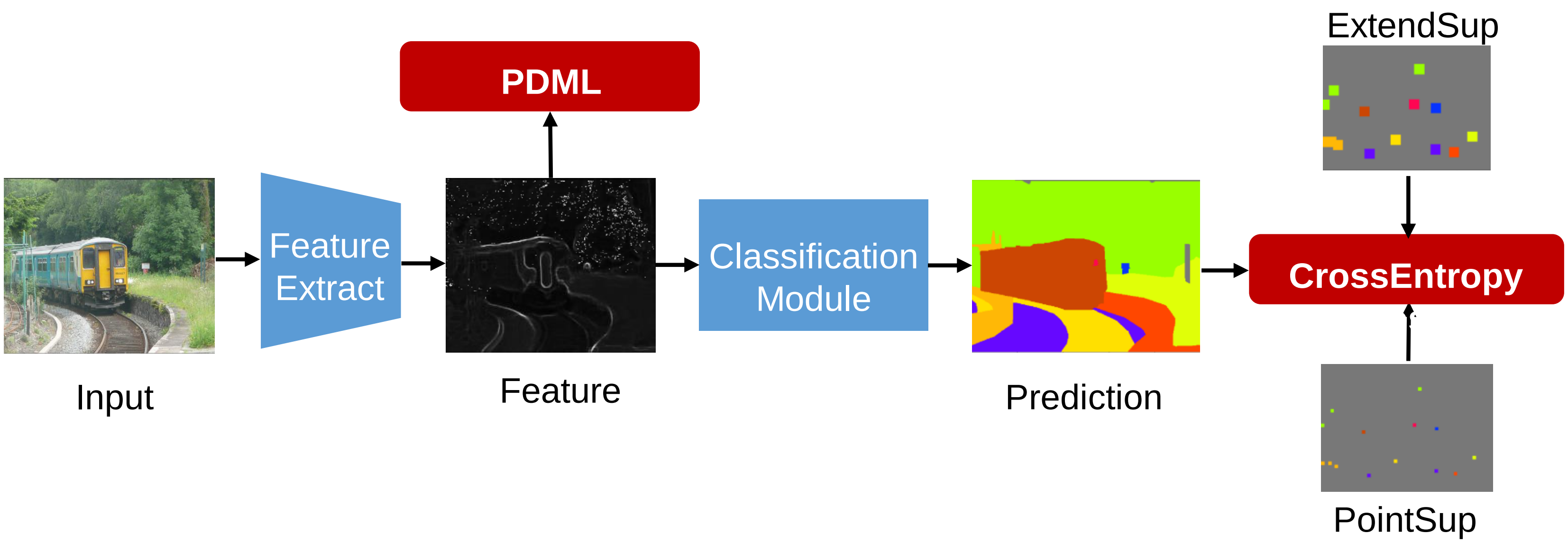
How to utilize limited annotation?

Cross-image semantic similarity!

Proposed method(1/5)

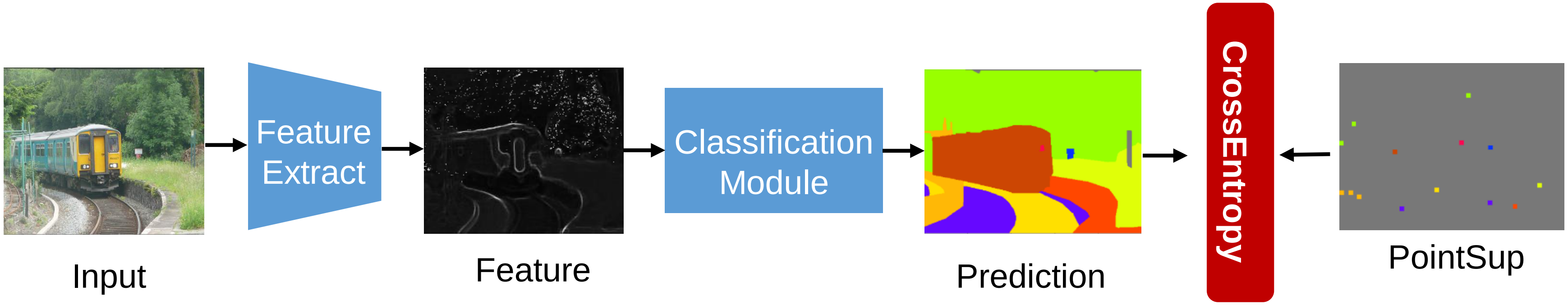
Overview

- Point-based distance metric learning(PDML)
- Point supervision(PointSup)
- Online extension supervision(ExtendSup)



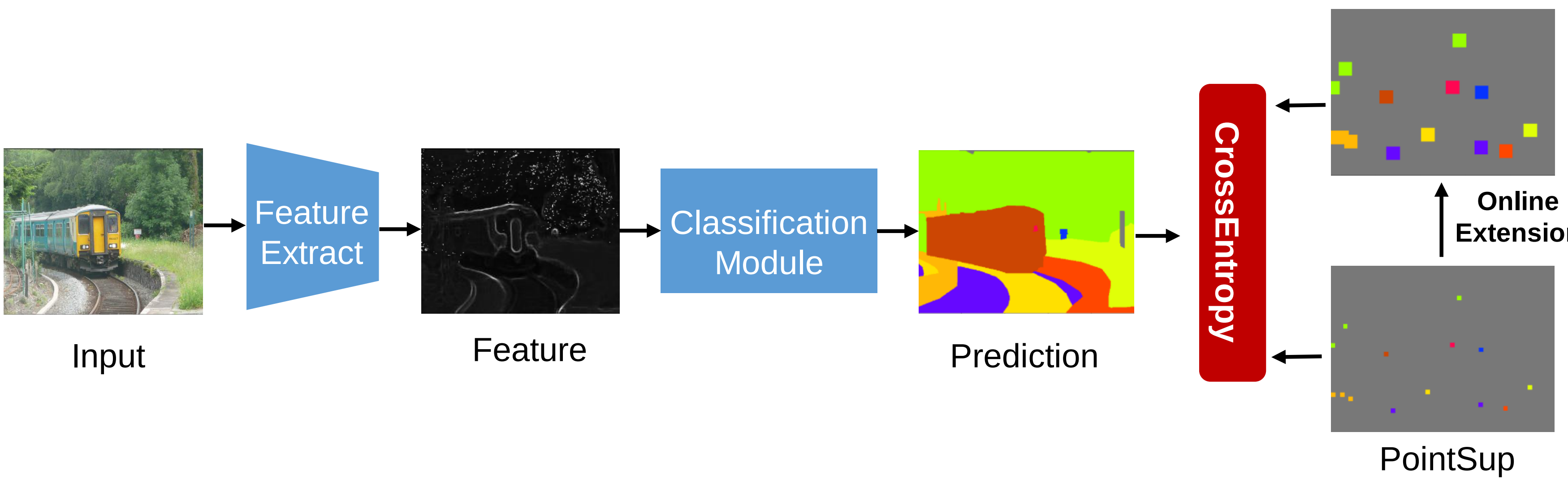
Proposed method(2/5)

- **Point supervision**
  - Only calculate cross-entropy loss on annotated pixels
  - Back propagate gradients accordingly
  - Optimize by stochastic gradient descent



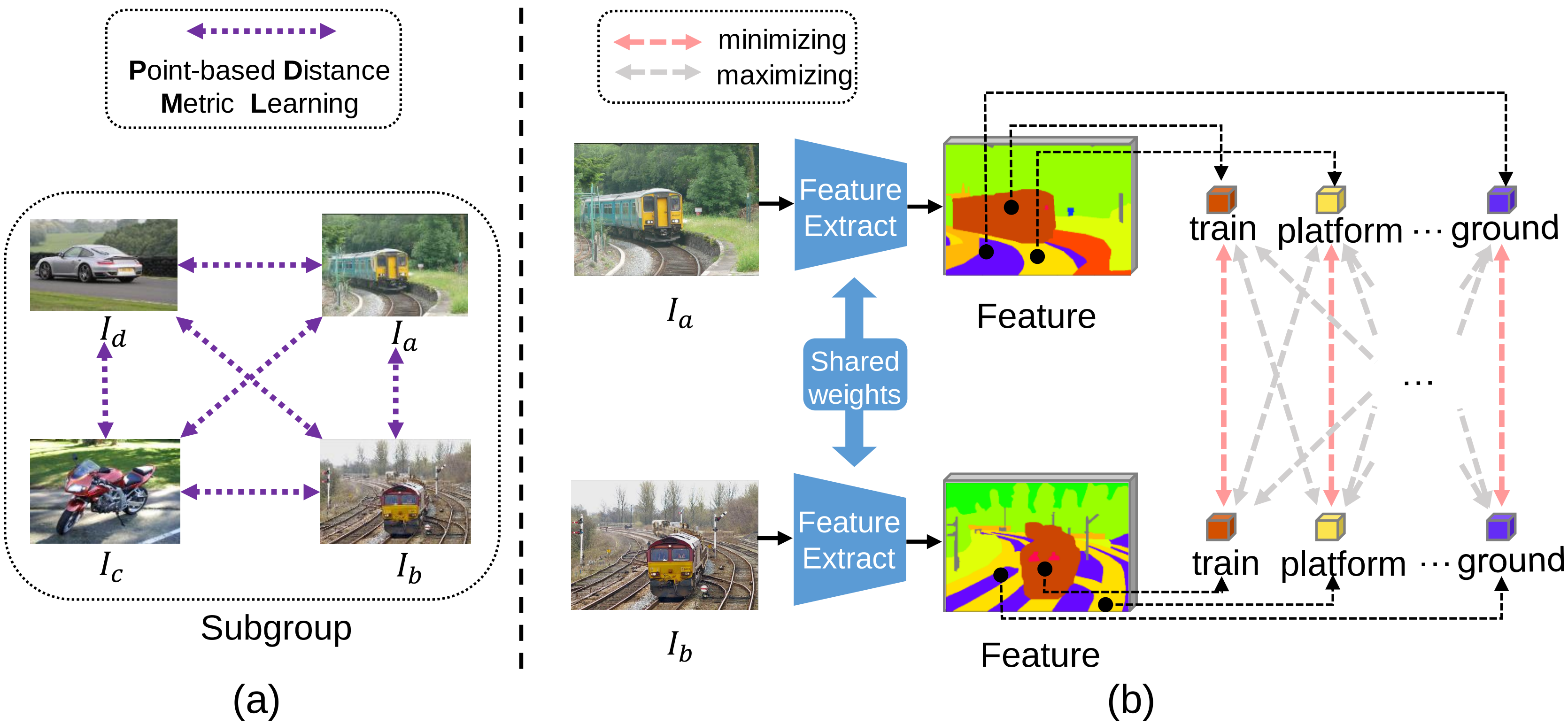
Proposed method(3/5)

- **Online extension supervision**
  - Extension method1 (region):
    - Select pixels in 5*5 square near the annotated ones
  - Extension method2 (score):
    - Select pixels with score over 0.7 in the prediction
  - Finally choose the intersection of two methods



Proposed method(4/5)

Point-based distance metric learning



Loss function of PDML

- For each image I_a, define the embedding vector set as E_a:
  - $E_a = \bigcup_{i=1}^{|M_a|}\{P_{ai}\}$
  - $|M_a|$ is the number of annotated pixel of I_a
  - P_{ai} is the feature vector of ith pixel
- We optimize in the triplet form of $\{P_{ai}, P_{bj}, P_{bk}\}$:
  - P_{ai} shares the same category with P_{bj}
  - P_{bk} shares different category with P_{ai}, P_{bj}
- We use the loss function of :

$$L_t(P_{ai}, P_{bj}, P_{bk}) = \alpha L_p(P_{ai}, P_{bj}) + \beta L_n(P_{ai}, P_{bj}, P_{bk})$$

  - $L_p(P_{ai}, P_{bj}) = ||P_{ai} - P_{bj}||_2$
  - $L_n(P_{ai}, P_{bj}, P_{bk}) = \max(||P_{ai} - P_{bj}||_2 - ||P_{ai} - P_{bk}||_2 + m, 0)$
  - α, β, m are hyper-params and are set to 0.8, 1, 20 in practice

- **Scene parsing datasets**
  - PASCAL-Context
  - ADE 20K

Dataset	#Training	#Evaluation	#Instance/Image
PASCAL-Context	4998	5105	12.26
ADE20K	20210	2000	13.96

Quantitative evaluation on PASCAL-Context

- The combination of three techniques is best
- We use only 0.007% annotated data but reached 75% of the full supervision performance!

Method				Metrics	
FullSup	PointSup	PDML	Online Ext.	mIoU	Pixel Acc
√				39.6	78.6%
	√			27.9	55.3%
	√	√		29.7	57.5%
	√	√	√	**30.0**	**57.6%**

Quantitative evaluation on ADE20K

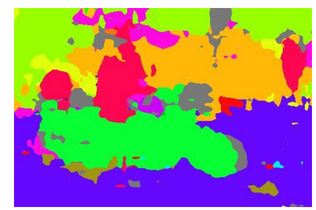
- The combination of three techniques is best
- Our method approaches the result SegNet under full supervision scheme

Method				Metrics	
FullSup	PointSup	PDML	Online Ext.	mIoU	Pixel Acc
√				33.9	75.8%
√(SegNet)				21.0	/
	√			17.7	58.0%
	√	√		19.0	59.0%
	√	√	√	**19.6**	**61.0%**

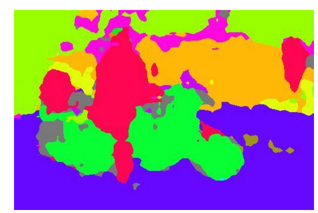
Subjective Evaluation

Image

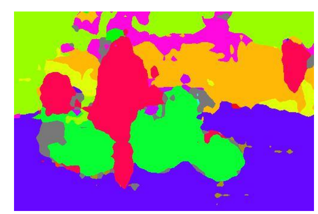
GT



PointSup

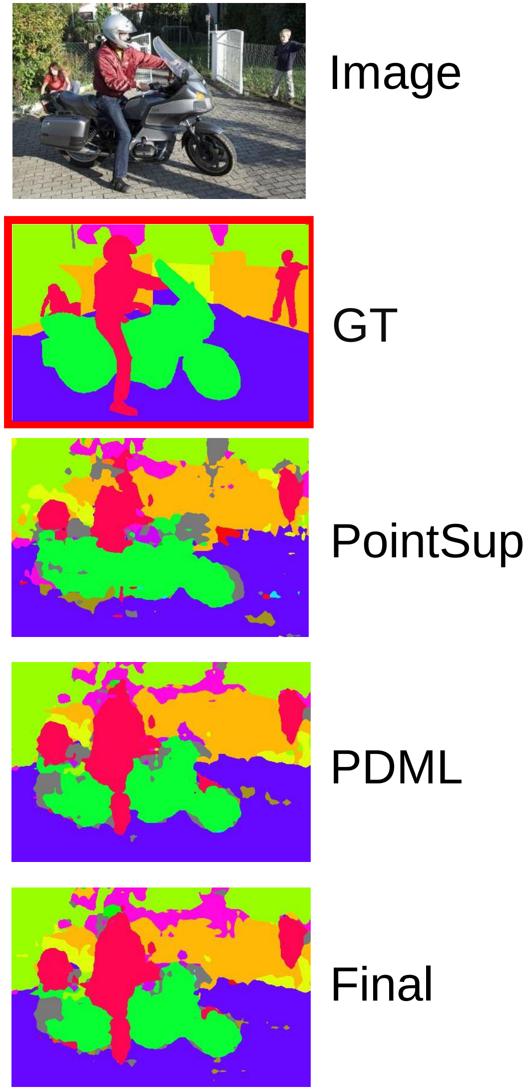


PDML

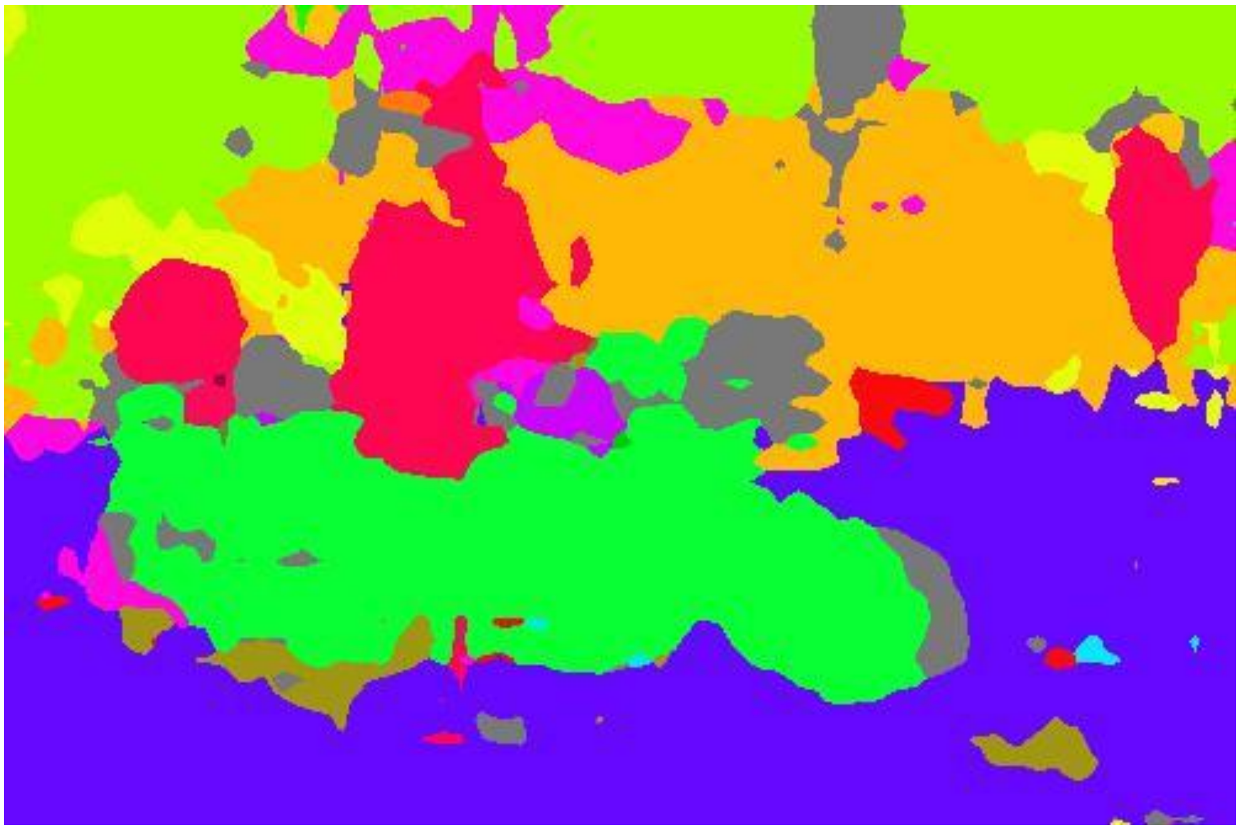


Final

Subjective Evaluation

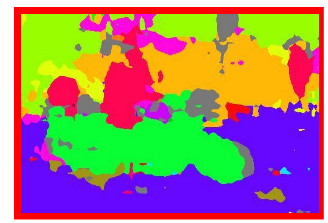


Subjective Evaluation

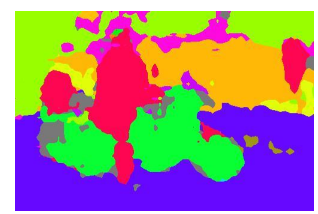


Image

GT



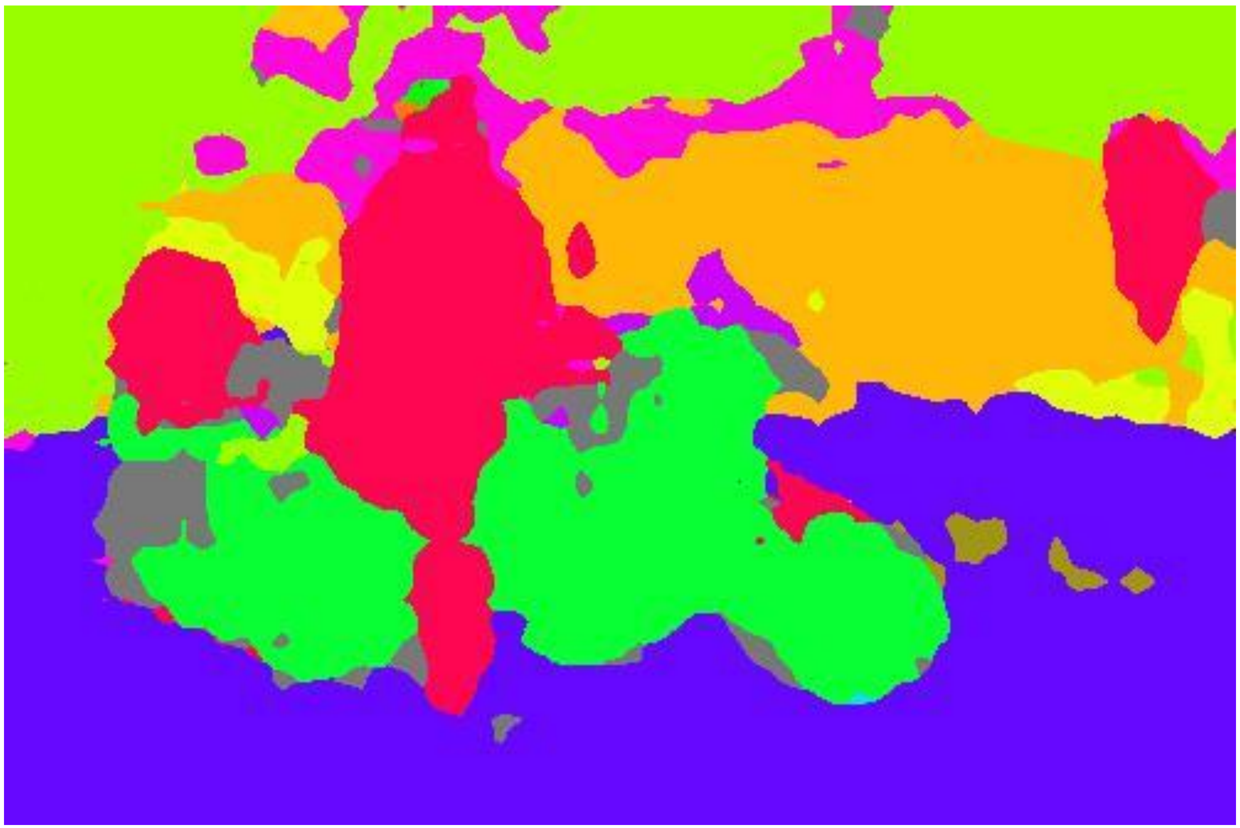
PointSup

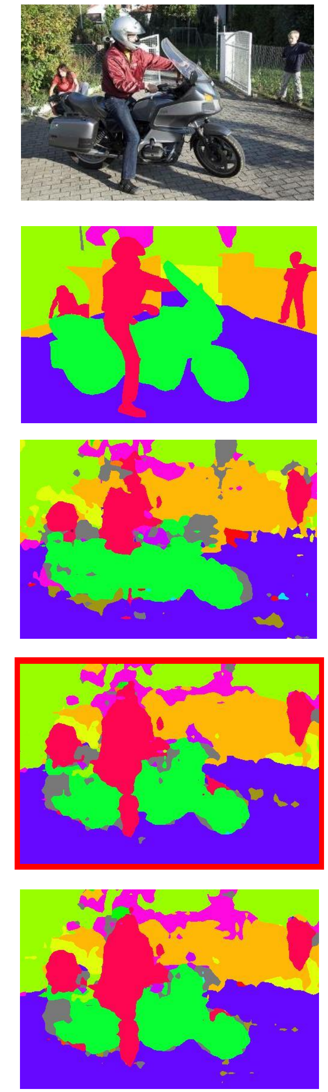


PDML

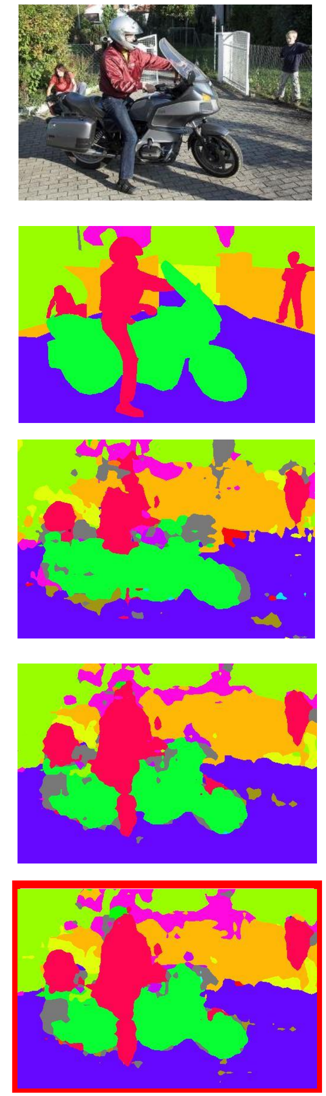
Final

Subjective Evaluation





Subjective Evaluation



Image

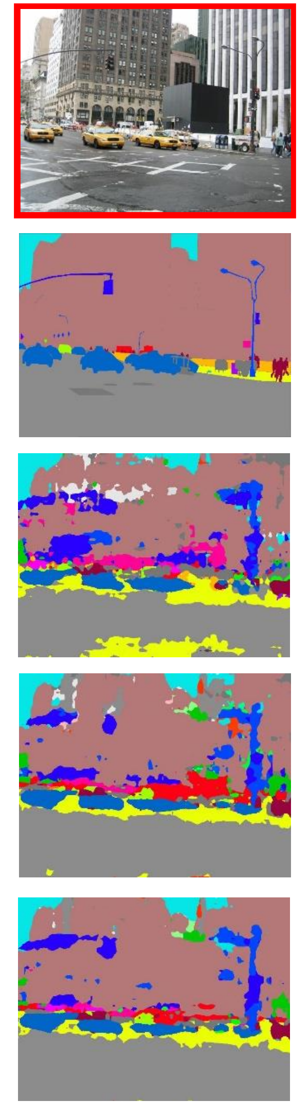
GT

PointSup

PDML

Final

Subjective Evaluation



Image

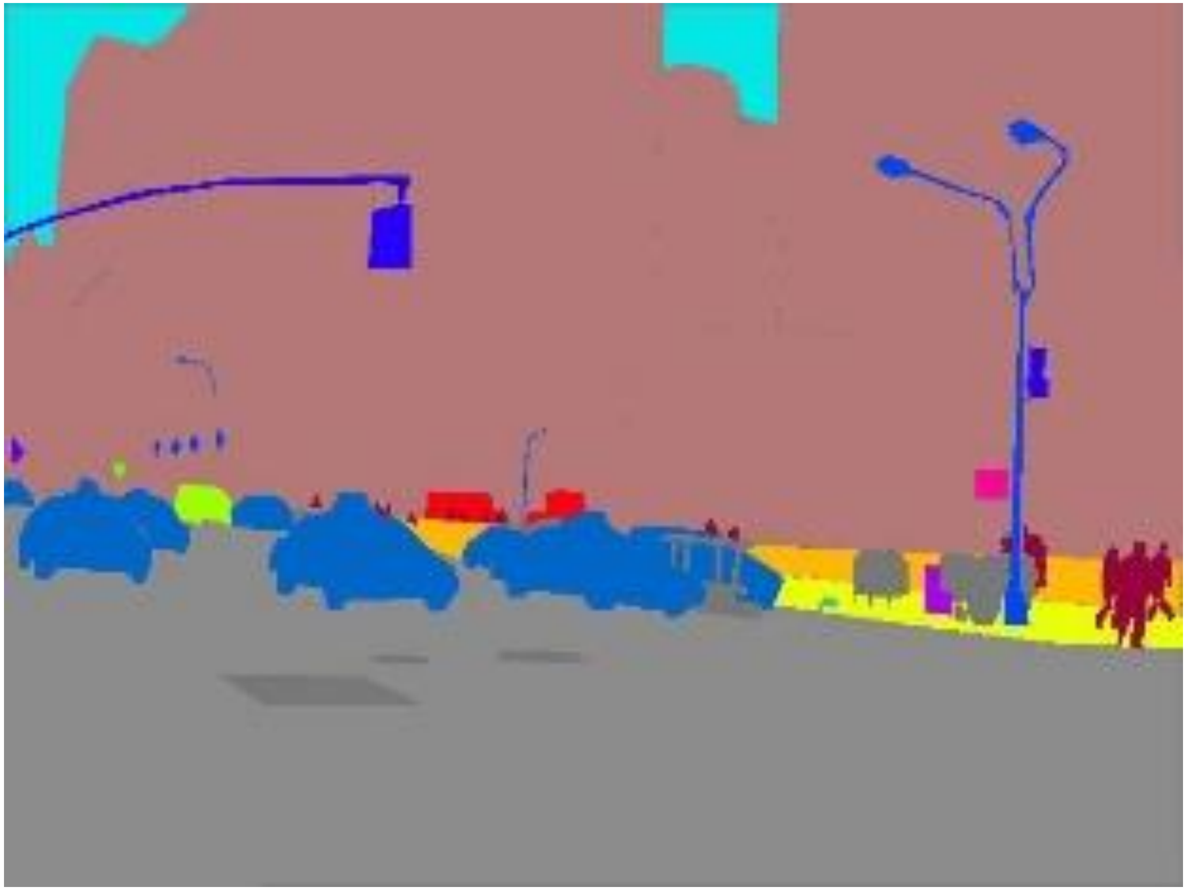
GT

PointSup

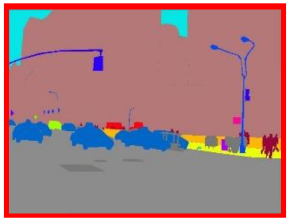
PDML

Final

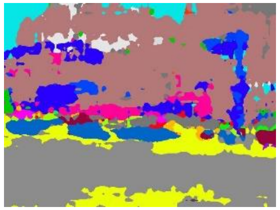
Subjective Evaluation



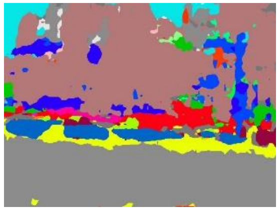
Image



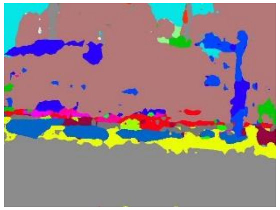
GT



PointSup

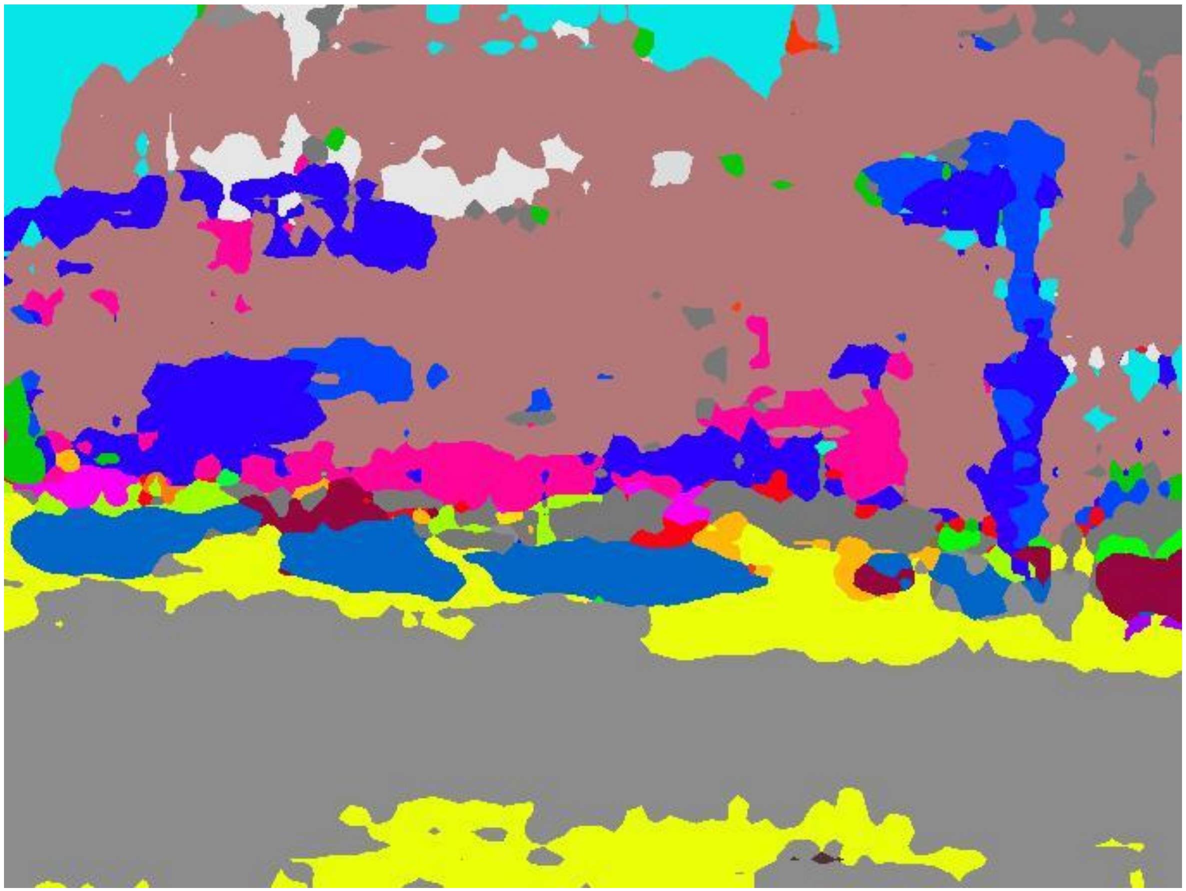


PDML

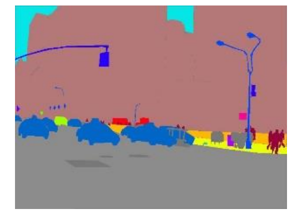


Final

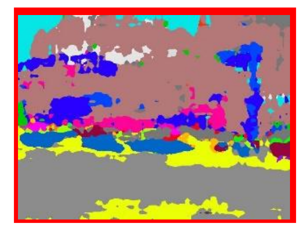
Subjective Evaluation



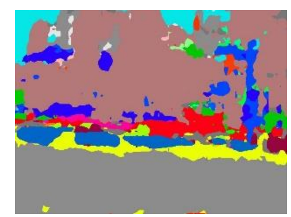
Image



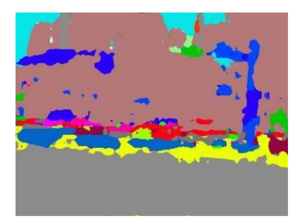
GT



PointSup

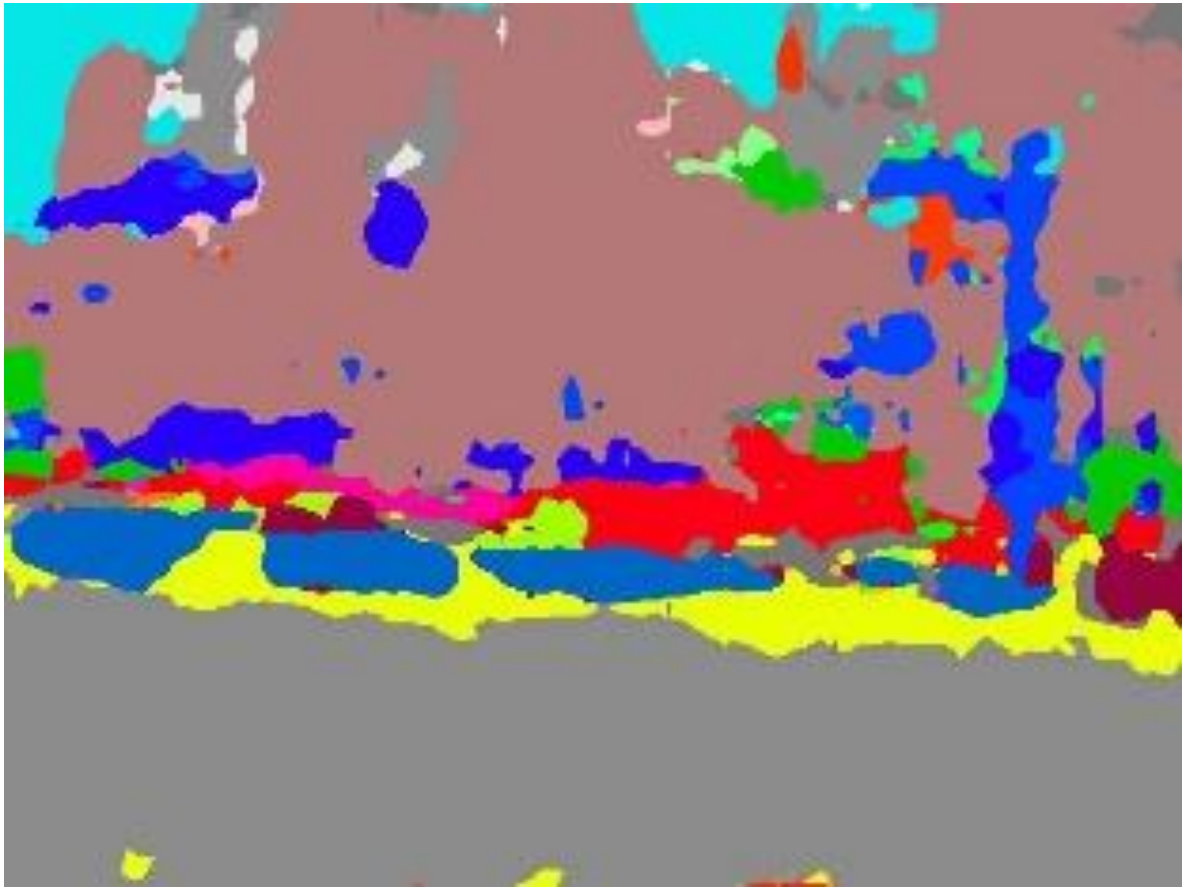


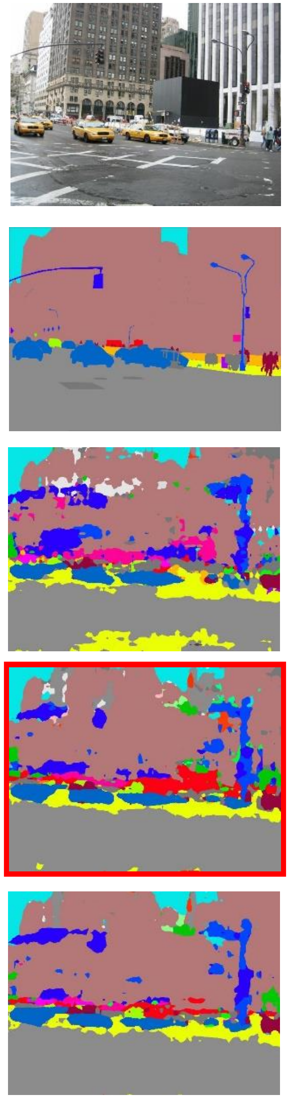
PDML



Final

Subjective Evaluation





Image

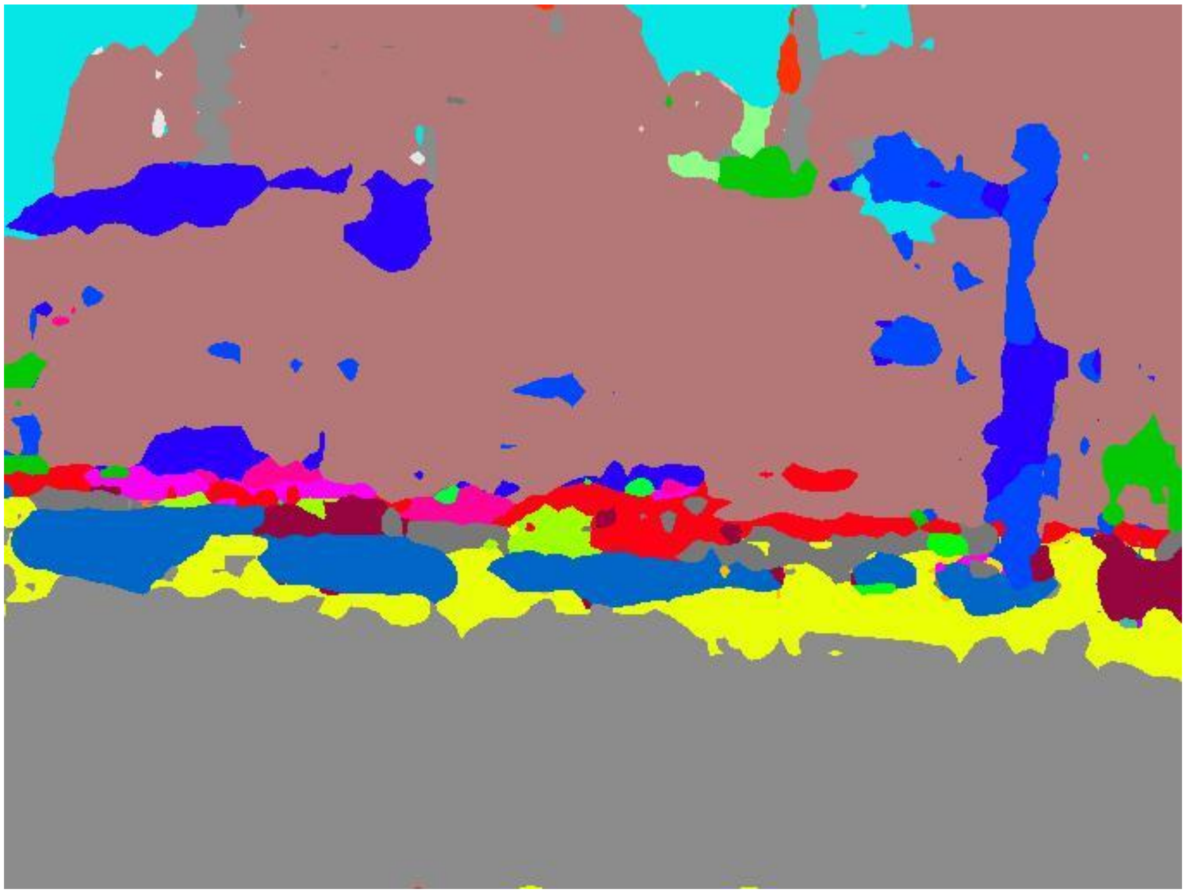
GT

PointSup

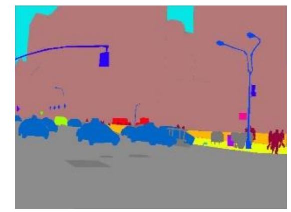
PDML

Final

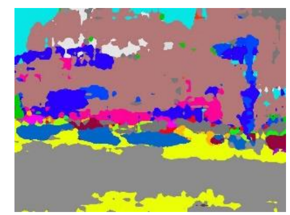
Subjective Evaluation



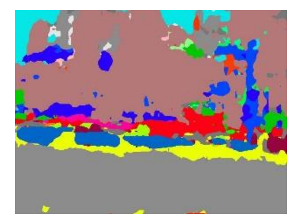
Image



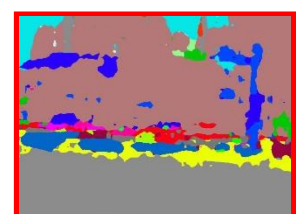
GT



PointSup



PDML

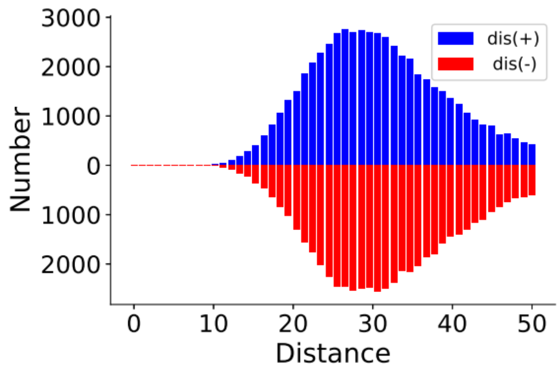


Final

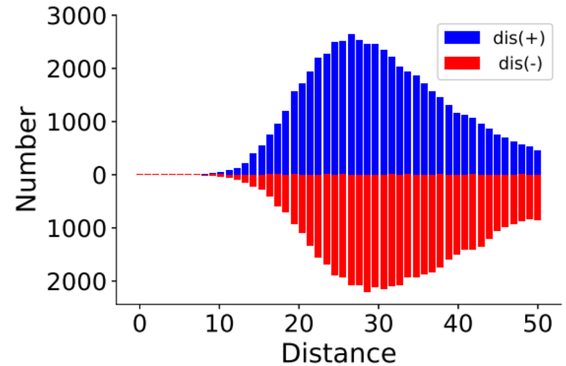
Discussion

Visualization the effect of PDML

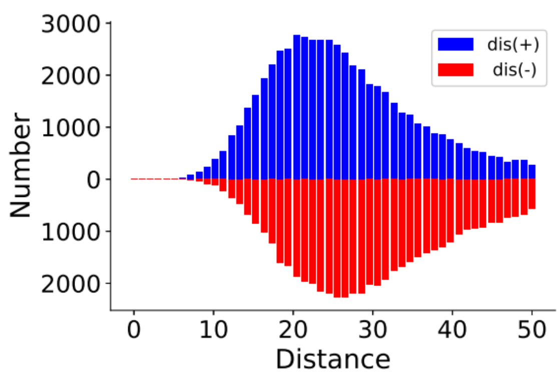
- dis(+): L2 norm distance between same-class feature vectors
- dis(-): L2 norm distance between different-class feature vectors



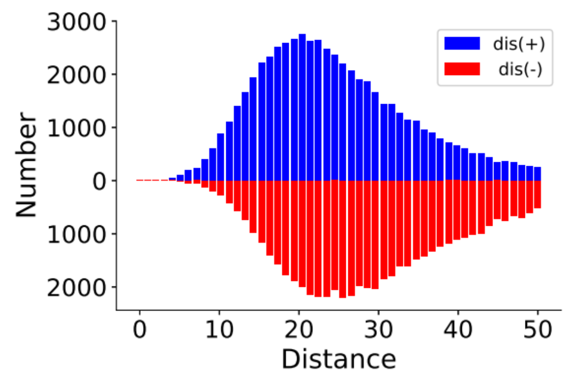
(a) Epoch = 1



(b) Epoch = 20



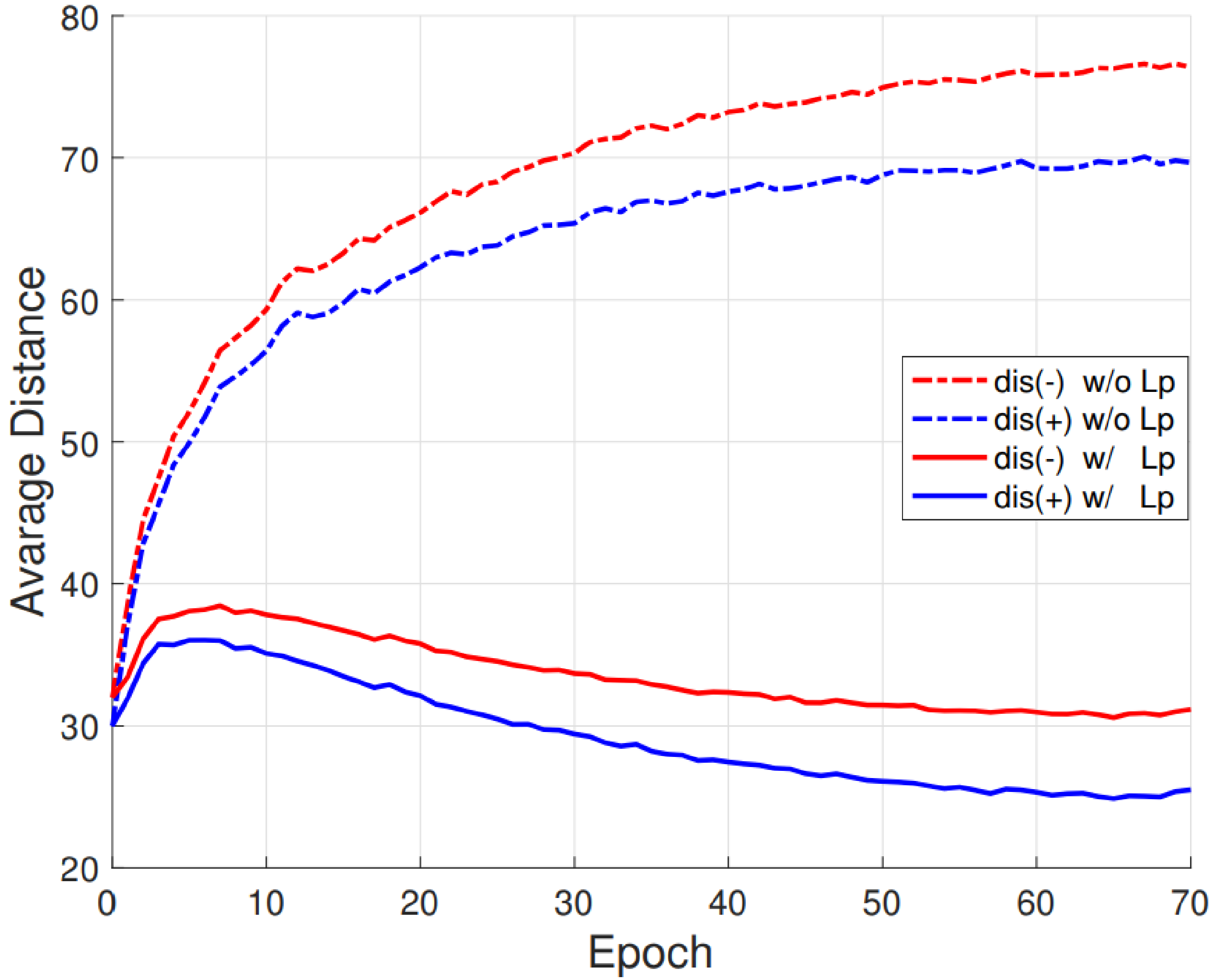
(c) Epoch = 40



(d) Epoch = 60

Ablation on the design of PDML loss function

- $L_t(P_{ai}, P_{bj}, P_{bk}) = \alpha L_p(P_{ai}, P_{bj}) + \beta L_n(P_{ai}, P_{bj}, P_{bk})$



Conclusion

- Problem
  - Point-guided scene parsing
- Point-based distance metric learning
  - Exploit semantic relationship across images
- Experimental results
  - Good performance both quantitatively and qualitatively

Conclusion

Weakly Supervised Learning for Real-World Computer Vision Applications & The 1st Learning from Imperfect Data (LID) Challenge

CVPR 2019 Workshop, Long Beach, CA

https://lidchallenge.github.io/

Task 1

Object Segmentation on ILSVRC DET (Image-level Supervision)

Task 2

Scene Parsing on ADE20K (Point Supervision)


Thank You!