香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Weakly-Supervised Semantic Segmentation

Jun Wei

2020.08.28

Paper List

- Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation
  CVPR2020 Oral

- Weakly-Supervised Semantic Segmentation via Sub-category Exploration
  CVPR2020

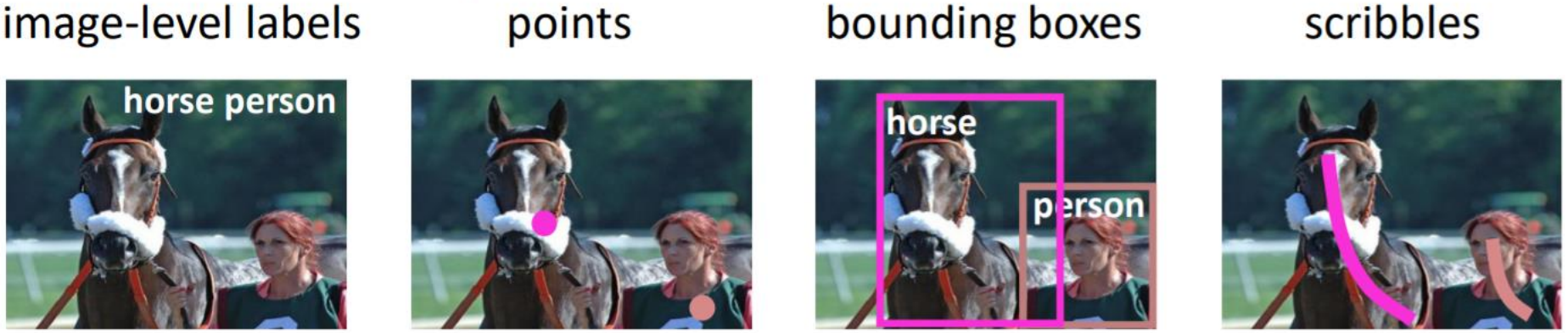
- Label Decoupling Framework for Salient Object Detection
  CVPR2020

Task Introduction

Semantic Segmentation

➢ The cost of pixel-level annotation is too heavy

- image-level
- video-level
- bounding box
- point-level
- scribble-based



Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu

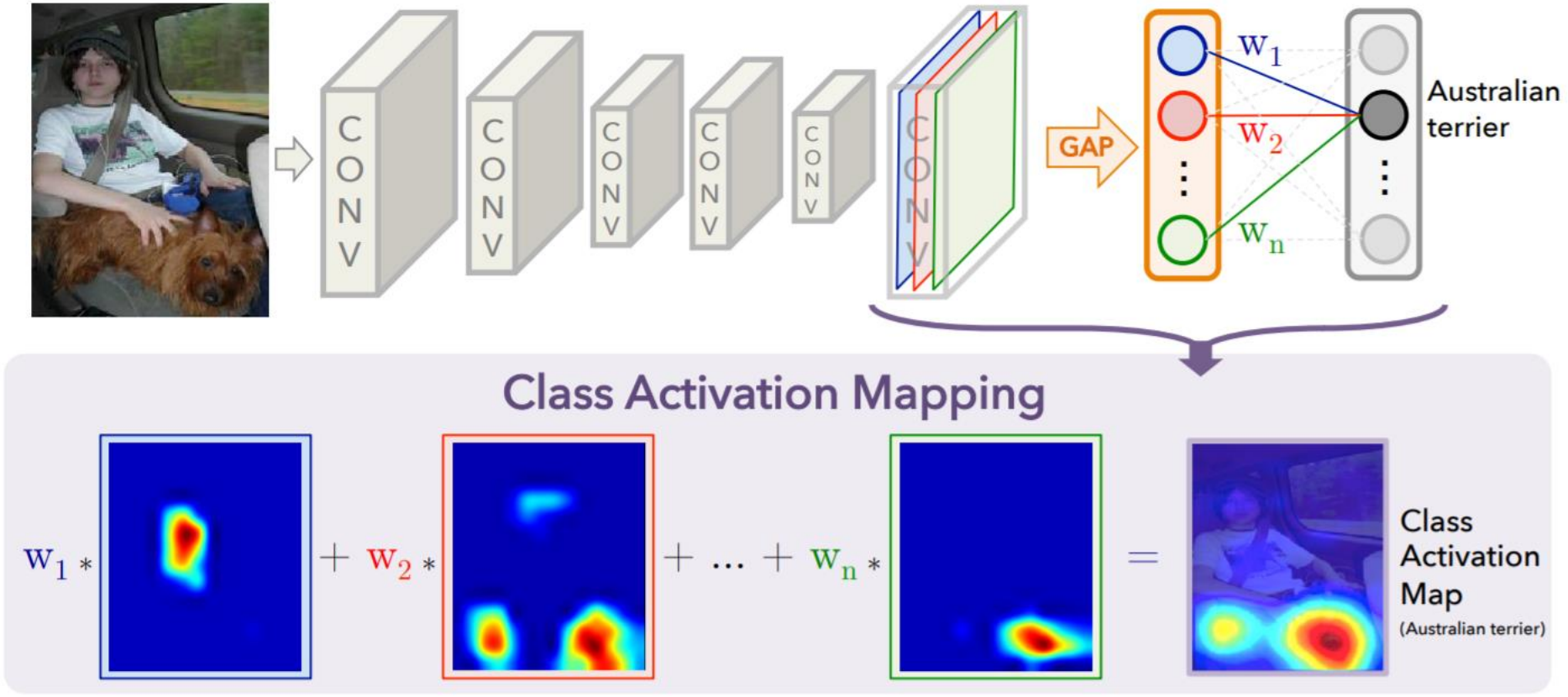
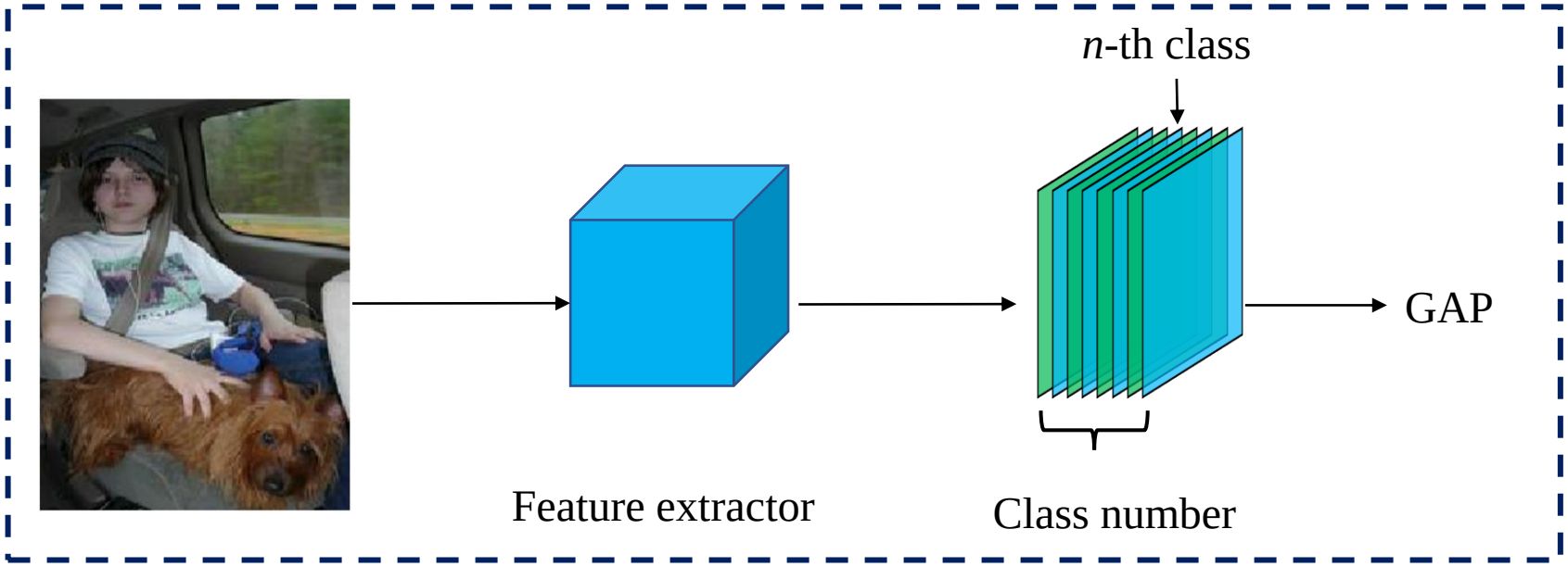


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

feature extraction + global average pooling + fully connection

CAM



Simplified CAM

- Only activate the most discriminative part

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation

Yude Wang[1,2], Jie Zhang[1,2], Meina Kan[1,2], Shiguang Shan[1,2,3], Xilin Chen[1,2]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, 200031, China
yude.wang@vipl.ict.ac.cn, {zhangjie, kanmeina, sgshan, xlchen}@ict.ac.cn

➢ Motivation

- CAMs only cover the most discriminative part of the object and incorrectly activate in background regions, which can be summarized as under-activation and over-activation respectively.
- CAMs are not consistent when images are augmented by affine transformations.
- The supervision gap between fully and weakly supervised semantic segmentation.

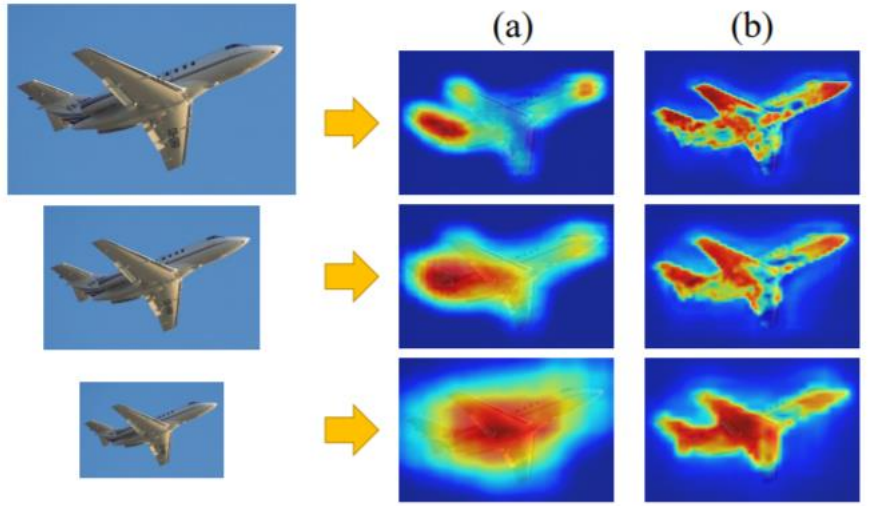


Figure 1. Comparisons of CAMs generated by input images with different scales. (a) Conventional CAMs. (b) CAMs predicted by our SEAM, which are more consistent over rescaling.

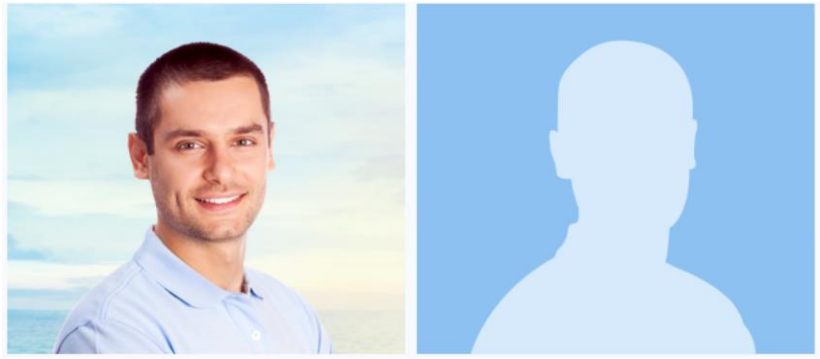


	Image	**label**
Fully-supervised	δx	δx
Weakly-supervised	δx	0

➢ Contribution

- We propose a self-supervised equivariant attention mechanism (SEAM), incorporating equivariant regularization with pixel correlation module (PCM), to narrow the supervision gap between fully and weakly supervised semantic segmentation.

- The design of siamese network architecture with equivariant cross regularization (ECR) loss efficiently couples the PCM and self-supervision, producing CAMs with both fewer over-activated and underactivated regions.

- Experiments on PASCAL VOC 2012 illustrate that our algorithm achieves state-of-the-art performance with only image-level annotations

➢ Framework

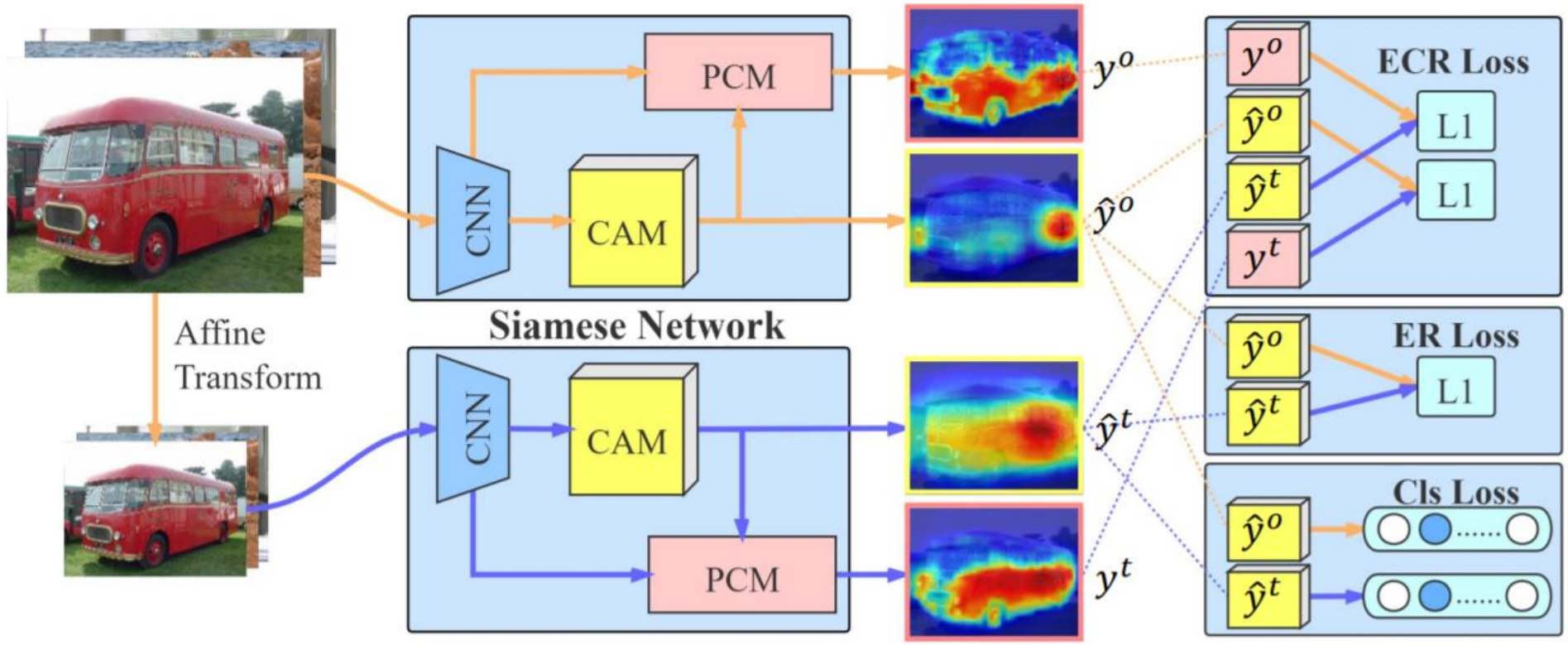
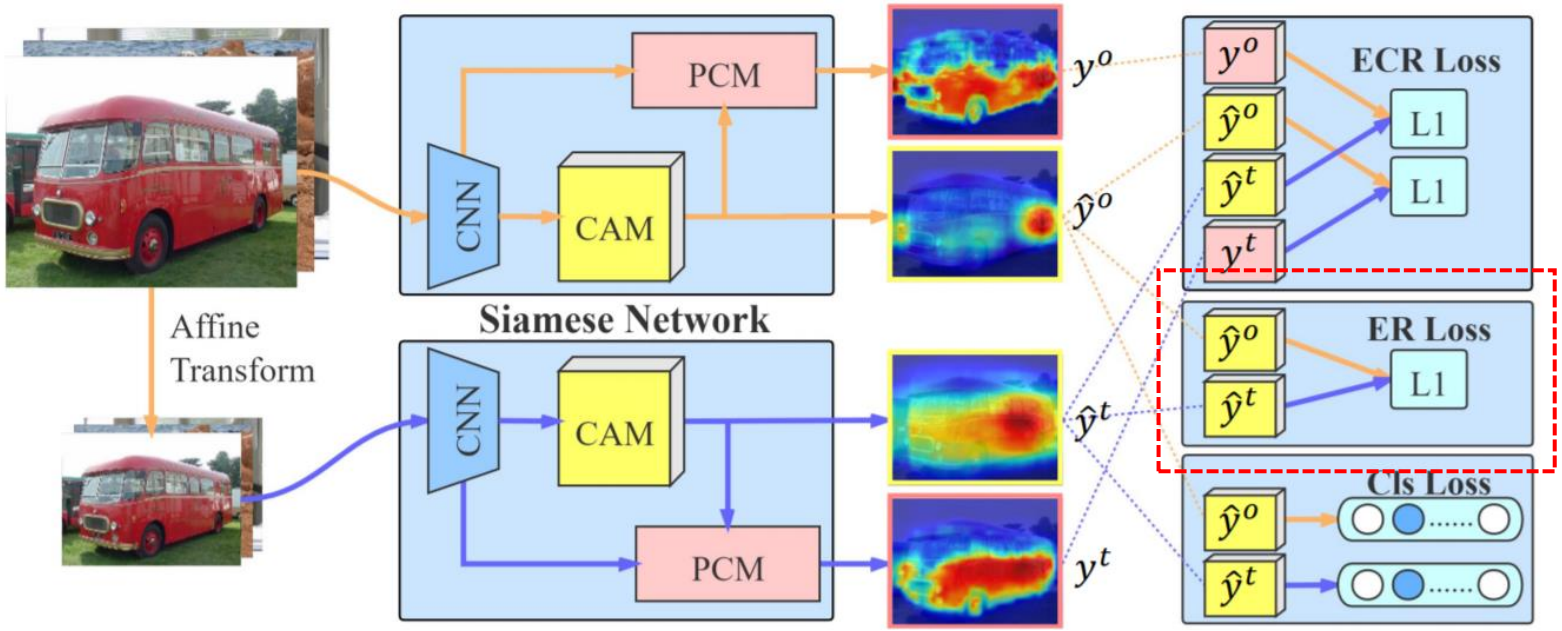


Figure 2. The siamese network architecture of our proposed SEAM method. The SEAM is the integration of equivariant regularization (ER) (Section. 3.2) and pixel correlation module (PCM) (Section. 3.3). With specially designed losses (Section 3.4), the revised CAMs not only keep consistent over affine transformation but also well fit the object contour.

➢ Equivariant Regularization (ER)

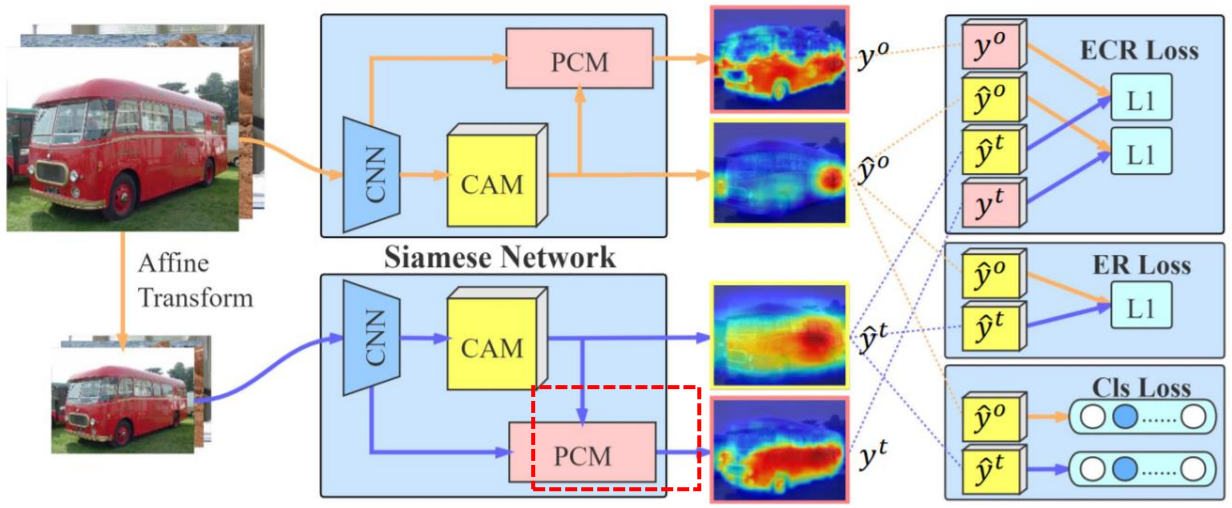


$$\mathcal{R}_{ER} = \|F(A(I)) - A(F(I))\|_1$$

I : input image
A: affine transformation
F: feedforward network

The output doesn't depend on the order of transformation and feedforwarding

➢ Pixel Correlation Module (PCM)



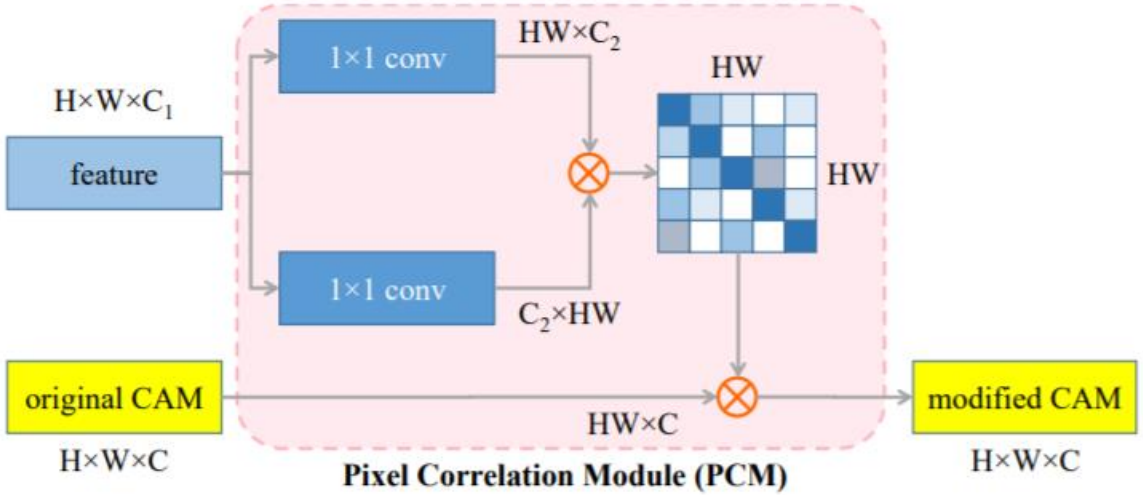
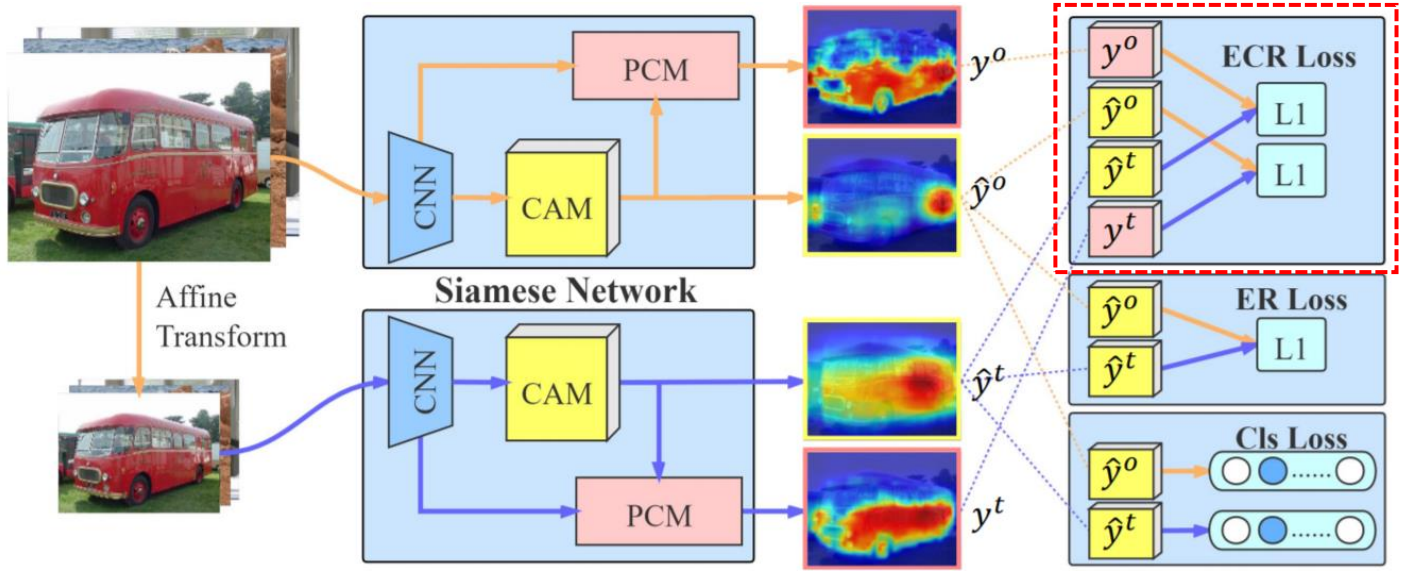


Figure 3. The structure of PCM, where $H, W, C/C_1/C_2$ denote height, width and channel numbers of feature maps respectively.

$$y_i = \frac{1}{\mathcal{C}(x_i)} \sum_{\forall j} \mathrm{ReLU}\left(\frac{\theta(x_i)^{\mathrm{T}}\theta(x_j)}{||\theta(x_i)|| \cdot ||\theta(x_j)||}\right)\hat{y}_j$$

To further refine original CAMs by context information

➢ Equivariant Cross Regularization (ECR)



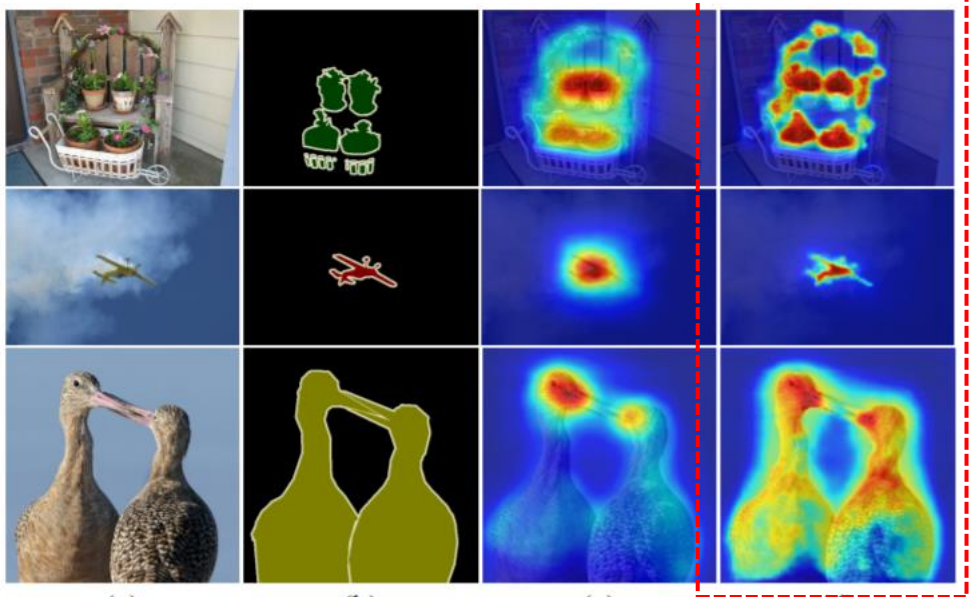
$$\mathcal{L}_{ECR} = ||A(\mathrm{y}^o) - \hat{\mathrm{y}}^t||_1 + ||A(\hat{\mathrm{y}}^o) - \mathrm{y}^t||_1$$

- The output maps of PCM fall into the local minimum quickly that all pixels in the image are predicted the same class.

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{ER} + \mathcal{L}_{ECR}$$

SEAM

➢ Experiments



baseline	ER	PCM	OHEM	CRF	mIoU
√					47.43%
√				√	52.40%
√	√				49.90%
√	√	√			55.08%
√	√	√	√		55.41%
√	√	√	√	√	56.83%

Table 1. The ablation study for each part of SEAM. **ER**: equivariant regularization. **PCM**: pixel correlation module. **OHEM**: online hard example mining. **CRF**: conditional random field.

Methods	Backbone	Saliency	*val*	*test*
CCNN [25]	VGG16		35.3	35.6
EM-Adapt [24]	VGG16		38.2	39.6
MIL+seg [27]	OverFeat		42.0	43.2
SEC [19]	VGG16		50.7	51.1
STC [33]	VGG16	√	49.8	51.2
AdvErasing [32]	VGG16	√	55.0	55.7
MDC [34]	VGG16	√	60.4	60.8
MCOF [36]	ResNet101	√	60.3	61.2
DCSP [4]	ResNet101	√	60.8	61.9
SeeNet [15]	ResNet101	√	63.1	62.8
DSRG [16]	ResNet101	√	61.4	63.2
AffinityNet [2]	ResNet38		61.7	63.7
CIAN [10]	ResNet101	√	64.1	64.7
IRNet [1]	ResNet50		63.5	64.8
FickleNet [21]	ResNet101	√	64.9	65.3
Our baseline	ResNet38		59.7	61.9
Our SEAM	ResNet38		64.5	65.7

Table 7. Performance comparisons of our method with other state-of-the-art WSSS methods on PASCAL VOC 2012 dataset.

➢ Experiments

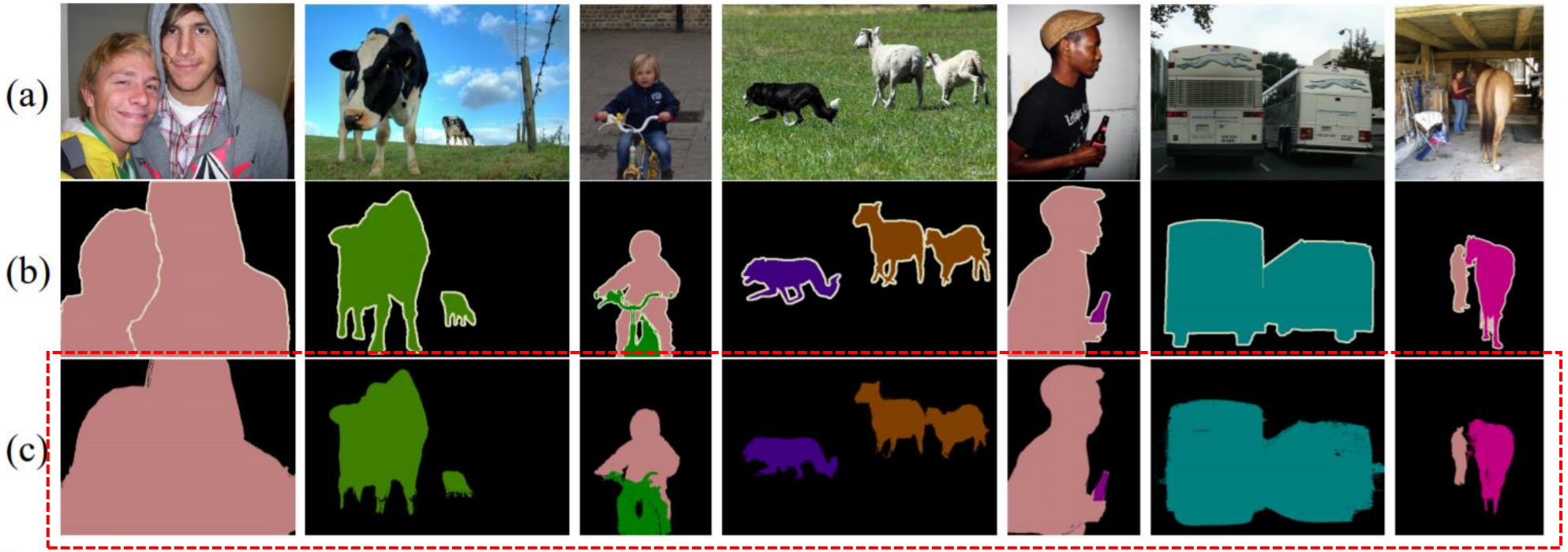
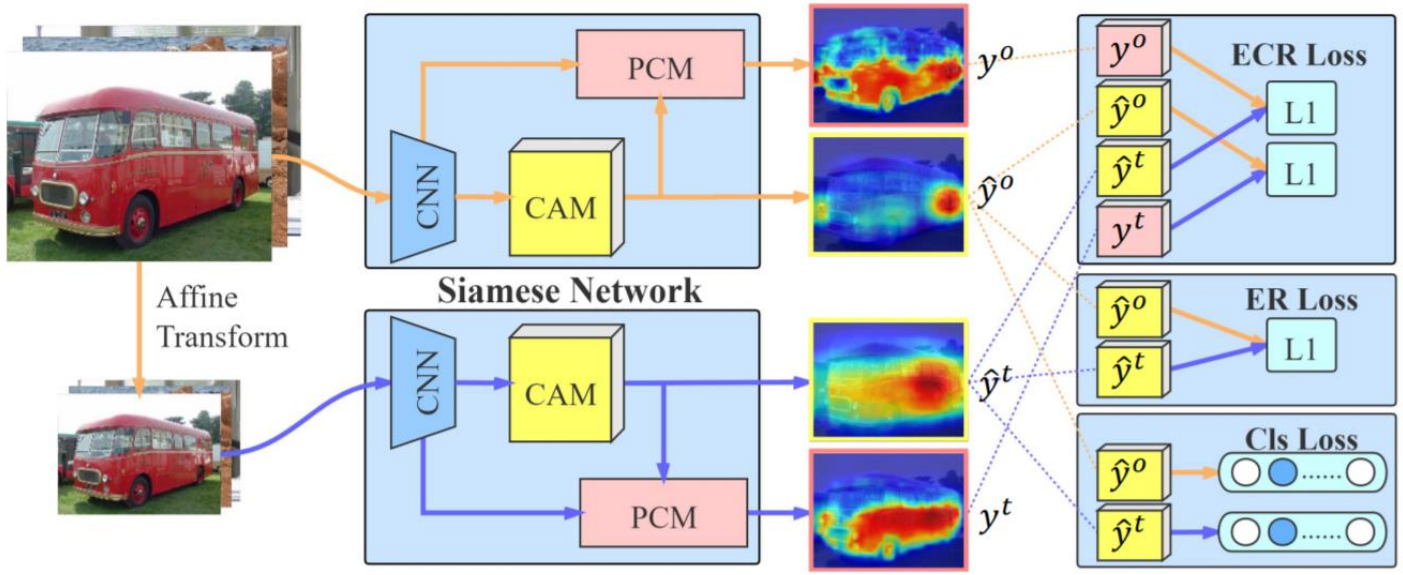


Figure 7. Qualitative segmentation results on PASCAL VOC 2012 *val* set. (a) Original images. (b) Ground truth. (c) Segmentation results predicted by DeepLab model retrained on our pseudo labels.

model	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
CCNN [25]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9	21.0	44.5	34.5	46.2	40.7	30.4	36.3	22.2	38.8	36.9	35.3
MIL+seg [27]	79.6	50.2	21.6	40.9	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
SEC [19]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
AdvErasing [32]	83.4	71.1	30.5	72.9	41.6	55.9	63.1	60.2	74.0	18.0	66.5	32.4	71.7	56.3	64.8	52.4	37.4	69.1	31.4	58.9	43.9	55.0
AffinityNet [2]	88.2	68.2	30.6	81.1	**49.6**	61.0	77.8	66.1	75.1	29.0	66.0	40.2	**80.4**	62.0	70.4	73.7	42.5	70.7	**42.6**	**68.1**	51.6	61.7
Our SEAM	**88.8**	**68.5**	**33.3**	**85.7**	40.4	**67.3**	**78.9**	**76.3**	**81.9**	**29.1**	**75.5**	**48.1**	79.9	**73.8**	**71.4**	**75.2**	**48.9**	**79.8**	40.9	58.2	**53.0**	**64.5**

Table 6. Category performance comparisons on PASCAL VOC 2012 *val* set with only image-level supervision.

➢ Summary



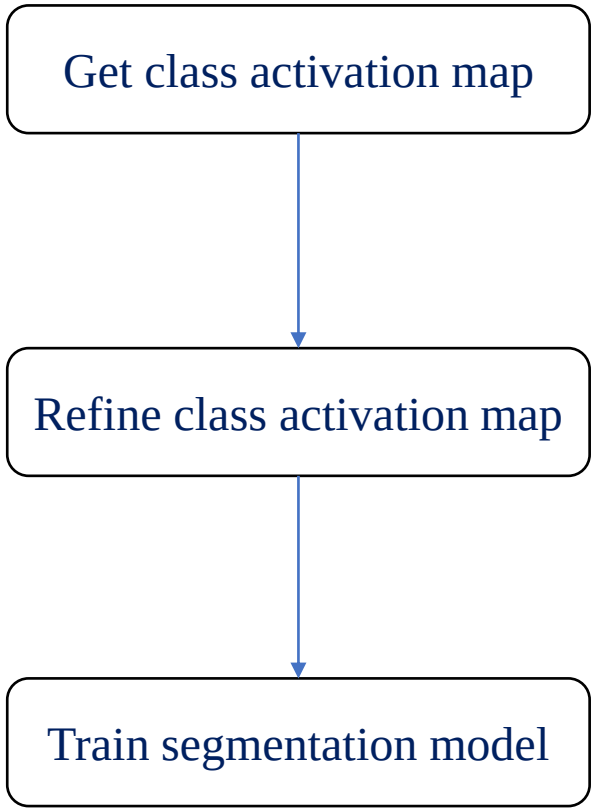
- Problem
  - Class label is not equivariant
- Solution
  - Equivariant Regularization (ER)
  - Pixel Correlation Module (PCM)

- Core Idea: Self-supervised Learning
- How to design better self-supervised mechanism?

Weakly-Supervised Semantic Segmentation via Sub-category Exploration

Yu-Ting Chang[1] Qiaosong Wang[2] Wei-Chih Hung[1] Robinson Piramuthu[2]
Yi-Hsuan Tsai[3] Ming-Hsuan Yang[1,4]
[1]UC Merced [2]eBay Inc. [3]NEC Labs America [4]Google Research

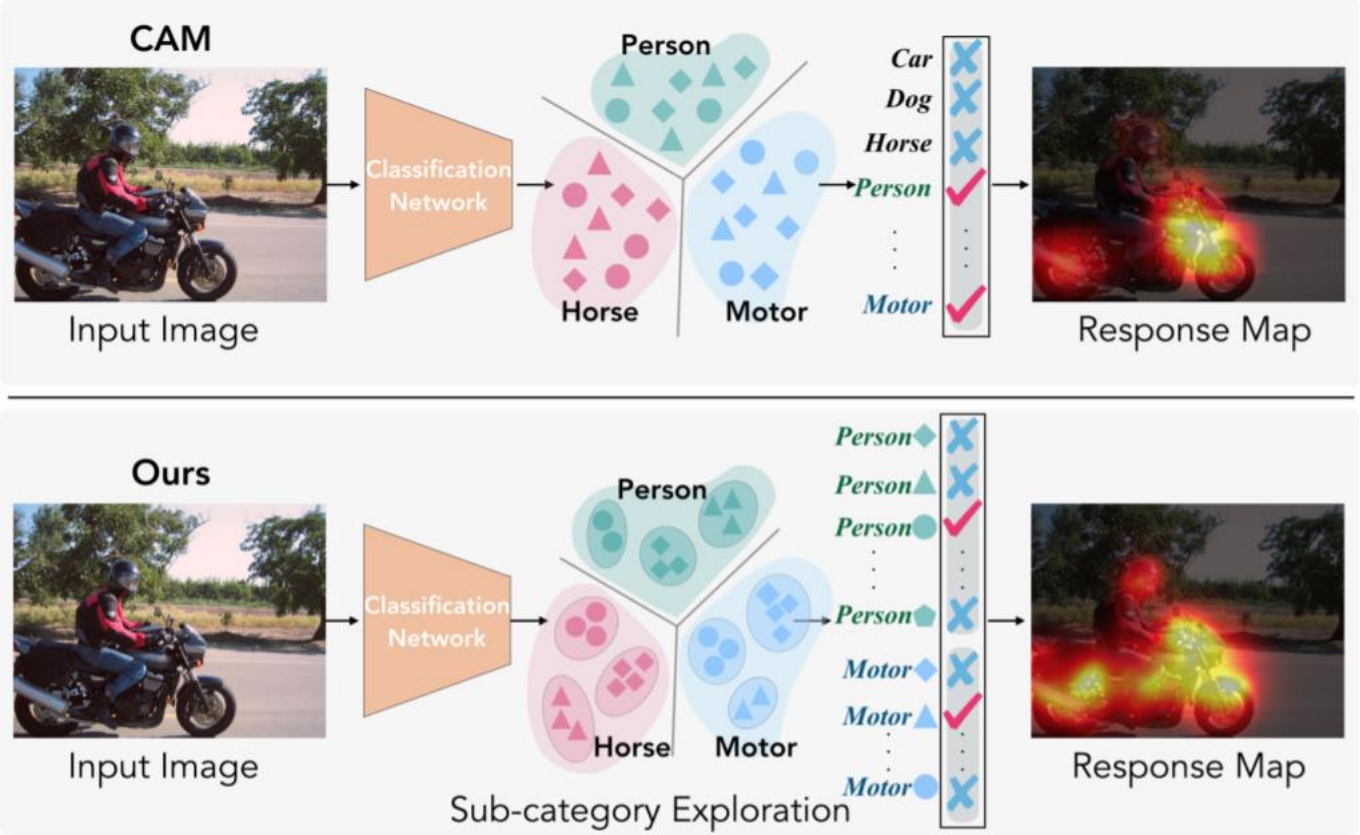
➢ Motivation



- Previous methods focus on last two steps
- A good class activation map is very helpful for sub-sequential steps

- Only discriminative parts are activated
  - Classification is a simple task, no localization and boundary
  - Introducing a more challenge task to force model to learn more

➢ Contribution



- People stand, people sit, people sleep
- Provide more information for classification

- Unsupervised clustering
- Iterative training

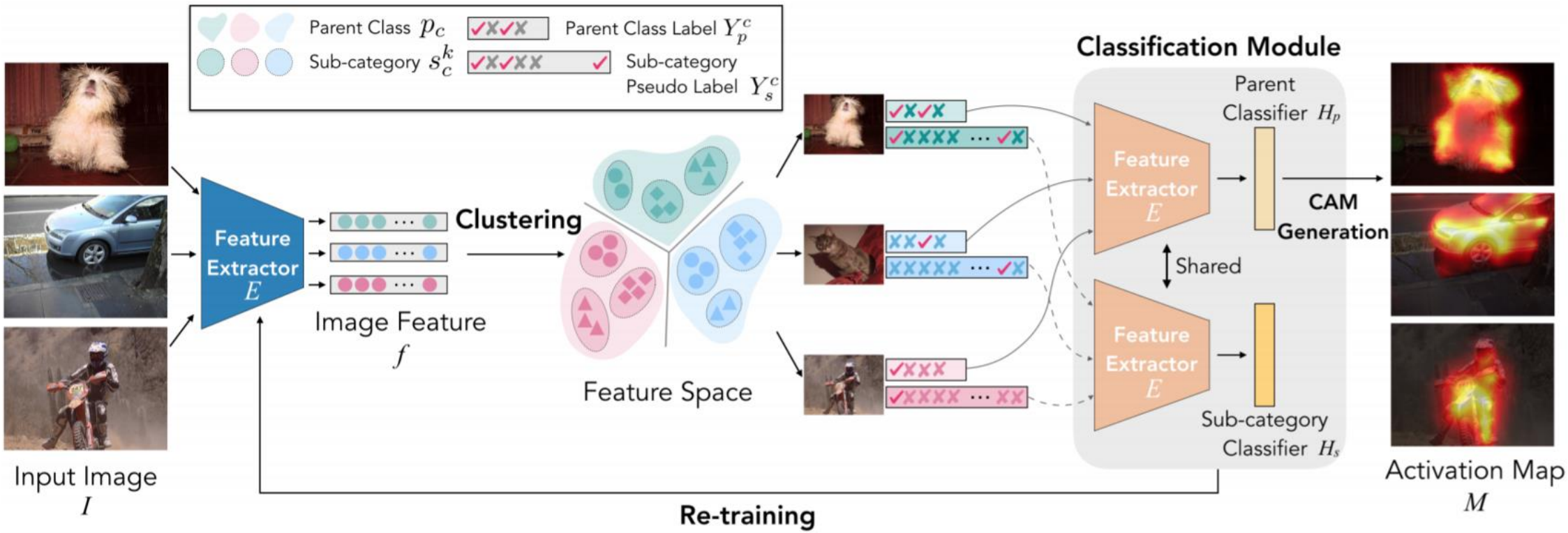


Figure 2: Proposed framework for generating the class activation map. Given input images I, we first feed them into a feature extractor E to obtain their features f. Then, we adopt unsupervised clustering on f and obtain sub-category pseudo labels Y_s for each image. Next, we train the classification network to jointly optimize the parent classifier H_p with ground truth labels Y_p for parent classes and the sub-category classifier H_s using the sub-category pseudo labels obtained in the clustering stage. By iteratively performing unsupervised clustering on image features and pseudo training the classification module, we use the jointly optimized classification network to produce the final activation map M.

➢ Sub-category Discovery

$$\min_{D\in\mathbb{R}^{d\times k}} \frac{1}{N^c}\sum_{i=1}^{N^c}\min_{Y_s^c}||f-TY_s^c||_2^2, \quad \text{s.t.}, Y_s^{c\top}1_k=1, \quad (2)$$

where T is a $D \times K$ centroid matrix, N^c is the number of images containing the class c, and $f = E(I) \in \mathbb{R}^D$ is the extracted feature. We use the clustering assignment Y_s^c for each image as the sub-category pseudo label to optimize $\mathcal{L}_s$.

- K-means
  - Calculate the shortest distance from the center point
  - Optimize the center points

- **Joint training**

Parent-class loss Sub-class loss

$$\min_{\theta_p,\theta_s} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_p(H_p(f_i), Y_p) + \lambda \mathcal{L}_s(H_s(f_i), Y_s), \quad (3)$$

where N is the total number of images and λ is weight to balance two loss functions. With this method, the parent classification learns a feature space through supervised training via $\mathcal{L}_p$, while the sub-category objective $\mathcal{L}_s$ explores the feature sub-space and provides additional gradients to enhance feature representations f, which is used to compute CAM via (1).

➢ Iterative training

Algorithm 1 Learning Sub-category Discovery for CAM

Input: Image I; Parent Label Y_p; Category Number C; Sub-category Number K
Output: Class Activation Map M^c

Model: Feature extractor E; Parent Classifier $(H_p; \theta_p)$; Sub-category Classifier $(H_s; \theta_s)$

Optimize $\{E, H_p\}$ with Y_p via $\mathcal{L}_p$
while Training **do**
  Extract features via $f = E(I)$
  for $c \leftarrow 1$ to C **do**
    Generate pseudo labels Y_s^c with f via (2)
  Optimize $\{E, H_p, H_s\}$ with $\{Y_p, Y_s\}$ via (3)
Compute M^c via (1)

- Extract features
- Cluster to get pseudo features
- Training classification model

Sub-category

➢ Ablation study

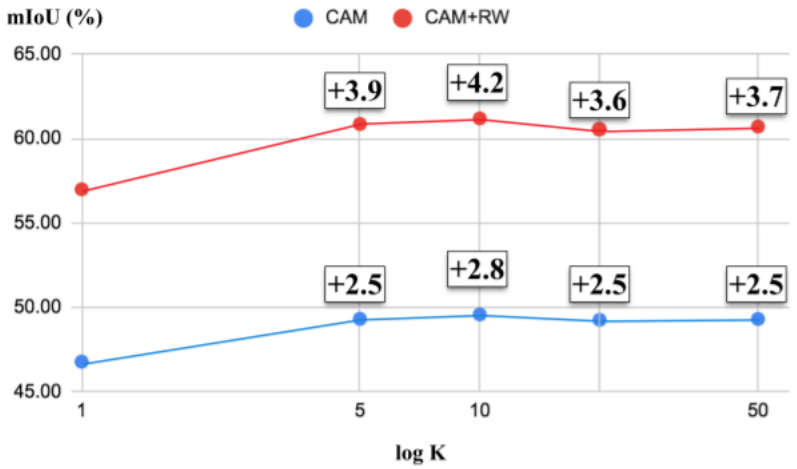


Figure 4: Ablation study for K. We show that the proposed method performs robustly with respect to K and is consistently better than the original CAM that did not apply clustering to discover sub-categories. We mark the value of mIoU of the original CAM at $K = 1$ and the improved mIoUs are presented.

- Performance is robust to K

Table 1: Performance comparison in mIoU (%) for evaluating activation maps on the PASCAL VOC training and validation sets.

	Training Set		Validation Set	
Method	CAM	CAM+RW	CAM	CAM+RW
AffinityNet [1]	48.0	58.1	46.8	57.0
Ours	50.9	63.4	49.6	61.2

Table 2: Segmentation quality of the initial response at different rounds of training on the PASCAL VOC 2012 validation set. We show there is a gradual improvement on both mIoU and F-Score metrics.

Round	mIoU (%) ↑	F-Score ↑
#0 (CAM)	46.8	65.1
#1	48.0	65.6
#2	48.7	66.6
#3	**49.6**	**67.0**

- The larger, the better

➢ Visualization

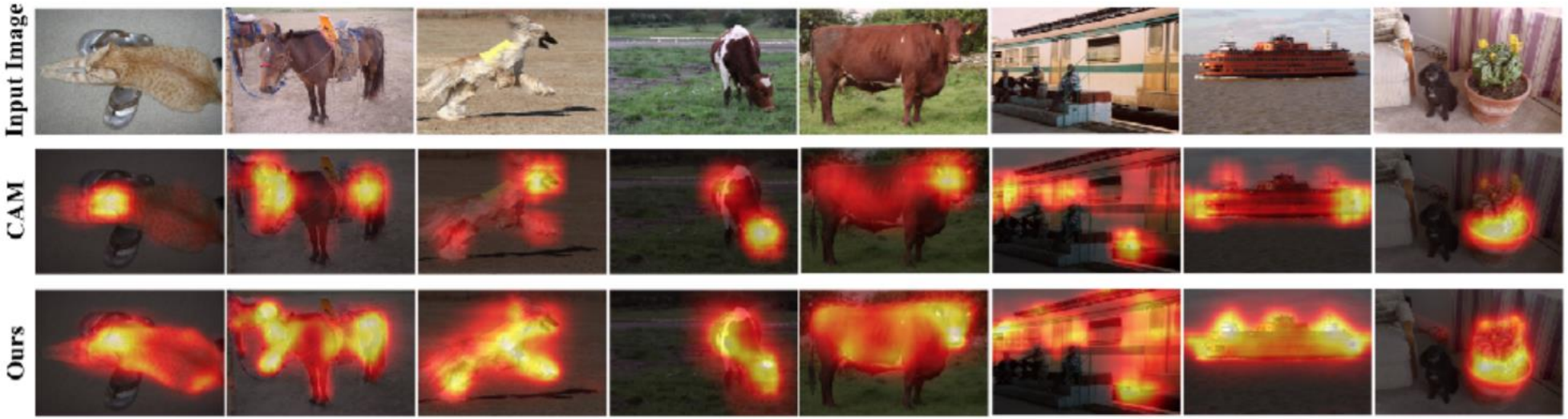


Figure 3: Sample results of initial responses. Our method often generates the response map that covers larger region of the object (i.e., attention on the body of the animal), while the response map produced by CAM [46] tends to highlight small discriminative parts.

Sub-category

➢ Visualization

Figure 5: Clustering results of the last round model (#3). We show 3 clusters for each parent class and demonstrate that our learned features are able to cluster objects based on their size (*Aeroplane, Bird, Cow*), context (*Aeroplane, Bird, Person*), type (*Boat, Bird*), pose (*Cow*), and interaction with other categories (*Person*).

- Sub-category has its own meaning

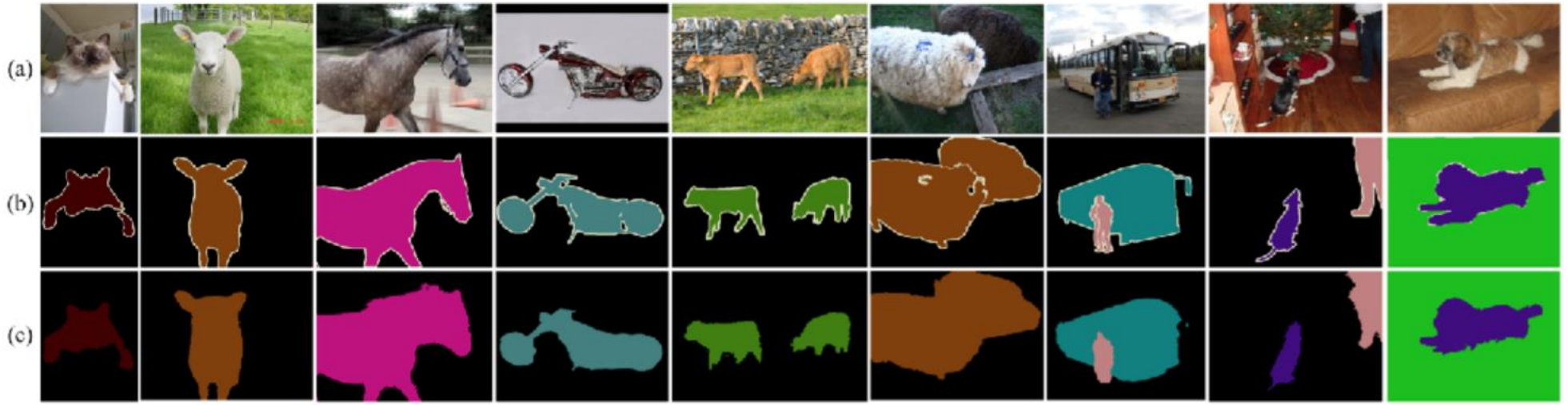


Figure 7: Qualitative results on the PASCAL VOC 2012 validation set. (a) Input images. (b) Ground truth. (c) Our results.

Table 3: Semantic segmentation performance on the PASCAL VOC 2012 validation set. Bottom group contains results with CRF refinement, while the top group is without CRF. Note that 11/20 classes obtain improvements using our approach w/ CRF. The best three results are in red, green and blue, respectively.

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
AffinityNet [1]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
Ours (w/o CRF)	88.1	49.6	30.0	79.8	51.9	74.6	87.7	73.7	85.1	31.0	77.6	53.2	80.3	76.3	69.6	69.7	40.7	75.7	42.6	66.1	58.2	64.8
MCOF [38]	87.0	78.4	29.4	68.0	44.0	67.3	80.3	74.1	82.2	21.1	70.7	28.2	73.2	71.5	67.2	53.0	47.7	74.5	32.4	71.0	45.8	60.3
Zeng et al. [43]	90.0	77.4	37.5	80.7	61.6	67.9	81.8	69.0	83.7	13.6	79.4	23.3	78.0	75.3	71.4	68.1	35.2	78.2	32.5	75.5	48.0	63.3
FickleNet [24]	89.5	76.6	32.6	74.6	51.5	71.1	83.4	74.4	83.6	24.1	73.4	47.4	78.2	74.0	68.8	73.2	47.8	79.9	37.0	57.3	64.6	64.9
Ours (w/ CRF)	88.8	51.6	30.3	82.9	53.0	75.8	88.6	74.8	86.6	32.4	79.9	53.8	82.3	78.5	70.4	71.2	40.2	78.3	42.9	66.8	58.8	66.1

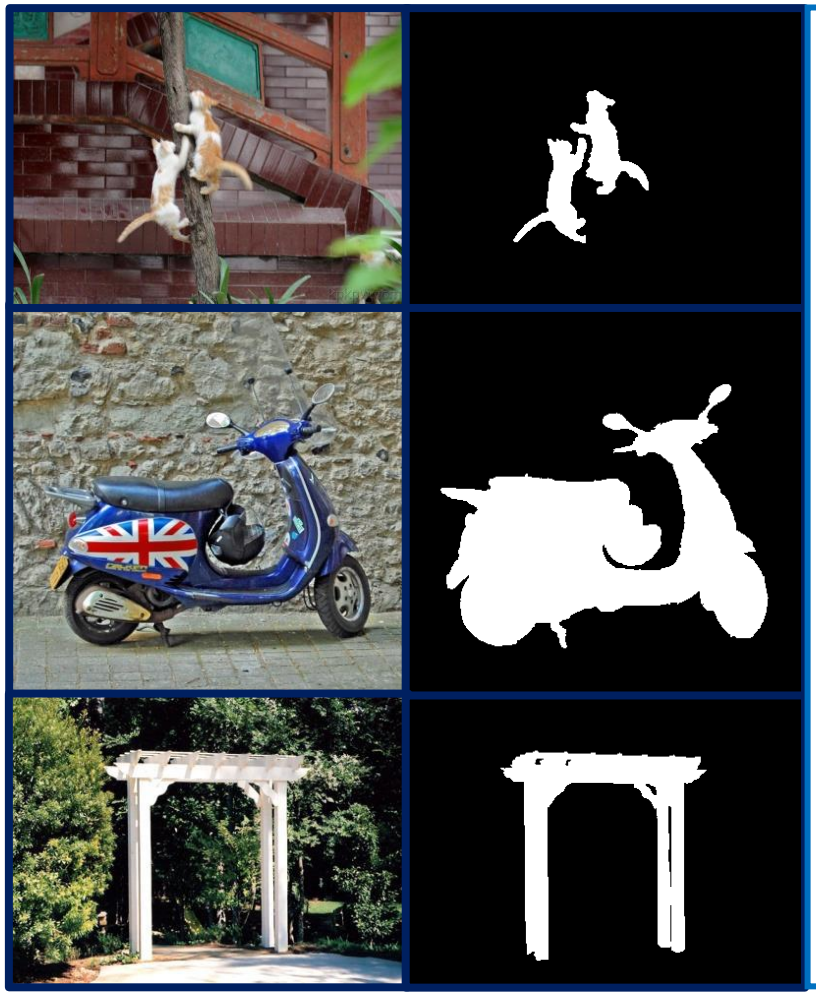
Table 4: Comparison of weakly-supervised semantic segmentation methods on the PASCAL VOC 2012 val and test sets. In addition, we present methods that aim to improve the initial response with ✓ in the "Init. Res." column.

Method	Backbone	Init. Res.	Val	Test
MCOF CVPR'18 [38]	ResNet-101		60.3	61.2
DCSP BMVC'17 [4]	ResNet-101		60.8	61.9
DSRG CVPR'18 [18]	ResNet-101		61.4	63.2
AffinityNet CVPR'18 [1]	Wide ResNet-38		61.7	63.7
SeeNet NIPS'18 [17]	ResNet-101	✓	63.1	62.8
Zeng *et al* ICCV'19 [43]	DenseNet-169		63.3	64.3
BDSSW ECCV'18 [13]	ResNet-101		63.6	64.5
OAA ICCV'19 [19]	ResNet-101	✓	63.9	65.6
CIAN CVPR'19 [12]	ResNet-101		64.1	64.7
FickleNet CVPR'19 [24]	ResNet-101	✓	64.9	65.3
Ours	ResNet101	✓	**66.1**	**65.9**

- Achieve the state-of-the-art

Label Decoupling Framework for Salient Object Detection

Jun Wei[1,2], Shuhui Wang[1]*, Zhe Wu[2,3], Chi Su[4], Qingming Huang[1,2,3], Qi Tian[5]
[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China [3]Peng Cheng Laboratory, Shenzhen, China
[4]Kingsoft Cloud, Beijing, China [5]Noah's Ark Lab, Huawei Technologies, China
jun.wei@vipl.ict.ac.cn, wangshuhui@ict.ac.cn, zhe.wu@vipl.ict.ac.cn
suchi@kingsoft.com, qmhuang@ucas.ac.cn, tian.qi1@huawei.com



➢ Salient Object Detection

Saliency

Integrity

Class-agnostic

➢ Similarity & Difference

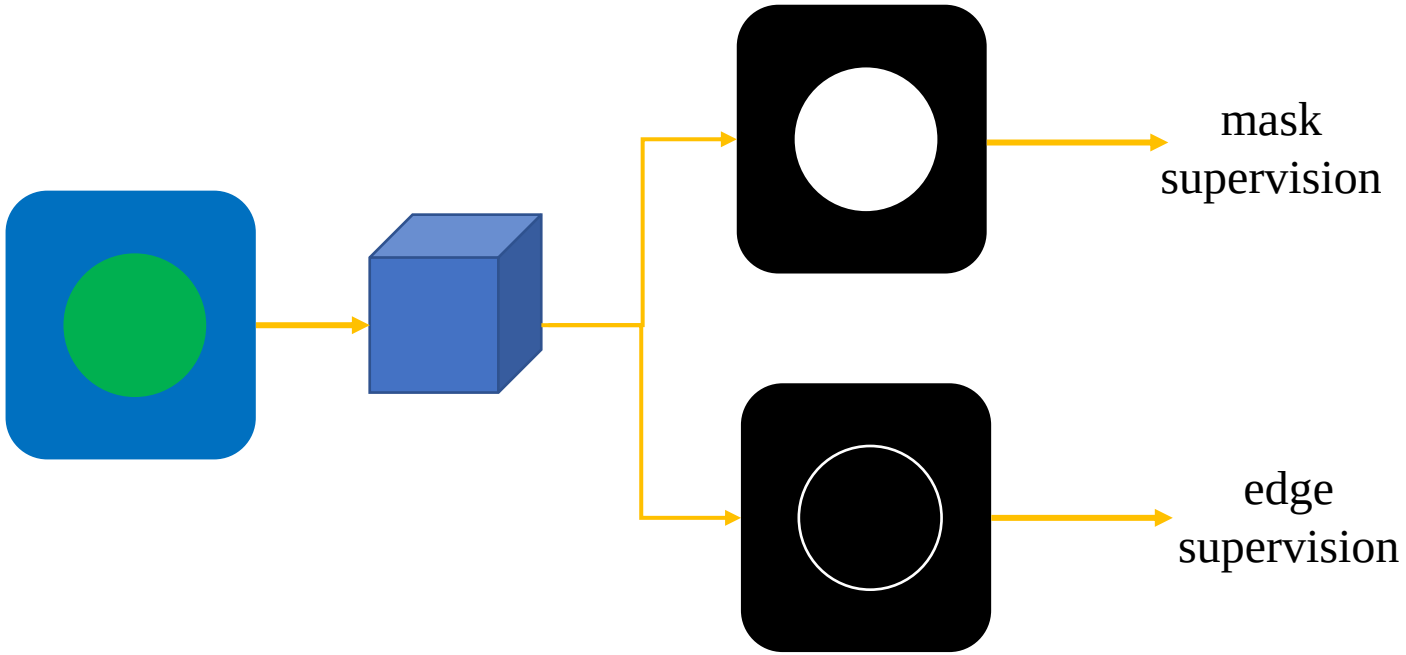
- Form : binary segmentation ***VS*** multiple segmentation
- Semantic : low-level ***VS*** high-level
- Application: less restrictive ***VS*** more restrictive

➢ Application: extract the subject target in the image or video

- image classification
- semantic segmentation
- background blur
- video & image compression
- advertising recommendation

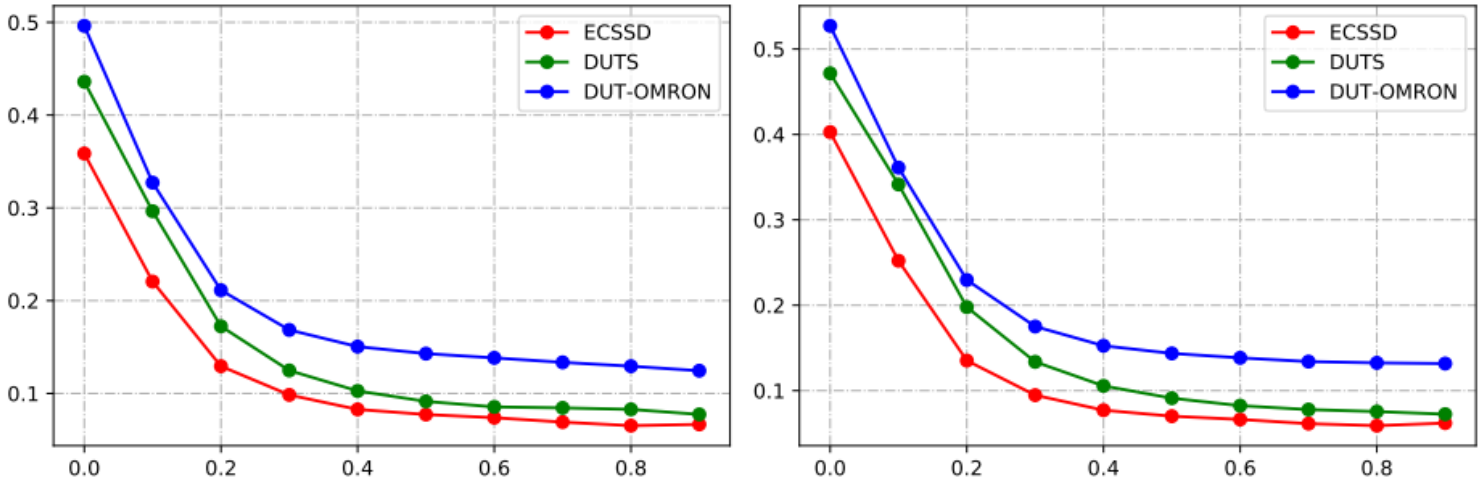
➢ Research Direction

- feature fusion
- context & attention
- real-time & acceleration
- weakly-supervised learning
- boundary-guided model



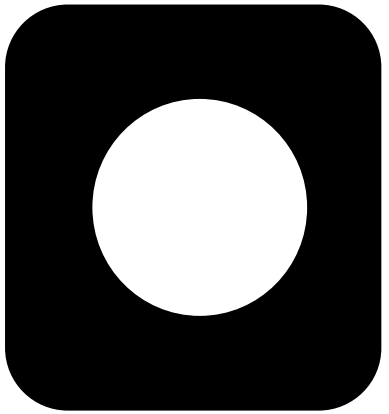
➢ Motivation

Relationship between prediction error and distance from edge



(a) EGNet [41]

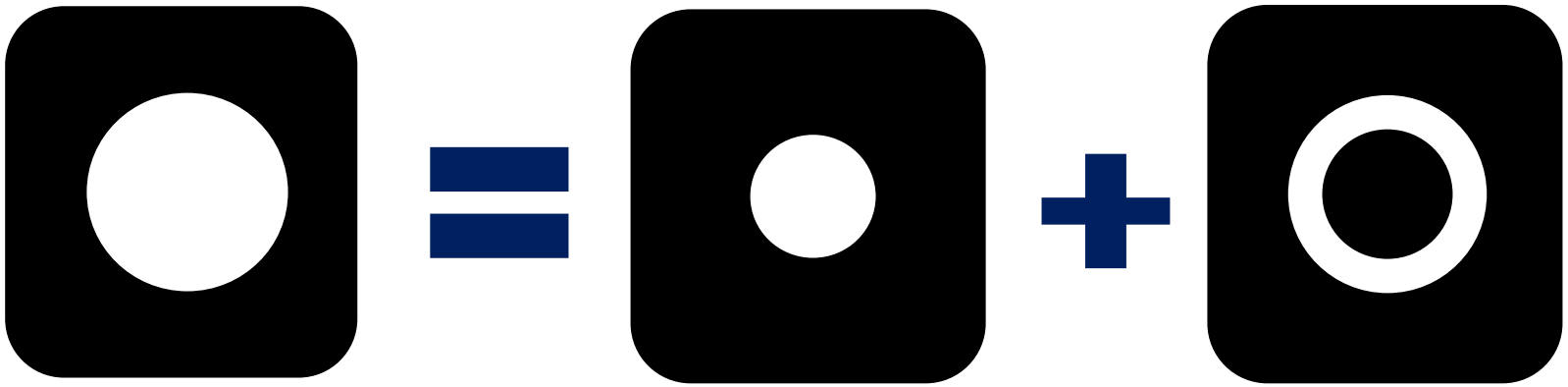
(b) SCRN [34]



label

- Edge pixels have the largest prediction error
- From center to edge, prediction error increases exponentially
- Previous methods ignore the pixels closed to edge
- Pixels of different difficulty are mixed together

- Solution: break the label into two parts



label = center area + edge area

- Need to determine where to start
- There are mutations near the interface

➢ Solution: break the label into two parts

label body detail

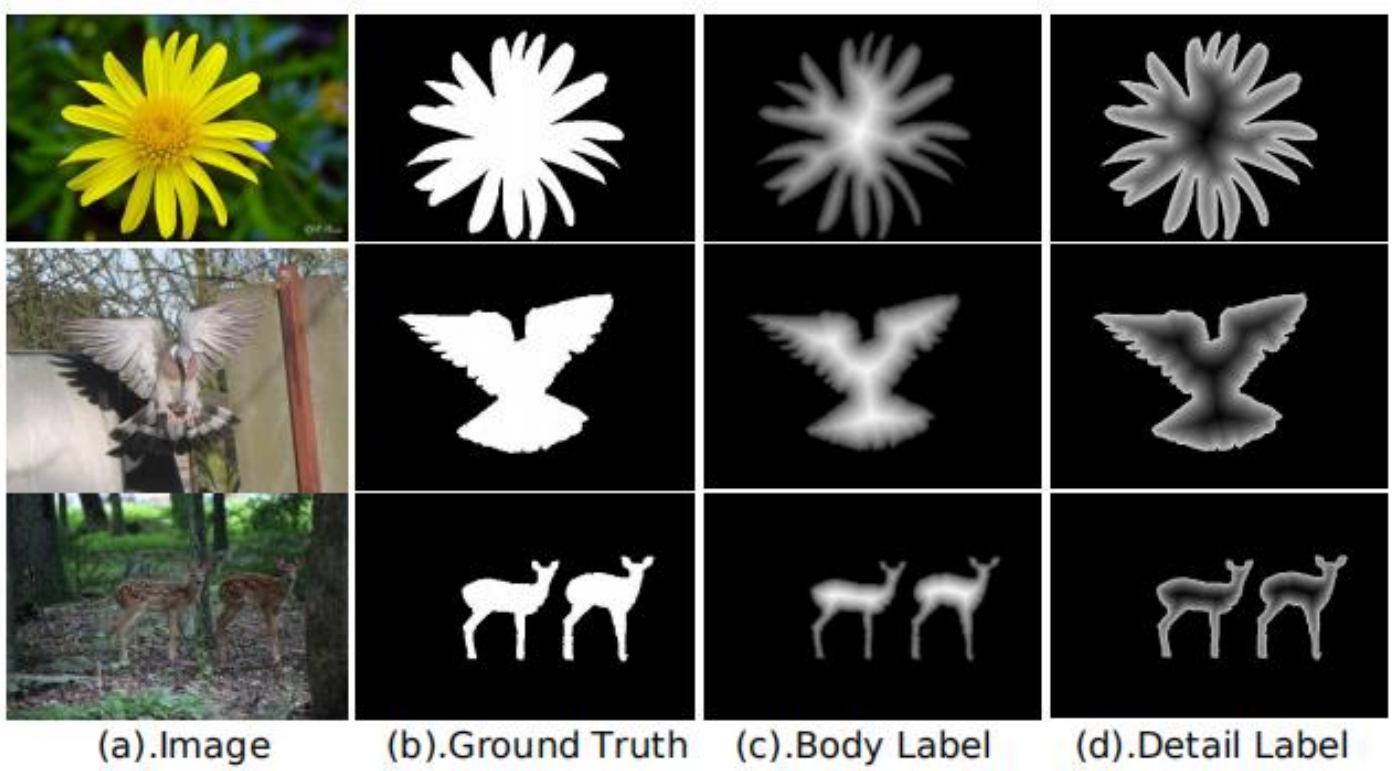
$$I^{'}(p) = \begin{cases} \min\limits_{q \in I_{bg}} f(p,q), & p \in I_{fg} \\ 0, & p \in I_{bg} \end{cases}$$

P belongs to foreground
Q belongs to background
f() represents distance function

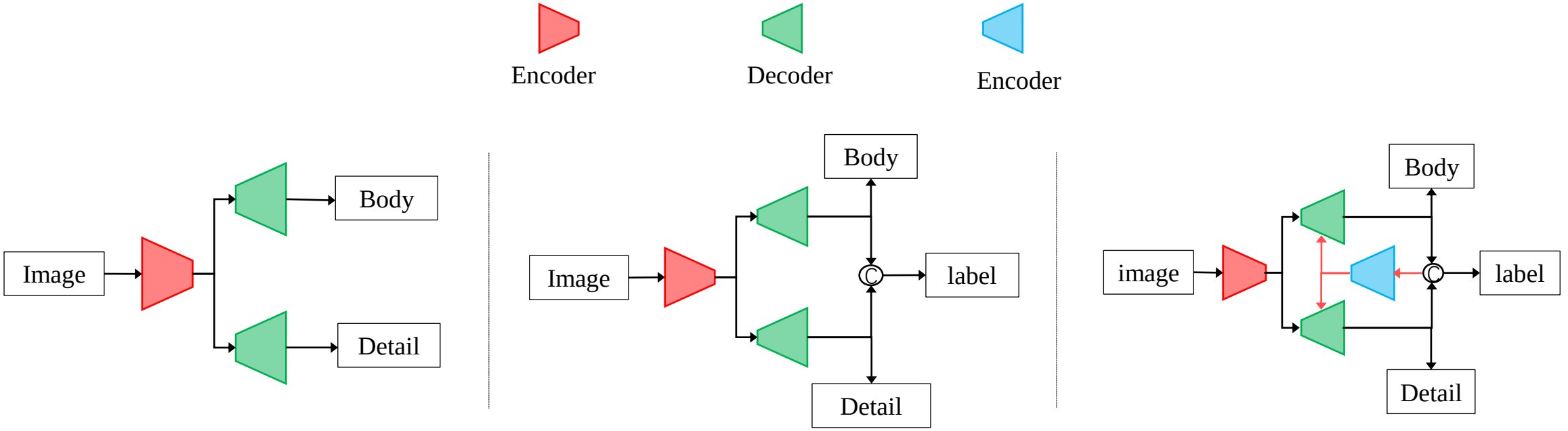
For any foreground pixel, find the shortest distance from the background

- The central region and the edge region are decoupled
- There is no need to determine where to start
- There are no mutations

- Some examples



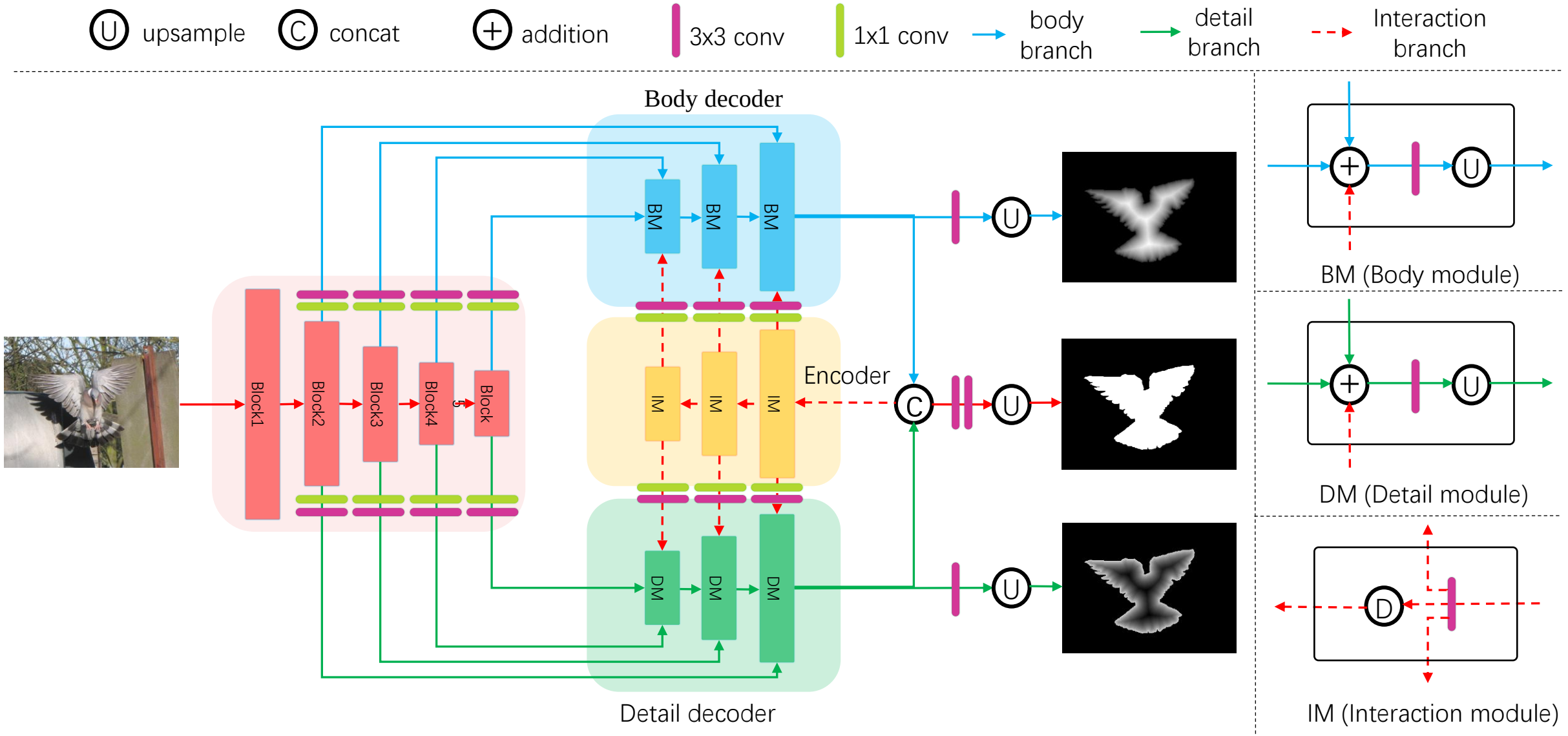
➢ How to utilize the decoupled maps?



Two-branch network

Two-branch network
Feature concatenation

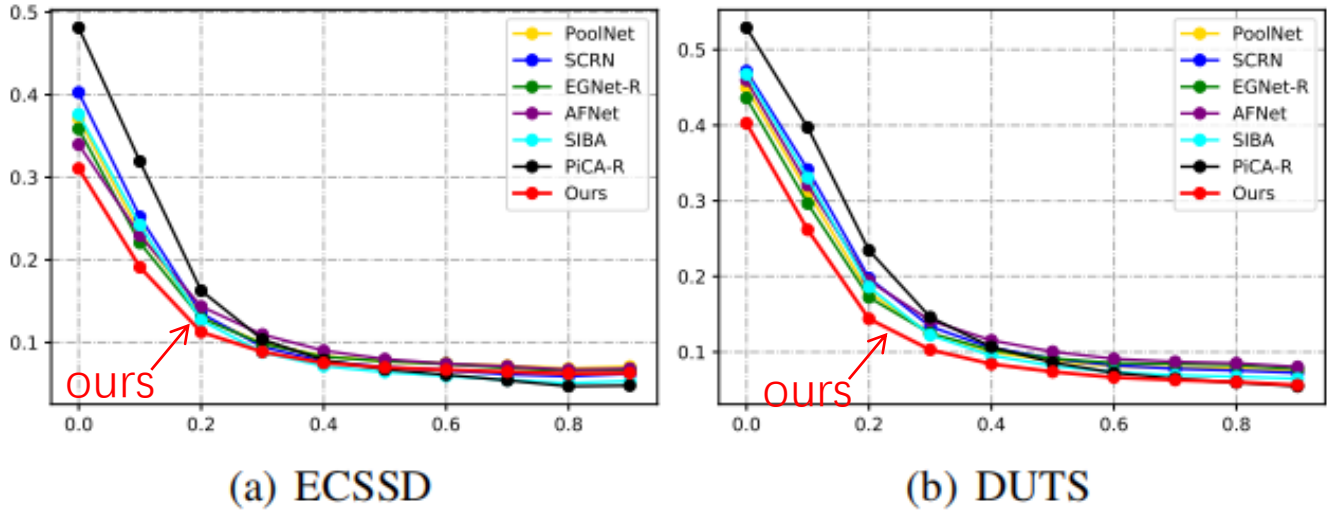
Two-branch network
Feature concatenation
Feature interaction



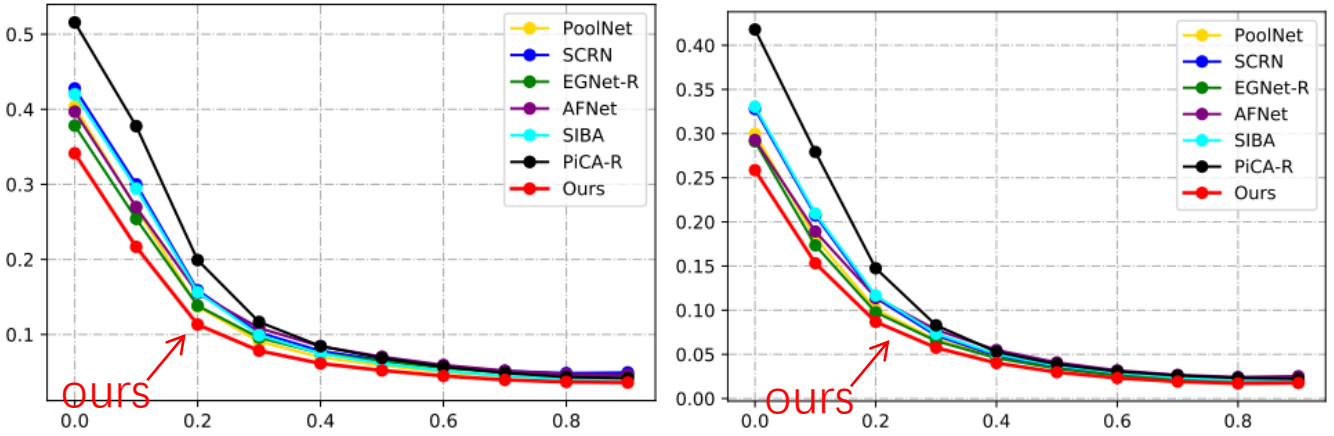
$$\ell^{(k)} = \ell^{(k)}_{body} + \ell^{(k)}_{detail} + \ell^{(k)}_{segm}$$

$$\ell_{bce} = -\sum_{(x,y)} [g(x,y)log(p(x,y)) + (1-g(x,y))log(1-p(x,y))]$$

Relationship between prediction error and distance



(a) ECSSD (b) DUTS



(c) HKU-IS (d) THUR15K

Table 5. Comparison on different combinations of supervision. Body, detail, saliency and edge maps are used, respectively.

Label	THUR15K			DUTS-TE		
	MAE	mF	E_ξ	MAE	mF	E_ξ
Body + Detail	0.064	0.764	0.842	0.034	0.855	0.910
Body + Edge	0.066	0.758	0.836	0.036	0.850	0.904
Sal + Detail	0.066	0.756	0.835	0.037	0.848	0.901
Sal + Edge	0.070	0.752	0.827	0.039	0.844	0.895

- Improvement of edge pixels is larger
- Mean prediction error is smaller
- Combination of body+detail gets better results

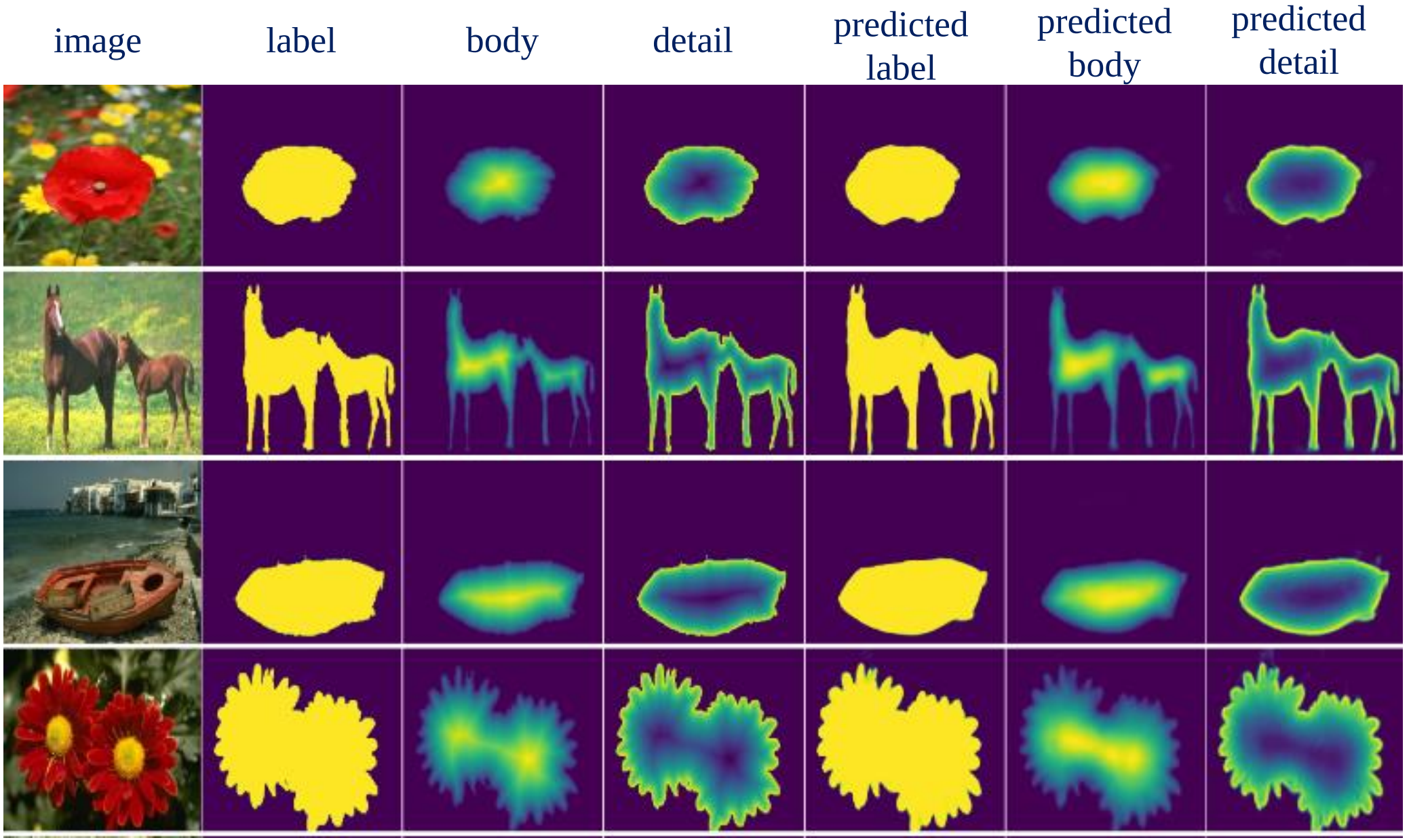


image
label
body
detail
predicted label
predicted body
predicted detail

Table 2. Performance comparison with state-of-the-art methods on six datasets. MAE (smaller is better), mean F-measure (mF, larger is better) and E-measure (E_ξ, larger is better) are used to measure the model performance. '-' means the author has not provided corresponding saliency maps. The best and the second best results are highlighted in red and blue respectively.

Algorithm	ECSSD 1,000 images			PASCAL-S 850 images			DUTS-TE 5,019 images			HKU-IS 4,447 images			DUT-OMRON 5,168 images			THUR15K 6,232 images		
	MAE	mF	E_ξ	MAE	mF	E_ξ	MAE	mF	E_ξ	MAE	mF	E_ξ	MAE	mF	E_ξ	MAE	mF	E_ξ
BMPM [38]	.044	.894	.914	.073	.803	.838	.049	.762	.859	.039	.875	.937	.063	.698	.839	.079	.704	.803
DGRL [28]	.043	.903	.917	.074	.807	.836	.051	.764	.863	.037	.881	.941	.063	.709	.843	.077	.716	.811
R^{3}Net [7]	.051	.883	.914	.101	.775	.824	.067	.716	.827	.047	.853	.921	.073	.690	.814	.078	.693	.803
RAS [4]	.055	.890	.916	.102	.782	.832	.060	.750	.861	.045	.874	.931	.063	.711	.843	.075	.707	.821
PiCA-R [21]	.046	.867	.913	.075	.776	.833	.051	.754	.862	.043	.840	.936	.065	.695	.841	.081	.690	.803
AFNet [13]	.042	.908	.918	.070	.821	.846	.046	.792	.879	.036	.888	.942	.057	.738	.853	.072	.730	.820
BASNet [23]	.037	.880	.921	.076	.775	.847	.048	.791	.884	.032	.895	.946	.056	.756	.869	.073	.733	.821
CPD-R [33]	.037	.917	.925	.072	.824	.849	.043	.805	.886	.034	.891	.944	.056	.747	.866	.068	.738	.829
EGNet-R [41]	.037	.920	.927	.074	.823	.849	.039	.815	.891	.032	.898	.948	.053	.755	.867	.067	.741	.829
PAGE [31]	.042	.906	.920	.077	.810	.841	.052	.777	.869	.037	.882	.940	.062	.736	.853	-	-	-
TDBU [30]	.041	.880	.922	.071	.779	.852	.048	.767	.879	.038	.878	.942	.061	.739	.854	-	-	-
SCRN [34]	.037	.918	.926	.064	.832	.857	.040	.808	.888	.034	.896	.949	.056	.746	.863	.066	.741	.833
SIBA [24]	.035	.923	.928	.070	.830	.855	.040	.815	.892	.032	.900	.950	.059	.746	.860	.068	.741	.832
~~PoolNet [20]~~	~~.039~~	~~.915~~	~~.924~~	~~.074~~	~~.822~~	~~.850~~	~~.040~~	~~.809~~	~~.889~~	~~.032~~	~~.899~~	~~.949~~	~~.056~~	~~.747~~	~~.863~~	~~.070~~	~~.732~~	~~.822~~
LDF(ours)	.034	.930	.925	.060	.848	.865	.034	.855	.910	.027	.914	.954	.051	.773	.873	.064	.764	.842

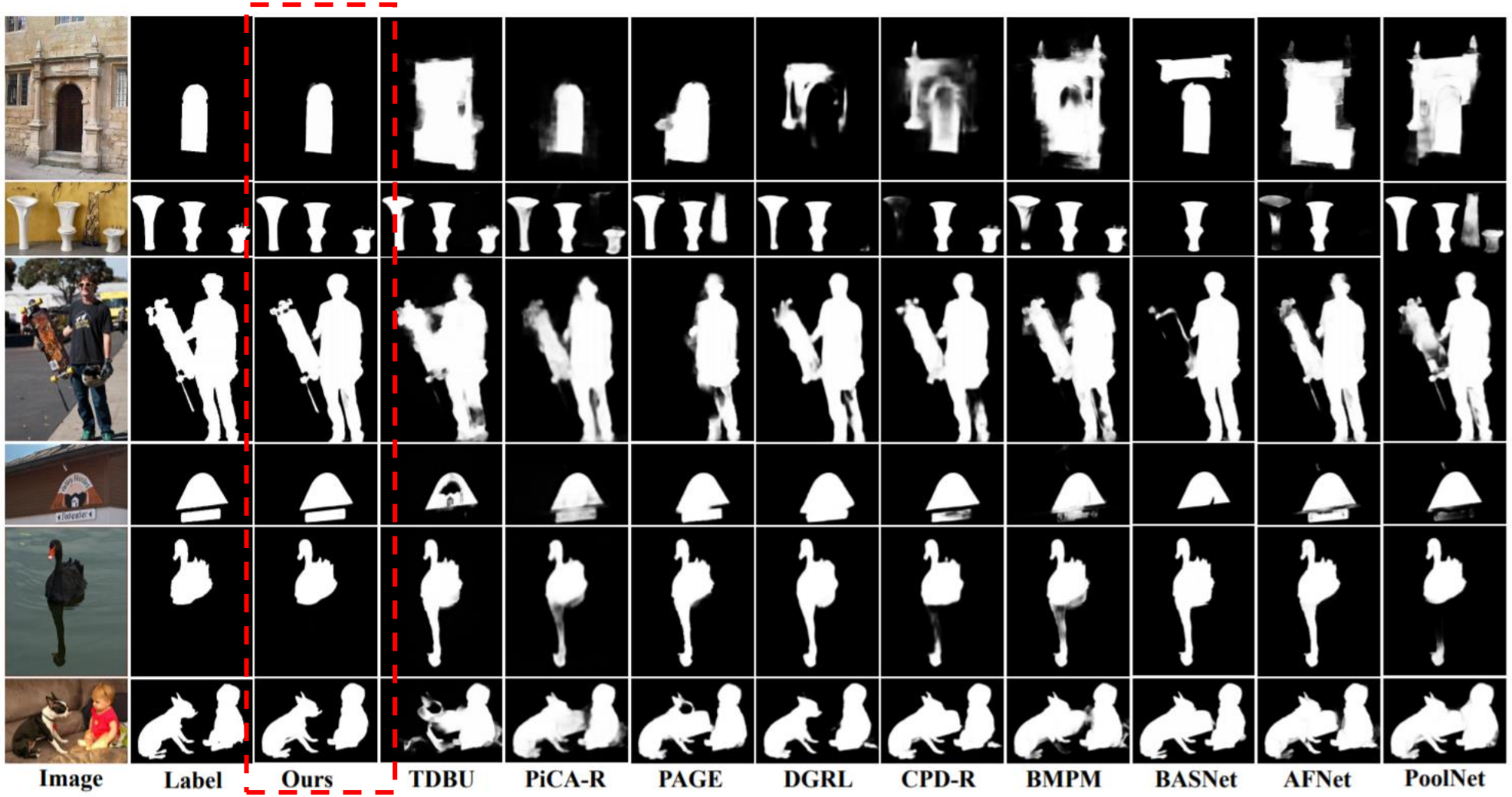


Figure 5. Visual comparison of different algorithms. Each row represents one image and corresponding saliency maps. Each column represents the predictions of one method. Apparently, our method is good at dealing with cluttered background and producing more accurate and clear saliency maps.

➢ Summary

- Analyze the relationship between prediction error and distance from edge
- Propose to decouple label into body and detail to supervise model training, respectively
- Propose feature interaction to promote mutual improvement among branches

Thank you!