RecurSeed and EdgePredictMix: Single-stage learning is sufficient for Weakly-Supervised Semantic Segmentation

Sanghyun Jo[*1], **In-Jae Yu**[*2], **Kyung-Su Kim**[†34]

[1] OGQ GYN, Seoul, Korea
[2] Samsung Electronics, Suwon, Korea
[3] Medical AI Research Center, Samsung Medical Center, Seoul, Korea
[4] Sungkyunkwan University School of Medicine, Seoul, Korea

Background [1/8]

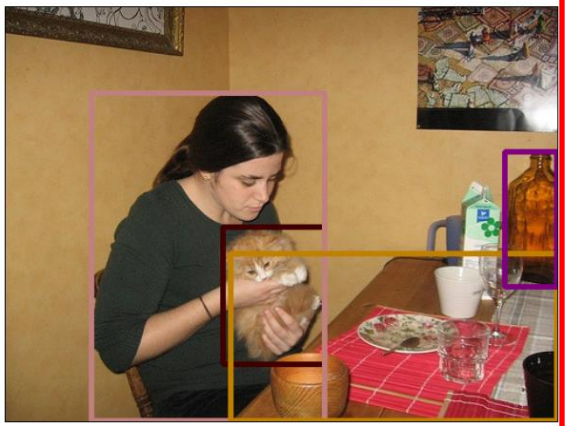
What is the labeling for AI?

✓ In deep learning, data labeling is the process of identifying raw data (images, videos, etc.) and adding more informative labels.

Image Classification

Labeling time: 20s

Object Detection



Semantic Segmentation

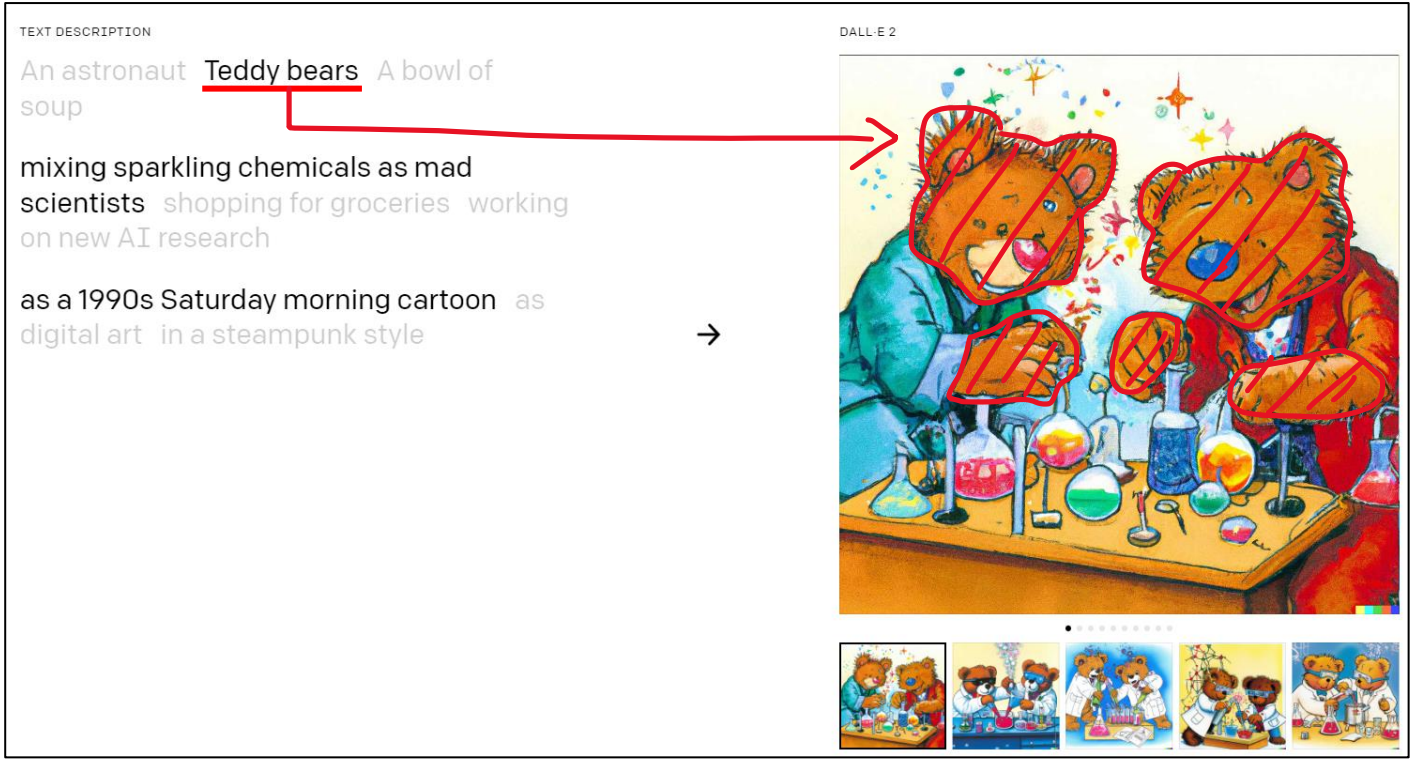
Labeling time: 239.7s

[1] Amy et al., What's the Point: Semantic Segmentation with Point Supervision (ECCV 2016)

Background [2/8]

Examples using semantic segmentation

- ✓ Image generation
- ✓ Self-driving cars
- ✓ Construction



https://openai.com/dall-e-2/

Background [3/8]

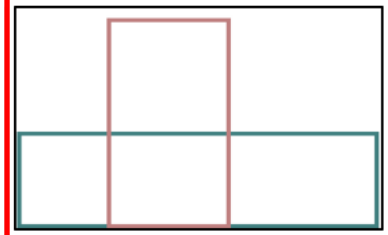
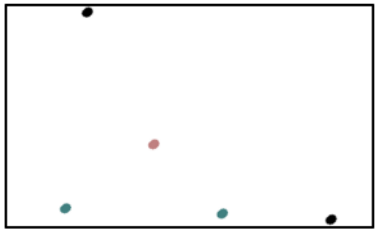
What is the specific tasks related to semantic segmentation?

- ✓ Fully-supervised semantic segmentation (Supervised learning)
- ✓ Unsupervised semantic segmentation (Self-supervised learning)
- ✓ Semi-supervised semantic segmentation
- ✓ Weakly-supervised semantic segmentation

Background [4/8]

Labeling for weakly-supervised semantic segmentation

✓ Tags, bounding boxes, points, and scribbles.

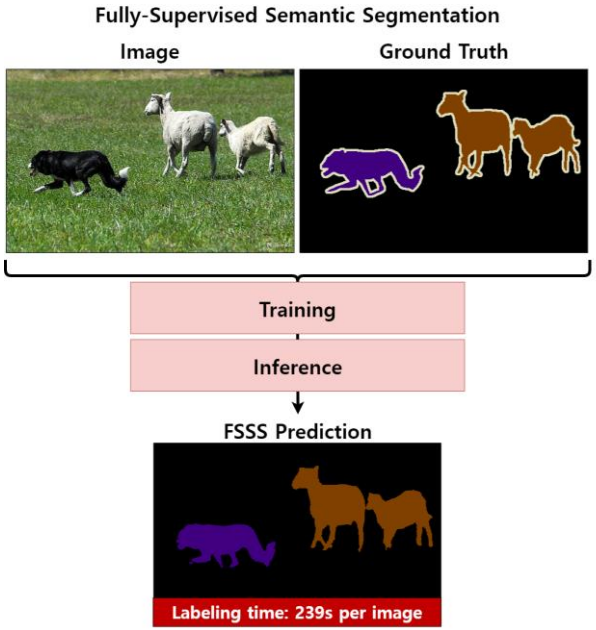
image	image tags	bounding boxes	labeled points	scribbles
	Person Motorbike			

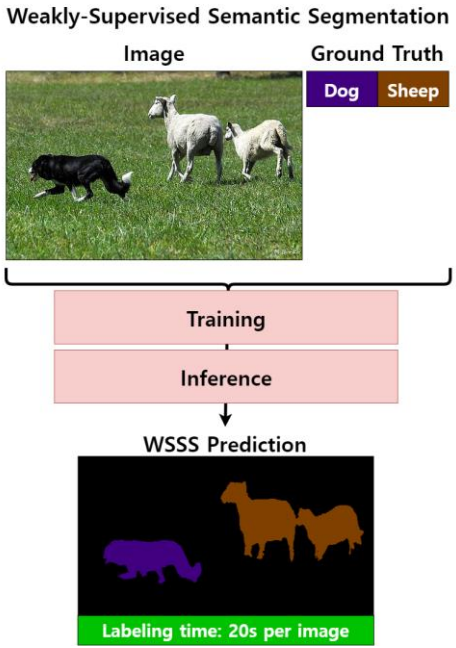
[1] Ke et al., Universal weakly supervised segmentation by pixel-to-segment contrastive learning (ICLR 2021)

Background [5/8]

What is the difference between FSSS and WSSS?

- ✓ Fully-supervised semantic segmentation requires pixel-wise annotations.
- ✓ Weakly-supervised semantic segmentation requires image-wise annotations (i.e., tags).

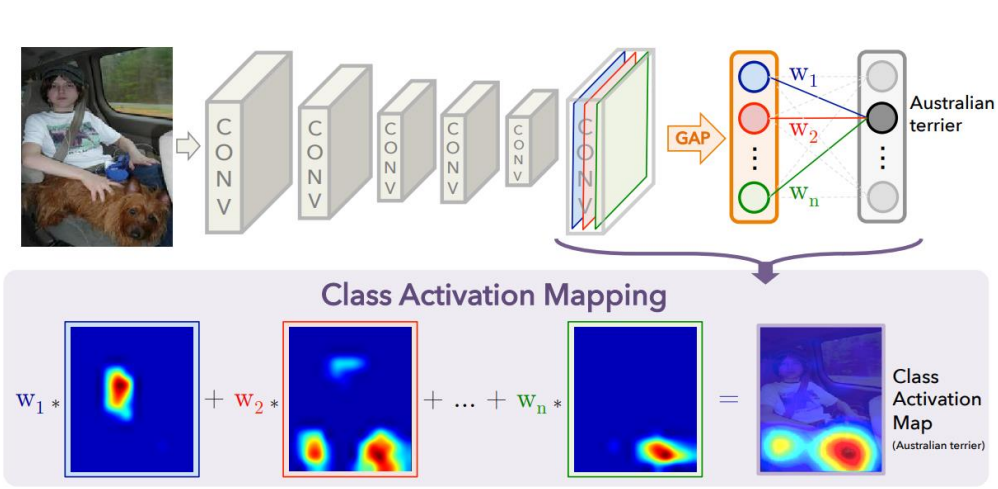


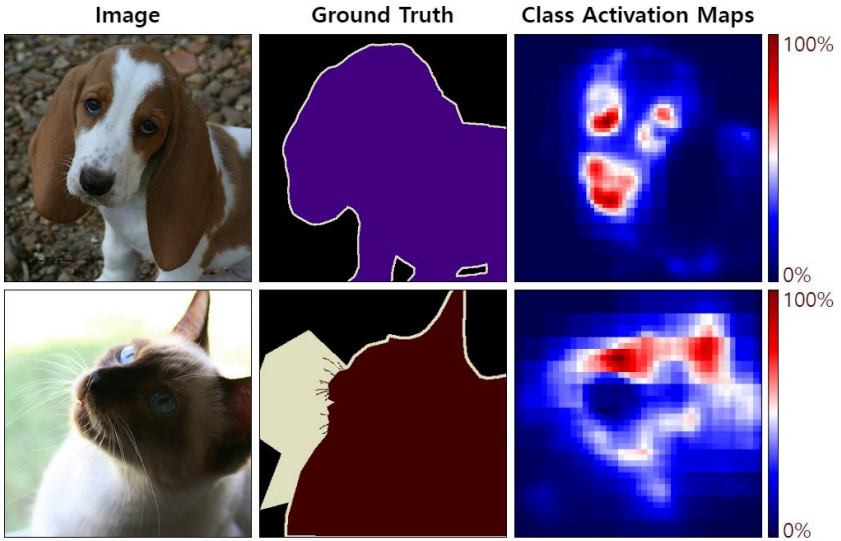


Background [6/8]

The limitations of a class activation maps[1] (CAM)

✓ CAM is designed to visualize the most discriminative part of an object.



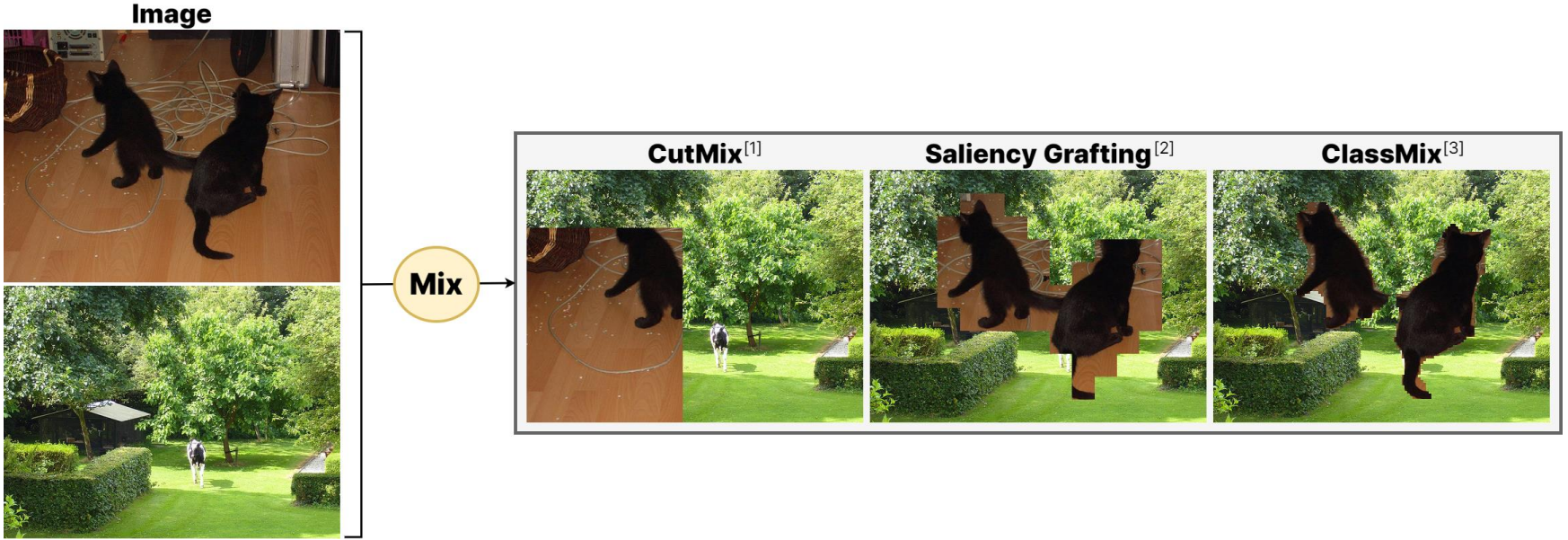


[1] Zhou et al., Learning Deep Features for Discriminative Localization (CVPR 2016)

Background [7/8]

Effect of using data augmentation

✓ Data augmentation (DA) is a simple method that significantly improves data efficiency.

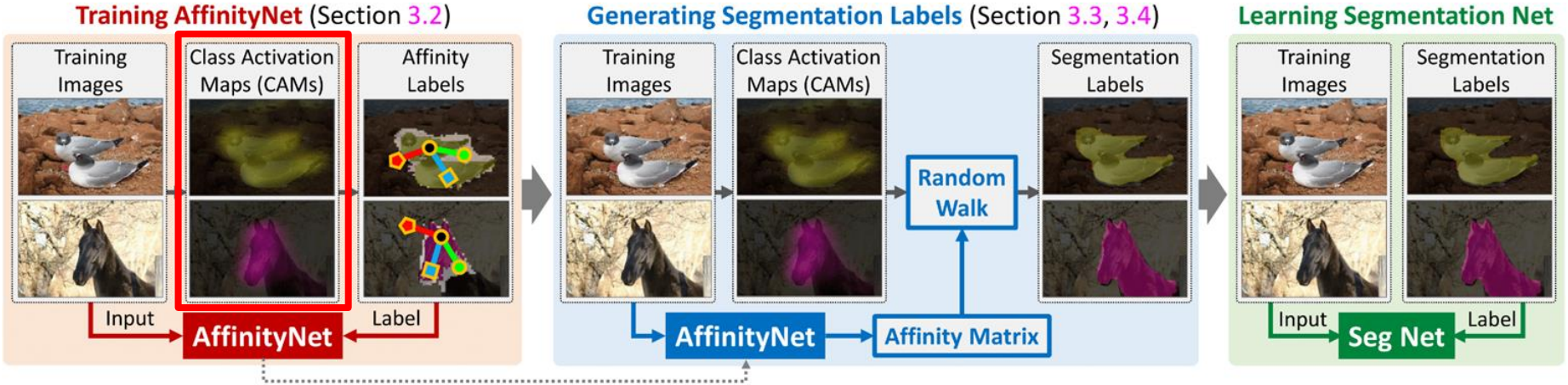


[1] Yun et al., Cutmix: Regularization strategy to train strong classifiers with localizable features (ICCV 2019)
[2] Park et al., Saliency Grafting: Innocuous Attribution-Guided Mixup with Calibrated Label Mixing (AAAI 2022)
[3] Olsson et al., Classmix: Segmentation-based data augmentation for semi-supervised learning. (WACV 2021)

Background [8/8]

Most WSSS methods use multi-stage learning frameworks (MLFs).

1. Training the classification model to produce the CAM.
2. Training the AffinityNet[1] or IRNet[2] to generate pseudo label using a random walk (RW).
3. Training the segmentation model with pseudo labels.



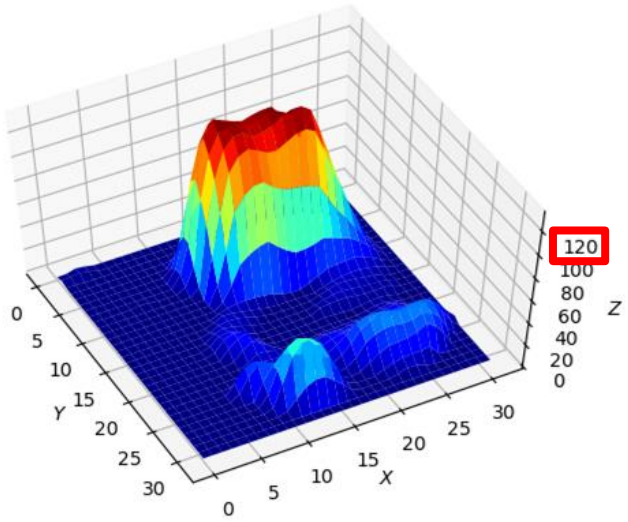
[1] Ahn et al., Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation (CVPR 2018)
[2] Ahn et al., Weakly supervised learning of instance segmentation with inter-pixel relations (CVPR 2019)

Puzzle-CAM / Motivation

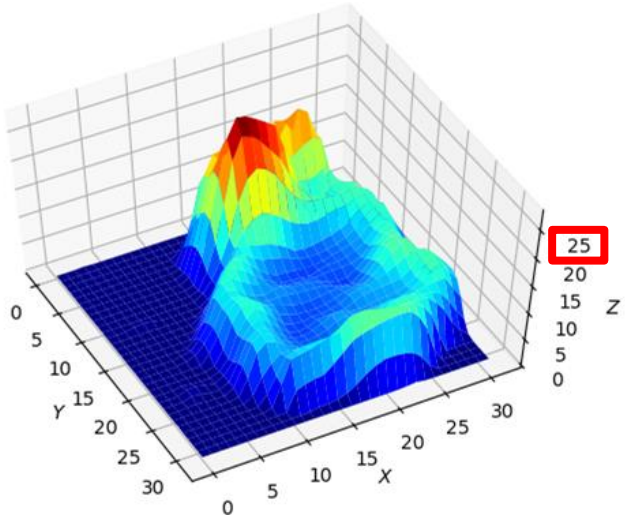
How to avoid detecting the discriminative region of the object?

✓ Puzzle-CAM suppresses the attention on the discriminative region of the object.

CAM



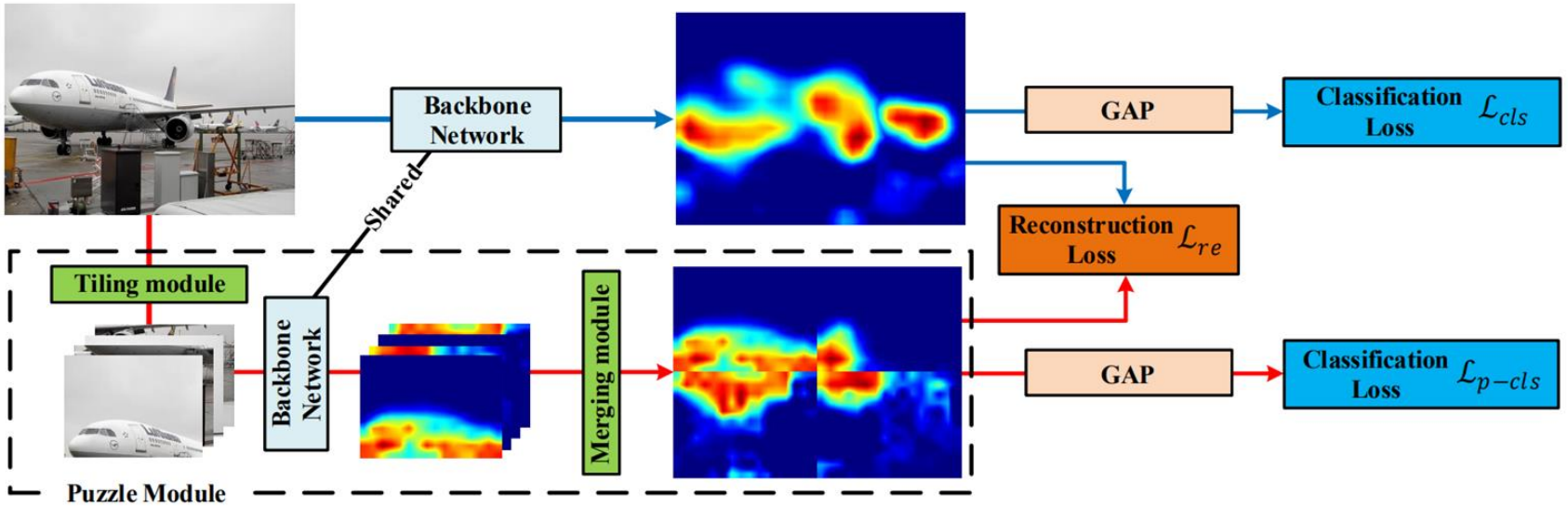
Puzzle-CAM (Ours)



Puzzle-CAM / Method [1/2]

How to learn the integral region of the object by using the tags?

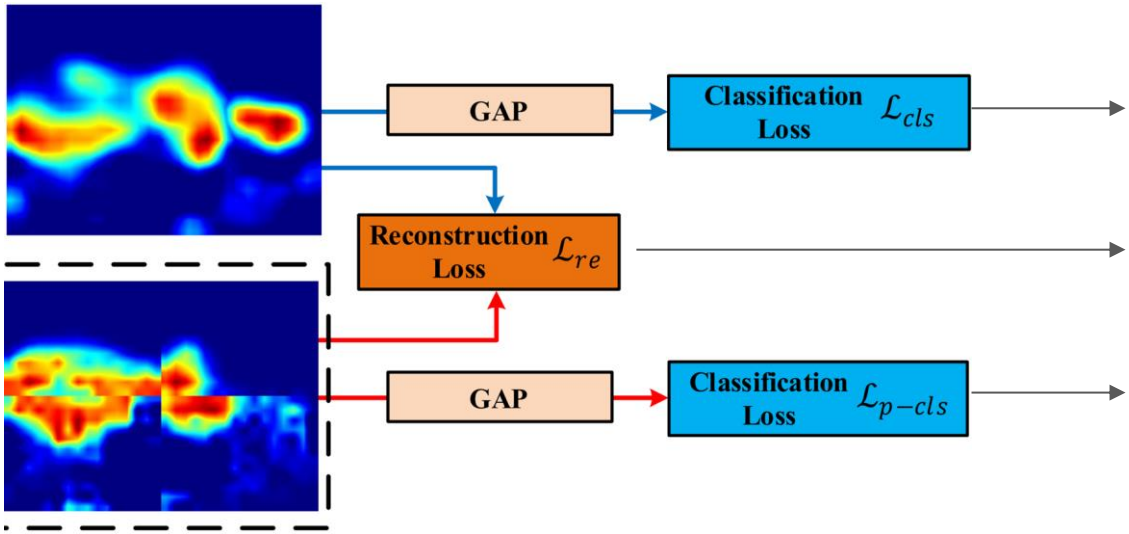
1. Tilling an image to image patches to divide into the attention.
2. Merging the feature maps from the network to produce the reconstructed features.
3. Matching partial and full features with reconstructing regularization.



Puzzle-CAM / Method [2/2]

How to narrow the gaps between the single and reconstructed features?

✓ α is the balance of the weights for the different losses.



$$\mathcal{L}_{cls} = \ell_{cls}(\hat{Y}^s, Y)$$

$$\mathcal{L}_{re} = \|A^s - A^{re}\|_1$$

$$\mathcal{L}_{p-cls} = \ell_{cls}(\hat{Y}^{re}, Y)$$

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{p-cls} + \alpha\mathcal{L}_{re}$$

Puzzle-CAM / Experimental Results [1/2]

Quantitative results

> Our approach outperforms existing state-of-the-art methods without additional supervisions on PASCAL VOC 2012 *val* and *test* sets.

Table 2: Quality of the pseudo semantic segmentation labels in mIoU, evaluated on the PASCAL VOC 2012 training set [14]. RW, random walk with AffinityNet [4]; dCRF, dense conditional random field [16].

Method	Backbone	CAM (%)	CAM +RW (%)	CAM+RW +dCRF (%)
AffinityNet [4]	ResNet-50	47.82	58.10	59.70
Puzzle-CAM	ResNet-50	51.53	64.16	64.70
Puzzle-CAM	ResNeSt-50	57.59	69.48	69.91
Puzzle-CAM	ResNeSt-101	61.85	71.92	72.46
Puzzle-CAM	ResNeSt-269	**62.45**	**74.14**	**74.67**

Table 3: Comparison of Puzzle-CAM and existing state-of-the-art methods on the PASCAL VOC 2012 val and $test$ datasets. $\mathcal{I}$, image-level labels; $\mathcal{S}$, external saliency models.

Method	Backbone	Supervision	val	test
AffinityNet [4]	Wide-ResNet-38	$\mathcal{I}$	61.7	63.7
DSRG [12]	ResNet-101	$\mathcal{I}+\mathcal{S}$	61.4	63.2
SeeNet [13]	ResNet-101	$\mathcal{I}+\mathcal{S}$	63.1	62.8
IRNet [4]	ResNet-50	$\mathcal{I}$	63.5	64.8
FickleNet [6]	ResNet-101	$\mathcal{I}+\mathcal{S}$	64.9	65.3
ICD [17]	ResNet-101	$\mathcal{I}$	64.1	64.3
SEAM [5]	Wide-ResNet-38	$\mathcal{I}$	64.5	65.7
Ours (Puzzle-CAM)	ResNeSt-101	$\mathcal{I}$	66.9	67.7
Ours (Puzzle-CAM)	ResNeSt-269	$\mathcal{I}$	**71.9**	**72.2**

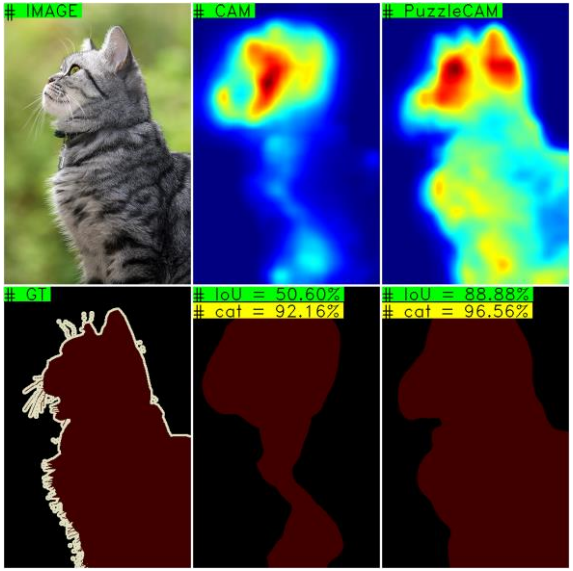
[4] Ahn et al., Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation (CVPR 2018)
[5] Wang et al., Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation (CVPR 2020)
[17] Fan et al., Learning Integral Objects With Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation (CVPR 2020)

Puzzle-CAM / Experimental Results [2/2]

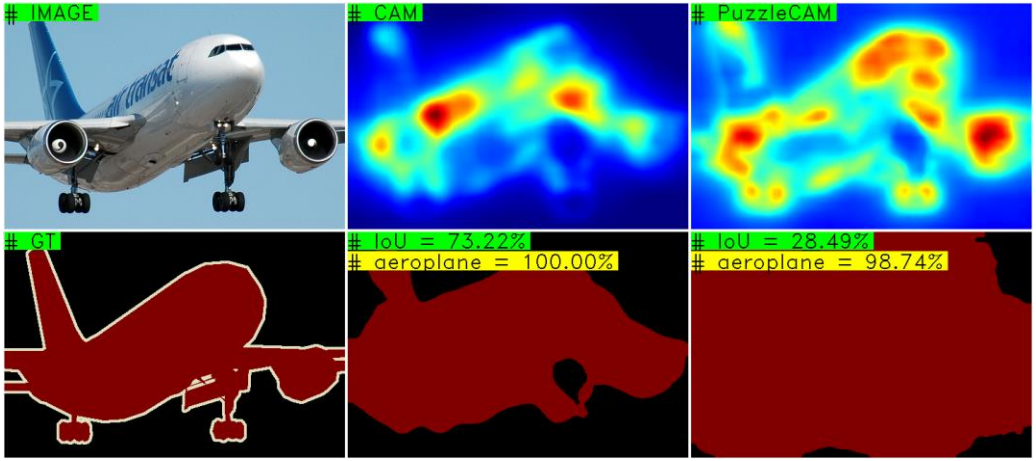
Additional qualitative results

✓ Our method produces high quality semantic segmentation results on large-scale objects. Sometimes, our method causes over-activated CAMs on similar background or small objects.

Successful case



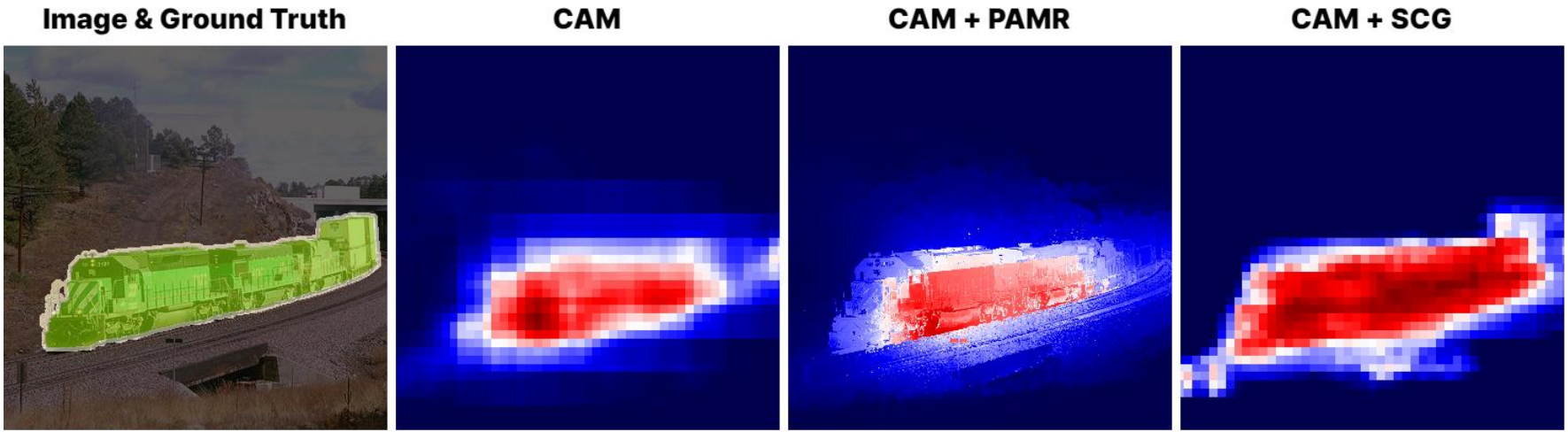
Typical failure case



RS+EPM / Motivation [1/2]

The drawbacks of SCG[1] and PAMR[2]

- ✓ SCG decreases the false negative of the CAM but increases the false positive of the CAM.
- ✓ PAMR decreases the false positive of the CAM but increases the false negative of the CAM.



[1] Pan et al., Unveiling the Potential of Structure Preserving for Weakly Supervised Object Localization (CVPR 2021)
[2] Araslanov and Roth, Single-Stage Semantic Segmentation from Image Labels (CVPR 2021)

RS+EPM / Motivation [2/2]

The shortcoming of existing DA approaches in the WSSS setup

- ✓ The simple synthesis without any refinements using predicted masks accelerates the ambiguity of mixed results due to insufficient quality of the mixed masks.

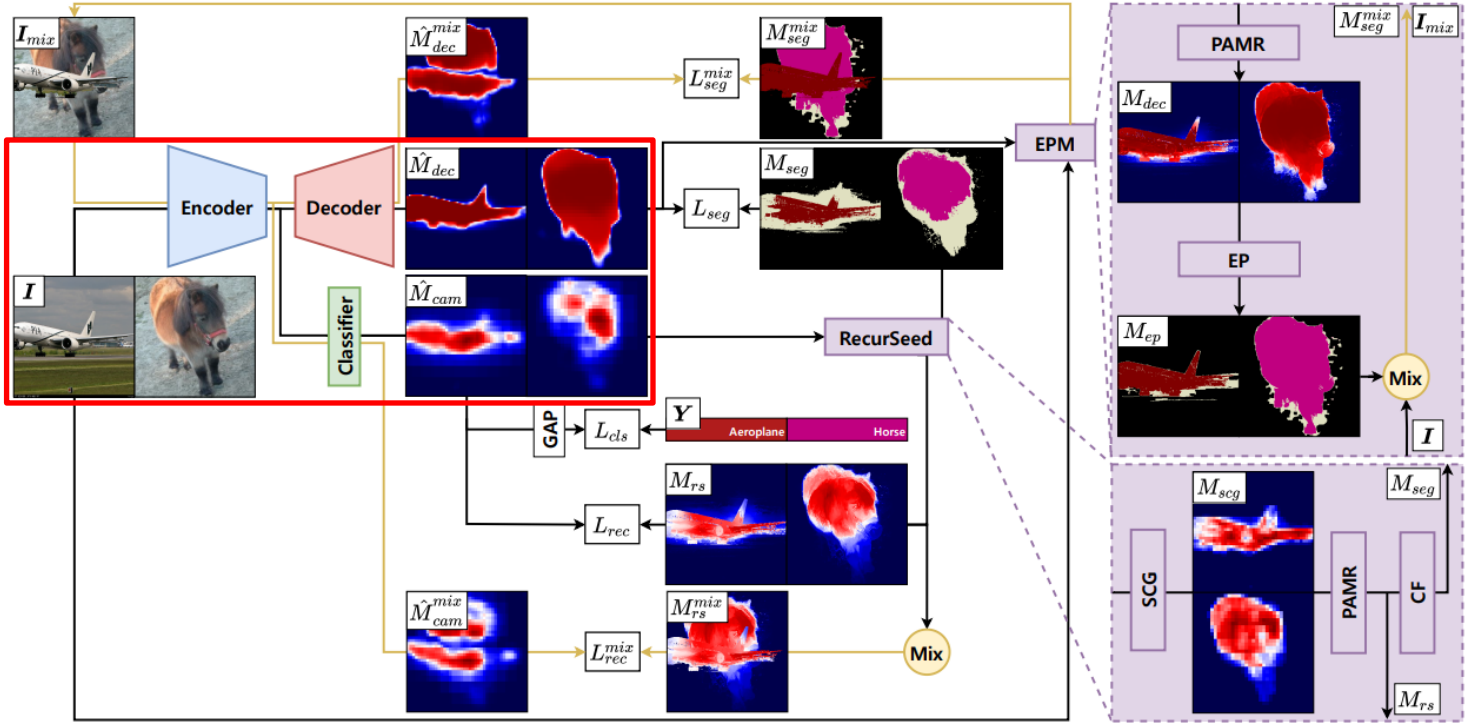
ClassMix [1]

[1] Olsson et al., Classmix: Segmentation-based data augmentation for semi-supervised learning. (WACV 2021)

RS+EPM / Method [1/5]

Overview of the single-stage learning framework with RS and EPM.

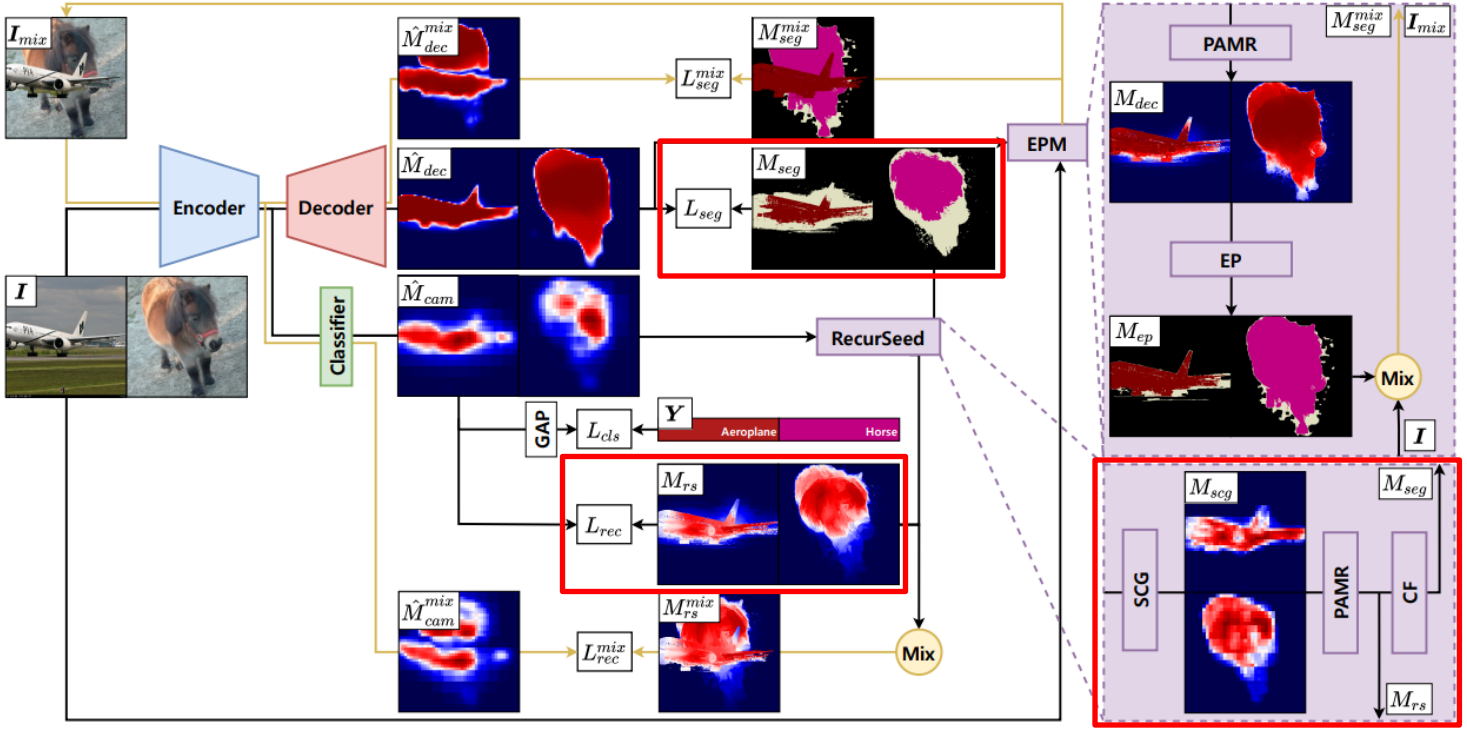
1. The classifier and decoder produce the CAMs and fine-grained segmentation masks



RS+EPM / Method [2/5]

Overview of the single-stage learning framework with RS and EPM.

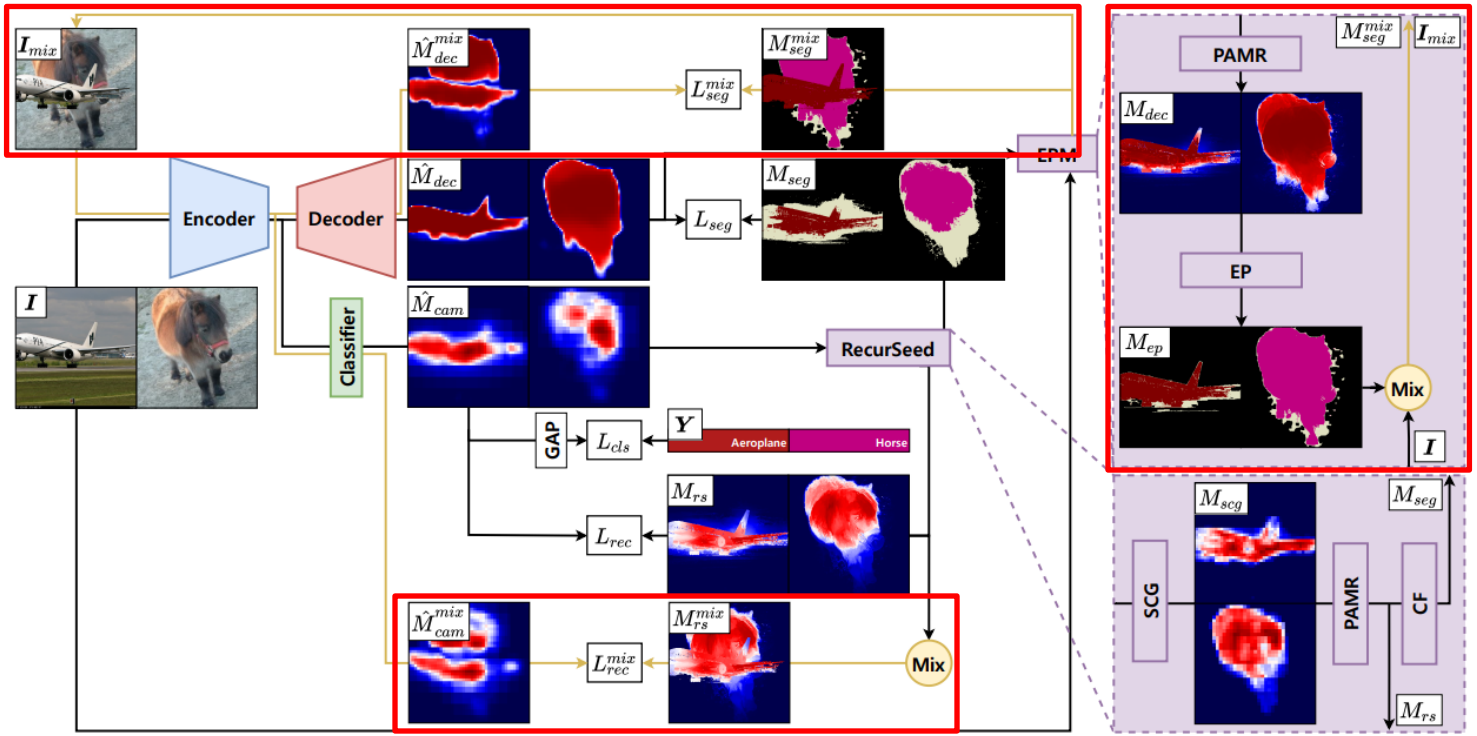
2. RS provides refined CAMs and pseudo masks by leveraging SCG and PAMR.



RS+EPM / Method [3/5]

Overview of the single-stage learning framework with RS and EPM.

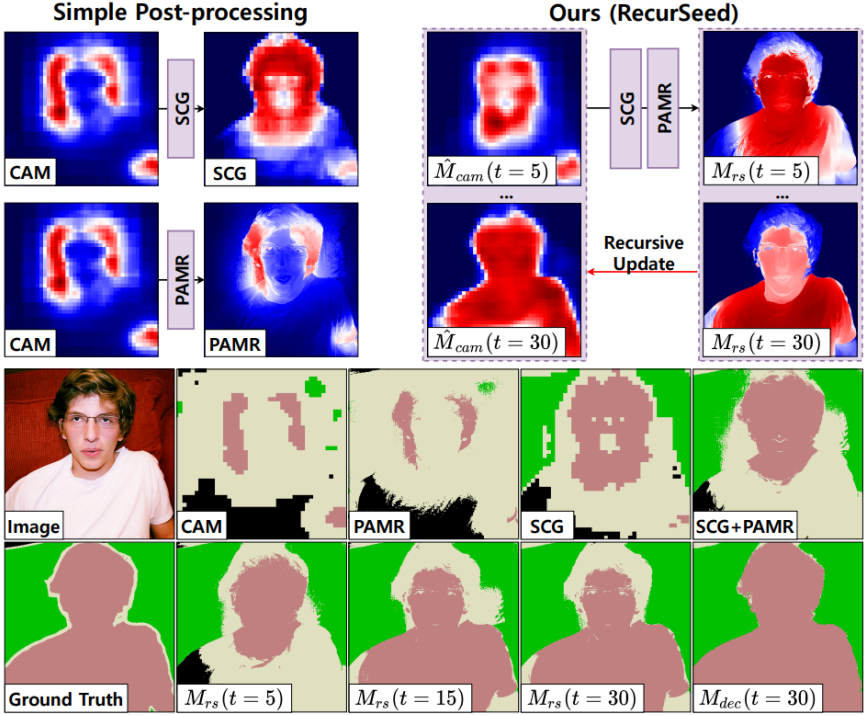
3. EPM mixes two images and pseudo masks refined by EP.



RS+EPM / Method [4/5]

The details of the proposed RecurSeed (RS).

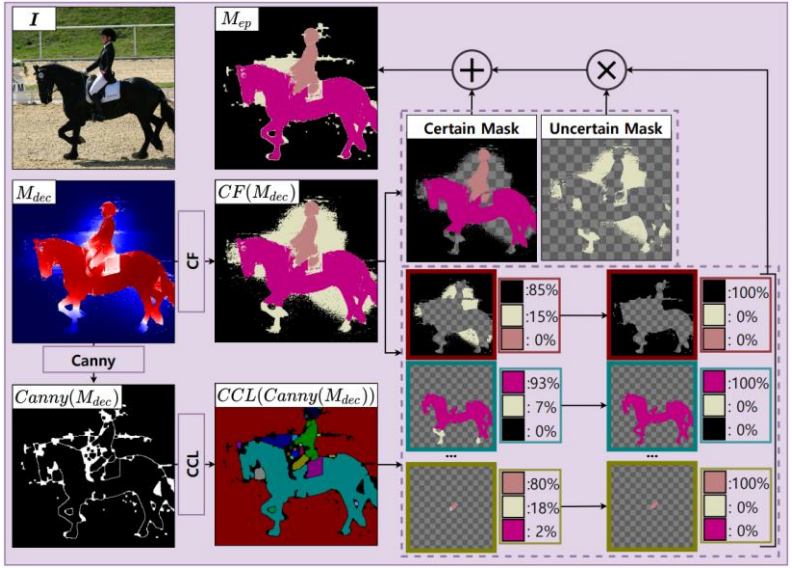
✓ We propose RecurSeed that recursively rectifies the CAM by leveraging SCG and PAMR, making a seed that minimizes both the false positive and false negative.



RS+EPM / Method [5/5]

The details of the proposed EdgePredictMix (EPM).

- ✓ We propose EdgePredictMix that remedies the uncertainty of mixed results by reviving edge information of target objects.
- ✓ Our EP refines predicted masks by utilizing the absolute (using CF) and relative (using Canny) per-pixel probability information.



Ours (EPM)

RS+EPM / Experimental Results [1/5]

Quantitative results

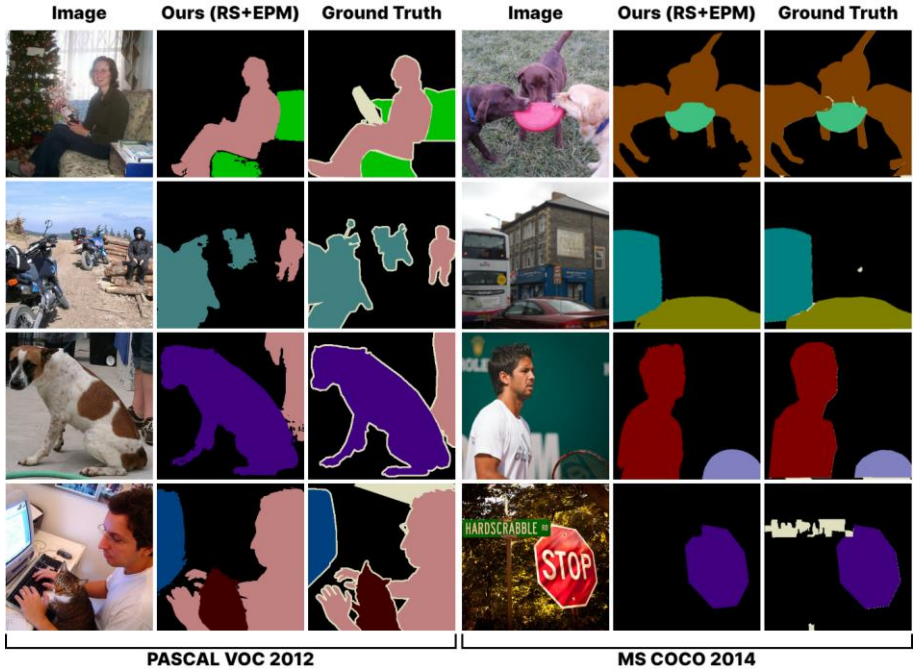
> Our SLF and MLF methods outperformed existing state-of-the-art methods without additional supervisions on PASCAL VOC 2012 and MS COCO 2014 sets.

Method	Backbone	Sup.	VOC *val*	VOC *test*	COCO *val*
Single stage:					
RRM (Zhang et al. 2020a)	WR38	$\mathcal{I}$	62.6	62.9	-
SSSS (Araslanov and Roth 2020)	WR38	$\mathcal{I}$	62.7	64.3	-
AFA (Ru et al. 2022)	MiT-B1	$\mathcal{I}$	66.0	66.3	38.9
Ours (single-stage, RS)	R50	$\mathcal{I}$	66.5	67.9	40.0
Ours (single-stage, RS+EPM)	R50	$\mathcal{I}$	**69.5**	**70.6**	**42.2**
Multiple stages:					
FickleNet (Lee et al. 2019)	R101	$\mathcal{I}$+$\mathcal{S}$	64.9	65.3	-
DRS (Kim, Han, and Kim 2021)	R101	$\mathcal{I}$+$\mathcal{S}$	71.2	71.4	-
RCA (Zhou et al. 2022)	R101	$\mathcal{I}$+$\mathcal{S}$	72.2	72.8	*36.8
CSE (Kweon et al. 2021)	WR38	$\mathcal{I}$	68.4	68.2	36.4
RIB (Lee et al. 2021a)	R101	$\mathcal{I}$	68.3	68.6	43.8
SIPE (Chen et al. 2022)	R101	$\mathcal{I}$	68.8	69.7	40.6
Ours (multi-stage, RS)	R101	$\mathcal{I}$	72.8	72.8	45.8
Ours (multi-stage, RS+EPM)	R101	$\mathcal{I}$	**74.4**	**73.6**	**46.4**

RS+EPM / Experimental Results [2/5]

Qualitative results

- ✓ Our method not only performs well for different complex scenes, small objects, or multiple instances but also can achieve a satisfactory segmentation performance for various challenging scenes.



RS+EPM / Experimental Results [3/5]

Novelty of RecurSeed

1. Without RS, the simple integration of SCG and PAMR shows a marginal improvements.
2. The proposed RS achieved the highest performance and decreased both FP and FN.
3. * denotes the decoder map result.

RS	SCG	PAMR	mIoU	FP	FN
✗	✓	✗	58.0	0.268	0.165
✗	✓	✓	59.3	0.225 (↓ 0.043)	0.194 (↑ 0.029)
✓	✓	✗	65.2	0.216	0.143
✓	✓	✓	65.9	0.210 (↓ 0.006)	0.141 (↓ 0.002)
✓	✓	✗	*67.4	*0.196	*0.141
✓	✓	✓	*70.7	*0.171 (↓ 0.025)	*0.134 (↓ 0.007)

RS+EPM / Experimental Results [4/5]

Novelty of EdgePredictMix

1. We implemented existing DA methods with the proposed RS for a fair comparison.
2. Our EPM further refines pseudo masks by leveraging the reconstituted edge from uncertain regions, resulting in the highest WSSS performance compared to other DA methods.

Method	Backbone	F_1	mIoU
RecurSeed	R50	94.7	70.7
RecurSeed + *CutMix (Yun et al. 2019)	R50	95.6	68.5
RecurSeed + *SaliencyGrafting (Park et al. 2021)	R50	**96.8**	68.6
RecurSeed + *CDA (Su et al. 2021)	R50	96.0	69.0
RecurSeed + *ClassMix (Olsson et al. 2021)	R50	94.6	71.2
RecurSeed + EdgePredictMix	R50	95.2	**75.2**

RS+EPM / Experimental Results [5/5]

Comparison with existing MLFs using RW

1. We compared the performance of the proposed SLF with existing MLFs by extending it to MLF with the same configuration of RW.
2. Our method outperformed the related works by achieving mIoUs of 75.2% and 76.7% without and with RW, respectively.

Method	Backbone	Seed	RW
SEAM (Wang et al. 2020)	WR38	55.4	63.6
IRNet (Ahn, Cho, and Kwak 2019)	R50	48.8	66.3
CDA (Su et al. 2021)	R50	50.8	67.7
CPN (Zhang et al. 2021)	WR38	57.4	67.8
CONTA (Zhang et al. 2020b)	R50	48.8	67.9
AMR (Qin et al. 2021)	R50	56.8	69.7
RIB (Lee et al. 2021a)	R50	56.5	70.6
Ours (RS)	R50	70.7	74.8
Ours (RS+EPM)	R50	**75.2**	**76.7**

RS+EPM / Conclusion & Significance

1. The proposed RS and EPM significantly outperformed previous state-of-the-art methods with the same level of supervision on the PASCAL VOC 2012 and MS COCO 2014 datasets.
2. Our SLF suffices to accomplish advanced performance on WSSS-IL without the complicated learning configuration.

Appendix

✓ Demonstrated the state-of-the-art performance of the proposed method.

Computer Vision

Weakly-Supervised Semantic Segmentation

84 papers with code • 3 benchmarks • 5 datasets

The semantic segmentation task is to assign a label from a label set to each pixel in an image. In the case of fully supervised setting, the dataset consists of images and their corresponding pixel-level class-specific annotations (expensive pixel-level annotations). However, in the weakly-supervised setting, the dataset consists of images and corresponding annotations that are relatively easy to obtain, such as tags/labels of objects present in the image.

(Image credit: Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing)

Benchmarks

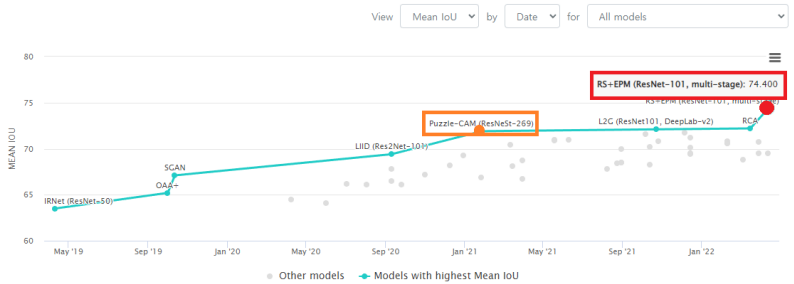
Add a Result

These leaderboards are used to track progress in Weakly-Supervised Semantic Segmentation

Trend	Dataset	Best Model	Paper	Code	Compare
	PASCAL VOC 2012 val	RS+EPM (ResNet-101, multi-stage)			See all
	PASCAL VOC 2012 test	RS+EPM (ResNet-101, multi-stage)			See all
	COCO 2014 val	RS+EPM (ResNet-101, multi-stage)			See all

Weakly-Supervised Semantic Segmentation on PASCAL VOC 2012 val

Leaderboard Dataset

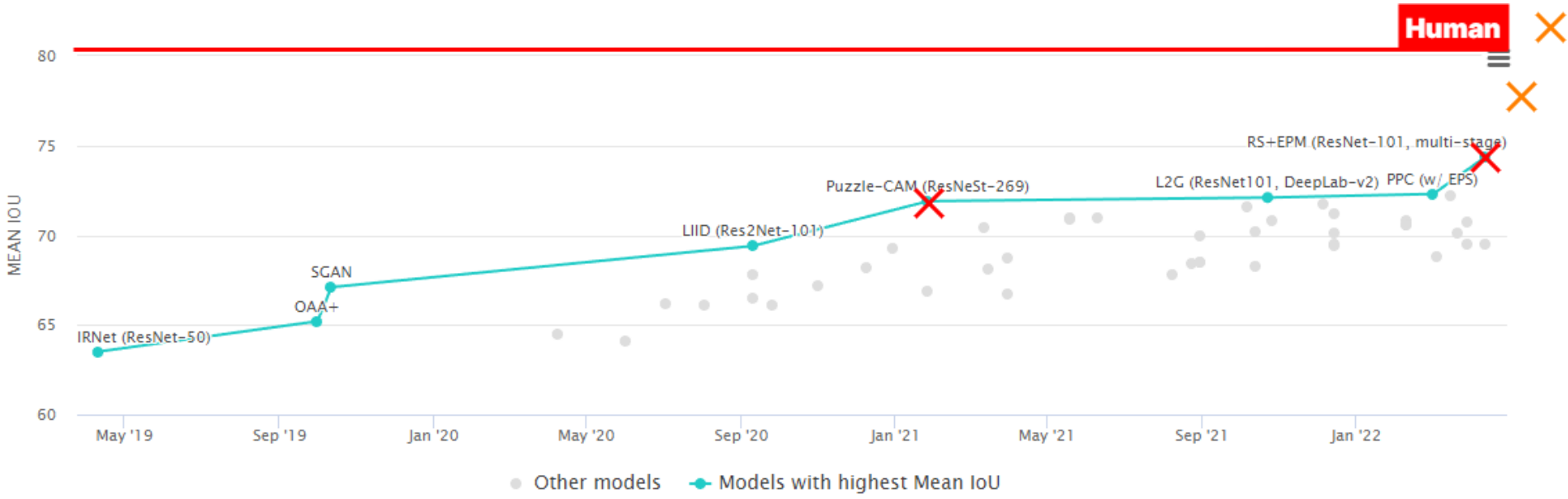


Weakly-Supervised Semantic Segmentation on COCO 2014 val

Leaderboard Dataset

https://paperswithcode.com/task/weakly-supervised-semantic-segmentation

Appendix


Human
80
75
70
65
60
MEAN IOU
RS+EPM (ResNet-101, multi-stage)
Puzzle-CAM (ResNeSt-269)
L2G (ResNet101, DeepLab-v2)
PPC (w/ EPS)
LIID (Res2Net-101)
SGAN
OAA+
IRNet (ResNet-50)
May '19
Sep '19
Jan '20
May '20
Sep '20
Jan '21
May '21
Sep '21
Jan '22
Other models
Models with highest Mean IoU

Q & A

Puzzle-CAM

RS+EPM

shjo.april@gmail.com