WED-AM-295

CVPR JUNE 18-22, 2023 VANCOUVER, CANADA

KAIST

Weakly Supervised Semantic Segmentation via Adversarial Learning of Classifier and Reconstructor

Hyeokjun Kweon*, Sung–Hoon Yoon*, and Kuk-Jin Yoon

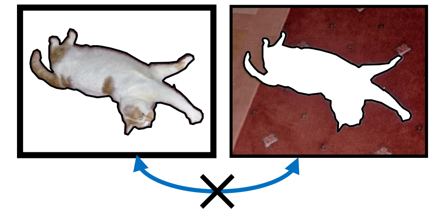
{0327june, yoon307, kjyoon}@kaist.ac.kr

Summary of the Paper

- Aims to obtain **pseudo-GT** using **weak (classification) labels** only, and trains segmentation model with it

Preciseness of semantic segmentation

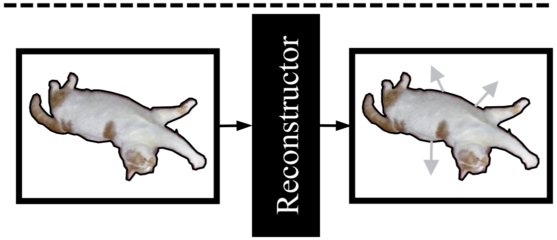
Measure of the **preciseness** of semantic segmentation
≅ ***Inferability*** between the segments



Cannot ***infer*** the color or texture of one segment from the other

Inferability between the segments

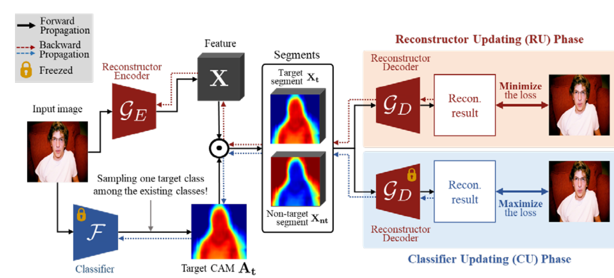
Inferability
≅ Error between the input image and the reconstructed image
≅ **Reconstruction capability**



IF Precise:
Fail to correctly reconstruct

Reconstruction capability between the segments

The **classifier learns to generate precise CAM**
that can fool the reconstructor while competing with each other

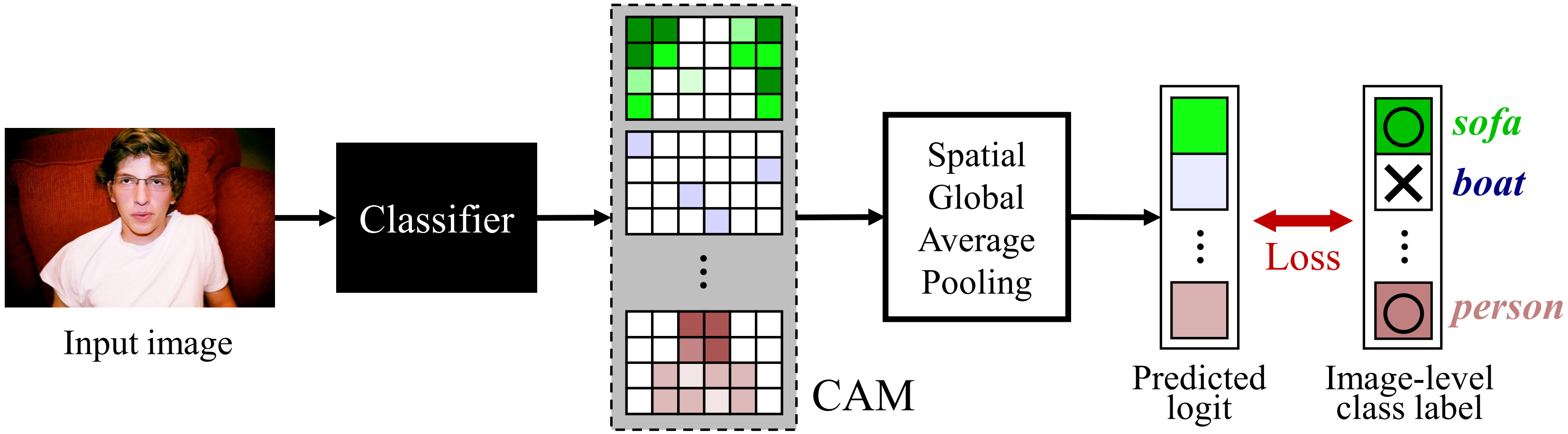


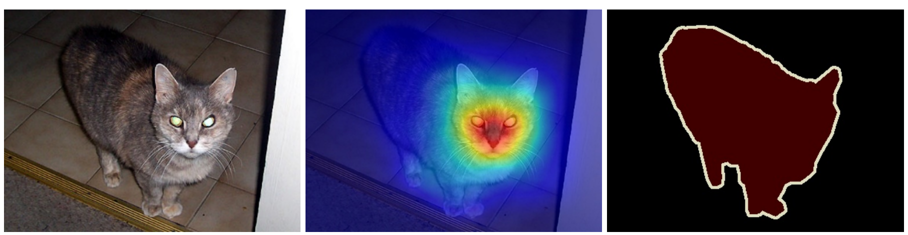
The proposed framework

Our method

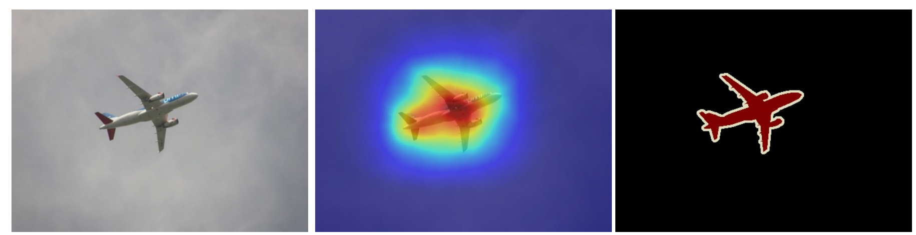
Introduction: Class Activation Map (CAM) and Issues

- WSSS methods have used a **CAM** [1] of classifier to obtain pseudo-GT from image-level class labels
- A classifier tends to focus on the **shared visual patterns** of each class as training proceeds
→ The resulting CAM highlights the **discriminative regions** of the corresponding class





(1) Incomplete activation



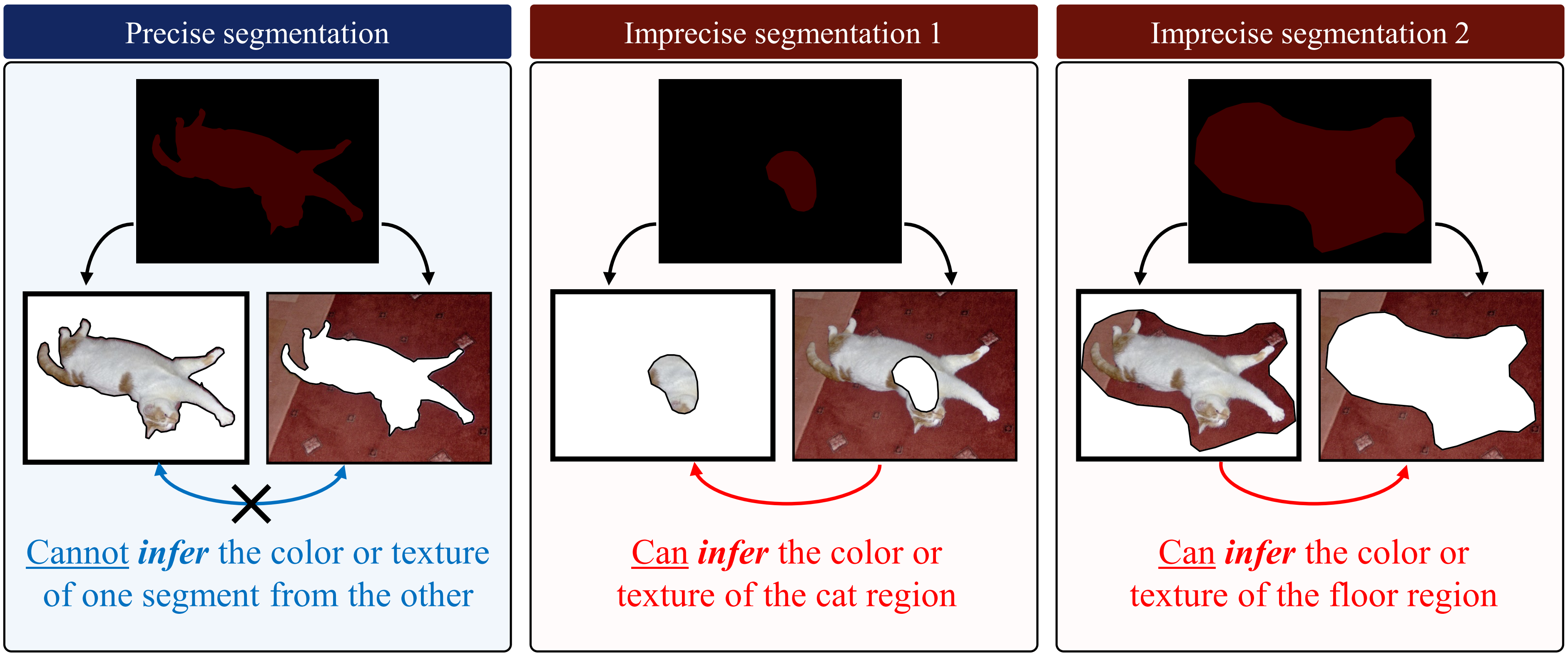
(2) Imprecise boundary

- CAM = a reasonable localizer extracting **spatial information from semantic supervision**
- However, from the perspective of semantic segmentation, CAM still has several issues

Methods: Key Idea 1 (Semantic Segmentation and Inferability)

- How can we measure the preciseness of semantic segmentation without using GT map?

Precise Segmentation → Low Inferability between the segments
Imprecise Segmentation → High Inferability between the segments



- ***Inferability*** between the segments = Measure of the **preciseness** of semantic segmentation
- Train CAM to segment the image into the segments that have low ***inferability*** between them
- But how can we **quantify** the inferability?

Methods: Key Idea 2 (Reconstruction Capability)

Preciseness of semantic segmentation

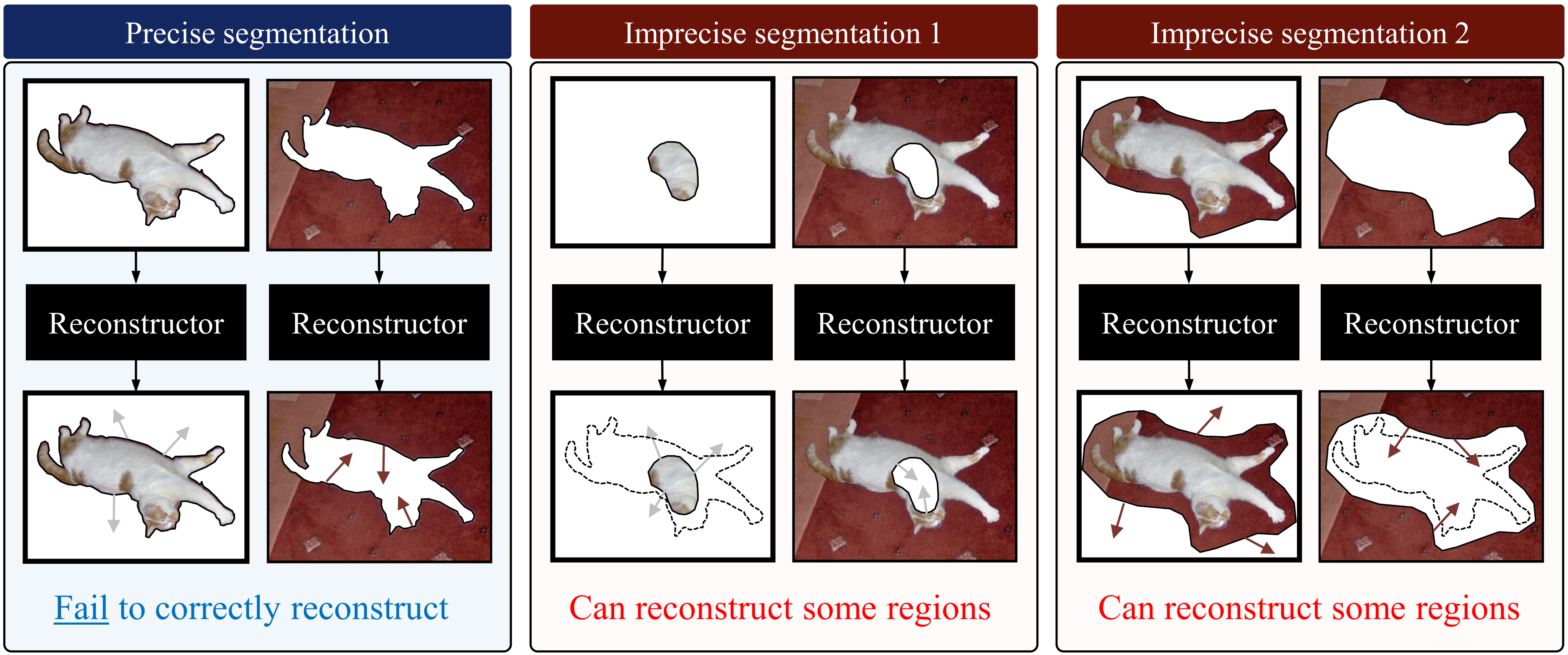
↓

Inferability between the segments

↓

Reconstruction capability between the segments

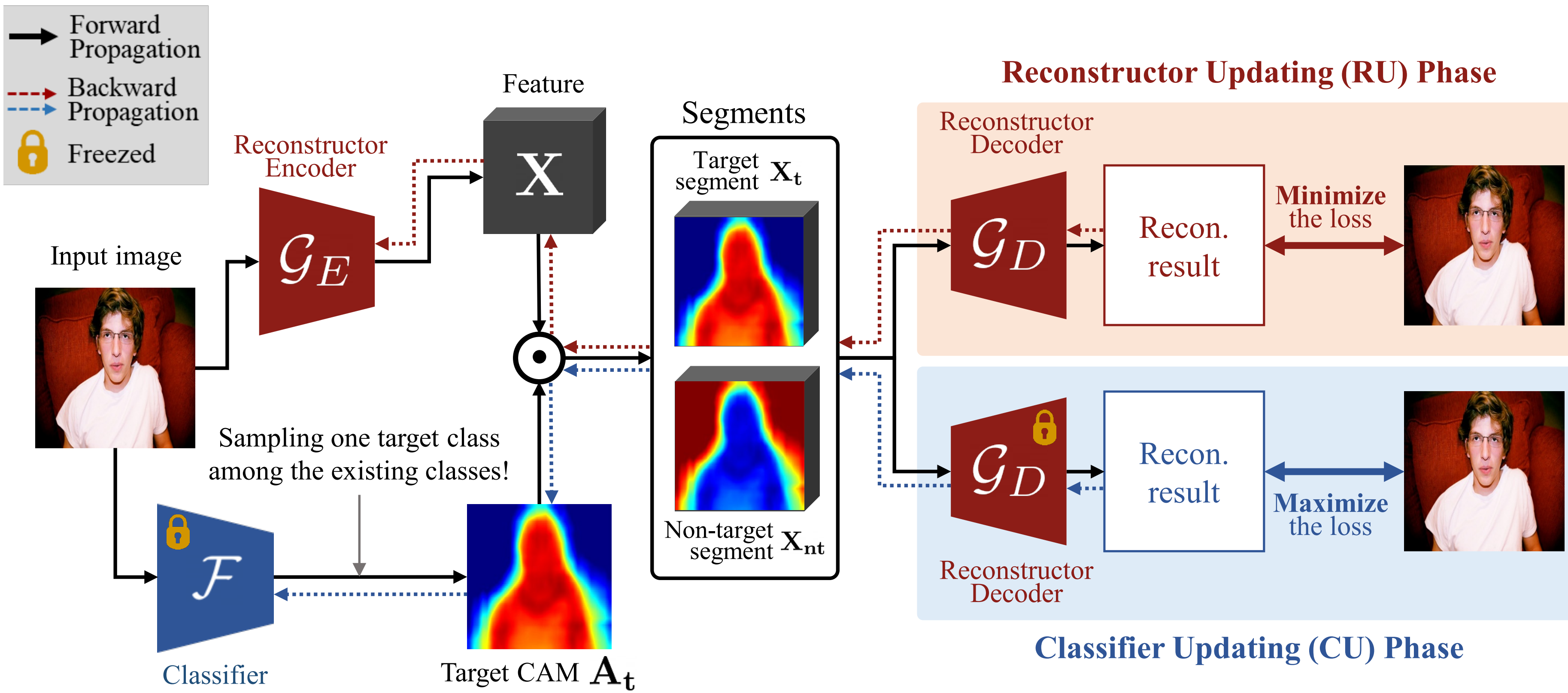
- If one can infer color or texture of a segment from the other segment → the segment could be **reconstructed** from the other segment.
- Image reconstructor (or impainter) relies on neighboring pixels to reconstruct the masked pixels → Reconstructor can use **remnants** as reconstruction cues!



- **Reconstruction capability** = Error between the input image and the reconstructed image = Inferability
- Train CAM to segment the image into the segments that **make reconstructor fail to impaint each other**
- But how can we obtain such reconstructor?

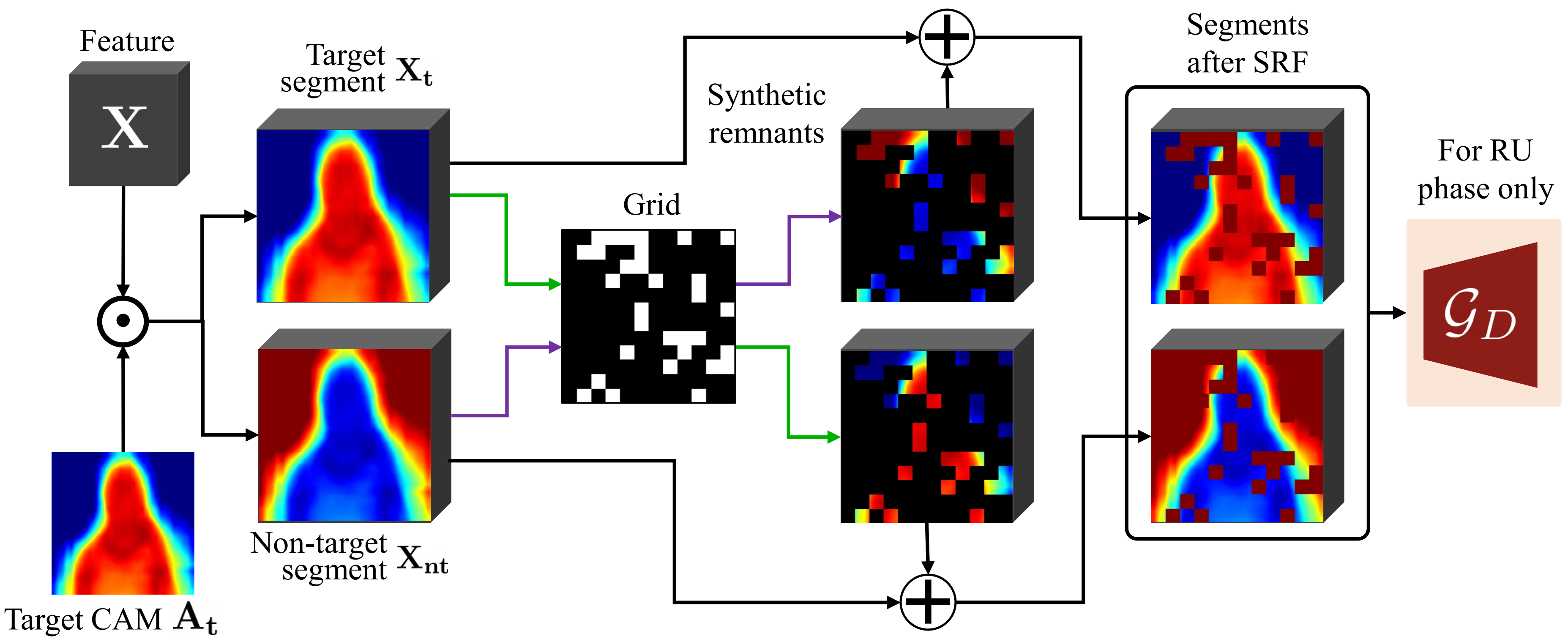
Methods: Adversarial Learning of Classifier and Reconstructor

- Similar to the GANs, a classifier and a reconstructor are jointly trained in an alternative manner
- The target CAM of the **classifier** decomposes a feature into two segments: target and non-target
- Goal of **reconstructor**: reconstructing the input image from the segments, by fully exploiting the **remnants**
- Goal of **classifier:** making the reconstructor fails, by providing good CAM that achieves precise segmentation.
- The **classifier learns to generate precise CAM** that can fool the reconstructor while competing with each other



Methods: Stochastic Random Feeding (SRF)

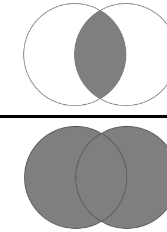
- The proposed adversarial learning has a risk of falling into an undesirable local optimum
- The classifier succeeds to provide precise CAM → The segments are perfectly segmented
- The reconstructor tends to ***generate*** the original image (rather than learning how to exploit the remnants)
- We devise **SRF technique** to assure the **<u>existence of remnants</u>** by feeding synthetic remnants
- SRF technique is applied to the RU phase only, not for the CU phase



Experimental Results: Settings and Qualitative Analysis

- Dataset: PASCAL VOC 2012 [13] and MS COCO 2014 [14]
- Metric: mean Intersection over Union (mIoU)
- Backbones: ResNet38d for both classifier and reconstructor
- Protocols:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



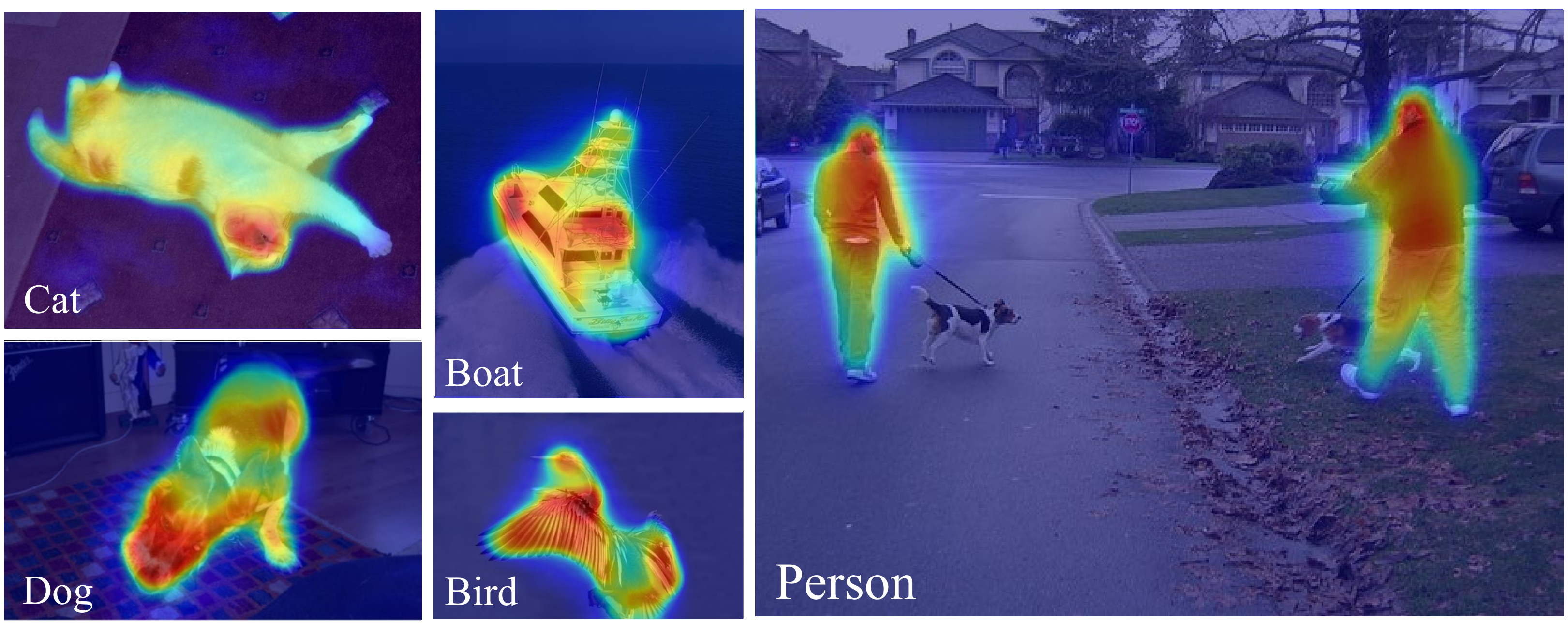
(1) Train the framework using *train* set.

(2) Obtain CAM of the images of *train* set.

(3) Postprocess the CAM into pseudo-labels using CRF, random walking, etc.

(4) Train semantic segmentation network on *train* set using the acquired pseudo-labels.

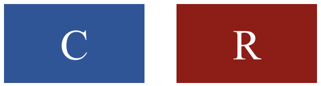
(5) Evaluate the network on *val* and *test* set. (We use the GT segmentation map here only.)



Experimental Results: Quantitative Analysis

- Ablation studies about learning strategy and the loss terms

(1) Adversarial strategy



Both networks are jointly trained in an alternative manner

(2) Pre-trained strategy

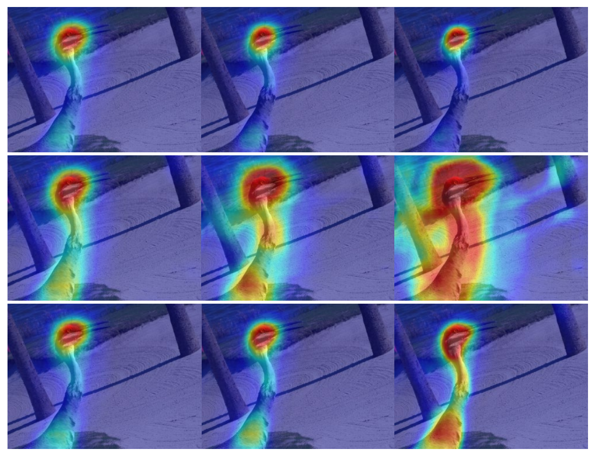
First, solely train the classifier without using reconstructor

Second, train the reconstructor while freezing the classifier

Finally, train the classifier while freezing the reconstructor

Learning strategy for reconstructor	SRF	mIoU (%)
Baseline		48.4
Pre-trained		52.9
Pre-trained	✓	54.6
Adversarial		55.8
Adversarial	✓	**60.3**

(a)
(b)
(c)



	Loss function			Metrics		
	$\mathcal{L}_{cls}^{CU}$	$\mathcal{L}_{t}^{CU}$	$\mathcal{L}_{nt}^{CU}$	Precision	Recall	mIoU (%)
	✓			0.61	0.72	48.4
(a)	✓	✓		0.73 (+0.12)	0.68 (-0.04)	53.4 (+5.0)
(b)	✓		✓	0.58 (-0.03)	**0.77 (+0.05)**	51.5 (+3.1)
(c)	✓	✓	✓	**0.75 (+0.14)**	0.76 (+0.04)	**60.3 (+13.6)**

- Comparisons with SoTAs (at CAM)

Methods	Backbone	seed	w/ CRF	Mask
CONTA [48] $_{NeurIPS20}$	WRN38	56.2	65.4	66.1
EDAM [39] $_{CVPR21}$	WRN38	52.8	58.2	68.1
AdvCAM [21] $_{CVPR21}$	RN50	55.6	62.1	68.0
ECS [35] $_{ICCV21}$	WRN38	56.6	58.6	-
OC-CSE [19] $_{ICCV21}$	WRN38	56.0	62.8	66.9
CDA [33] $_{ICCV21}$	WRN38	58.4	-	66.4
PMM [28] $_{ICCV21}$	WRN38	58.2	61.5	61.0
RIB [20] $_{NeurIPS}$	RN50	56.5	62.9	70.6
AMR [32] $_{AAAI22}$	RN50	56.8	-	69.7
ReCAM [8] $_{CVPR22}$	RN50	54.8	-	70.5
SIPE [7] $_{CVPR22}$	RN50	58.6	<u>64.7</u>	-
CLIMS [41] $_{CVPR22}$	WRN38	56.6	-	70.5
W-OoD [22] $_{CVPR22}$	RN50	53.3	58.4	
PPC [10] $_{CVPR22}$	WRN38	**61.5**	64.0	70.1
AEFT [46] $_{ECCV22}$	WRN38	56.0	63.5	<u>71.0</u>
Ours (ACR)	WRN38	<u>60.3</u>	**65.9**	**72.3**
MCT [44] $_{CVPR22}$	ViT	61.7	-	69.1
Ours (ACR + ViT [9])	ViT	**65.5**	-	**70.9**

- Comparisons with SoTAs (at SS)

Methods	*Backbone*	VOC *val*	VOC *test*	COCO *val*
AffinityNet [2] $_{CVPR18}$	WRN38	61.7	63.7	-
IRNet [1] $_{CVPR19}$	RN50	63.5	64.8	41.4
SEAM [37] $_{CVPR20}$	WRN38	64.5	65.7	31.9
OC-CSE [19] $_{ICCV21}$	WRN38	68.4	68.2	36.4
CPN [49] $_{ICCV21}$	WRN38	67.8	68.5	-
RIB [20] $_{NeurIPS21}$	RN101	68.3	68.6	43.8
PMM [28] $_{ICCV21}$	WRN38	68.5	69.0	36.7
ReCAM [8] $_{CVPR22}$	RN101	68.5	68.4	42.9
SIPE [7] $_{CVPR22}$	RN101	68.8	69.7	-
SIPE [7] $_{CVPR22}$	WRN38	-	-	43.6
CLIMS [41] $_{CVPR22}$	RN50	69.3	68.7	
W-OoD [22] $_{CVPR22}$	WRN38	70.7	70.1	
Spatial-BCE [38] $_{ECCV22}$	RN101	70.0	71.3	-
Spatial-BCE [38] $_{ECCV22}$	VGG16	-	-	35.2
AEFT [46] $_{ECCV22}$	WRN38	70.9	71.7	44.8
Ours (ACR)	WRN38	**71.9**	**71.9**	**45.3**
MCT [44] $_{CVPR22}$	WRN38	71.9	71.6	42.0
Ours (ACR + ViT [9])	WRN38	**72.4**	**72.4**	**-**