Self-supervised difference detection for weakly supervised segmentation

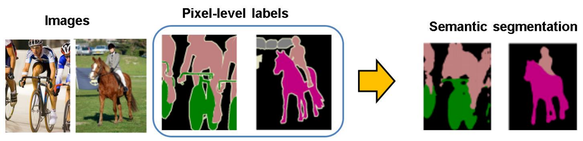
Wataru Shimoda **Keiji Yanai** The University of Electro-Communications, Tokyo, Japan

Objective

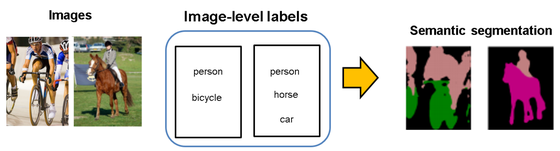
Weakly-supervised segmentation

- Use only image-level annotation and generate segmentation masks

Fully-supervised segmentation

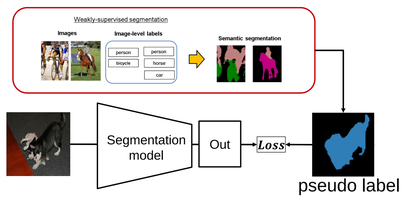


Weakly-supervised segmentation

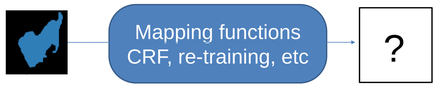


How to improve pseudo labels?

Recent weakly-supervised segmentation methods generate pseudo labels in advance and train a segmentation model with them.

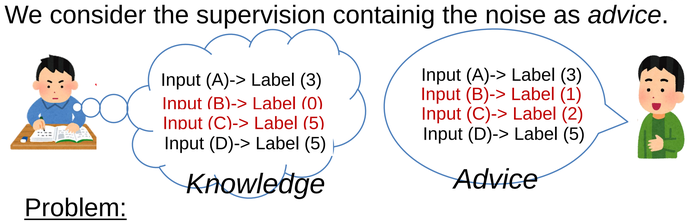


We can not evaluate the pseudo labels in weakly supervised setting
Most previous approaches based on heuristic knowledge.
There are noise in the mapped results



Key idea

We denote the inputs of the mapping functions as *knowledge*,
We consider the supervision containig the noise as *advice*.

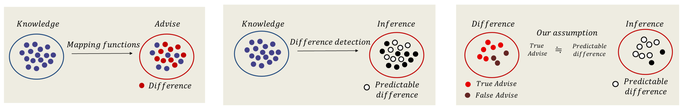


Problem:

- It is unclear which advice is useful
- We want to detect good *advice*

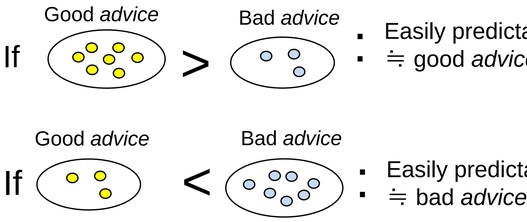
Idea

- Different opinions from the adviser are important
- Predict the important *advice* in advance
- Use the prediction for detecting good *advice*



Assumption

The number of the training sample is related to the degree of difficulty of the inference in difference detection



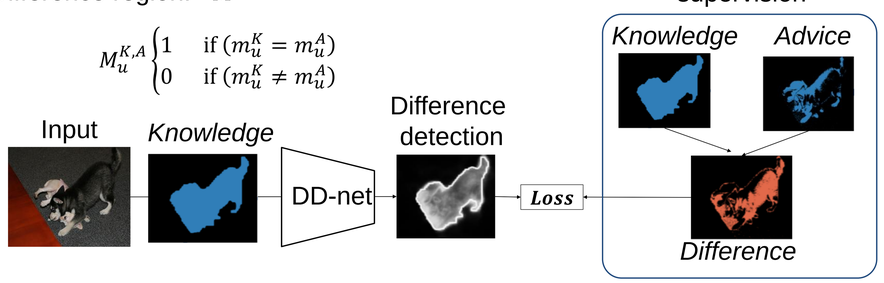
Self-supervised difference detection(SSDD)

The definition of difference detection

Estimate difference between *knowledge* and *advice*

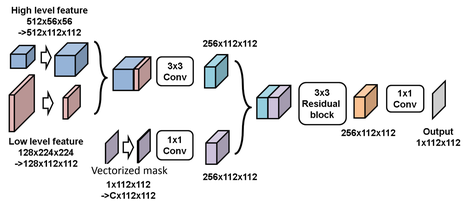
Difference region: $M^{K,A}$

$$M_u^{K,A}\begin{cases}1 & \text{if } (m_u^K = m_u^A)\\ 0 & \text{if } (m_u^K \neq m_u^A)\end{cases}$$

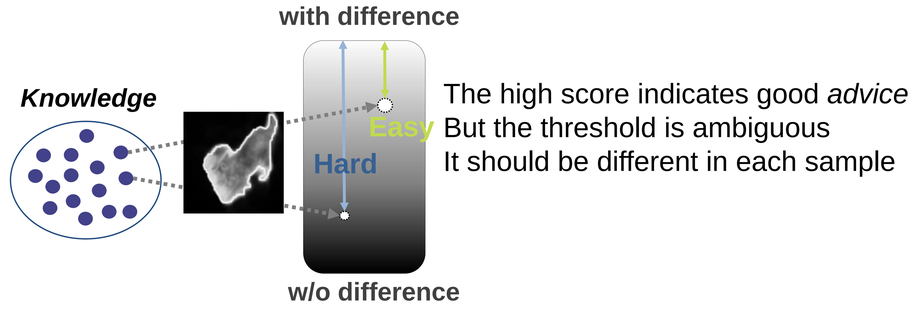


Difference detection network (DD-net)

- Input: mask , features $(e^h(x), e^l(x))$
- Output: Probability map
- $d = DDnet(e^h(x), e^l(x), \hat{m}) \in \mathbb{R}^{H\times W}$
  - $\hat{m}$ is the vectorized mask

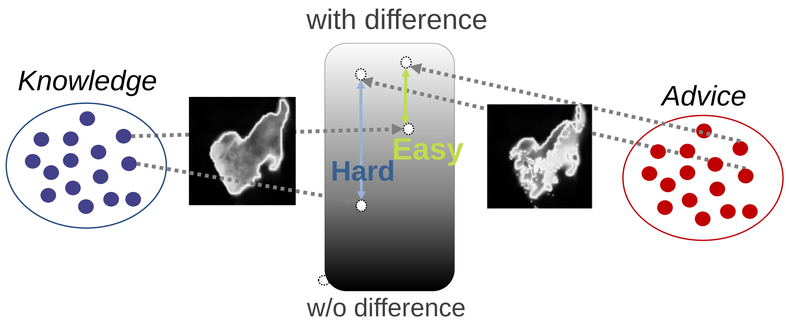


Define the border of the good *advice*



The high score indicates good *advice*
But the threshold is ambiguous
It should be different in each sample

For a flexible measurement, we also estimate the difference from *advice*.
We use the distance of the outputs of DDnet for the measurement



Mask integration

The computation of the confidence scores

$$\mathrm{w_u} = d_u^K - d_u^A + bias_u$$

$$bias_u = \begin{cases} b_{dd} \pm b_{class} & \text{if } m_u^A \text{ or } m_u^k \text{ belongs to } \hat{\mathcal{C}} \\ b_{dd} & \text{if otherwise} \end{cases}$$

$$\forall c \in \hat{\mathcal{C}} \text{ satisfy } \frac{|S_c^{m^A}|}{|S_c^{m^K}|} < 0.5 \text{ and } c \in \mathcal{C}$$

Easy

We integrate m^K and m^A using the confidence scores

$$m_u^D\begin{cases} m_u^K & \text{if } (w_u \geq 0)\\ m_u^A & \text{if } (w_u < 0)\end{cases}$$

We denote this integration process as SSDD module

$$\mathrm{m}^D = SSDD(e(x;\theta_e), m^K, m^A; \theta_d)$$

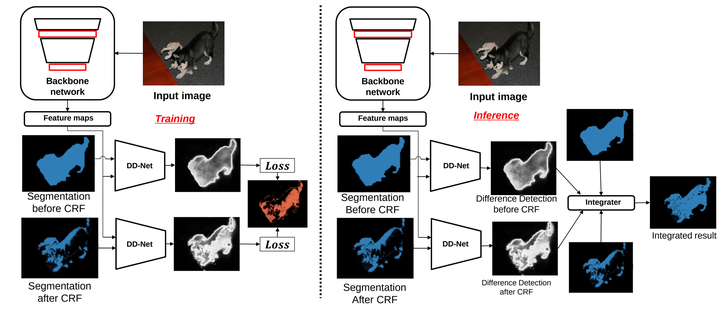
SSDD training/inference

-Training

Train a difference detection model using the difference of the *Knowledge* and *advice*

-Inference

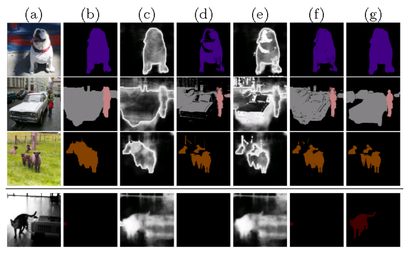
Integrate a pair of the mask using DD-net outputs

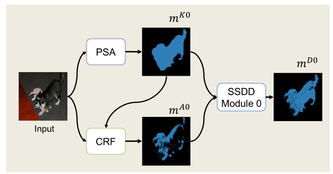


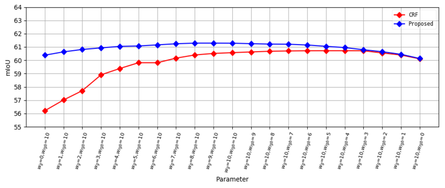
Static region refinement [A]

***Knowledge* : PSA[1]**

***Advice* : PSA[1] + CRF**





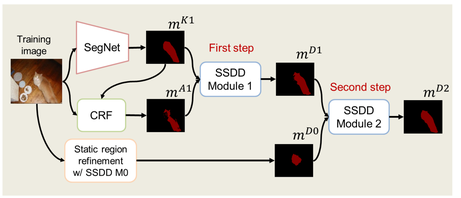


Dynamic region refinement [B]

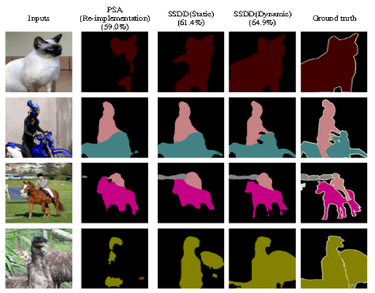
***Knowledge* : Pseudo labels**

***Advice* : The outputs of the segmentation model**

We attempt to adapt the re-training scheme[2] to SSDD module



- Dataset: Pascal VOC 2012 dataset
- Evaluation metric: mean IoU



The comparison with WSS methods w/o additional supervision.

Methods	Val set	Test set
FCN-MIL(ICLR2015)	25.7	24.9
CCNN(ICCV2015)	35.3	35.6
EM-Adapt(ICCV2015)	38.2	39.6
DCSM(ECCV2016)	44.1	45.1
BFBP(ECCV2016)	46.6	48.0
SEC (ECCV2016)	50.7	51.7
TPL(ICCV2016)	53.1	53.8
CBTS(CVPR2017)	52.8	53.7
PSA(CVPR2018)	61.7	63.7
SSDD(proposed)	**64.9**	**65.5**

The comparison with WSS methods w/additional supervision.

Methods	Additional information	Val set	Test set
MIL-seg(CVPR2015)	Saliency mask + Imagenet images	42.0	40.6
STC (PAMI2017)	Saliency mask + Web images	49.8	51.2
AE-PSL(CVPR2017)	Saliency mask	55.0	55.7
Hong et al. CVPR2017	Web videos	58.1	58.7
DSRG (CVPR2018)	Saliency mask	61.4	63.2
Shen et al. (CVPR2018)	Web images	63.0	63.9
SeeNet(NIPS2018)	Saliency mask	63.1	62.8
AISI(ECCV2018)	Instance saliency mask	63.6	64.5
SSDD(proposed)	-	**64.9**	**65.5**

The detailed results on PASCAL VOC 2012 val set.

Methods	BG	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	chair	Cow	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mIoU
PSA[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
SSDD	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	87.0	34.0	83.7	47.6	84.1	77.0	73.9	69.6	29.8	84.0	43.2	68.0	53.4	64.9
Gain	*+0.8*	*-5.7*	*-1.7*	*+2.6*	*+3.3*	*-1.5*	*-0.2*	*+7.6*	*+11.9*	*+5.0*	*+17.7*	*+7.4*	*+3.7*	*+15.0*	*+3.5*	*-4.1*	***-12.7***	*+13.3*	*+0.6*	*-0.1*	*+1.8*	*+3.2*

References and source codes

[1] Pixel-level semantic affinity, Ahn et al., CVPR 2018, arXiv:1803.10464
[2] STC: A Simple to Complex Framework, Wei et al., TPAMI 2016, arXiv:1509.03150

Source codes: https:github.com/shimoda-uec/ssdd