Self-supervised Difference Detection for Refinement CRF and Seed Interpolation

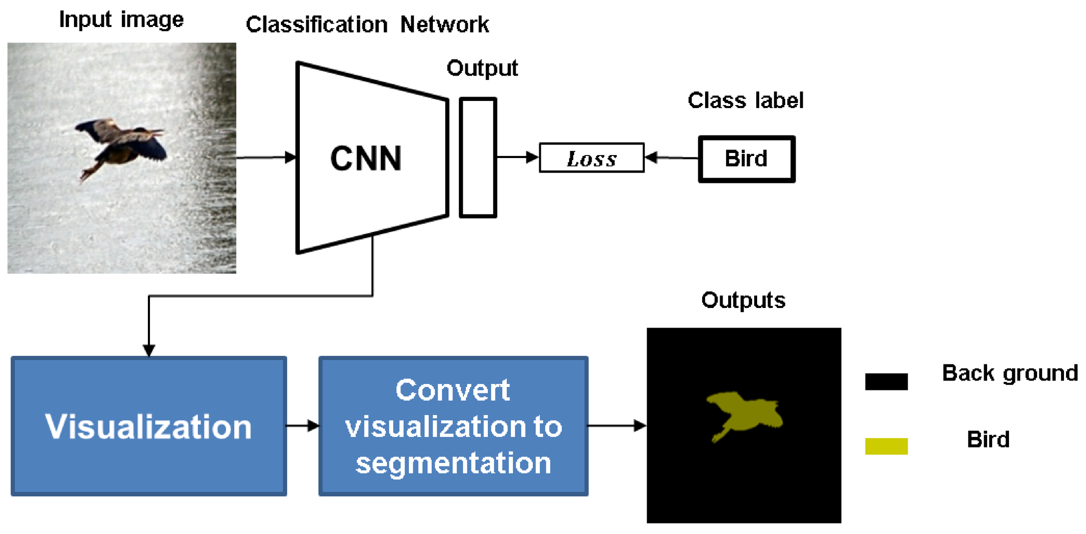
Wataru Shimoda   Keiji Yanai

The University of Electro-Communications, Tokyo, Japan

Objective

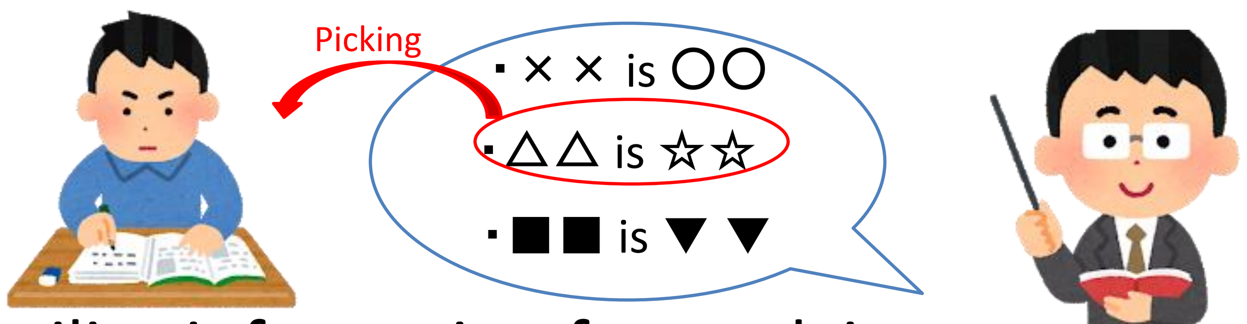
Weakly supervised segmentation

- Use only image-level annotation and generate segmentation masks



Motivation

- CRF is a good refinement method but it often degrades the results due to unstable unary terms in weakly supervised segmentation
- Our motivation is to use CRF results as not teacher but adviser
- In our situation, we suppose advisers give us noisy information



To utilize information from advisers

-We consider that many people first determine whether it is against their principles, and utilize opinions of other advisors for problems that are difficult to judge

-We model this scheme by difference detection task

Overview

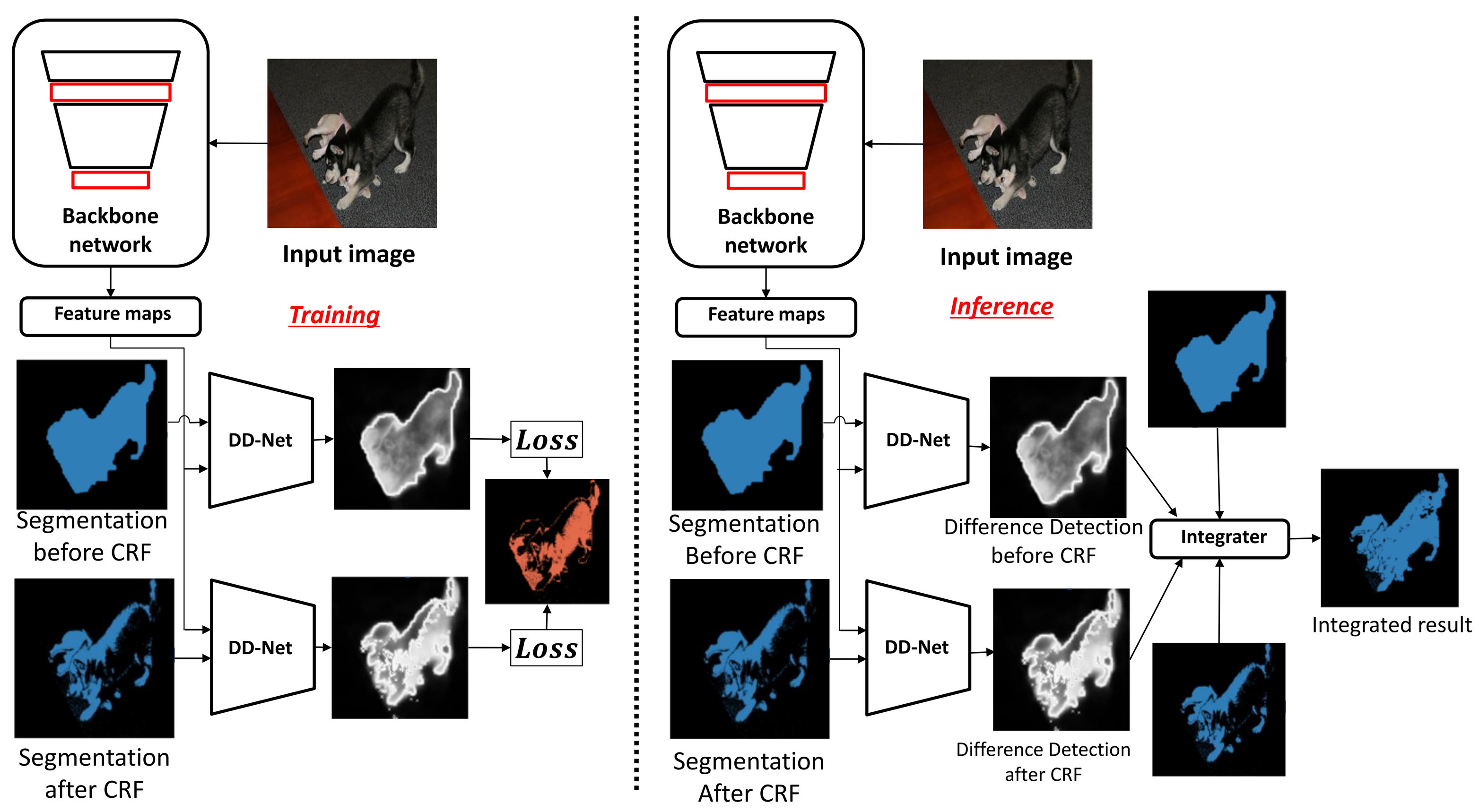
Overview of Difference Detection Network(DD-Net)

-Training

Train a difference detection model using difference regions between raw segmentation masks generated by PSA[1] and its CRF results

-Inference

Integrate a pair of mask using DD-net outputs



Difference detection network

Definition of difference detection task

- Input: Pair of mask

$$(m^{before}, m^{after})$$

- Output: Difference region

$$M^{before,after}$$

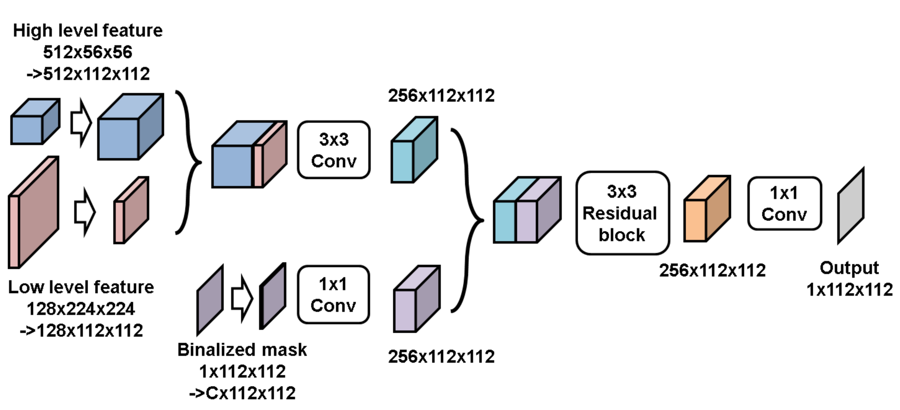
$$M_u^{before,after} \begin{cases} 1 & \text{if } (m_u^{before} = m_u^{after}) \\ 0 & \text{if } (m_u^{before} \neq m_u^{after}) \end{cases}$$

Definition of difference detection network (DD-net)

- Input: One of the pair of mask $(m^{before}$ or $m^{after})$ and feature maps $(e^h(x), e^l(x))$
- Output: Probability map $(d^{before}, d^{after}) \in \mathbb{R}^{H\times W}$

$$\mathrm{d}^{\mathrm{before}} = DDnet(e^h(x), e^l(x), m_u^{before})$$

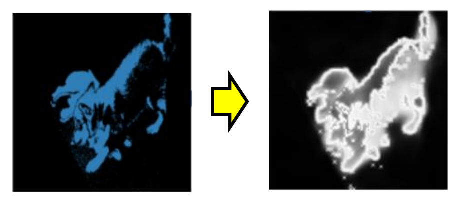
$$\mathrm{d}^{\mathrm{after}} = DDnet(e^h(x), e^l(x), m_u^{before})$$



Evaluation of input masks for integrations

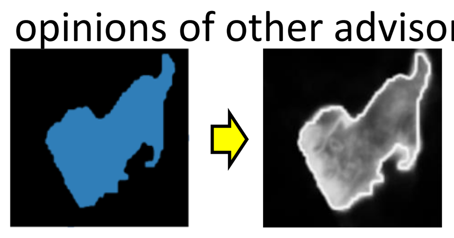
Modeling the decision process about whether to use the adviser's opinion

Decision by own principles



Make low confidence on pixels predicted by large differences between the input feature maps and segmentation masks

Decision taking into consideration the opinions of other advisors



Make low confidence on pixels predicted by learning similar training samples

-In both of decisions, high value (highlighted by white) indicates low confidence

Calculation of confidence scores

- Bias $\hat{b}$ is for making gap between decisions
- Bias b_m is for missing class labels in results

$$\mathrm{w}_{\mathrm{u}}^{\mathrm{before}} = \left(d_u^{after} + b_{m_u^{after}}\right) - \left(d_u^{before} + b_{m_u^{before}}\right) + \hat{b}$$

$\hat{b}$ = 0.4

b_{m_u} = 1.0 if (N_c^{after}/N_c^{before}>0.5, $\mathrm{m_u} = \mathrm{c}$)

N is number of pixels which belongs to category $c \in label$ in an image.

Mask integration based on the confidence scores

$$m_u^{refine} \begin{cases} \mathrm{m}_\mathrm{u}^{\mathrm{before}} & \text{if } (w_u^{before,after} \geq 0) \\ m_u^{after} & \text{if } (w_u^{before,after} < 0) \end{cases}$$

We denote this integration process as SSDD module by a below equation

$$\mathrm{m}^{refine} = SSDD\left(e(x;\theta_e), m^{before}, m^{after}; \theta_d\right)$$

Static region refinement

Loss for difference detection in PSA and CRF

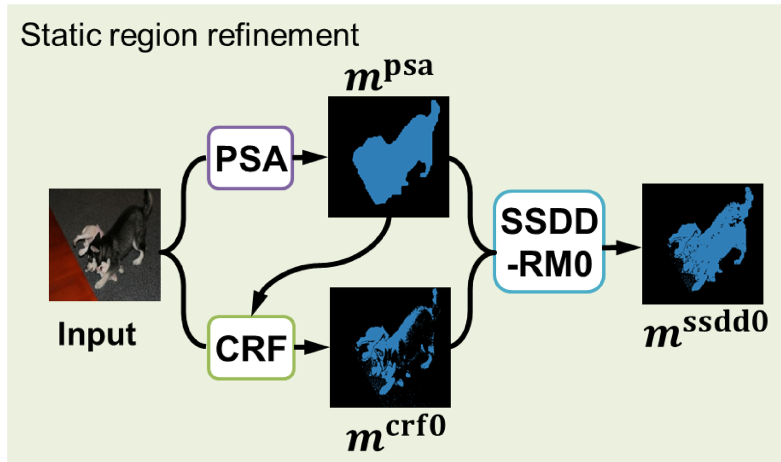
$$L_{change} = \frac{1}{|S|}\sum_{u\in S}\left(J\left(M^{psa,crf0}, d^{psa}, u\right) + J(M^{psa,crf0}, d^{crf0}, u)\right)$$

Loss for segmentation network for obtaining good representation from backbone network

$$L_{seg} = L_{mask}\left(m^{psa}; \theta_{seg}\right), \quad \mathrm{L}_{mask}(\mathrm{m}^{\mathrm{mask}}; \theta) = \frac{1}{\sum_{k\in C}|S^{mask}|}\sum_{k\in C_u}\sum_{u\in S_k^{mask}} \log(h_u^k(x;\theta))$$

Final loss

$$L_{static} = L_{seg} + L_{change}$$



Dynamic region refinement

Losses for difference detection

$$L_{dd-crf} = \frac{1}{|S|}\sum_{u\in S}\left(J\left(M^{seg,crf1}, d^{seg}, u\right) + J(M^{seg,crf1}, d^{crf1}, u)\right)$$

$$L_{dd-seed} = \frac{1}{|S|}\sum_{u\in S}\left(J\left(M^{ssdd0,sub}, d^{ssdd0}, u\right) + J(M^{ssdd1,sub}, d^{ssdd1}, u)\right)$$

Losses for segmentation network

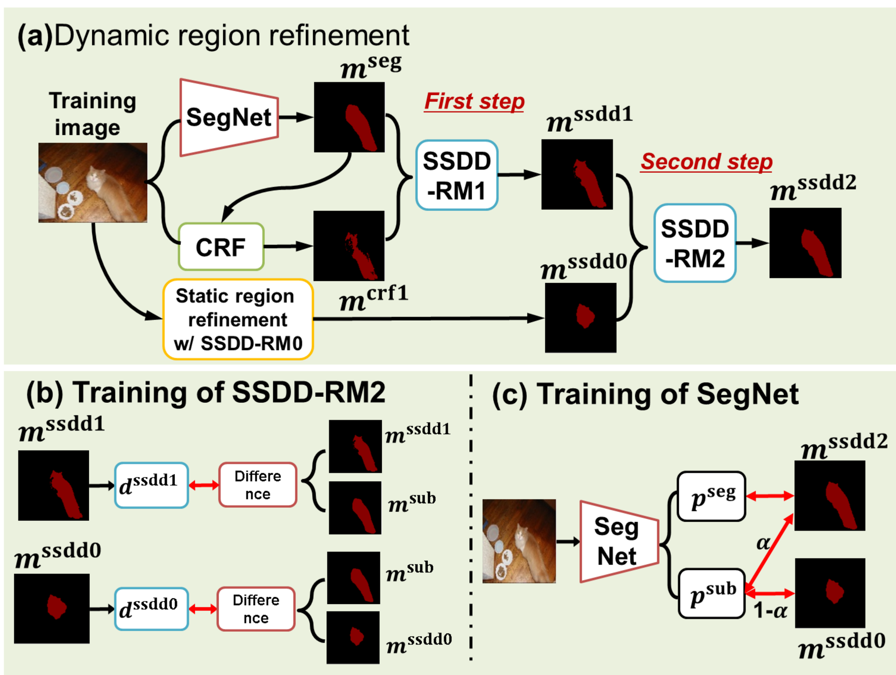
$$L_{seg-main} = L_{mask}\left(m^{ssdd2}; \theta_{s0}\right)$$

$$L_{seg-sub} = \alpha L^{ssdd0} + (1-\alpha)L^{ssdd2}$$

$$L^{ssdd0} = L_{mask}\left(m^{ssdd0}; \theta_{s1}\right), \quad L^{ssdd2} = L_{mask}\left(m^{ssdd2}; \theta_{s1}\right)$$

Final loss

$$L_{dynamic} = L_{dd-crf} + L_{dd-seed} + L_{seg-main} + L_{seg-sub}$$



Experiments

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

Comparison with WSS methods w/o additional supervision.

Method	Raw seed (Train set)	Trained seg model (Val set)
PSA (re-implementation)	52.5	58.4
PSA+CRF(re-implementation)	48.0	59.0
Static region refinement	53.4	61.4

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
Static region refinement	61.4	-
Dynamic region refinement	**64.9**	**65.5**

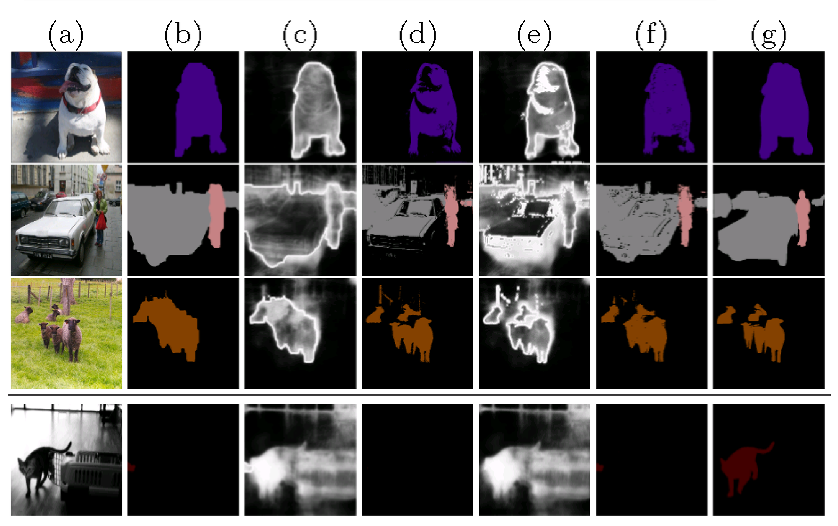
Results on PASCAL VOC 2012 val set.

Methods	BG	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	chair	Cow
PSA[1]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0
SSDD	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	87.0	34.0	83.7
Gain	*+0.8*	*-5.7*	*-1.7*	*+2.6*	*+3.3*	*-1.5*	*-0.2*	*+7.6*	*+11.9*	*+5.0*	*+17.7*

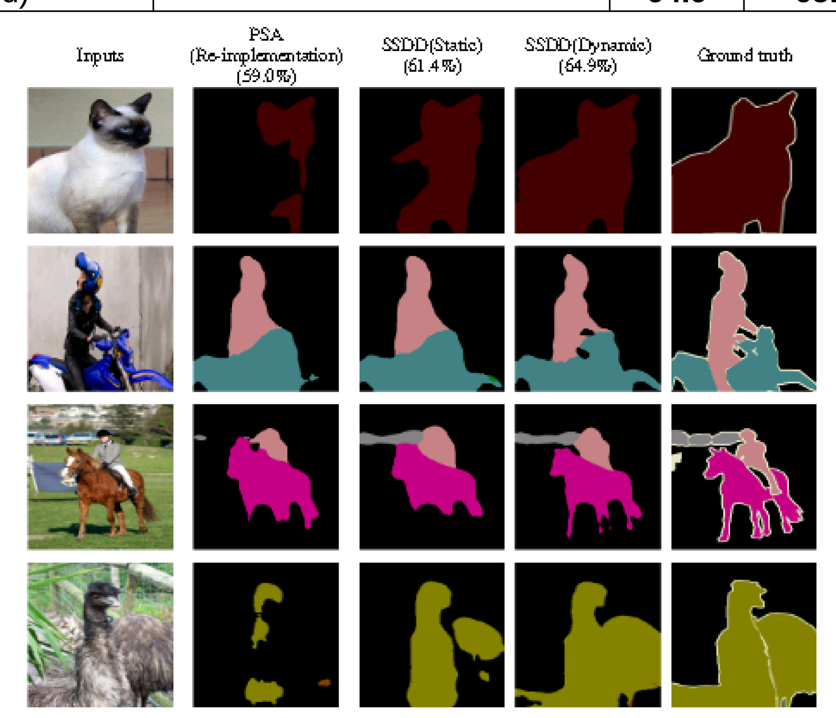
	Table	Dog	Horse	Motor	Person	Plant	Sheep	Sofa	Train	Tv	mIoU
PSA[1]	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
SSDD	47.6	84.1	77.0	73.9	69.6	29.8	84.0	43.2	68.0	53.4	64.9
Gain	*+7.4*	*+3.7*	*+15.0*	*+3.5*	*-4.1*	***-12.7***	*+13.3*	*+0.6*	*-0.1*	*+1.8*	*+3.2*

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**



For each row, from the left, (a) input images, (b) Raw PSA segmentation masks, (c) Difference detection maps of (b), (d) CRF masks of (b), (e) Difference detection maps of (d), (f) Refined segmentation masks by the proposed method and (g) Ground truth masks. Two bottom rows show failure cases.



References

[1] Jiwoon Ahn, Suha Kwak : **Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation**, CVPR 2018