Semi-supervised Semantic Segmentation with Directional Context-aware Consistency

Xin Lai[1*] Zhuotao Tian[1*] Li Jiang[1] Shu Liu[2]
Hengshuang Zhao[3] Liwei Wang[1] Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong [2]SmartMore [3]University of Oxford
{xinlai,zttian,lijiang,lwwang,leojia}@cse.cuhk.edu.hk sliu@smartmore.com hengshuang.zhao@eng.ox.ac.uk

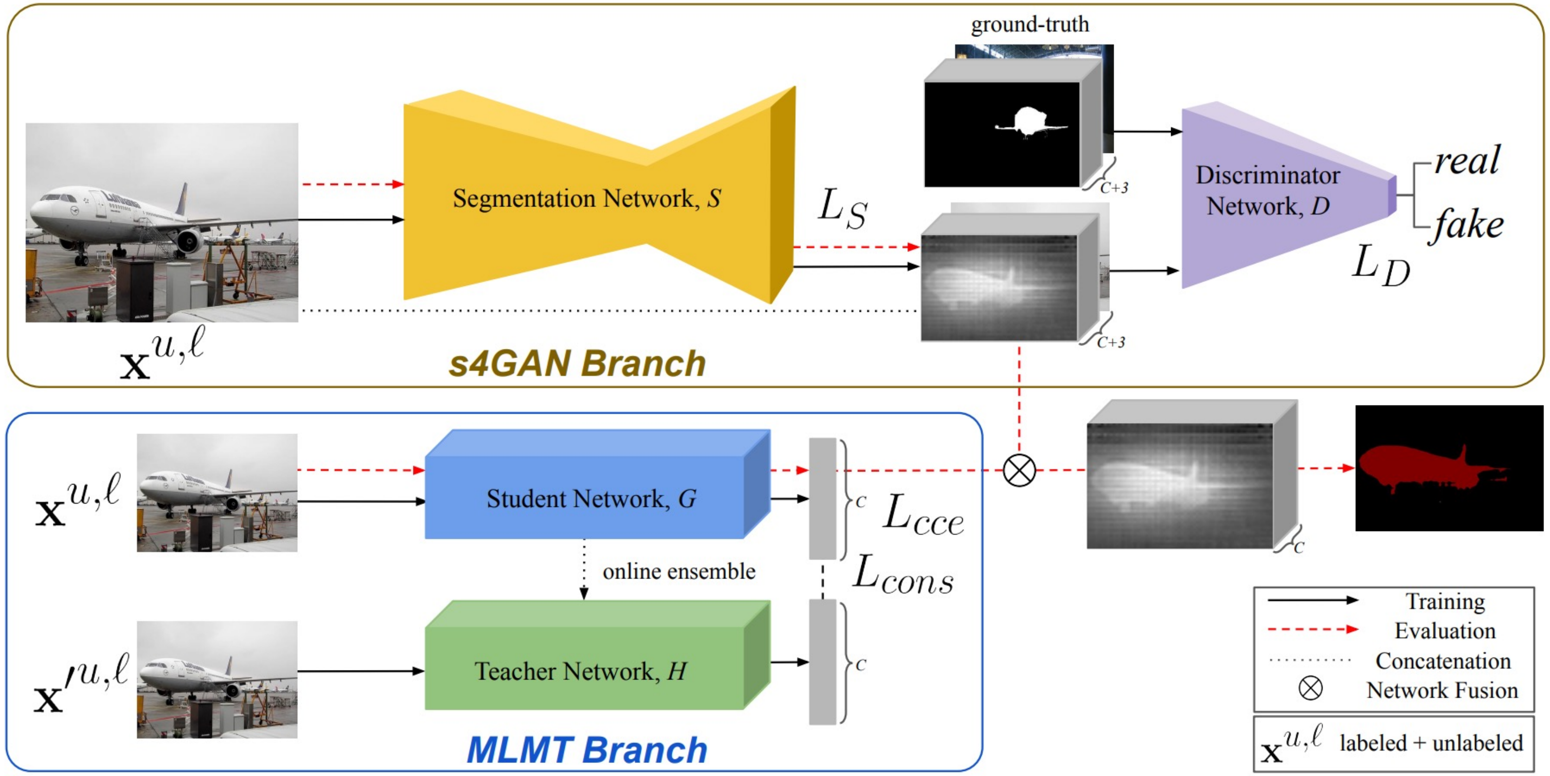
semi-supervised

Semi-supervised learning aims to exploit unlabel data to futher improve the representation learning given limited labeled data.

labeled data: pixel level annotation
unlabeled data: data without any annotation
weakly labeled data: bounding box, image-level labels, scribbles

Semi-supervised semantic segmentation

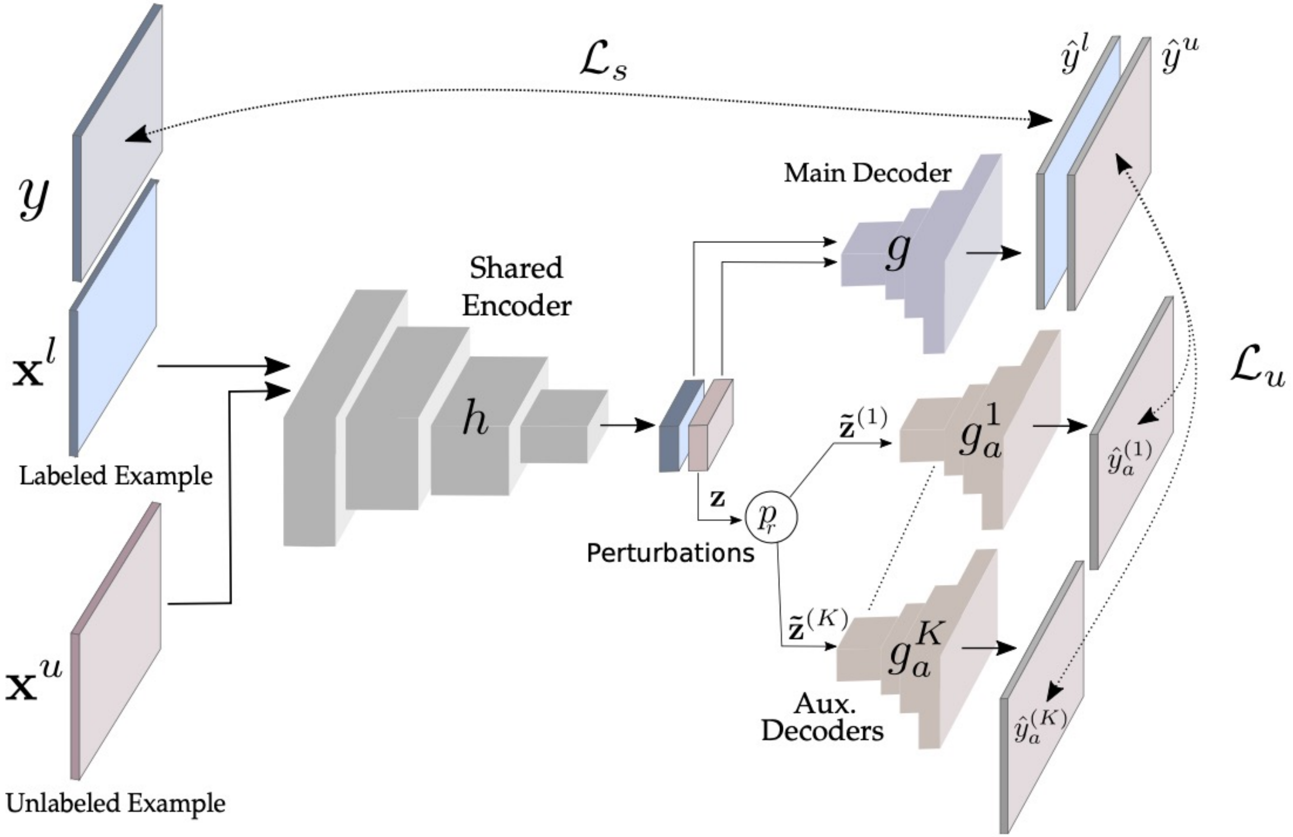
adversarial learning



Semi-supervised semantic segmentation with high- and low-level consistency. TPAMI, 2019.

Semi-supervised semantic segmentation

consistency training



Semi- supervised semantic segmentation with cross-consistency training. In CVPR, 2020 motivation

Prior **consistency-base** methods simply apply low-level data augmentations and constrain the perturbed ones to be consistent. However, model could not produce consistent embedding distribution under **different contexts**.

Consistency with **contextual augmentation** cloud be an additional constraint supplying low-level augmentations.

contribution

To alleviate the overfitting problem, we propose to maintain ***context-aware consistency*** between pixels under different environments.

To accomplish contextual alignment, we design the ***Directional Contrastive Loss***, what applies the constrastive learning in a pixel-wise manner. Also, two effective **sampling strategies** are proposed to further improve performance.

visualize

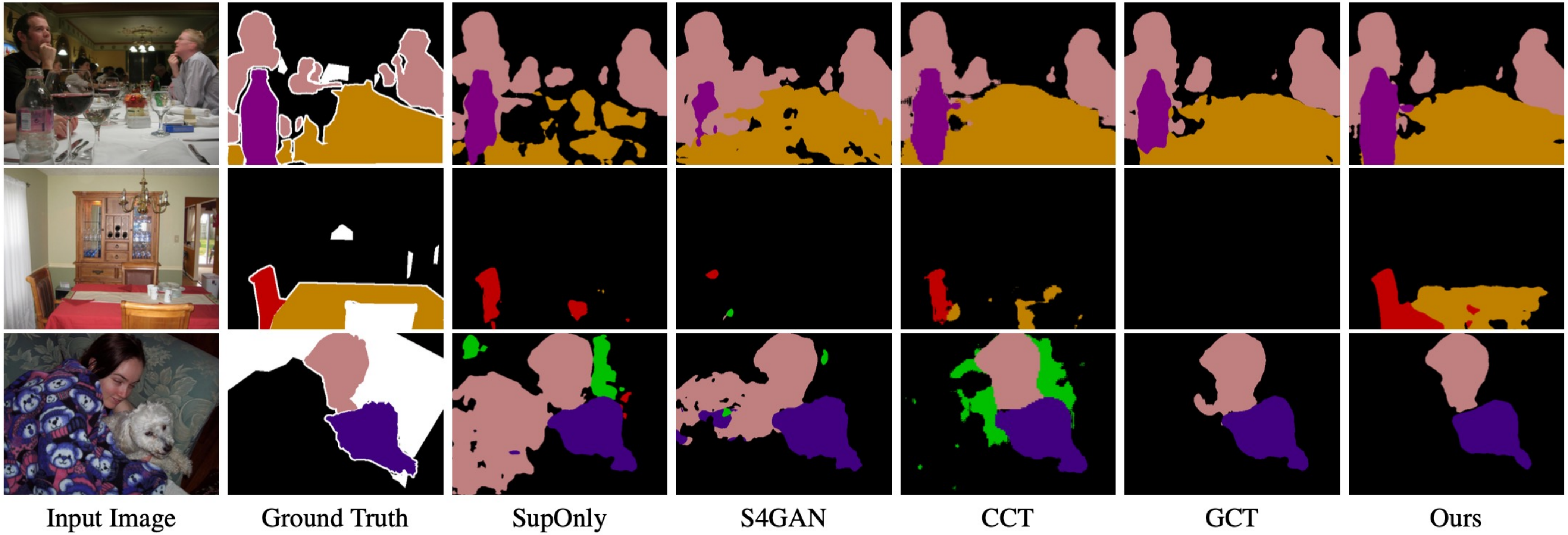


Figure 8. Visual comparison between SupOnly (*i.e.*, trained with only supervised loss) and current state-of-the-art methods with ours.

Overview

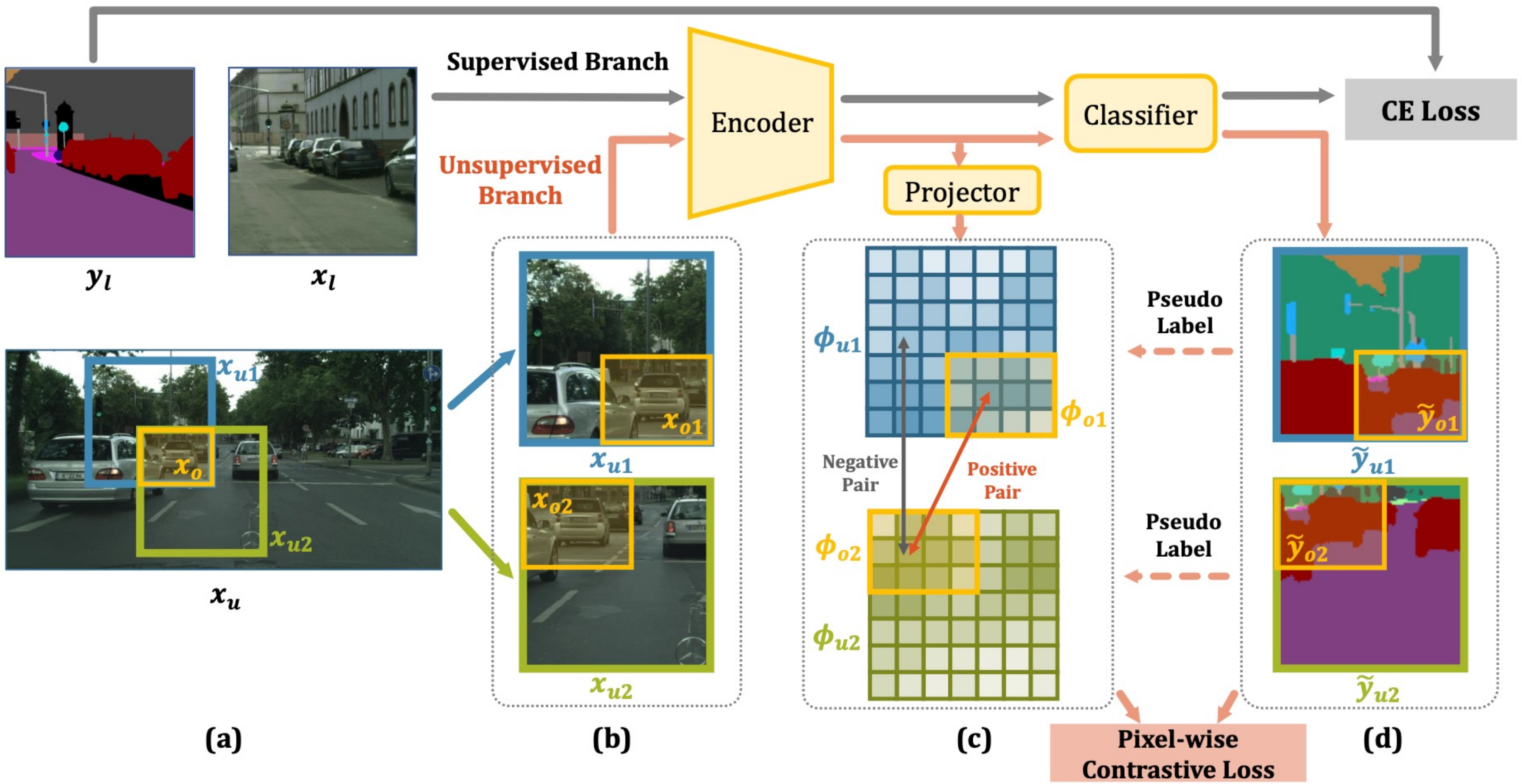
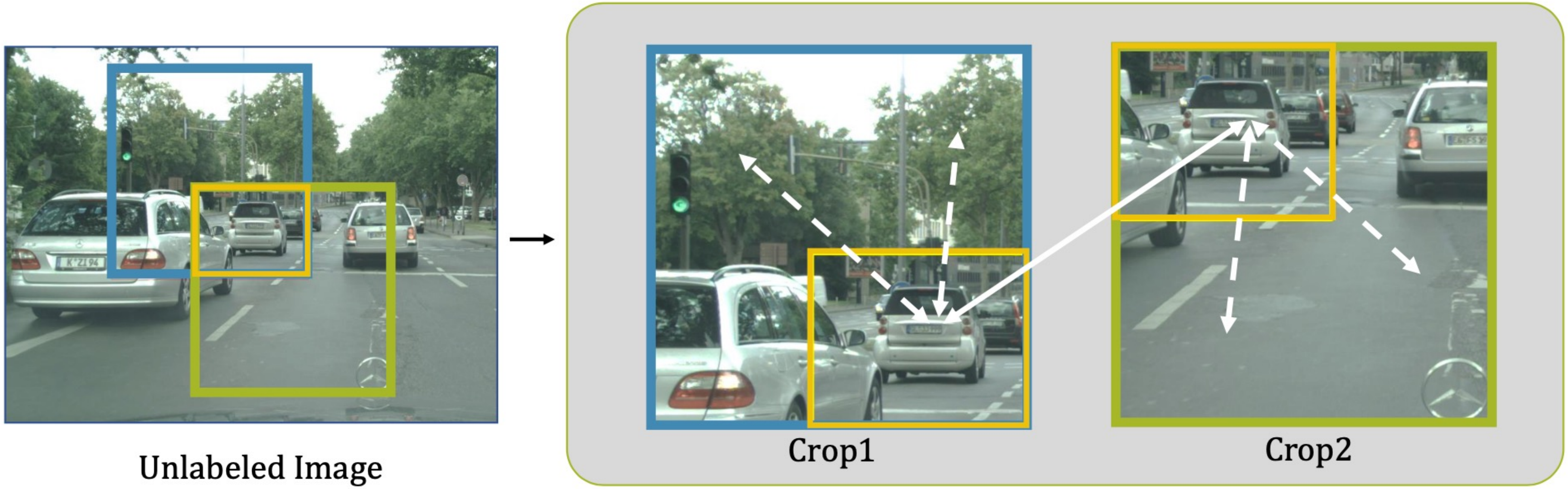


Figure 4. Overview of our framework. In the unsupervised branch, two patches are randomly cropped from the same image with a partially overlapping region. We aim to maintain a pixel-to-pixel consistency between the feature maps corresponding to the overlapping region.

context-aware consistency



Make the representations more robust to the changing environments.

Directional contrastive loss

base loss

$$l_{dc}^{b}(\phi_{o1}, \phi_{o2}) =$$

$$-\frac{1}{N}\sum_{h,w}\mathcal{M}_d^{h,w} \cdot \log \frac{r(\phi_{o1}^{h,w}, \phi_{o2}^{h,w})}{r(\phi_{o1}^{h,w}, \phi_{o2}^{h,w}) + \sum_{\phi_n \in \mathcal{F}_u} r(\phi_{o1}^{h,w}, \phi_n)} \qquad (1)$$

$$\mathcal{M}_d^{h,w} = \mathbf{1}\{\max \mathcal{C}(f_{o1}^{h,w}) < \max \mathcal{C}(f_{o2}^{h,w})\} \qquad (2)$$

$$\mathcal{L}_{dc}^{b} = l_{dc}^{b}(\phi_{o1}, \phi_{o2}) + l_{dc}^{b}(\phi_{o2}, \phi_{o1}) \qquad (3)$$

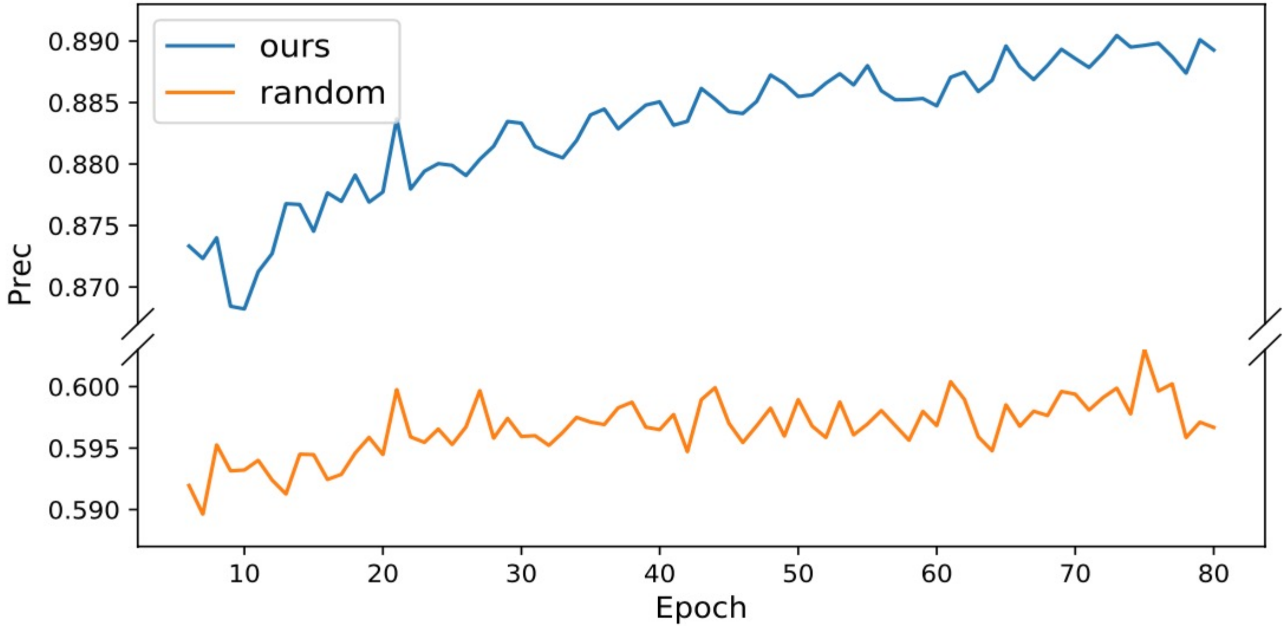
$$r(\phi_1, \phi_2) = \exp\left(s(\phi_1, \phi_2)/\tau\right)$$

$l_{dc}^{b}(\phi_{o1}, \phi_{o2})$ only back propogate to $\phi_{o1}^{h,w}$

negative sampling -- filter out false negative samples

$$l_{dc}^{b,ns}(\phi_{o1}, \phi_{o2}) = - \frac{1}{N} \sum_{h,w} \mathcal{M}_d^{h,w} \cdot \log \frac{r(\phi_{o1}^{h,w}, \phi_{o2}^{h,w})}{r(\phi_{o1}^{h,w}, \phi_{o2}^{h,w}) + \sum_{\phi_n \in \mathcal{F}_u} \mathcal{M}_{n,1}^{h,w} \cdot r(\phi_{o1}^{h,w}, \phi_n)} \tag{5}$$

$$\mathcal{M}_{n,1}^{h,w} = \mathbf{1}\{\tilde{y}_{o1}^{h,w} \neq \tilde{y}_n\}$$



positive filtering -- filter out low low confidence positive samples

$$l_{dc}^{b,ns,pf}(\phi_{o1},\phi_{o2}) = - \frac{1}{N}\sum_{h,w}\mathcal{M}_{d,pf}^{h,w}\cdot\log\frac{r(\phi_{o1}^{h,w},\phi_{o2}^{h,w})}{r(\phi_{o1}^{h,w},\phi_{o2}^{h,w})+\sum\limits_{\phi_n\in\mathcal{F}_u}\mathcal{M}_{n,1}^{h,w}\cdot r(\phi_{o1}^{h,w},\phi_n)} \tag{6}$$

$$\mathcal{M}_{d,pf}^{h,w} = \mathcal{M}_{d}^{h,w}\cdot\mathbf{1}\{\max\mathcal{C}(f_{o2}^{h,w})>\gamma\} \tag{7}$$

γ threshold to filter positive samples with low confidence , 0.75 in experiments total loss

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{dc}^{ns,pf}$$

$$\mathcal{L}_{dc}^{ns,pf} = \frac{1}{B} \sum_{b=1}^{B} (l_{dc}^{b,ns,pf}(\phi_{o1}, \phi_{o2}) + l_{dc}^{b,ns,pf}(\phi_{o2}, \phi_{o1}))$$

λ balance weigth for unsupervised loss, 30 in experiment

supervised only:

L = Lce

unspervised experiments

Method	SegNet	Backbone	1/16	1/8	1/4	Full
SupOnly	PSPNet	ResNet50	57.4	65.0	68.3	75.1
CCT [41]	PSPNet	ResNet50	62.2	68.8	71.2	75.3
Ours	PSPNet	ResNet50	**67.1**	**71.3**	**72.5**	**76.4**
SupOnly	DeepLabv3+	ResNet50	63.9	68.3	71.2	76.3
ECS [37]	DeepLabv3+	ResNet50	-	70.2	72.6	76.3
Ours	DeepLabv3+	ResNet50	**70.1**	**72.4**	**74.0**	**76.5**
SupOnly	DeepLabv3+	ResNet101	66.4	71.0	73.5	77.7
S4GAN [38]	DeepLabv3+	ResNet101	69.1	72.4	74.5	77.3
GCT [25]	DeepLabv3+	ResNet101	67.2	72.5	75.1	77.5
Ours	DeepLabv3+	ResNet101	**72.4**	**74.6**	**76.3**	**78.2**

pascal voc

Methods	1/8	1/4	Full
SupOnly	66.0	70.7	**77.7**
Ours	**69.7**	**72.7**	77.5

cityscapes

SupOnly: Only with supervised loss
ECS: Semi-supervised segmentation based on error-correcting supervision. In ECCV, 2020

ablation experiments

ID	Proj	Context	CL	Dir	NS	PF	mIoU
SupOnly							64.7
ST							66.3
I	✓	✓					64.2
II	✓	✓	✓				56.4
III	✓	✓	✓	✓			64.8
IV	✓	✓	✓	✓	✓		71.6
V	✓	✓	✓		✓	✓	71.2
VI	✓		✓	✓	✓	✓	70.5
VII		✓	✓	✓	✓	✓	61.5
VIII	✓	✓	✓	✓	✓	✓	**72.4**

Table 3. Ablation Study. Exp.I uses ℓ_2 loss to align positive feature pairs. **ST**: Self-Training. **Proj**: Non-linear Projector Φ. **Context**: Context-aware Consistency. **CL**: Vanilla Contrastive Loss. **Dir**: Directional Mask $\mathcal{M}_d^{h,w}$ defined in Eq. (2). **NS**: Negative Sampling. **PF**: Positive Filtering.

weakly experiment

Methods	Backbone	Semi	Weakly
WSSN [42]	VGG-16	-	64.6
GAIN [33]	VGG-16	-	60.5
MDC [56]	VGG-16	-	65.7
DSRG [22]	VGG-16	-	64.3
Souly *et al.* [49]	VGG-16	64.1	65.8
FickleNet [31]	ResNet-101	-	65.8
CCT [41]	ResNet-50	69.4	73.2
Ours	VGG-16	68.7	69.3
CCT‡	ResNet-50	72.8	74.6
Ours	ResNet-50	**74.5**	**76.1**

Table 5. Results with extra image-level annotations. CCT‡: Reproduced with the same setting as ours. Semi: Semi-supervised setting. Weakly: the setting with extra image-level labels.

experiment settings
dataset: Pascal Voc
1464 pixel level
9118 image level(from SBD)

implement:
extra classifier Cw for weakly data

loss:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{dc}^{ns,pf} + \lambda_w \mathcal{L}_w \tag{10}$$

$$\mathcal{L}_w = \frac{1}{2} \cdot (CE(\mathcal{C}_w(f_{u1}), y_p) + CE(\mathcal{C}_w(f_{u2}), y_p)) \tag{11}$$