Rfc 0.0. Jc_plgle dpmk Gkncpdcar B_r_ &JGB' Af_jjlec+ ATNP 0.0. Umpiqfmnq

Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation

Guolei Sun , Wenguan Wang , Luc Van Gool
ETH Zurich, Switzerland
{sunguolei.kaust, wenguanwang.ai}@gmail.com

Abstract

This paper studies the problem of learning weakly supervised semantic segmentation from image-level supervision only. Current popular solutions leverage object localization maps from classifiers as supervision for semantic segmentation learning, and struggle to make the localization maps capture more complete object content. Rather than previous efforts that primarily focus on intra-image information, we address the value of cross-image semantic relations for comprehensive object pattern mining. To achieve this, two neural co-attentions are incorporated into the classifier to complimentarily capture cross-image semantic similarities and differences. In particular, given a pair of training images, one co-attention enforces the classifier to recognize the common semantics from co-attentive objects, while the other one, called contrastive co-attention, drives the classifier to identify the unshared semantics from the rest, uncommon objects. This helps the classifier discover more object patterns and better ground semantics in image regions. By these careful designs, our approach ranked 1^{st} place in the Weakly-Supervised Semantic Segmentation Track of CVPR2020 Learning from Imperfect Data (LID) Challenge.

1. Introduction

This work focuses on weakly supervised semantic segmentation (WSSS) with only image-level labels. Current popular solutions are based on network visualization techniques [10], which discover discriminative regions that are activated for classification. They use image-level labels to train a classifier network, from which class-activation maps are derived as pseudo ground-truths for further supervising pixel-level semantics learning. However, it is commonly evidenced that the trained classifier over-addresses the most discriminative parts rather than entire objects, which becomes the focus of this area. Diverse solutions are explored, such as *image-level* operations [3], *regions growing* strategies [5], and *feature-level* enhancements [7].

However, as shown in Fig. 1(a), previous efforts typically use only single-image information for object pattern discovering, ignoring the rich semantic context among the weakly annotated data. For example, with the image-level labels, not only the semantics of each individual image can be identified, the cross-image semantic relations, *i.e.*, two images whether sharing certain semantics, are also given and should be used as cues for object pattern mining. Inspired by this, rather than relying on *intra-image* information only, we further address the value of *cross-image* semantic correlations for complete object pattern learning and effective class-activation map inference (Fig. 1(b)). In particular, our classifier is equipped with a differentiable co-attention mechanism that addresses semantic homogeneity and difference understanding across training *image pairs*. More specifically, two kinds of co-attentions are learned in the classifier. The former one aims to capture cross-image common semantics, which enables the classifier to better ground the common semantic labels over the co-attentive regions. The latter one, called contrastive co-attention, focuses on the rest, unshared semantics, which helps the classifier better separate semantic patterns of different objects. These

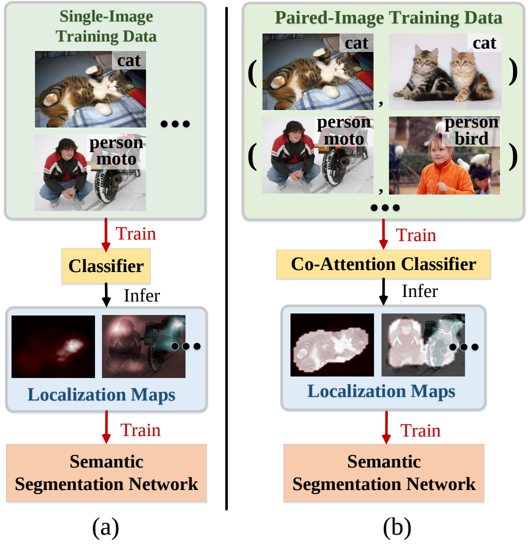
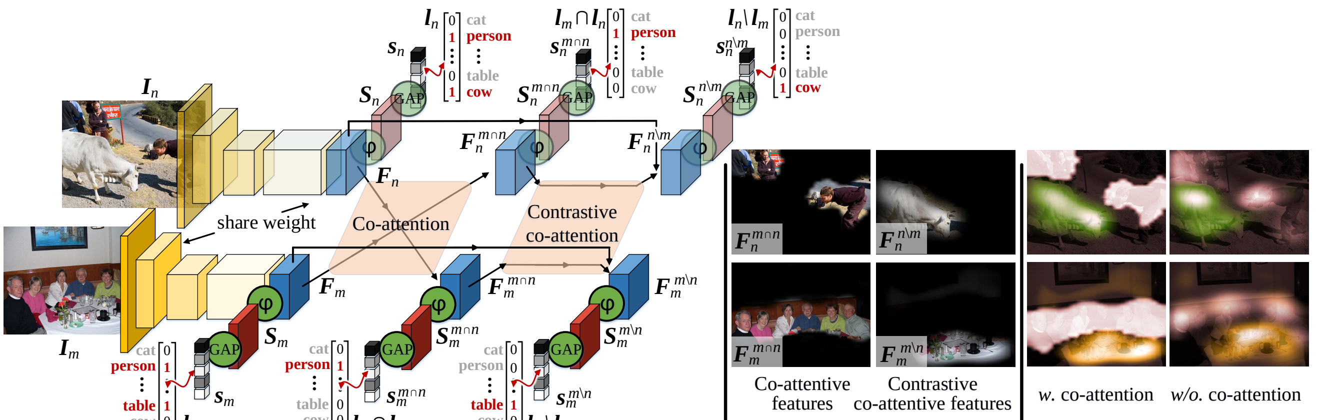


Figure 1: (a) Current WSSS methods only use single-image information for object pattern discovering. (b) Our co-attention classifier leverages cross-image semantics as class-level context to benefit object pattern learning and localization map inference.


$\boldsymbol{I}_n$
$\boldsymbol{I}_m$
share weight
Co-attention
Contrastive co-attention
$\boldsymbol{F}_n$
$\boldsymbol{F}_m$
$\boldsymbol{F}_n^{m\cap n}$
$\boldsymbol{F}_n^{n\setminus m}$
$\boldsymbol{F}_m^{m\cap n}$
$\boldsymbol{F}_m^{m\setminus n}$
φ
GAP
$\boldsymbol{S}_n$
$\boldsymbol{S}_n^{m\cap n}$
$\boldsymbol{S}_n^{n\setminus m}$
$\boldsymbol{S}_m$
$\boldsymbol{S}_m^{m\cap n}$
$\boldsymbol{S}_m^{m\setminus n}$
$\boldsymbol{s}_n$
$\boldsymbol{s}_n^{m\cap n}$
$\boldsymbol{s}_n^{n\setminus m}$
$\boldsymbol{s}_m$
$\boldsymbol{s}_m^{m\cap n}$
$\boldsymbol{s}_m^{m\setminus n}$
$\boldsymbol{l}_n$
$\boldsymbol{l}_m\cap\boldsymbol{l}_n$
$\boldsymbol{l}_n\setminus\boldsymbol{l}_m$
$\boldsymbol{l}_m$
$\boldsymbol{l}_m\setminus\boldsymbol{l}_n$
cat
person
table
cow
0
1
$\boldsymbol{F}_n^{m\cap n}$
$\boldsymbol{F}_n^{n\setminus m}$
$\boldsymbol{F}_m^{m\cap n}$
$\boldsymbol{F}_m^{m\setminus n}$
Co-attentive features
Contrastive co-attentive features
w. co-attention
w/o. co-attention

(a) Overview of our co-attention classifier during training phase

(b)Visualization of (contrastive) co-attentive features

(c) Object localization maps *w.* and *w/o.* co-attention

Figure 2: **(a)** In addition to mining object semantics from single-image labels, semantic similarities and differences between paired training images are both leveraged for supervising object pattern learning. **(b)** Co-attentive and contrastive co-attentive features complimentarily capture the shared and unshared objects. **(c)** Our co-attention classifier is able to learn object patterns more comprehensively.

two co-attentions work in a cooperative and complimentary manner, together making the classifier understand object patterns more comprehensively. Another advantage is that our co-attention based classifier learning paradigm brings an efficient data augmentation strategy, due to the use of training image pairs. With above efforts, our method ranked 1^{st} place in the Weakly-supervised Semantic Segmentation Track of CVPR2020 Learning from Imperfect Data (LID) Challenge [8] (LID$_{20}$), outperforming other competitors by large margins.

2. Our Algorithm

2.1. Co-attention Classification Network

Let us denote the training data as $\mathcal{I}=\{(\boldsymbol{I}_n,\boldsymbol{l}_n)\}_n$, where $\boldsymbol{I}_n$ is the n^{th} training image, and $\boldsymbol{l}_n\in\{0,1\}^K$ is the associated *ground-truth* image label for K semantic categories. As shown in Fig. 2(a), image pairs, *i.e.*, $(\boldsymbol{I}_m,\boldsymbol{I}_n)$, are sampled from $\mathcal{I}$ for training the classifier. After feeding $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ into the conv part of the classifier, corresponding feature maps, $\boldsymbol{F}_m\in\mathbb{R}^{C\times H\times W}$ and $\boldsymbol{F}_n\in\mathbb{R}^{C\times H\times W}$, are obtained, each with $H\times W$ spatial dimension and C channels. Then we can first separately pass $\boldsymbol{F}_m$ and $\boldsymbol{F}_n$ to a *class-aware fully convolutional layer* $\varphi(\cdot)$ to generate *class-aware activation maps*, *i.e.*, $\boldsymbol{S}_m=\varphi(\boldsymbol{F}_m)\in\mathbb{R}^{K\times H\times W}$ and $\boldsymbol{S}_n=\varphi(\boldsymbol{F}_n)\in\mathbb{R}^{K\times H\times W}$, respectively. Then, we apply *global average pooling* (GAP) over $\boldsymbol{S}_m$ and $\boldsymbol{S}_n$ to obtain class score vectors $\boldsymbol{s}_m\in\mathbb{R}^K$ and $\boldsymbol{s}_n\in\mathbb{R}^K$ for $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, respectively. Finally, the *sigmoid cross entropy* (CE) loss is used for supervision:

$$\begin{aligned}\mathcal{L}^{mn}_{\text{basic}}\big((\boldsymbol{I}_m,\boldsymbol{I}_n),(\boldsymbol{l}_m,\boldsymbol{l}_n)\big) &= \mathcal{L}_{\text{CE}}(\boldsymbol{s}_m,\boldsymbol{l}_m)+\mathcal{L}_{\text{CE}}(\boldsymbol{s}_n,\boldsymbol{l}_n),\\ &= \mathcal{L}_{\text{CE}}\big(\text{GAP}(\varphi(\boldsymbol{F}_m)),\boldsymbol{l}_m\big)+\\ &\quad\ \mathcal{L}_{\text{CE}}\big(\text{GAP}(\varphi(\boldsymbol{F}_n)),\boldsymbol{l}_n\big).\end{aligned} \tag{1}$$

Next we will endow the classifier with a co-attention mechanism for further mining cross-image semantics and eventually better localizing objects.

Co-Attention for Cross-Image Common Semantics Mining. Our co-attention attends to the two images, *i.e.*, $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, simultaneously, and captures their correlations. We first compute the affinity matrix $\boldsymbol{P}$ between $\boldsymbol{F}_m$ and $\boldsymbol{F}_n$ [4]:

$$\boldsymbol{P}=\boldsymbol{F}_m^{\top}\boldsymbol{W}_{\boldsymbol{P}}\boldsymbol{F}_n\in\mathbb{R}^{HW\times HW}, \tag{2}$$

where $\boldsymbol{W}_{\boldsymbol{P}}\in\mathbb{R}^{C\times C}$ is a learnable matrix. $\boldsymbol{P}$ stores similarity scores corresponding to all pairs of positions in $\boldsymbol{F}_m$ and $\boldsymbol{F}_n$. Then, $\boldsymbol{P}$ is normalized column-wise to derive attention maps across $\boldsymbol{F}_m$ for each position in $\boldsymbol{F}_n$, and row-wise to derive attention maps across $\boldsymbol{F}_n$ for each position in $\boldsymbol{F}_m$:

$$\begin{aligned}\boldsymbol{A}_m&=\text{softmax}(\boldsymbol{P})\ \ \in[0,1]^{HW\times HW},\\ \boldsymbol{A}_n&=\text{softmax}(\boldsymbol{P}^{\top})\in[0,1]^{HW\times HW},\end{aligned} \tag{3}$$

where softmax is performed column-wise. In this way, $\boldsymbol{A}_n$ and $\boldsymbol{A}_m$ store the co-attention maps in their columns. Next, we can compute attention summaries of $\boldsymbol{F}_m$ ($\boldsymbol{F}_n$) in light of each position of $\boldsymbol{F}_n$ ($\boldsymbol{F}_m$):

$$\boldsymbol{F}_m^{m\cap n}=\boldsymbol{F}_n\boldsymbol{A}_n\in\mathbb{R}^{C\times H\times W},\quad \boldsymbol{F}_n^{m\cap n}=\boldsymbol{F}_m\boldsymbol{A}_m\in\mathbb{R}^{C\times H\times W}. \tag{4}$$

Co-attentive feature $\boldsymbol{F}_m^{m\cap n}$, derived from $\boldsymbol{F}_n$, preserves the common semantics between $\boldsymbol{F}_m$ and $\boldsymbol{F}_n$ and locate the common objects in $\boldsymbol{F}_m$. Thus we can expect only the common semantics $\boldsymbol{l}_m\cap\boldsymbol{l}_n$[1] can be safely derived from $\boldsymbol{F}_m^{m\cap n}$, and the same goes for $\boldsymbol{F}_n^{m\cap n}$. Such co-attention based common semantic classification can let the classifier understand the object patterns more completely and precisely.

To make things intuitive, consider the example in Fig. 2, where $\boldsymbol{I}_m$ contains **Table** and **Person**, and $\boldsymbol{I}_n$ has **Cow** and **Person**. As the co-attention is essentially the affinity computation between all the position pairs between $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, only the semantics of the common objects, **Person**, will be preserved in the co-attentive features, *i.e.*, $\boldsymbol{F}_m^{m\cap n}$ and $\boldsymbol{F}_n^{m\cap n}$ (Fig.2(b)). If we feed $\boldsymbol{F}_m^{m\cap n}$ and $\boldsymbol{F}_n^{m\cap n}$ into the class-aware

[1] The set operation '∩' is extended here to represent bitwise-and.

fully convolutional layer φ, the generated class-aware activation maps, *i.e.*, $\boldsymbol{S}_m^{m\cap n}=\varphi(\boldsymbol{F}_m^{m\cap n})$ and $\boldsymbol{S}_n^{m\cap n}=\varphi(\boldsymbol{F}_n^{m\cap n})$, are able to locate the common object **Person** in $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, respectively. After GAP, the predicted semantic classes $\boldsymbol{s}_m^{m\cap n}$ and $\boldsymbol{s}_n^{m\cap n}$ should be the common semantic labels $\boldsymbol{l}_m\cap\boldsymbol{l}_n$, *i.e.*, **Person**. Through co-attention computation, not only the human face, the most discriminative part of **Person**, but also other parts, such as legs and arms, are highlighted in $\boldsymbol{F}_m^{m\cap n}$ and $\boldsymbol{F}_n^{m\cap n}$ (Fig. 2(b)). When we set the common class labels, *i.e.*, **Person**, as the supervision signal, the classifier would realize that the semantics preserved in $\boldsymbol{F}_m^{m\cap n}$ and $\boldsymbol{F}_n^{m\cap n}$ are related and can be used to recognize **Person**. Thus, the co-attention, computed across two related images, *explicitly* helps the classifier associate semantic labels and corresponding object regions and better understand the relations between different object parts. It makes full use of the context across training data.

For co-attention based common semantic classification, the common labels $\boldsymbol{l}_m\cap\boldsymbol{l}_n$ are used to supervise learning:

$$\begin{aligned}\mathcal{L}^{mn}_{\text{co-att}}\big((\boldsymbol{I}_m,\boldsymbol{I}_n),(\boldsymbol{l}_m,\boldsymbol{l}_n)\big)=&\mathcal{L}_{\text{CE}}(\boldsymbol{s}_m^{m\cap n},\boldsymbol{l}_m\cap\boldsymbol{l}_n)+\\ &\mathcal{L}_{\text{CE}}(\boldsymbol{s}_n^{m\cap n},\boldsymbol{l}_m\cap\boldsymbol{l}_n),\\ =&\mathcal{L}_{\text{CE}}\big(\text{GAP}(\varphi(\boldsymbol{F}_m^{m\cap n})),\boldsymbol{l}_m\cap\boldsymbol{l}_n\big)+\\ &\mathcal{L}_{\text{CE}}\big(\text{GAP}(\varphi(\boldsymbol{F}_n^{m\cap n})),\boldsymbol{l}_m\cap\boldsymbol{l}_n\big).\end{aligned}\tag{5}$$

Contrastive Co-Attention for Cross-Image Exclusive Semantics Mining. Aside from the co-attention described above that explores cross-image common semantics, we propose a contrastive co-attention that mines semantic differences between paired images. The co-attention and contrastive co-attention complementarily help the classifier better understand the concept of the objects.

As shown in Fig. 2(a), for $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, we first derive *class-agnostic co-attentions* from their co-attentive features, *i.e.*, $\boldsymbol{F}_m^{m\cap n}$ and $\boldsymbol{F}_n^{m\cap n}$, respectively:

$$\begin{aligned}\boldsymbol{B}_m^{m\cap n}&=\sigma(\boldsymbol{W_B}\boldsymbol{F}_m^{m\cap n})\in[0,1]^{H\times W},\\ \boldsymbol{B}_n^{m\cap n}&=\sigma(\boldsymbol{W_B}\boldsymbol{F}_n^{m\cap n})\in[0,1]^{H\times W},\end{aligned}\tag{6}$$

where $\sigma(\cdot)$ is the *sigmoid* activation function, and the parameter matrix $\boldsymbol{W_B}\in\mathbb{R}^{1\times C}$ learns for common semantics collection and is implemented by a conv layer with 1×1 kernel. $\boldsymbol{B}_m^{m\cap n}$ and $\boldsymbol{B}_n^{m\cap n}$ are class-agnostic and highlight all the common object regions in $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, respectively, based on which we derive contrastive co-attentions:

$$\begin{aligned}\boldsymbol{A}_m^{m\setminus n}&=\boldsymbol{1}-\boldsymbol{B}_m^{m\cap n}\in[0,1]^{H\times W},\\ \boldsymbol{A}_n^{n\setminus m}&=\boldsymbol{1}-\boldsymbol{B}_n^{m\cap n}\in[0,1]^{H\times W}.\end{aligned}\tag{7}$$

The contrastive co-attention $\boldsymbol{A}_m^{m\setminus n}$ of $\boldsymbol{I}_m$ addresses those *unshared* object regions that are only of $\boldsymbol{I}_m$, but not of $\boldsymbol{I}_n$, and the same goes for $\boldsymbol{A}_n^{n\setminus m}$. Then we get *contrastive co-attentive features*, *i.e.*, unshared semantics in each images:

$$\begin{aligned}\boldsymbol{F}_m^{m\setminus n}&=\boldsymbol{F}_m\otimes\boldsymbol{A}_m^{m\setminus n}\in\mathbb{R}^{C\times H\times W},\\ \boldsymbol{F}_n^{n\setminus m}&=\boldsymbol{F}_n\otimes\boldsymbol{A}_n^{n\setminus m}\in\mathbb{R}^{C\times H\times W}.\end{aligned}\tag{8}$$

'$\otimes$' denotes element-wise multiplication, where the attention values are copied along the channel dimension. Next, we can sequentially get class-aware activation maps, *i.e.*, $\boldsymbol{S}_m^{m\setminus n}=\varphi(\boldsymbol{F}_m^{m\setminus n})$ and $\boldsymbol{S}_n^{n\setminus m}=\varphi(\boldsymbol{F}_n^{n\setminus m})$, and semantic scores, *i.e.*, $\boldsymbol{s}_m^{m\setminus n}=\text{GAP}(\boldsymbol{S}_m^{m\setminus n})$ and $\boldsymbol{s}_n^{n\setminus m}=\text{GAP}(\boldsymbol{S}_n^{n\setminus m})$. For $\boldsymbol{s}_m^{m\setminus n}$ and $\boldsymbol{s}_n^{n\setminus m}$, they are expected to identify the categories of the unshared objects, *i.e.*, $\boldsymbol{l}_m\setminus\boldsymbol{l}_n$ and $\boldsymbol{l}_n\setminus\boldsymbol{l}_m$[2].

Compared with the co-attention that investigates common semantics as informative cues for boosting object patterns mining, the contrastive co-attention addresses complementary knowledge from the semantic differences between paired images. Fig. 2(b) gives an intuitive example. After computing the contrastive co-attentions between $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ (Eq. 7), **Table** and **Cow**, which are unique in their original images, are highlighted. Based on the contrastive co-attentive features, *i.e.*, $\boldsymbol{F}_m^{m\setminus n}$ and $\boldsymbol{F}_n^{n\setminus m}$, the classifier is required to accurately recognize **Table** and **Cow** classes, respectively. When the common objects are filtered out by the contrastive co-attentions, the classifier has a chance to focus more on the rest image regions and mine the unshared semantics more consciously. This also helps the classifier better discriminate the semantics of different objects, as the semantics of common objects and unshared ones are disentangled by the contrastive co-attention. For example, if some parts of **Cow** are wrongly recognized as **Person**-related, the contrastive co-attention will discard these parts in $\boldsymbol{F}_n^{n\setminus m}$. However, the rest semantics in $\boldsymbol{F}_n^{n\setminus m}$ may be not sufficient enough for recognizing **Cow**. This will enforce the classifier to better discriminate different objects.

For the contrastive co-attention based unshared semantic classification, the supervision loss is designed as:

$$\begin{aligned}\mathcal{L}^{mn}_{\overline{\text{co-att}}}\big((\boldsymbol{I}_m,\boldsymbol{I}_n),(\boldsymbol{l}_m,\boldsymbol{l}_n)\big)=&\mathcal{L}_{\text{CE}}(\boldsymbol{s}_m^{m\setminus n},\boldsymbol{l}_m\setminus\boldsymbol{l}_n)+\\ &\mathcal{L}_{\text{CE}}(\boldsymbol{s}_n^{n\setminus m},\boldsymbol{l}_n\setminus\boldsymbol{l}_m),\\ =&\mathcal{L}_{\text{CE}}\big(\text{GAP}\big(\varphi(\boldsymbol{F}_m^{m\setminus n})\big),\boldsymbol{l}_m\setminus\boldsymbol{l}_n\big)+\\ &\mathcal{L}_{\text{CE}}\big(\text{GAP}\big(\varphi(\boldsymbol{F}_n^{n\setminus m}\big),\boldsymbol{l}_n\setminus\boldsymbol{l}_m\big).\end{aligned}\tag{9}$$

2.2. Co-Attention Guided WSSS Learning

Training Co-Attention Classifier. The overall training loss for our co-attention classifier ensembles the three terms defined in Eqs. 1, 5, and 9:

$$\mathcal{L}=\sum\nolimits_{m,n}\mathcal{L}^{mn}_{\text{basic}}+\mathcal{L}^{mn}_{\text{co-att}}+\mathcal{L}^{mn}_{\overline{\text{co-att}}}.\tag{10}$$

During training, to fully leverage the co-attention to mine the common semantics, we sample two images $(\boldsymbol{I}_m,\boldsymbol{I}_n)$ with at least one common class, *i.e.*, $\boldsymbol{l}_m\cap\boldsymbol{l}_n\neq\boldsymbol{0}$. Our image classifier is based on ResNet-38 [9], pretrained on ImageNet. The training parameters are set as: initial learning rate (0.005) and the poly policy based training schedule:

[2] The set operation '$\setminus$' is slightly extend here, *i.e.*, $\boldsymbol{l}_n\setminus\boldsymbol{l}_m=\boldsymbol{l}_n-\boldsymbol{l}_n\cap\boldsymbol{l}_m$.

Table 1: Ablation study with mIoU metric (%).

Techniques	Training Images (#)	Val
Random sample	20K	31
Balance sample	20K	33
Balance sample + WCE loss	20K	36
Balance sample + WCE loss + label refinement	20K	38
Balance sample + WCE loss + label refinement	Full (300K+)	46

$lr = lr_{init} \times (1 - \frac{iter}{max\_iter})^{\gamma}$ with γ(0.9), batch size (8), weight decay (0.0005), and max epoch (15). During training, the equivariant attention [6] is also adopted. Our classifier is trained on 2 NVIDIA Tesla V100 GPUs.

Generating Object Localization Maps. Once our image classifier is trained, for each training image $\boldsymbol{I}_n \in \mathcal{I}$, we run the classifier and directly use its class-aware activation map (*i.e.*, $\boldsymbol{S}_n$) as the object localization map $\boldsymbol{L}_n$. We also use integral attention learning [2] to refine localization maps.

Learning Semantic Segmentation Network. After obtaining high-quality localization maps, we generate pseudo pixel-wise labels for all the training samples $\mathcal{I}$. Specifically, we follow [1]: localization maps are first used to train an AffinityNet model, which is then used to generate pseudo ground truth masks and background threshold is set as 0.2. Note that no saliency maps are used. For the semantic segmentation network, we choose ResNet-101 based DeepLab-V3. The parameters are set as below: initial learning rate (0.007) with poly schedule, batch size (48), max epoch (100), and weight decay (0.0001). The segmentation model is trained on 4 Tesla V100 GPUs. During testing, results from multi scales are averaged, with CRF refinement.

3. Experiment

Dataset: The dataset of LID$_{20}$ WSSS track [8] is built upon ImageNet. It contains 349,319 images with image-level labels from 200 classes. Evaluations are conducted on the *val* and *test* sets, which have 4,690 and 10,000 images, respectively. In this challenge, the standard mean intersection over union (mIoU) criterion is used to rank competitors.

Ablation Study. We found three challenges in the dataset: **1)** Huge data imbalance between different classes. For three most common classes: *bird*, *dog*, and *person*, number of images is more than 20,000 while most other classes only have ∼1,000 images. **2)** Imbalance between negative and positive samples in such a multi-label classification setting, due to the sparsity of label matrix and large number of classes. Sigmoid cross entropy loss, which is used most common, does not work well. **3)** Noisy labels (especially for *person*). We solved those problems by developing following techniques: **1)** sampling images for each class in a balanced way; **2)** using the weighted sigmoid cross entropy (WCE) loss; and **3)** first training a strong classifier and then using its output to refine labels. To study the efficacy of our above techniques, we retain our model on 20K randomly sampled training images, by gradually adding techniques. The performance is reported on the *val* set, as shown in Table 1.

Table 2: Results on *val* and *test* sets of LID$_{20}$ WSSS track.

Team	Val	Test
play-njupt	22.07	31.90
IOnlyHaveSevenDays	39.00	36.24
UCU & SoftServe	39.65	37.34
VL-task1	40.08	37.73
CVL (**ours**)	**46.29**	**45.18**

Main Results. The final results on LID$_{20}$ WSSS track is shown in Table 2. As can be seen, our approach largely outperforms other methods on both val and test sets.

4. Conclusion

We propose a co-attention classification network to discover integral object regions by addressing cross-image semantics. Our method ranked 1^{st} place in the weakly-supervised semantic segmentation track of LID$_{20}$ challenge.

Acknowledgements This work was sponsored by Zhejiang Lab's Open Fund (No. 2019KD0AB04), Zhejiang Lab's International Talent Fund for Young Professionals, and CCF-Tencent Open Fund.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 2018. 4

[2] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In *ICCV*, 2019. 4

[3] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 1

[4] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 2

[5] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018. 1

[6] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020. 4

[7] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, 2018. 1

[8] Yunchao Wei, Shuai Zheng, Ming-Ming Cheng, and etc. Zhao, Hang. Lid 2020: The learning from imperfect data challenge results. 2020. 2, 4

[9] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 3

[10] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1