Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation

Guolei Sun[1], ✉Wenguan Wang[1], Jifeng Dai[2], and Luc Van Gool[1]

[1] ETH Zurich [2] SenseTime Research
[3] Qing Yuan Research Institute, Shanghai Jiao Tong University
https://github.com/GuoleiSun/MCIS_wsss

Abstract. This paper studies the problem of learning semantic segmentation from image-level supervision only. Current popular solutions leverage object localization maps from classifiers as supervision signals, 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. In addition to boosting object pattern learning, the co-attention can leverage context from other related images to improve localization map inference, hence eventually benefiting semantic segmentation learning. More essentially, our algorithm provides a unified framework that handles well different WSSS settings, *i.e.*, learning WSSS with (1) precise image-level supervision only, (2) extra simple single-label data, and (3) extra noisy web data. It sets new state-of-the-arts on all these settings, demonstrating well its efficacy and generalizability. Moreover, our approach ranked 1^{st} place in the Weakly-Supervised Semantic Segmentation Track of CVPR2020 Learning from Imperfect Data Challenge.

Keywords: Semantic Segmentation, Weakly Supervised Learning

1 Introduction

Recently, modern deep learning based semantic segmentation models[5,6], trained with massive manually labeled data, achieve far better performance than before. However, the fully supervised learning paradigm has the main limitation of requiring intensive manual labeling effort, which is particularly expensive for annotating pixel-wise ground-truth for semantic segmentation. Numerous efforts are

✉ Corresponding author: Wenguan Wang (wenguanwang.ai@gmail.com).

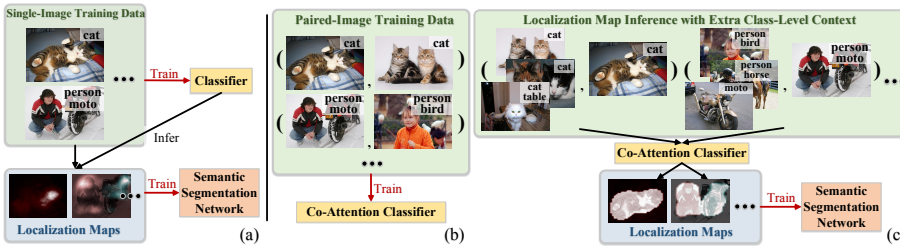

Fig. 1: (a) Current WSSS methods only use single-image information for object pattern discovering. (b-c) Our co-attention classifier leverages cross-image semantics as class-level context to benefit object pattern learning and localization map inference.

motivated to develop semantic segmentation with weaker forms of supervision, such as bounding boxes [45], scribbles [36], points [3], and image-level labels [46], *etc.* Among them, a prominent and appealing trend is using only image-level labels to achieve weakly supervised semantic segmentation (WSSS), which demands the least annotation efforts and is followed in this work.

To tackle the task of WSSS with only image-level labels, current popular methods are based on network visualization techniques [80, 86], which discover discriminative regions that are activated for classification. These methods 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 tends to over-address the most discriminative parts rather than entire objects, which becomes the focus of this area. Diverse solutions are explored, typically adopting: *image-level* operations, such as region hiding and erasing [30, 69], *regions growing* strategies that expand the initial activated regions [27, 62], and *feature-level* enhancements that collect multi-scale context from deep features [33, 71].

These efforts generally achieve promising results, which demonstrates the importance of discriminative object pattern mining for WSSS. However, as shown in Fig. 1(a), they 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 (see Fig. 1(b-c)). 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 two co-attentions work in a cooperative and complimentary manner, together making the classifier understand object patterns more comprehensively.

In addition to benefiting object pattern learning, our co-attention provides an efficient tool for precise localization map inference (see Fig. 1(c)). Given a training image, a set of related images (*i.e.*, sharing certain common semantics) are utilized by the co-attention for capture richer context and generate more accurate localization maps. 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. Overall, our co-attention boosts object discovering during both the classifier's training phase as well as localization map inference stage. This provides the possibility of obtaining more accurate pseudo pixel-level annotations, which facilitate final semantic segmentation learning.

Our algorithm is a unified and elegant framework, which generalizes well different WSSS settings. Recently, to overcome the inherent limitation in WSSS without additional human supervision, some efforts resort to extra image-level supervision from simple single-class data readily available from other existing datasets [35,48], or cheap web-crawled data [18,53,54,70]. Although they improve the performance to some extent, complicated techniques, such as energy function optimization [18, 58], heuristic constraints [54], and curriculum learning [70], are needed to handle the challenges of domain gap and data noise, restricting their utility. However, due to the use of paired image data for classifier training and object map inference, our method has good tolerance to noise. In addition, our method also handles domain gap naturally, as the co-attention effectively addresses domain-shared object pattern learning and achieves domain adaption as a part of co-attention parameter learning. We conduct extensive experiments on PASCAL VOC 2012 [10], under three WSSS settings, *i.e.*, learning WSSS with **(1)** PASCAL VOC image-level supervision only, **(2)** extra simple single-label data, and **(3)** extra web data. Our algorithm sets state-of-the-art on each case, verifying its effectiveness and generalizability. Our method also ranked 1^{st} place in the Weakly-supervised Semantic Segmentation Track of CVPR2020 Learning from Imperfect Data (LID) Challenge[72] (LID_{20}), outperforming other competitors by large margins.

Our contributions are three-fold. **(1)** We address the value of cross-image semantic correlations for complete object pattern learning as well as object location inference, which is achieved by a co-attention classifier that works over paired training samples. **(2)** Our co-attention classifier mines semantic cues in a more comprehensive manner. In addition to single-image semantics, it mines complimentary supervision from cross-image semantic similarities and differences by the co-attention and contrastive co-attention, respectively. **(3)** Our approach is general enough to learn WSSS with precise image-level supervision, or with extra simple single-label, or even noisy web-crawled data. It solves inherent challenges of different WSSS settings elegantly, and shows promising results consistently.

2 Related Work

Weakly Supervised Semantic Segmentation. Recently, lots of WSSS methods have been proposed to alleviate labeling cost. Various weak supervision forms have been explored, such as bounding boxes [9, 45], scribbles [36], point supervision [3], *etc.* Among them, image-level supervision, due to its less annotation demand, gains most attention and is also adopted in our approach.

Current popular solutions for WSSS with image-level supervision rely on network visualization techniques [80, 86], especially the Class Activation Map (CAM) [86], which discovers image pixels that are informative for classification. However, CAM typically only identifies small discriminative parts of objects, making it not an ideal proxy ground-truth for semantic segmentation training. Therefore, numerous efforts are made towards expanding the CAM-highlighted regions to the whole objects. In particular, some representative approaches make use of *image-level* hiding and erasing operations to drive a classifier to focus on different parts of objects [30, 34, 69]. A few ones instead resort to a *regions growing* strategy, *i.e.*, view the CAM-activated regions as initial "seeds" and gradually grow the seed regions until cover the complete objects [2, 22, 27, 62]. Meanwhile, some researchers investigate to directly enhance the activated regions on *feature-level* [31, 33, 71]. When constructing CAMs, they collect multi-scale context, which is achieved by dilated convolution [71], multi-layer feature fusion [33], saliency-guided iterative training [62], or stochastic feature selection [31]. Some others accumulate CAMs from multiple training phases [23], or self-train a difference detection network to complete the CAMs with trustable information [55]. In addition, a recent trend is to utilize class-agnostic saliency cues to filter out background responses [11, 22, 31, 34, 62, 69, 71] during localization map inference.

Since the supervision provided in above problem setting is so weak, another category of approaches explores to leverage more image-level supervision from other sources. There are mainly two types: (1) exploring simple and single-label examples [35, 48] (*e.g.*, images from existing datasets [15, 51]); or (2) utilizing near-infinite yet noisy web-sourced image [18, 53, 54, 70] or video [18, 32, 58] data (also referred as *webly supervised semantic segmentation* [24]). In addition to the common challenge of domain gap between the extra data and target semantic segmentation dataset, the second-type methods need to handle data noise.

Past efforts only consider each image individually, while only few exceptions [11, 53] address cross-image information. [53] simply applies off-the-shelf co-segmentation [25] over the web images to generate foreground priors, instead of ours encoding the semantic relations into network learning and inference. For [11], although also exploiting correlations within image pairs, the core idea is to use extra information from a support image to supplement current visual representations. Thus the two images are expected to better contain the same semantics, and unmatched semantics would bring negative influences. In contrast, we view both semantic homogeneity and difference as informative cues, driving our classifier to more explicitly identify the common as well as unshared objects, respectively. Moreover, [11] only utilizes single image to infer the activated objects, but our method comprehensively leverages the cross-image semantics in

both classifier training and localization map inference stages. More essentially, our framework is neat and flexible, which is not only able to learn WSSS from clean image-level supervision, but general enough to naturally make use of extra noisy web-crawled or simple single-label data, contrarily to previous efforts which are limited to specific training settings and largely dependent on complicated optimization methods [18,58] or heuristic constraints [54].

Deterministic Neural Attention. Differentiable attention mechanisms enable a neural network to focus more on relevant elements of the input than on irrelevant parts. With their popularity in the field of natural language processing [7,37,41,47,59], attention modeling is rapidly adopted in various computer vision tasks, such as image recognition [12,21,57,64,73], domain adaptation [65,84], human pose estimation [8, 61, 78], reasoning among objects [52, 87], and image generation [77, 82, 88]. Further, co-attention mechanisms become an essential tool in many vision-language applications and sequential modeling tasks, such as visual question answering [39,42,76,79], visual dialog [74,85], vision-language navigation [66], and video segmentation [40,60], showing its effectiveness in capturing the underlying relations between different entities. Inspired by the general idea of attention mechanisms, this work leverages co-attention to mine semantic relations within training image pairs, which helps the classifier network learn complete object patterns and generate precise object localization maps.

3 Methodology

Problem Setup. Here we follow current popular WSSS pipelines: given a set of training images with image-level labels, a *classification network* is first trained to discover corresponding discriminative object regions. The resulting *object localization maps* over the training samples are refined as pseudo ground-truth masks to further supervise the learning of a *semantic segmentation network*.

Our Idea. Unlike most previous efforts that treat each training image *individually*, we explore cross-image semantic relations as class-level context for understanding object patterns more *comprehensively*. To achieve this, two neural co-attentions are designed. The first one drives the classifier to learn common semantics from the co-attentive object regions, while the other one enforces the classifier to focus on the rest objects for unshared semantics classification.

3.1 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 convolutional embedding 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.

As in [23, 31, 32], 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 $\boldsymbol{s}_m \in \mathbb{R}^K$

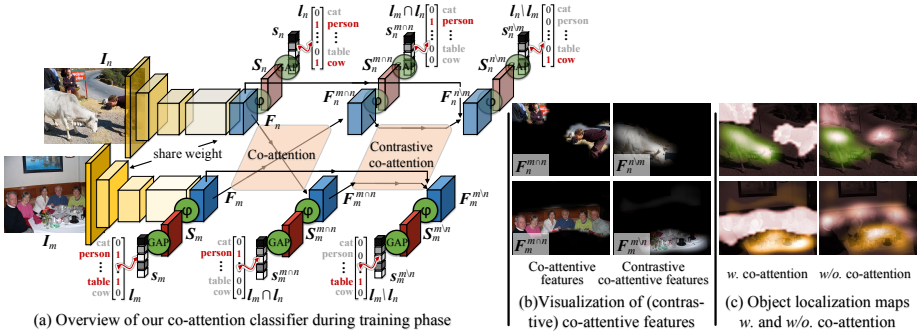

Fig. 2: **(a)** In addition to mining object semantics from single-image labels, semantic similarities and differences between paired training images are both leveraged for supervising object pattern learning. **(b)** Co-attentive and contrastive co-attentive features complimentarily capture the shared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

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:

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

So far the classifier is learned in a standard manner, *i.e.*, only individual-image information is used for semantic learning. One can directly use the activation maps to supervise next-stage semantic segmentation learning, as done in [22, 32]. Differently, our classifier additionally utilizes 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 $\boldsymbol{F}_n$:

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

where $\boldsymbol{F}_m \in \mathbb{R}^{C\times HW}$ and $\boldsymbol{F}_n \in \mathbb{R}^{C\times HW}$ are flattened into matrix formats, and $\boldsymbol{W}_{\boldsymbol{P}} \in \mathbb{R}^{C\times C}$ is a learnable matrix. The affinity matrix $\boldsymbol{P}$ stores similarity scores corresponding to all pairs of positions in $\boldsymbol{F}_m$ and $\boldsymbol{F}_n$, *i.e.*, the $(i,j)^{th}$ element of $\boldsymbol{P}$ gives the similarity between i^{th} location in $\boldsymbol{F}_m$ and j^{th} location in $\boldsymbol{F}_n$.

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 $\boldsymbol{F}_m$:

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

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 $\boldsymbol{F}_n$ ($\boldsymbol{F}_m$):

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

where $\boldsymbol{F}_m^{m\cap n}$ and $\boldsymbol{F}_n^{m\cap n}$ are reshaped into $\mathbb{R}^{C\times W\times H}$. 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 $\boldsymbol{l}_m \cap \boldsymbol{l}_n$ 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 $\boldsymbol{F}_n^{m\cap n}$ (see Fig. 2(b)). If we feed $\boldsymbol{F}_m^{m\cap n}$ and $\boldsymbol{F}_n^{m\cap n}$ into the class-aware fully convolutional layer φ, the generated class-aware activation maps, *i.e.*, $\boldsymbol{S}_m^{m\cap n}=\varphi(\boldsymbol{F}_m^{m\cap n})\in\mathbb{R}^{K\times H\times W}$ and $\boldsymbol{S}_n^{m\cap n}=\varphi(\boldsymbol{F}_n^{m\cap n})\in\mathbb{R}^{K\times H\times W}$, are able to locate the common object **Person** in $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, respectively. After GAP, the predicted semantic classes (scores) $\boldsymbol{s}_m^{m\cap n}\in\mathbb{R}^K$ and $\boldsymbol{s}_n^{m\cap n}\in\mathbb{R}^K$ should be the common semantic labels $\boldsymbol{l}_m\cap\boldsymbol{l}_n$ of $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$, *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 $\boldsymbol{F}_n^{m\cap n}$ (see 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**. Therefore, 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 essentially makes full use of the context across training data.

Intuitively, for the co-attention based common semantic classification, the labels $\boldsymbol{l}_m\cap\boldsymbol{l}_n$ shared between $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ are used to supervise learning:

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

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:

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

The set operation '$\cap$' is slightly extended here to represent bitwise-and.

where $\sigma(\cdot)$ is the *sigmoid* activation function, and the parameter matrix $\boldsymbol{W}_{\boldsymbol{B}} \in \mathbb{R}^{1\times C}$ learns for common semantics collection and is implemented by a convolutional 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:

$$\boldsymbol{A}_m^{m\backslash n} = \mathbf{1} - \boldsymbol{B}_m^{m\cap n} \in [0,1]^{H\times W}, \quad \boldsymbol{A}_n^{n\backslash m} = \mathbf{1} - \boldsymbol{B}_n^{m\cap n} \in [0,1]^{H\times W}. \tag{7}$$

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

$$\boldsymbol{F}_m^{m\backslash n} = \boldsymbol{F}_m \otimes \boldsymbol{A}_m^{m\backslash n} \in \mathbb{R}^{C\times H\times W}, \quad \boldsymbol{F}_n^{n\backslash m} = \boldsymbol{F}_n \otimes \boldsymbol{A}_n^{n\backslash m} \in \mathbb{R}^{C\times H\times W}. \tag{8}$$

'$\otimes$' 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.*, $\boldsymbol{S}_m^{m\backslash n} = \varphi(\boldsymbol{F}_m^{m\backslash n}) \in \mathbb{R}^{K\times H\times W}$ and $\boldsymbol{S}_n^{n\backslash m} = \varphi(\boldsymbol{F}_n^{n\backslash m}) \in \mathbb{R}^{K\times H\times W}$, and semantic scores, *i.e.*, $\boldsymbol{s}_m^{m\backslash n} = \mathrm{GAP}(\boldsymbol{S}_m^{m\backslash n}) \in \mathbb{R}^K$ and $\boldsymbol{s}_n^{n\backslash m} = \mathrm{GAP}(\boldsymbol{S}_n^{n\backslash m}) \in \mathbb{R}^K$. For $\boldsymbol{s}_m^{m\backslash n}$ and $\boldsymbol{s}_n^{n\backslash m}$, they are expected to identify the categories of the unshared objects, *i.e.*, $\boldsymbol{l}_m \backslash \boldsymbol{l}_n$ and $\boldsymbol{l}_n \backslash \boldsymbol{l}_m$.

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.*, $\boldsymbol{F}_m^{m\backslash n}$ and $\boldsymbol{F}_n^{n\backslash m}$, 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 $\boldsymbol{F}_n^{n\backslash m}$. However, the rest semantics in $\boldsymbol{F}_n^{n\backslash m}$ 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:

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

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

More In-Depth Discussion. One can interpret our co-attention classifier from a view of *auxiliary-task learning* [14,43], which is investigated in self-supervised learning field to improve data efficiency and robustness, by exploring auxiliary tasks from inherent data structures. In our case, rather than the task of single-image semantic recognition which has been extensively studied in conventional WSSS methods, we explore two auxiliary tasks, *i.e.*, predicting the common and uncommon semantics from image pairs, for fully mining supervision signals from weak supervision. The classifier is driven to better understand the cross-image semantics by attending to (contrastive) co-attentive features, instead of only relying on intra-image information (see Fig. 2(c)). In addition, such strategy shares a spirit of *image co-segmentation*. Since the image-level semantics of training set are given, the knowledge about some images share or unshare certain semantics should be used as a cue, or supervision signal, to better locate corresponding objects. Our co-attention based learning pipeline also provides an *efficient data augmentation* strategy, due to the use of paired samples, whose amount is near the square of the number of single training images.

3.2 Co-Attention Classifier Guided WSSS Learning

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

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

The coefficients of different loss terms are set as 1 in our all experiments. 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.*, $\boldsymbol{l}_m \cap \boldsymbol{l}_n \neq \mathbf{0}$.

Generating Object Localization Maps. Once our image classifier is trained, we apply it over the training data $\mathcal{I} = \{(\boldsymbol{I}_n, \boldsymbol{l}_n)\}_n$ to produce corresponding object localization maps, which are essential for semantic segmentation network training. We explore two different strategies to generate localization maps.

- *Single-round feed-forward prediction*, made over each training image individually. For each training image $\boldsymbol{I}_n$, running the classifier and directly using its class-aware activation map (*i.e.*, $\boldsymbol{S}_n \in \mathbb{R}^{K \times H \times W}$) as the object localization map $\boldsymbol{L}_n$, as most previous network visualization based methods [23,32,54] done.
- *Multi-round co-attentive prediction with extra reference information*, which is achieved by considering extra information from other related training images (see Fig. 1(c)). Specifically, given a training image $\boldsymbol{I}_n$ and its associated label vector $\boldsymbol{l}_n$, we generate its localization map $\boldsymbol{L}_n$ in a *class-wise* manner. For each semantic class $k \in \{1, \cdots, K\}$ labeled for $\boldsymbol{I}_n$, *i.e.*, $l_{n,k} = 1$ and $l_{n,k}$ is the k^{th} element of $\boldsymbol{l}_n$, we sample a set of related images $\mathcal{R} = \{\boldsymbol{I}_r\}_r$ from $\mathcal{I}$, which are also annotated with label k, *i.e.*, $l_{r,k} = 1$. Then we compute the co-attentive feature $\boldsymbol{F}^{m \cap r}_n$ from each related image $\boldsymbol{I}_r \in \mathcal{R}$ to $\boldsymbol{I}_n$, and get the co-attention based class-aware activation map $\boldsymbol{S}^{m \cap r}_n$. Given all the class-aware activation maps $\{\boldsymbol{S}^{m \cap r}_n\}_r$ from $\mathcal{R}$, they are integrated to infer the localization map *only* for class k, *i.e.*, $L_{n,k} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} S^{m \cap r}_{n,k}$. Here $L_{n,k} \in \mathbb{R}^{H \times W}$ and $S^{(\cdot)}_{n,k} \in \mathbb{R}^{H \times W}$

indicate the feature map at k^{th} channel of $\boldsymbol{L}_n \in \mathbb{R}^{K\times H\times W}$ and $\boldsymbol{S}_n^{(\cdot)} \in \mathbb{R}^{K\times H\times W}$, respectively. '$|\cdot|$' numerates the elements. After inferring the localization maps for all the annotated semantic classes of $\boldsymbol{I}_n$, we can get $\boldsymbol{L}_n$.

These two localization map generation strategies are studied in our experiments (§4.5), and the last one is more favored, as it uses both intra- and inter-image semantics for object inference, and shares a similar data distribution of the training phase. One may notice that the contrastive co-attention is not used here. This is because contrastive co-attentive feature (Eq. 8) is from its original image, which is effective for boosting feature representation learning during classifier training, while contributes little for localization maps inference (with limited cross-image information). Related experiments can be found at §4.5.

Learning Semantic Segmentation Network. After obtaining high-quality localization maps, we generate pseudo pixel-wise labels for all the training samples $\mathcal{I}$, which can be used to train arbitrary semantic segmentation network. For pseudo groundtruth generation, we follow current popular pipeline [20, 22, 23, 31, 32, 81], that uses localization maps to extract class-specific object cues and adopts saliency maps [19, 38] to get background cues. For the semantic segmentation network, as in [20, 23, 31, 32], we choose DeepLab-LargeFOV [5].

Learning with Extra Simple Single-Label Images. Some recent efforts [35, 48] are made towards exploring extra simple single-label images from other existing datasets [15, 51] for further boosting WSSS. Though impressive, specific network designs are desired, due to the issue of domain gap between additionally used data and the target complex multi-label dataset, *i.e.*, PASCAL VOC 2012 [10]. Interestingly, our co-attention based WSSS algorithm provides an alternate that addresses the challenge of domain gap naturally. Here we revisit the computation of co-attention in Eq. 2. When $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ are from different domains, the parameter matrix $\boldsymbol{W_P}$, in essence, learns to map them into a unified *common semantic space* [44] and the co-attentive features can capture domain-shared semantics. Therefore, for such setting, we learn three different parameter matrixes for $\boldsymbol{W_P}$, for the cases where $\boldsymbol{I}_m$ and $\boldsymbol{I}_n$ are from (1) the target semantic segmentation domain, (2) the one-label image domain, and (3) two different domains, respectively. Thus the domain adaption is efficiently achieved as a part of co-attention learning. We conduct related experiments in §4.2.

Learning with Extra Web Images. Another trend of methods [18, 24, 54, 70] address webly supervised semantic segmentation, *i.e.*, leveraging web images as extra training samples. Though cheaper, web data are typically noisy. To handle this, previous arts propose diverse effective yet sophisticated solutions, such as multi-stage training [24] and self-paced learning [70]. Our co-attention based WSSS algorithm can be easily extended to this setting and solve data noise elegantly. As our co-attention classifier is trained with paired images, instead of previous methods only relying on each image individually, our model provides a more robust training paradigm. In addition, during localization map inference, a set of extra related images are considered, which provides more comprehensive and accurate cues, and further improves the robustness. We experimentally demonstrate the effectiveness of our method in such a setting in §4.3.

3.3 Detailed Network Architecture

Network Configuration. In line with conventions [23,71,83], our image classifier is based on ImageNet [29] pre-trained VGG-16 [56]. For VGG-16 network, the last three fully-connected layers are replaced with three convolutional layers with 512 channels and kernel size 3×3, as done in [23,83]. For the semantic segmentation network, for fair comparison with current top-leading methods [2,23,31,55], we adopt the ResNet-101 [17] version Deeplab-LargeFOV architecture.

Training Phases of the Co-Attention Classifier and Semantic Segmentation Network. Our co-attention classifier is fully end-to-end trained by minimizing the loss defined in Eq. 10. The training parameters are set as: initial learning rate (0.001) which is reduced by 0.1 after every 6 epochs, batch size (5), weight decay (0.0002), and momentum (0.9). Once the classifier is trained, we generate localization maps and pseudo segmentation masks over all the training samples (see §3.2). Then, with the masks, the semantic segmentation network is trained in a standard way [23] using the hyper-parameter setting in [5].

Inference Phase of the Semantic Segmentation Network. Given an *unseen* test image, our segmentation network works in the *standard* semantic segmentation pipeline [5], *i.e.*, directly generating segments without using any other images. Then CRF [28] post-processing is performed to refine predicted masks.

Note that above settings are used in traditional WSSS datasets (*i.e.*, §4.1, §4.2, §4.3). Due to the specific task setup in LID_{20} [72], corresponding training and testing settings will be detailed in §4.4.

4 Experiment

Overview. Experiments are first conducted over *three* different WSSS settings: **(1)** The most standard paradigm [22, 23, 55, 69] that only allows image-level supervision from PASCAL VOC 2012 [10] (see §4.1). **(2)** Following [35, 48], additional single-label images can be used, yet bringing the challenge of domain gap (see §4.2). **(3)** Webly supervised semantic segmentation paradigm [24, 32, 54], where extra web data can be accessed (see §4.3). Then, in §4.4, we show the results in WSSS track of LID_{20}, where our method achieves the champion. Finally, in §4.5, ablation studies are made to assess the effectiveness of essential parts of our algorithm.

Evaluation Metric. In our experiments, the standard intersection over union (IoU) criterion is reported on the val and test sets of PASCAL VOC 2012 [10]. The scores on test set are obtained from official PASCAL VOC evaluation server.

4.1 Experiment 1: Learn WSSS only from PASCAL VOC [10] Data

Experimental Setup: We first conduct experiment following the most standard setting that learns WSSS with only image-level labels [22, 23, 55, 69], *i.e.*, only image-level supervision from PASCAL VOC 2012 [10] is accessible. PASCAL VOC 2012 contains a total of 20 object categories. As in [5, 69], augmented training data from [16] are also used. Finally, our model is trained on totally 10,582 samples with only image-level annotations. Evaluations are conducted on the val and test sets, which have 1,449 and 1,456 images, respectively.

Table 1: Experimental results for WSSS under three different settings. **(a)** Standard setting where only PASCAL VOC 2012 images are used (§4.1). **(b)** Additional single-label images are used (§4.2). **(c)** Additional web-crawled images are used (§4.3). *: VGG backbone. †: ResNet backbone.

Methods	Publication	Val	Test
Using PASCAL VOC data only			
*DCSM [68]	ECCV16	44.1	45.1
*SEC [27]	ECCV16	50.7	51.7
*AFF [49]	ECCV16	54.3	55.5
†DCSP [4]	BMVC17	60.8	61.9
*CBTS [50]	CVPR17	52.8	53.7
*AE-PSL [69]	CVPR17	55.0	55.7
*Oh *et al.* [18]	CVPR17	55.7	56.7
*TPL [26]	ICCV17	53.1	53.8
*MEFF [13]	CVPR18	-	55.6
*GAIN [34]	CVPR18	55.3	56.8
*MDC [71]	CVPR18	60.4	60.8
†MCOF [63]	CVPR18	60.3	61.2
†DSRG [22]	CVPR18	61.4	63.2
†PSA [2]	CVPR18	61.7	63.7
†SeeNet [20]	NIPS18	63.1	62.8
†IRN [1]	CVPR19	63.5	64.8
†FickleNet [31]	CVPR19	64.9	65.3
†SSDD [55]	ICCV19	64.9	65.5
†OAA+ [23]	ICCV19	65.2	66.4
***Ours** (VGG)	-	63.5	63.6
†**Ours** (ResNet)	-	**66.2**	**66.9**

(a)

Methods	Publication	Val	Test
Using extra simple single-label images			
*MCNN [58]	ICCV15	-	36.9
*MIL-ILP [48]	CVPR15	32.6	-
*MIL-sppxl [48]	CVPR15	36.6	35.8
*MIL-bb [48]	CVPR15	37.8	37.0
*MIL-seg [48]	CVPR15	42.0	40.6
*AttnBN [35]	ICCV19	62.1	63.0
***Ours** (VGG)	-	64.6	64.6
†**Ours** (ResNet)	-	**67.1**	**67.2**

(b)

Methods	Publication	Val	Test
Using extra noisy web images/videos			
*MCNN [58]	ICCV15	38.1	39.8
†Shen *et al.* [53]	BMVC17	56.4	56.9
*STC [70]	PAMI17	49.8	51.2
*Hong *et al.* [18]	CVPR17	58.1	58.7
*WebS-i1 [24]	CVPR17	51.6	-
*WebS-i2 [24]	CVPR17	53.4	55.3
†Shen *et al.* [54]	CVPR18	63.0	63.9
***Ours** (VGG)	-	65.0	64.7
†**Ours** (ResNet)	-	**67.7**	**67.5**

(c)

Experimental Results: Table 1a compares our approach and current top-leading WSSS methods (highest mIoU is used for comparison) with image-level supervision, on both PASCAL VOC12 val and test sets. Additionally, we show some segmentation results in Fig. 3. We can observe that our method achieves mIoU scores of 66.2 and 66.9 on val and test sets respectively, outperforming all the competitors. The performance of our method is 87% of the DeepLab-LargeFOV [5] trained with fully annotated data, which achieved an mIoU of 76.3 on val set. When compared to OAA+ [23], current best-performing method, our approach obtains the improvement of 1.0% on val set. This well demonstrates that the localization maps produced by our co-attention classifier effectively detect more complete semantic regions towards the whole target objects. Note that our network is elegantly trained end-to-end in a single phase. In contrast, many other recent approaches including OAA+ [23] and SSDD [55], use extra networks [2, 23, 55] to learn auxiliary information (*e.g.*, integral attention [23], pixel-wise semantic affinity [55], *etc.*), or adopt multi-step training [1, 69, 71].

4.2 Experiment 2: Learn WSSS with Extra Simple Single-Label Data

Experimental Setup: Following [35, 48], we train our co-attention classifier and segmentation network with PASCAL images and extra single-label images. The extra single-label images are borrowed from the subsets of Caltech-256 [15] and ImageNet CLS-LOC [51], and whose annotations are within 20 VOC object categories. There are a total of 20,057 extra single-label images.

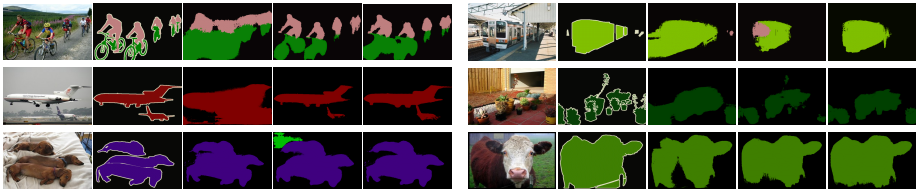

Fig. 3: Visual comparison results on PASCAL VOC12 val set. From *left* to *right*: input image, ground truth, results for PSA [2], OAA+ [23], and our method.

Table 2: Ablation study for different object localization map generate strategies, reported on PASCAL VOC12 val set. See §4.5 for details.

Method	Inference Mode	Input Image(s)	Val
Basic Classifier	Single-round feed-forward	Test image *only*	61.7
Our Variant	Single-round feed-forward	Test image *only*	64.7
	Multi-round co-attention and contrastive co-attention	Test image and other related images	66.2
Full Model	Multi-round co-attention	Test image and other related images	**66.2**

Experimental Results: The comparisons are shown in Table 1b. Our method significantly improves the most recent method (*i.e.*, AttnBN [35]) in this setting by 5.0% and 4.2% in val and test sets, respectively. With the fact that objects of the same category but from different domains share similar visual patterns [35], our co-attention provides an end-to-end strategy that efficiently captures the common, cross-domain semantics, and learns domain adaption naturally. Even AttnBN is specifically designed for addressing such setting by knowledge transfer, our method still suppresses it by a large margin. Compared with the setting in §4.1 where only PASCAL images are used for training, our method obtains improvements on both val and test sets, verifying that it successfully mines knowledge from extra simple single-label data and copes with domain gap well.

4.3 Experiment 3: Learn WSSS with Extra Web-Sourced Data

Experimental Setup: We also conduct experiments using both PASCAL VOC images and webly craweled images as training data. We use the web data provided by [54], which are retrieved from Bing based on class names. The final dataset contains 76,683 images across 20 PASCAL VOC classes.

Experimental Results: Table 1c shows the performance comparisons between our method and the previous webly supervised segmentation methods. It shows that our method outperforms all other approaches and sets new state-of-the-arts with mIoU score of 67.7 and 67.5 on PASCAL VOC 2012 val and test sets, respectively. Among the compared methods, Hong *et al.* [18] utilize richer information of the temporal dynamics provided by additional large-scale videos. In contrast, although only using static image data, our method still outperforms it on the val and test sets by 9.6% and 8.8%, respectively. Compared with Shen *et al.* [54] which uses the same web data as ours, our method substantially improves it by a clear margin of 3.6% on the test set.

Table 3: Results on *val* and *test* sets of both LID_{19} and LID_{20} WSSS track.

Year	Team	Extra Saliency Annotation	Val	Test
LID_{19}	T.T (T.T)	✓	-	8.1
	LEAP_DEXIN	✓	20.7	19.6
	MVN	✓	41.0	40.0
LID_{20}	play-njupt	✗	22.1	31.9
	IOnlyHaveSevenDays	✗	39.0	36.2
	UCU & SoftServe	✗	39.7	37.3
	VL-task1	✗	40.1	37.7
	CVL (**ours**)	✗	**46.2**	**45.1**

4.4 Experiment 4: Performance on WSSS Track of LID_{20} Challenge

Experimental Setup: The challenge dataset [72] is built upon ImageNet [51]. 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, our co-attention image classifier is built upon ResNet-38 [75], as the dataset has 200 classes and a stronger backbone can better learn subtle semantics between classes. The training parameters are set as: initial learning rate (0.005) and the poly policy based training schedule: $lr = lr_{init} \times (1 - \frac{iter}{max\_iter})^{\gamma}$ with $\gamma(0.9)$, batch size (8), weight decay (0.0005), and max epoch (15). During training, the equivariant attention [67] is also adopted. Once our image classifier is trained, we run the classifier and directly use its class-aware activation map (*i.e.*, $\boldsymbol{S}_n$) as the object localization map $\boldsymbol{L}_n$. Then we generate pseudo pixel-wise labels for all the training samples $\mathcal{I}$. Since only image tags can be used, we follow [2]: 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. For better segmentation results, 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 multiple scales are averaged, with CRF refinement.

Experimental Results: The final results with the standard mean intersection over union (mIoU) criterion for WSSS track of both LID_{19} and LID_{20} challenges are shown in Table 3. Both LID_{19} and LID_{20} challenge use the same data. In LID_{19}, competitors can use extra saliency annotations to learn saliency models and refine pseudo ground truths. However, in LID_{20}, only image tags can be accessed. For methods shown in the table, top performing methods are included. As can be seen from Table 3, our approach not only outperforms the champion team in LID_{19}, which can use deep learning based saliency models, but also achieves the best performance in LID_{20} and sets a new state-of-the-art (*i.e.*, mIoU of 46.2 and 45.1 in val and test sets, respectively).

4.5 Ablation Studies

Inference Strategies. Table 2 shows mIoU scores on PASCAL VOC 2012 val set *w.r.t.* different inference modes (see §3.2). When using the traditional inference mode "single-round feed-forward", our method substantially suppresses

Table 4: Ablation study for our co-attention and contrastive co-attention mechanisms for training, reported on PASCAL VOC12 val set. See §4.5 for details.

Method	(Contrastive) Co-Attention	Training Loss	Val
Basic Classifier	-	$\mathcal{L}_{\text{basic}}$ (Eq. 1)	61.7
Our Variant	co-attention *only*	$\mathcal{L}_{\text{basic}}$ (Eq. 1)+$\mathcal{L}_{\text{co-att}}$ (Eq. 5)	65.5
Full Model	co-attention +contrastive co-attention	$\mathcal{L}_{\text{basic}}$ (Eq. 1)+$\mathcal{L}_{\text{co-att}}$ (Eq. 5)+$\mathcal{L}_{\overline{\text{co-att}}}$ (Eq. 9) = $\mathcal{L}$ (Eq. 10)	**66.2**

basic classifier, by improving mIoU score from 61.7 to 64.7. This evidences that co-attention mechanism (trained in an end-to-end manner) in our classifier improves the underlying feature representations and more object regions are identified by the network. We can observe that by using more images to generate localization maps, our method obtains consistent improvement from "Test image *only*" (64.7), to "Test images and other related images" (66.2). This is because more semantic context are exploited during localization map inference. In addition, using contrastive co-attention for localization map inference doesn't boost performance (66.2). This is because the contrastive co-attentive features for one image are derived from the image itself. In contrast, co-attentive features are from the other related image, thus can be effective in the inference stage.

(Contrastive) Co-Attention. As seen in Table 4, by only using co-attention (Eq. 5), we already largely suppress the basic classifier (Eq. 1) by 3.8%. When adding additional contrastive co-attention (Eq. 9), we obtain mIoU improvement of 0.7%. Above analysis verify our two co-attentions indeed boost performance.

Number of Related Images for Localization Map Inference. For localization map generation, we use 3 extra related images (§3.2). Here, we study how the number of reference images affect the performance. From Table 5, it is easily observed that when increasing the number of related images from 0 to 3, the performance gets boosted consistently. However, when further using more images, the performance degrades. This can be attributed to the trade-off between useful semantic information and noise brought by related images. From 0 to 3 reference images, more semantic information is used and more integral regions for objects are mined. When further using more related images, useful information reaches its bottleneck and noise, caused by imperfect localization of the classifier, takes over, decreasing performance.

Table 5: Ablation study for using different numbers of related images during object localization map generation, reported on PASCAL VOC12 val set (see §4.5).

Method	Extra Related Images (#)	Val
Our Variant	0	64.7
	1	65.9
	2	66.0
	4	66.1
	5	66.0
Full Model	3	**66.2**

5 Conclusion

This work proposes a co-attention classification network to discover integral object regions by addressing cross-image semantics. With this regard, a co-attention is exploited to mine the common semantics within paired samples, while a contrastive co-attention is utilized to focus on the exclusive and unshared ones

for capturing complimentary supervision cues. Additionally, by leveraging extra context from other related images, the co-attention boosts localization map inference. Further, by exploiting additional single-label images and web images, our approach is proven to generalize well under domain gap and data noise. Experiments over three WSSS settings consistently show promising results. Our method also ranked 1^{st} place in the weakly-supervised semantic segmentation track of LID_{20} challenge.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: CVPR (2019) 12
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: CVPR (2018) 4, 11, 12, 13, 14
3. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: ECCV (2016) 2, 4
4. Chaudhry, A., Dokania, P.K., Torr, P.H.: Discovering class-specific pixels for weakly-supervised semantic segmentation. In: BMVC (2017) 12
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI **40**(4), 834–848 (2017) 1, 10, 11, 12
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018) 1
7. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. In: EMNLP (2016) 5
8. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: CVPR (2017) 5
9. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV (2015) 4
10. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111**(1), 98–136 (2015) 3, 10, 11
11. Fan, J., Zhang, Z., Tan, T.: Cian: Cross-image affinity net for weakly supervised semantic segmentation. In: AAAI (2020) 4
12. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR (2019) 5
13. Ge, W., Yang, S., Yu, Y.: Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In: CVPR (2018) 12
14. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018) 9
15. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007) 4, 10, 12
16. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV (2011) 11
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 11

18. Hong, S., Yeo, D., Kwak, S., Lee, H., Han, B.: Weakly supervised semantic segmentation using web-crawled videos. In: CVPR (2017) 3, 4, 5, 10, 12, 13
19. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. TPAMI **41**(4), 815–828 (2019) 10
20. Hou, Q., Jiang, P., Wei, Y., Cheng, M.M.: Self-erasing network for integral object attention. In: NeurIPS (2018) 10, 12
21. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018) 5
22. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: CVPR (2018) 4, 6, 10, 11, 12
23. Jiang, P.T., Hou, Q., Cao, Y., Cheng, M.M., Wei, Y., Xiong, H.K.: Integral object mining via online attention accumulation. In: ICCV (2019) 4, 5, 9, 10, 11, 12, 13
24. Jin, B., Ortiz Segovia, M.V., Susstrunk, S.: Webly supervised semantic segmentation. In: ICCV (2017) 4, 10, 11, 12
25. Joulin, A., Bach, F., Ponce, J.: Discriminative clustering for image co-segmentation. In: CVPR (2010) 4
26. Kim, D., Cho, D., Yoo, D., So Kweon, I.: Two-phase learning for weakly supervised object localization. In: ICCV (2017) 12
27. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: ECCV (2016) 2, 4, 12
28. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NeurIPS (2011) 11
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012) 11
30. Kumar Singh, K., Jae Lee, Y.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: ICCV (2017) 2, 4
31. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: CVPR (2019) 4, 5, 10, 11, 12
32. Lee, J., Kim, E., Lee, S., Lee, J., Yoon, S.: Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In: ICCV (2019) 4, 5, 6, 9, 10, 11
33. Lee, S., Lee, J., Lee, J., Park, C.K., Yoon, S.: Robust tumor localization with pyramid grad-cam. arXiv preprint (2018) 2, 4
34. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: CVPR (2018) 4, 12
35. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Attention bridging network for knowledge transfer. In: ICCV (2019) 3, 4, 10, 11, 12, 13
36. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: CVPR (2016) 2, 4
37. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. In: ICLR (2017) 5
38. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR (2019) 10
39. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: NeurIPS (2016) 5
40. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR (2019) 5
41. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: EMNLP (2015) 5

42. Nguyen, D.K., Okatani, T.: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: CVPR (2018) 5
43. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: ICML (2017) 9
44. Pan, B., Cao, Z., Adeli, E., Niebles, J.C.: Adversarial cross-domain action recognition with co-attention. In: AAAI (2020) 10
45. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: ICCV (2015) 2, 4
46. Pathak, D., Shelhamer, E., Long, J., Darrell, T.: Fully convolutional multi-class multiple instance learning. arXiv preprint (2014) 2
47. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In: ICLR (2018) 5
48. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: CVPR (2015) 3, 4, 10, 11, 12
49. Qi, X., Liu, Z., Shi, J., Zhao, H., Jia, J.: Augmented feedback in semantic segmentation under image level supervision. In: ECCV (2016) 12
50. Roy, A., Todorovic, S.: Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In: CVPR (2017) 12
51. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015) 4, 10, 12, 14
52. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In: NeurIPS (2017) 5
53. Shen, T., Lin, G., Liu, L., Shen, C., Reid, I.: Weakly supervised semantic segmentation based on web image co-segmentation. In: BMVC (2017) 3, 4, 12
54. Shen, T., Lin, G., Shen, C., Reid, I.: Bootstrapping the performance of webly supervised semantic segmentation. In: CVPR (2018) 3, 4, 5, 9, 10, 11, 12, 13
55. Shimoda, W., Yanai, K.: Self-supervised difference detection for weakly-supervised semantic segmentation. In: ICCV (2019) 4, 11, 12
56. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint (2014) 11
57. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: ECCV (2018) 5
58. Tokmakov, P., Alahari, K., Schmid, C.: Weakly-supervised semantic segmentation using motion cues. In: ECCV (2016) 3, 4, 5, 12
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 5
60. Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L.: Zero-shot video object segmentation via attentive graph neural networks. In: ICCV (2019) 5
61. Wang, W., Zhu, H., Dai, J., Pang, Y., Shen, J., Shao, L.: Hierarchical human parsing with typed part-relation reasoning. In: CVPR (2020) 5
62. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: CVPR (2018) 2, 4
63. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: CVPR (2018) 12
64. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018) 5

65. Wang, X., Li, L., Ye, W., Long, M., Wang, J.: Transferable attention for domain adaptation. In: AAAI (2019) 5
66. Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: CVPR (2019) 5
67. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: CVPR (2020) 14
68. Wataru, S., Keiji, Y.: Distinct class saliency maps for weakly supervised semantic segmentation. In: ECCV (2016) 12
69. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: CVPR (2017) 2, 4, 11, 12
70. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. TPAMI **39**(11), 2314–2320 (2016) 3, 4, 10, 12
71. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: CVPR (2018) 2, 4, 11, 12
72. Wei, Y., Zheng, S., Cheng, M.M., Zhao, Hang, e.: Lid 2020: The learning from imperfect data challenge results (2020) 3, 11, 14
73. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: ECCV (2018) 5
74. Wu, Q., Wang, P., Shen, C., Reid, I., Van Den Hengel, A.: Are you talking to me? reasoned visual dialog generation through adversarial learning. In: CVPR (2018) 5
75. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition **90**, 119–133 (2019) 14
76. Xiong, C., Zhong, V., Socher, R.: Dynamic coattention networks for question answering. In: ICLR (2017) 5
77. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018) 5
78. Ye, Q., Yuan, S., Kim, T.K.: Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In: ECCV (2016) 5
79. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: CVPR (2019) 5
80. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014) 2, 4
81. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L.: Joint learning of saliency detection and weakly supervised semantic segmentation. In: ICCV (2019) 10
82. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: ICML (2019) 5
83. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: CVPR (2018) 11
84. Zhang, Y., Nie, S., Liu, W., Xu, X., Zhang, D., Shen, H.T.: Sequence-to-sequence domain adaptation network for robust text image recognition. In: CVPR (2019) 5
85. Zheng, Z., Wang, W., Qi, S., Zhu, S.C.: Reasoning visual dialogs with structural and partial observations. In: CVPR (2019) 5
86. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016) 2, 4

87. Zhou, T., Wang, W., Qi, S., Ling, H., Shen, J.: Cascaded human-object interaction recognition. In: CVPR (2020) 5
88. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019) 5