Regional Semantic Contrast and Aggregation for Weakly Supervised Semantic Segmentation

Tianfei Zhou[1,*], Meijie Zhang[2,*], Fang Zhao[3], Jianwu Li[2,†]

[1] Computer Vision Lab, ETH Zurich [2] Beijing Institute of Technology [3] Inception Institute of AI

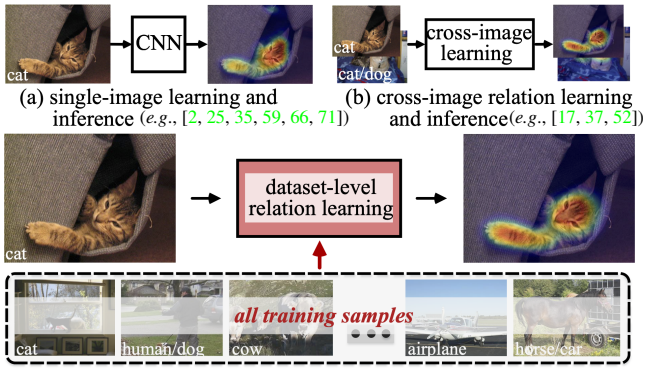
https://github.com/maeve07/RCA.git

Abstract

*Learning semantic segmentation from weakly-labeled (*e.g.*, image tags only) data is challenging since it is hard to infer dense object regions from sparse semantic tags. Despite being broadly studied, most current efforts directly learn from limited semantic annotations carried by individual image or image pairs, and struggle to obtain integral localization maps. Our work alleviates this from a novel perspective, by exploring rich semantic contexts synergistically among abundant weakly-labeled training data for network learning* ***and*** *inference. In particular, we propose regional semantic contrast and aggregation (RCA). RCA is equipped with a regional memory bank to store massive, diverse object patterns appearing in training data, which acts as strong support for exploration of dataset-level semantic structure. Particularly, we propose* ***i) semantic contrast*** *to drive network learning by contrasting massive categorical object regions, leading to a more holistic object pattern understanding, and* ***ii) semantic aggregation*** *to gather diverse relational contexts in the memory to enrich semantic representations. In this manner, RCA earns a strong capability of fine-grained semantic understanding, and eventually establishes new state-of-the-art results on two popular benchmarks,* i.e.*, PASCAL VOC 2012 and COCO 2014.*

1. Introduction

Semantic segmentation continues to be a fundamental task in computer vision, with numerous applications in autonomous driving, robotics, human-computer interactions and medical imaging analysis. While fully supervised systems have achieved tremendous progress, they are limited by the availability of pixel-level annotations, often harvested at great cost, even with smart interfaces [3]. Weakly supervised semantic segmentation (WSSS) alternatively investigates whether this task can be adequately addressed with efficient and weak supervisory signals (*e.g.*, image labels [2, 25, 37, 66], scribbles [39, 40, 54], bounding boxes [14, 34, 44, 51]). This work studies the form of image-level labels which can be obtained effortlessly, and thus have been widely embraced in mainstream approaches.



(a) single-image learning and inference (*e.g.*, [2, 25, 35, 59, 66, 71])
(b) cross-image relation learning and inference(*e.g.*, [17, 37, 52])
(c) dataset-level relation learning and inference

Figure 1. **The main idea** promoted throughout the paper is that semantic contexts subserve localization of individual objects in WSSS. Our RCA thus performs dataset-level relation learning (c) to mine rich contextual knowledge from massive (ideally all) training samples, rather than from an individual image (a) or image pair (b). This enables our model to procure in-depth semantic pattern understanding, improving object localization eventually.

In the absence of the true "image label" to "object region" correspondence in training data, learning to map visual concepts to pixel regions is particularly challenging. The seminal work, *i.e.*, class activation mapping (CAM) [81], solves this by mining regions from internal activation of an image classifier. However, the technique is prone to give sparse and incomplete object estimations, since the classifier is only driven to activate a small proportion of features with strong discriminative capability. To address this, a prevalent of subsequent efforts strive to learn more complete object regions by, *e.g.*, region growing to expand initial responses [24, 30, 60], adversarial erasing in a hide-and-seek fashion [23, 32, 33, 64], feature enrichment to collect within-image contexts [66, 72], seeking auxiliary saliency supervisions [35, 71, 74], or self-supervised learning with pre-designed pretext tasks [6, 47, 63].

Though impressive, these methods use only single-image information for object localization (Fig. 1 (a)), neglecting inter-image contextual information. Image-level labels not only tell the categories appearing in each individual image, but also unveil the semantic structure of all

* Equal contributions; † Corresponding author: Jianwu Li.

images in the dataset. For each concept (*i.e.*, `cat` in Fig. 1), the dataset contains numerous semantically similar but visually different instances; for any two different concepts (*e.g.*, `cat` and `dog`), all their instances are semantically different, even though some may look very similar with each other. This *a priori* knowledge should be exploited to gain more accurate semantic pattern understanding. Though some preliminary attempts [17, 52, 80, 83] have been made towards this (Fig. 1(b)), they focus on pairwise [17, 52, 80] or quadruplet [83] context modeling in a *limited* number of images, and thus cannot guarantee a sufficient understanding of holistic semantic patterns in the entire dataset. In addition, all these methods favor pixel-wise relation modeling, which is rather difficult due to the lack of proper supervisory signal and causes prohibitive computation cost.

Motivated by above analysis, we propose regional semantic contrast and aggregation (RCA) to maximally exploit contextual knowledge in visual data (Fig. 1 (c)), aiming for comprehensive object pattern learning as well as effective CAM inference. In lieu of pixel-level relation modeling in [17, 52, 83], RCA prefers *region-aware* representations that are more efficient and robust to noises. In particular, for each mini-batch image, we divide it into categorical *pseudo* regions according to an intermediate, coarse CAM, which is learned under the supervision of its single-image label. For each pseudo region, RCA establishes its relations to regions in all other images to facilitate dataset-level semantic context learning. For feasible computations, we associate RCA with a continuously-updated memory bank, which collects and preserves meaningful region semantics in the dataset as the training goes, and is applicable to both network learning and inference phases. During training, RCA explores semantic relations of regions in each mini-batch and the memory bank from two novel perspectives:

- *Semantic contrast*, which lets the model learn to discriminate all possible object regions in the dataset, promoting more holistic object pattern understanding. Particularly, for each pseudo region, semantic contrast enforces the network to pull its embedding close to memory embeddings of the same category and push apart those of different. Such a contrastive property well complements the classification objective (for each single image) to improve object representation learning.
- *Semantic aggregation*, which allows the model to gather dataset-level contextual knowledge to yield more meaningful object representations. This is achieved via a non-parametric attention module which summarizes memory representations for each image independently. In comparison with conventional *intra-image* context learning schemes [12, 73], our semantic aggregation focuses on *inter-image* context mining, and thus is able to capture more informative dataset-level semantics.

These two context modeling schemes are indispensable to our model. Semantic contrast helps the network to learn more structured object embedding space from a holistic view, while semantic aggregation focuses on improving feature representations of each image by collecting diverse semantic contexts. In addition, semantic contrast is essential to maintain unique and informative memory embeddings, which is a prerequisite to yield reliable semantic aggregation. These two components work together to make RCA a powerful WSSS model (see Table 1). Our RCA is flexible and can be easily incorporated into existing WSSS models. It shows consistently improved segmentation performance on challenging datasets (*i.e.*, PASCAL VOC 2012 [15] and COCO 2014 [41]), on top of state-of-the-art WSSS models (*i.e.*, OAA$^+$ [25], EPS [35]).

Main Contributions. **i)** We study an essential yet long-ignored problem in WSSS to explore rich contexts among weakly labeled training data for network learning. This essentially narrows the gap between image-level semantic concepts and pixel-level object regions. Technically, **ii)** we introduce a robust contrastive learning algorithm for semantic contrast, which is able to learn effective representations from imperfect, pseudo region features, as well as **iii)** a non-parametric attention model for semantic aggregation to collect rich contextual knowledge from the entire dataset .

2. Related Work

Weakly Supervised Semantic Segmentation is gaining popularity due to its practical value in reducing the burden of collecting pixel-level annotations at a large scale required by its fully-supervised counterparts [56, 57, 82, 84, 85] Here weak supervision may come in diverse forms, *e.g.*, image-level labels [7, 55, 65, 75, 83], scribbles [40, 54], bounding boxes [14, 28, 44, 51], point clicks [3, 27]. Among them, image-level labels gain the most attention due to its minimal annotation demand. However, since only the presence or absence of particular semantics is indicated, the task becomes extremely challenging. The pioneering work of [81] proposes to obtain coarse object localization maps (*i.e.*, CAMs) from CNN-based image classifiers as seeds to generate pixel-level pseudo segmentation labels. Follow-up works expand coarse CAMs to obtain full extents of object regions by region growing [24, 30, 66], using stochastic inference [33], incorporating self-supervised learning [6, 47, 63], exploring boundary constraints [8, 35], or alternatively mining and erasing object regions [23, 36, 64].

Past efforts only consider each image individually, ignoring the rich semantic context across different training images. Recent works [17, 52] address cross-image semantic mining by computing semantic co-attention between each pair of images, while [83] further enables high-order semantic mining from more images through a graph neural network architecture. Though impressive, these approaches still consider limited semantic context within a *small* num-

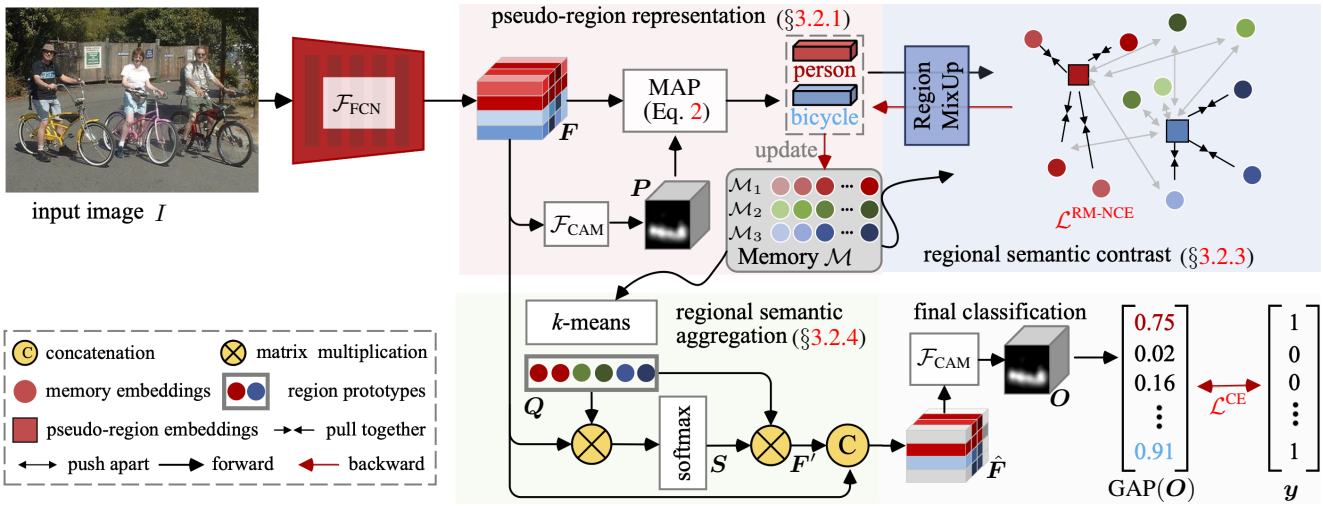


Figure 2. Detailed illustration of **regional semantic contrast and aggregation**. See §3 for more details.

ber of images (*i.e.*, 2 in [17, 52] and 4 in [83]). In contrast, our approach takes a further step to explore the learning of *rich* relations from a *large* number of weakly annotated data. It is equipped with a *pseudo-region* memory bank to store region-level semantic embedding for each category, which enables region-aware semantic contrast and aggregation for more comprehensive object pattern mining.

Contrastive Representation Learning is becoming increasingly attractive due to its great potential for *un-/self-supervised* representation learning [10, 21, 45, 50, 53, 68]. These approaches learn to compare samples in order to push apart dissimilar (or *negative*) data pairs while pulling together similar (or *positive*) pairs. Some approaches [4, 11, 19] even achieve compelling performance without using any negative pairs. Beyond image-level instance discrimination, recent efforts [5, 62, 70] explore pixel- or patch-level discrimination to learn visual representations that generalize better to downstream dense prediction tasks (*e.g.*, semantic segmentation, object detection). Furthermore, *supervised* contrastive learning has been studied in [29] for image recognition and in [58] for supervised semantic segmentation. These methods extend the self-supervised setup (by leveraging label information) to contrast the set of all samples from the same class as positives against the negatives from other classes. Inspired by these advances, our approach performs dense contrative learning to improve object localization ability of neural networks, using *weakly supervised* annotations. Our approach is naturally distinguished from the above dense representation learning methods which either neglect any annotations [5, 62, 70] or require pixel-level supervisions [58].

Relational Context Learning is popular in image and video segmentation to augment feature embedding of each pixel by gathering useful representations from its contextual pixels [18, 78] or regions [12, 73]. However, these methods are limited to capturing local contexts within each individual image, ignoring potential semantic contexts across different images. In sharp contrast, our semantic aggregation mines relational semantics across all images of the entire dataset to gain more informative context learning.

Non-Parametric Memory Bank has been found feasible to remember a massive number of samples for learning good representations [21, 43, 58, 61, 68]. Our memory bank is inspired by these efforts, which however, is unique in that *i)* it stores consistent and expressive region-level semantics inferred from image-level labels; *ii)* more importantly, it is also kept alive in the inference phase to provide holistic contextual knowledge for network inference.

3. Our Approach

3.1. Problem Statement

Task Setup. Following the standard setup, each training image $I \in \mathbb{R}^{w \times h \times 3}$ in the dataset $\mathcal{I}$ is associated with only an image-level label vector $\boldsymbol{y} = [y_1, y_2, \ldots, y_L] \in \{0,1\}^L$ for L pre-specified categories. Here, $y_l = 1$ indicates the presence of class l in I and 0 otherwise. Given such coarse annotations, most current solutions follow a two-phase pipeline to solve the task "*from classification to segmentation*", *i.e.*, training a *classification network* first for identifying object regions corresponding to each category, which are then refined to produce pseudo segmentation labels as the supervision of a *semantic segmentation network*.

Previous Solutions to WSSS. Recent approaches [25, 35, 79], in general, derive class-aware attention maps directly from a fully-convolutional network (FCN), which is proven to produce localization maps with the same quality as CAM [79]. Particularly, for a mini-batch image I, its class-aware attention map $\boldsymbol{P}$ is generated as follows:

$$\boldsymbol{F} = \mathcal{F}_{\text{FCN}}(I) \in \mathbb{R}^{W \times H \times D}, \quad \boldsymbol{P} = \mathcal{F}_{\text{CAM}}(\boldsymbol{F}) \in \mathbb{R}^{W \times H \times L}. \quad (1)$$

Here, $\mathcal{F}_{\text{FCN}}$ is an FCN network, typically corresponding to the convolutional part of a standard classifier (*e.g.*, VGG [49], ResNet [22]). $\boldsymbol{F}$ is the dense embedding of I, with D channels and $W\times H$ spatial size. $\mathcal{F}_{\text{CAM}}$ is a class-aware convolutional layer to produce $\boldsymbol{P}=[\boldsymbol{P}_1,\cdots,\boldsymbol{P}_L]$, with each map $\boldsymbol{P}_l\in\mathbb{R}^{W\times H}$ denoting network activation of the l-th class. Next, a score vector $\boldsymbol{p}=[p_1,p_2,\cdots,p_L]\in\mathbb{R}^L$ is derived from $\boldsymbol{P}$ via a global average pooling (GAP) layer, with $p_l=\text{GAP}(\boldsymbol{P}_l)$ being the un-normalized score of the l-th class. Finally, $\boldsymbol{p}$ is used for multi-label classification.

Our Main Idea. With above descriptions of existing WSSS solutions, we find that they only exploit limited contextual cues in individual images, causing difficulties for more complete understanding of diverse semantic patterns. To compensate for this limitation, we introduce a novel method, *i.e.*, RCA, to perform semantic contrast and semantic aggregation over *pseudo regions* of a large number of images (ideally the entire dataset). Both semantic contrast and semantic aggregation are supported by an external pseudo-region memory bank. Next, we will first describe the way to build initial pseudo-region representations (§3.2.1) as well as to construct the memory bank (§3.2.2). Then, we elaborate on semantic contrast (§3.2.3) and semantic aggregation (§3.2.4). The overall pipeline of RCA is illustrated in Fig. 2.

3.2. Regional Semantic Contrast and Aggregation

3.2.1 Pseudo-Region Representation

For each mini-batch sample I, we convert its dense embedding $\boldsymbol{F}$ (Eq. 1) into a set of categorical region representations based on $\boldsymbol{P}$ (Eq. 1). Particularly, for the l-th category that appears in I (*i.e.*, $y_l=1$), its region-level semantic information is summarized to a compact embedding vector $\boldsymbol{f}_l\in\mathbb{R}^D$ by masked average pooling (MAP) [48]:

$$\boldsymbol{f}_l=\frac{\sum_{x=1,y=1}^{W,H}\boldsymbol{M}_l(x,y)\boldsymbol{F}(x,y)}{\sum_{x=1,y=1}^{W,H}\boldsymbol{M}_l(x,y)}\in\mathbb{R}^D, \tag{2}$$

where $\boldsymbol{M}_l=\mathbb{1}(\boldsymbol{P}_l>\mu)\in\{0,1\}^{W\times H}$ is a binary mask, highlighting only strongly-activated pixels of class l in its activation map (*i.e.*, $\boldsymbol{P}_l\in\mathbb{R}^{W\times H}$). $\mathbb{1}(\cdot)$ is an indicator function, and the threshold μ is set to the mean value of $\boldsymbol{P}_l$. Here $\boldsymbol{f}_l$ is compact and lightweight, allowing for feasible exploration of its relations with a massive number of pseudo regions mining from other samples.

3.2.2 Pseudo-Region Memory Bank

Taking the inspiration from [68, 69], we setup a non-parametric and dynamic memory bank for RCA to store dataset-level regional semantic information. In particular, the memory bank $\mathcal{M}$ consists of L dictionaries, *i.e.*, $\mathcal{M}=\{\mathcal{M}_1,\mathcal{M}_2,\cdots,\mathcal{M}_L\}$, each for one category. Each entry of $\mathcal{M}_l$ denotes a holistic region-aware representation $\boldsymbol{m}_l\in\mathbb{R}^D$ of the l-th category in image I observed in the whole learning phase. At each training step, the memory bank will be updated during backward propagation to involve new observations. In particular, the current feature vector $\boldsymbol{f}_l$ (Eq. 2) of image I will be smoothly updated into the memory representation $\boldsymbol{m}_l$ as follows:

$$\boldsymbol{m}_l\leftarrow\gamma\boldsymbol{m}_l+(1-\gamma)\boldsymbol{f}_l, \tag{3}$$

where γ is the momentum for memory evolution. We update $\boldsymbol{m}_l$ when the l-th class appears in I (*i.e.*, $y_l=1$) and its classification score is higher than a threshold ν, *i.e.*, $p_l>\nu$. Otherwise, we keep $\boldsymbol{m}_l$ as it was.

Memory Mechanism Discussion. Though memory bank has been widely utilized in recent methods [21, 68, 69], ours shows several unique and appealing characteristics that could lift more advantages to the task of WSSS. **First**, the memory is compartmentalized enough to compress each potential semantic hypothesis (*i.e.* pseudo-region embedding) in each training sample individually and is able to well encode diverse semantic patterns of each category within weakly-labelled visual data; **Second**, the momentum updating scheme (Eq. 3) not only helps to gain *consistent* memory features for semantic contrast (§3.2.3) as [21, 68], but more crucially, offers *comprehensive* representations that can accurately describe object semantics. More concretely, Eq. 3 accumulates all intermediate states (*e.g.*, $\{\boldsymbol{f}_l\}$) of each object region produced by the image classifier at different training epochs. These states have shown to be well complementary with each other [25], and as a result of Eq. 3, each memory feature $\boldsymbol{m}_l$ will be gradually promoted to capture a more complete object region as the training goes. This eventually results in informative memory representations after training, which can be leveraged as reliable contexts for semantic aggregation (§3.2.4).

3.2.3 Regional Semantic Contrast (RSC)

We perform semantic contrast over *pseudo*-region semantics for learning more discriminative dense representations. For each categorical pseudo-region embedding $\boldsymbol{f}_l$ (Eq. 2) in image I, our objective is to increase its similarities to memory features $\{\boldsymbol{m}_l^+\in\mathcal{M}_l\}$ of the same class, while reducing the similarities to features $\{\boldsymbol{m}_l^-\in\mathcal{M}\setminus\mathcal{M}_l\}$ of different classes. We achieve this via a region-aware contrastive loss:

$$\mathcal{L}_l^{\text{NCE}}(\boldsymbol{f}_l,y_l)=\frac{1}{|\mathcal{M}_l|}\sum_{\boldsymbol{m}_l^+\in\mathcal{M}_l}-\log\frac{e^{\text{sim}(\boldsymbol{f}_l,\boldsymbol{m}_l^+)/\tau}}{e^{\text{sim}(\boldsymbol{f}_l,\boldsymbol{m}_l^+)/\tau}+\sum\limits_{\boldsymbol{m}_l^-\in\mathcal{M}\setminus\mathcal{M}_l}e^{\text{sim}(\boldsymbol{f}_l,\boldsymbol{m}_l^-)/\tau}}, \tag{4}$$

where τ is a temperature hyper-parameter scaling the distribution of distances, and $\text{sim}(\boldsymbol{i},\boldsymbol{j})=\frac{\boldsymbol{i}\cdot\boldsymbol{j}}{\|\boldsymbol{i}\|_2\|\boldsymbol{j}\|_2}$ is the dot product between ℓ_2-normalized $\boldsymbol{i}$ and $\boldsymbol{j}$ (*i.e.*, cosine similarity).

Eq. 4 falls into the regime of supervised contrastive learning [29], *i.e.*, the labels of $\boldsymbol{f}_l/\boldsymbol{m}_l^+/\boldsymbol{m}_l^-$ are given. Differently, in our context, the labels are weak and noisy, pos-

ing great challenges to learn robust representations. To alleviate this problem, we develop *region mixup* to regularize Eq. 4 to learn effective region representations, even from noisy samples. More specifically, for each region l in I, we create a mixed region by linearly combining it with a region l^- in another mini-batch image. Here we assume that regions l and l^- are from different categories, *i.e.*, $y_l \neq y_{l^-}$. The embedding of the mixed region is computed as:

$$\hat{\boldsymbol{f}}_l = \omega \boldsymbol{f}_l + (1-\omega)\boldsymbol{f}_{l^-}, \tag{5}$$

where the coefficient $\omega \sim \mathcal{B}(\beta, \beta)$ follows a Beta distribution $\mathcal{B}(\cdot,\cdot)$ with two shape parameters set to a same β [77]. Then, we define a new region mixup contrastive loss:

$$\mathcal{L}_l^{\text{RM-NCE}} = \omega \mathcal{L}_l^{\text{NCE}}(\hat{\boldsymbol{f}}_l, y_l) + (1-\omega)\mathcal{L}_l^{\text{NCE}}(\hat{\boldsymbol{f}}_l, y_{l^-}). \tag{6}$$

It computes two $\mathcal{L}_l^{\text{NCE}}$ losses with respect to y_l and y_{l^-}, which are combined by the same weight ω used for region mixup (Eq. 5). Eq. 6 encourages the network to learn relative similarities for mixed regions, regularizing the model to learn robust representations from label-imperfect samples.

3.2.4 Regional Semantic Aggregation (RSA)

Context is widely recognized to be significant for pixel understanding [26, 73, 78], but prior approaches focus on intra-image context modeling, ignoring rich and valuable inter-image contexts. To alleviate this, we devise semantic aggregation to exploit dataset-level context cues in the memory bank for enhancing semantic understanding. As stated in §3.2.2, our memory bank offers massive signatures of semantic regions. While a large-scale memory bank could benefit semantic contrast [21], it contains overcomplete (or redundant) representations and some are even noisy, making accurate context learning difficult. In addition, directly aggregating large-scale representations is computationally expensive, and will greatly slow down the learning and inference procedures.

To address these problems, we compress the overcomplete memory representations into a compact set of representative prototypes. For each class l, we do k-means clustering over all features in $\mathcal{M}_l$ to obtain K prototype vectors (*i.e.*, class centroids), organized in a matrix form $\boldsymbol{Q}_l \in \mathbb{R}^{K\times D}$. Here we use multiple prototypes (*i.e.*, $K>1$) for each class to account for significant intra-class variations. Next, all the categorical prototypes derived from the memroy bank $\mathcal{M}$ are concatenated together, delivering a holistic prototypical representation $\boldsymbol{Q} = [\boldsymbol{Q}_1, \cdots, \boldsymbol{Q}_L] \in \mathbb{R}^{K\times D\times L}$. Then, for each mini-batch image I with feature $\boldsymbol{F} \in \mathbb{R}^{W\times H\times D}$ (Eq. 1), we calculate its affinity matrix $\boldsymbol{S}$ with the prototypical representation $\boldsymbol{Q}$ as follows:

$$\boldsymbol{S} = \texttt{softmax}(\boldsymbol{F} \otimes \boldsymbol{Q}^\top) \quad \in \mathbb{R}^{(WH)\times(LK)}, \tag{7}$$

where $\boldsymbol{F} \in \mathbb{R}^{(WH)\times D}$ and $\boldsymbol{Q} \in \mathbb{R}^{(LK)\times D}$ are flattened into matrix representations for computational convenience. $\otimes$ indicates matrix multiplication. $\texttt{softmax}(\cdot)$ normalizes each row of the input. Each entry in $\boldsymbol{S}$ reflects the normalized similarity between each row (*i.e.*, feature) in $\boldsymbol{F}$ and each column (*i.e.*, prototype) in $\boldsymbol{Q}^\top$. Based on the affinity matrix, the contextual summaries for the feature embedding $\boldsymbol{F}$ w.r.t. the prototypical representation $\boldsymbol{Q}$ can be computed:

$$\boldsymbol{F}' = \boldsymbol{S} \otimes \boldsymbol{Q} \quad \in \mathbb{R}^{(WH)\times D}, \tag{8}$$

where $\boldsymbol{F}'$ denotes an enriched feature representation of $\boldsymbol{F}$, which is further reshaped into $\mathbb{R}^{W\times H\times D}$. Finally, we concatenate $\boldsymbol{F}'$ and the original feature $\boldsymbol{F}$ together:

$$\hat{\boldsymbol{F}} = [\boldsymbol{F}, \boldsymbol{F}'] \quad \in \mathbb{R}^{W\times H\times 2D}. \tag{9}$$

Here, $\hat{\boldsymbol{F}}$ not only encodes intra-image local contexts in $\boldsymbol{F}$, but also captures inter-image global contexts in $\boldsymbol{F}'$, thus enriching the representability for semantic understanding.

3.2.5 Class Activation Map Prediction

Finally, $\hat{\boldsymbol{F}}$ is fed into another class-aware convolutional layer $\mathcal{F}_{\text{CAM}}$ (Eq. 1) to produce the final activation maps $\boldsymbol{O}$:

$$\boldsymbol{O} = \mathcal{F}_{\text{CAM}}(\hat{\boldsymbol{F}}) \quad \in \mathbb{R}^{W\times H\times L}. \tag{10}$$

3.3. Detailed Network Architecture

Our classifier is comprised of four major components: **i)** The backbone network $\mathcal{F}_{\text{FCN}}$ (Eq. 1) maps an input image I into a convolutional representation $\boldsymbol{F}$. Any FCN network can be used here, and we use two popular ones, *i.e.*, VGG16 [49] and ResNet38 [22], for fair comparison with existing approaches. **ii)** The class-wise convolutional layer $\mathcal{F}_{\text{CAM}}$ (Eq. 1) produces a class-aware attention map from feature embeddings. In our network, two independent $\mathcal{F}_{\text{CAM}}$ are used in Eq. 1 and Eq. 10, respectively. Each is implemented as a 1×1 convolutional layer. **iii)** The memory bank $\mathcal{M}$ stores all region patterns in training data. Note that the memory bank is removed at the inference phase, with only compressed global prototypical representations kept instead. This reduces the cost to maintain a large memory bank during model deployment. **iv)** The loss function of our classifier is as follows:

$$\mathcal{L} = \sum_I \alpha_1 \mathcal{L}^{\text{RM-NCE}} + \alpha_2 \mathcal{L}^{\text{CE}}(\text{GAP}(\boldsymbol{P}), \boldsymbol{y}) + \mathcal{L}^{\text{CE}}(\text{GAP}(\boldsymbol{O}), \boldsymbol{y}), \tag{11}$$

where each image I is supervised by the combination of three losses. The first term $\mathcal{L}^{\text{RM-NCE}}$ is the region mixup contrastive loss (Eq. 6), which is computed as the average loss of all regions appearing in I. The second one is an auxiliary cross-entropy loss $\mathcal{L}^{\text{CE}}$ for supervising the intermediate CAM prediction $\boldsymbol{P}$ (Eq. 1), while the third loss is the main cross-entropy loss imposing on the final CAM prediction $\boldsymbol{O}$ (Eq. 10). The coefficients α_1 and α_2 balance the three terms.

variant	mIoU (%)	
	pseudo label (train)	segmentation (val)
OAA^{+}	-	65.2
OAA^{++}	68.2	67.7
w/ RSC (§3.2.3)	**69.5** ↑ 1.3	**69.3** ↑ 1.6
w/ RSA (§3.2.4)	**68.5** ↑ 0.3	**68.6** ↑ 0.9
w/ RSC *and* RSA (full model)	**71.4** ↑ 3.2	**70.6** ↑ 2.9

Table 1. **Ablation study** on VOC 2012 [15]. "pseudo label": generated pseudo labels on the train set; "segmentation": segmentation results on the val set.

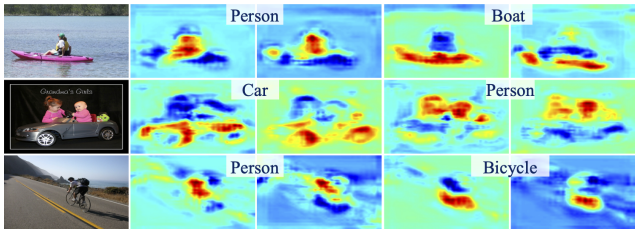


Figure 3. **Visualization of affinity $\boldsymbol{S}$ (Eq. 7).** Each heatmap correspondes to a column in matrix $\boldsymbol{S}$, which is a dot-product between a particular prototype with image feature $\boldsymbol{F}$. See §4.2 for details.

4. Experiment

4.1. Experimental Setting

Dataset. The experiments are conducted on two datasets:

- **PASCAL VOC 2012** [15] is a gold standard benchmark for WSSS. It contains 4,369 images, which are split into 1,464/1,449/1,456 for train/val/test, respectively. It provides pixel-level annotations for 21 categories. As common practices [24, 33, 75], we use additional 10,582 images [20] for training.
- **COCO 2014** [41] is a more challenging dataset, containing complex contextual interactions of 80 object classes, which attracts interests to verify the performance of our model in this dataset. We follow the official setting to use 80K images for train and 40K images for val.

Evaluation Protocol. We evaluate RCA in terms of **i)** semantic segmentation on VOC 2012 val/test and COCO 2014 val, and **ii)** quality of generated pseudo segmentation labels on VOC 2012 train. As conventions [25, 35], mean intersection-over-union (mIoU) is used as the metric in both cases. The scores on VOC 2012 test are obtained from the official evaluation server.

Implementation Details. As stated in §3.3, we test two commonly used backbones (*i.e.*, VGG16 [49], ResNet38 [22]) in RCA for the experiments. The weights of the backbones are loaded from ImageNet pre-trained weights. RCA is trained using the SGD optimizer with batch size 8, momentum 0.9 and weight decay 5e-4. The initial learning rates are set to 1e-3 for the backbone and 1e-2 for other components, which are reduced by 0.1 per five epochs. We warm up the network in the first epoch by using the cross-entropy losses only in Eq. 11, *i.e.*, $\alpha_1 = 0$. The network is trained for 30 epochs in total. For VOC 2012, we use an adaptive memory size for each class to store all region embeddings in the dataset, while for COCO 2014, the per-class memory size is set to 500 to avoid significant memory consumption. The k-means prototype clustering in §3.2.4 is performed only once at the beginning of each epoch, and the per-class prototype number is set to $K = 10$ by default. For the hyper-parameters, we empirically set the threshold ν, momentum γ, shape parameter β, weights α_1 and α_2 to 0.7, 0.99, 8, 0.01 and 0.4, respectively.

Once the classifier is well trained, we generate class-aware attention maps $\boldsymbol{O}$ (Eq. 10) for each training image and regard them as foreground seeds. In line with [25, 35, 37, 67, 71], we also compute a saliency map for each image using off-the-shelf models to estimate background cues. The final pseudo labels are obtained by combining the foreground and background cues together [25, 37]. Finally, with the pseudo masks as the supervision, we train DeepLabV2 [9] using the default hyper-parameter setting in [9]. Dense CRF [31] is used as a post-processing routine to refine segmentation boundaries, as in [35, 38, 67, 71, 76].

Baselines. RCA is flexible and can be easily incorporated into most WSSS models. In the experiments, we evaluate RCA based on two baselines, *i.e.*, OAA^{+} [25] (due to its popularity) and EPS [35] (due to its overall best performance). For the conventional OAA^{+}, we build a stronger baseline OAA^{++}, by replacing its saliency model with [42], which is widely-used by recent approaches [67, 71]. EPS is the leading WSSS model nowadays; we use it to validate the efficacy of RCA, even with a strong baseline.

Reproducibility. Our network is implemented in PyTorch and trained on four NVIDIA V100 cards. Testing is conducted on a single NVIDIA RTX2080Ti card.

4.2. Diagnostic Experiment

We first ablate the core designs of RCA in terms of pseudo label quality on VOC 2012 train. VGG16 is used as the classification backbone by default.

Semantic Contrast and Semantic Aggregation. We investigate the necessity to learn dataset-level visual contexts for WSSS. Table 1 summarizes the results. **First**, the variant "w/ RSC" significantly improves against OAA^{++} in both pseudo label (*i.e.*, **1.3%**) and segmentation (*i.e.*, **1.6%**) performance, proving that by contrasting massive object regions, our model fulfills the goal of more comprehensive object pattern understanding. **Second**, "w/ RSA" only achieves marginal performance gains. However, when integrating it with RSC together, our full model (*i.e.*, "w/ RSC *and* RSA") achieves remarkable improvements in comparison with "w/ RSC" (69.5% vs **71.4%** for pseudo label, 69.3% vs **70.6%** for segmentation). This reveals that RSC, which helps to obtain informative memory representations,

is essential for RSA to perform reliable context aggregation.

To gain more insights into RSA, we visualize feature-prototype affinity $\boldsymbol{S}$ (Eq. 7) in Fig. 3. We see that our prototypes are able to attend to semantically meaningful regions, which could benefit object localization.

Region Mixup. The following table ablates the design of region mixup in §3.2.3:

variant	w/o region mixup (Eq. 4)	w/ region mixup (Eq. 6)
mIoU (%)	70.6	71.4

We find that after dropping region mixup, the mIoU score reduces by 0.8%. This result reveals that region mixup indeed helps the model learn more robust representations from noisy data (*i.e.*, pseudo regions), leading to more accurate semantic understanding.

Memory Updating Coefficient γ. The table below shows accuracy of generated pseudo segmentation labels with different updating coefficients (Eq. 3):

coefficient γ	0	0.5	0.8	0.9	0.99	0.999
mIoU (%)	69.9	70.9	71.2	71.2	71.4	70.9

The optimal value is $\gamma = 0.99$ (our default). Moreover, RCA is robust when γ is in $0.8 \sim 0.99$, showing that it is beneficial to update the memory in a relatively slow speed, but not too slow (*i.e.*, $\gamma = 0.999$). When γ is too small, the performance degrades; at the extreme of *no momentum* (*i.e.*, $\gamma = 0$), the model significantly degrades. These results support our discussions in §3.2.2 that momentum updating helps to earn more consistent and comprehensive memory representations, providing powerful assistance for both semantic contrast and semantic aggregation.

Prototype Number K. The following table ablates the role of prototype number K in semantic aggregation (§3.2.4):

K	1	10	20	50	100	all
mIoU (%)	70.4	71.4	71.1	71.1	71.3	70.0

Note that for $K = 1$, we average all the embeddings in each dictionary to obtain a single prototype vector for each category; for the setting "all", we use all memory embeddings as the prototypes without clustering. As seen from the table, RCA shows stable performance when K is in $10 \sim 100$. At extreme cases, the model degrades due to severe information loss ($K = 1$) or too many noisy embeddings ("all").

Memory Size. By default, our memory bank stores all pseudo regions in the dataset. However, the following table shows that our model is not sensitive to this setting:

memory size	100	500	all
mIoU (%)	70.8	71.2	71.4

By storing only 100 or 500 region embeddings per class, the performance only degrades very slightly. This reveals that our model is scalable to larger-scale datasets (*e.g.*, COCO 2014), for which we cannot afford caching all embeddings.

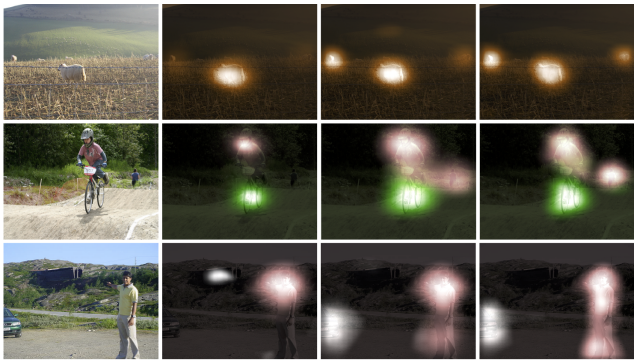


Figure 4. **Visualization of class activation maps** on VOC 2012 `train`. From left to right: input images, results of OAA^{++}, results from $\boldsymbol{P}$ (Eq. 1) and $\boldsymbol{O}$ (Eq. 10) of our full model.

method	backbone	mIoU (%)
SS-WSSS [CVPR20] [2]	ResNet38	62.2
ICD [CVPR20] [16]	VGG16	62.2
SubCat [CVPR20] [6]	ResNet38	63.4
CONTA [NeurIPS20] [75]	ResNet38	65.4
GroupWSSS [TIP21] [83]	VGG16	65.7
IRNet [CVPR19] [1]	ResNet50	66.5
BES [ECCV20] [8]	ResNet50	67.2
EDAM [CVPR21] [67]	ResNet38	68.1
OAA^{++}	VGG16	68.2
RCA+OAA^{++}		**71.4** ↑ 3.2
OAA^{++}	ResNet38	69.4
RCA+OAA^{++}		**73.2** ↑ 3.8
EPS [CVPR21] [35]	ResNet38	71.4
RCA+EPS		**74.1** ↑ 2.7

Table 2. **Quantitative performance of pseudo segmentation labels** on VOC 2012 [15] `train`.

4.3. Comparison with Prior Art

Object Localization. Table 2 reports the results of generated pseudo segmentation labels on VOC 2012 `train`. Notably, RCA improves OAA^{++} by **3.2%** and **3.8%** when using VGG16 and ResNet38 as the classifier backbones, respectively. It also yields a solid improvement against EPS (71.4% vs **74.1%**). These results confirm the strong localization capability of our approach.

Semantic Segmentation. Table 3 provides the comparison of RCA against representative methods on VOC 2012 `val` and `test`. As seen, RCA brings solid gains over the two baselines (*i.e.*, OAA^{++} and EPS). Using VGG16 (or ResNet38) as the classification backbone, RCA improves OAA^{++} by **2.9%** (**3.0%**) on `val`, **3.6%** (**3.4%**) on `test`. Consistent improvements (**1.3%**/**2.0%**) are also seen for EPS. In addition, RCA+EPS sets a new state-of-the-art.

Table 4 summarizes the segmentation results on COCO 2014 [41]. We observe that RCA surpasses OAA^{+} and EPS by **2.1%** and **1.1%**, respectively. Remarkably, RCA+EPS, which employs VGG16 as the backbone, outperforms many ResNet-based models (*e.g.*, AuxSegNet [71]).

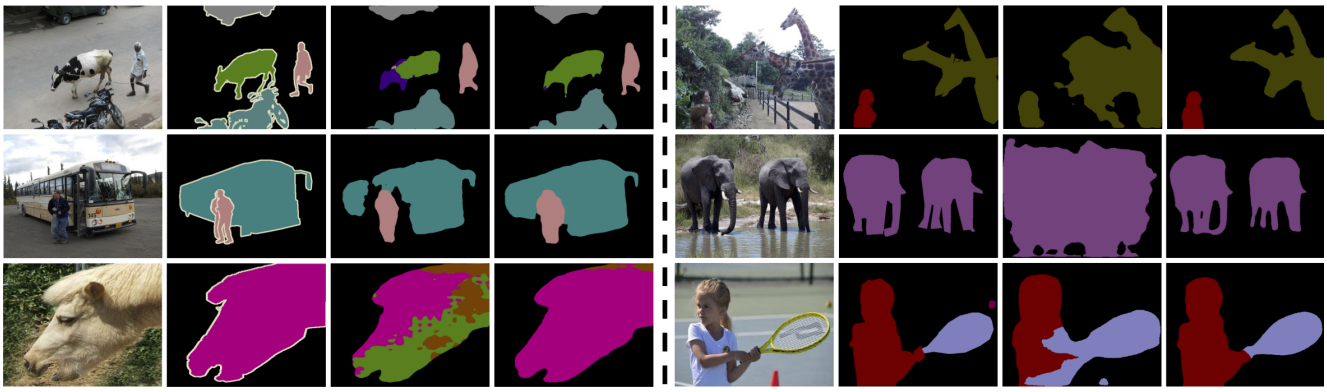


Figure 5. **Qualitative segmentation results** on VOC 2012 `val` (left) and COCO 2014 `val` (right). From left to right: input images, ground-truths, segmentation results of OAA^{++} as well as our RCA.

method	mIoU (%)	
	`val`	`test`
‡SSNet [ICCV19] [74]	63.3	64.3
‡RNet [CVPR19] [1]	63.5	64.8
*CIAN [AAAI20] [17]	64.3	65.3
*FickleNet [CVPR19] [33]	64.9	65.3
†SSDD [ICCV19] [47]	64.9	65.5
†SEAM [CVPR20] [63]	64.5	65.7
†SubCat [CVPR20] [6]	66.1	65.9
*OAA^{+} [ICCV19] [25]	65.2	66.4
‡BES [ECCV20] [8]	65.7	66.6
†CONTA [NeurIPS20] [75]	66.1	66.7
*MCIS [ECCV20] [52]	66.2	66.9
*ICD [CVPR20] [16]	67.8	68.0
†CPN [ICCV21] [76]	67.8	68.5
*NSROM [CVPR21] [72]	68.3	68.5
†AuxSegNet [ICCV21] [71]	69.0	68.6
‡PMM [ICCV21] [38]	68.5	69.0
*GroupWSSS [TIP21] [83]	68.7	69.0
†EDAM [CVPR21] [67]	70.9	70.6
†SPML [ICLR21] [27]	69.5	71.6
*OAA^{++}	67.7	67.4
***RCA**+OAA^{++}	**70.6** ↑ 2.9	**71.0** ↑ 3.6
†OAA^{++}	68.1	68.2
†**RCA**+OAA^{++}	**71.1** ↑ 3.0	**71.6** ↑ 3.4
†EPS [CVPR21] [35]	70.9	70.8
†**RCA**+EPS [CVPR21] [35]	**72.2** ↑ 1.3	**72.8** ↑ 2.0

Table 3. **Quantitative performance** on VOC 2012 [15] `val` and `test`. All models use ResNet as the segmentation backbone. *, † and ‡ denote models using VGG16, ResNet38 or ResNet50 as the classification backbone, respectively.

4.4. Visualization Result

Object Localization. Fig. 4 depicts some representative CAM predictions of OAA^{++} and RCA for training samples in PASCAL VOC 2012. As observed, our RCA is able to produce more integral object localization results across various challenging situations (*e.g.*, tiny objects, scale variations). In addition, the final CAM predictions (Eq. 10) are more accurate than the intermediate ones (Eq. 1), demonstrating the effectiveness of our core designs.

Semantic Segmentation. Fig. 5 illustrates some qualitative segmentation results of OAA^{++} and RCA on VOC 2012 `val` and COCO 2014 `val`. We find that RCA achieves more accurate segmentation results than OAA^{++}, showing remarkable capabilities in handling complex scenes, such as small/large objects, multiple instances, occlusions.

method	backbone	mIoU (%)
BFBP [ECCV16] [46]	VGG16	20.4
SEC [ECCV16] [30]	VGG16	22.4
DSRG [CVPR18] [24]	VGG16	26.0
IAL [IJCV20] [59]	VGG16	27.7
GroupWSSS [TIP21] [83]	VGG16	28.7
ADL [PAMI20] [13]	VGG16	30.8
SEAM [CVPR20] [63]	ResNet38	32.8
CONTA [NeurIPS20] [75]	ResNet38	32.8
AuxSegNet [ICCV21] [71]	ResNet38	33.9
OAA^{+} [ICCV19] [25]	VGG16	24.6
RCA+OAA^{+} [ICCV19] [25]	VGG16	**26.7** ↑ 2.1
EPS [CVPR21] [35]	VGG16	35.7
RCA+EPS [CVPR21] [35]	VGG16	**36.8** ↑ 1.1

Table 4. **Quantitative performance** on COCO 2014 [41] `val`.

5. Conclusion

In this work, we present a novel approach, RCA, to learn semantic segmentation using image-level supervision only. To alleviate the limited available knowledge carried by image labels, our approach explores the possibility to discover rich semantic contexts from weakly-labeled training data for learning. In particular, RCA is equipped with a continuously updated memory bank for storing massive historical pseudo-region features. The semantic relations between memory contents and mini-batch training samples are sufficiently exploited as additional supervisory signals (by semantic contrast) or holistic contextual cues (by semantic aggregation) to improve network learning and inference. Our approach is effective and principled, with extensive experiments manifesting its leading performance on popular benchmarks, *i.e.*, PASCAL VOC 2012 and COCO 2014.

Acknowledgements This work was supported by the Beijing Natural Science Foundation (L191004) and CCF-Baidu Open Fund.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019.

[2] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, 2020.

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *ECCV*, 2016.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[5] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *NeurIPS*, 2020.

[6] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *CVPR*, 2020.

[7] Arslan Chaudhry, Puneet K Dokania, and Philip HS Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1707.05821*, 2017.

[8] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *ECCV*, 2020.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020.

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[12] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A2-nets: Double attention networks. In *NeurIPS*, 2018.

[13] Junsuk Choe, Seungho Lee, and Hyunjung Shim. Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. *IEEE TPAMI*, 2020.

[14] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[16] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *CVPR*, 2020.

[17] Junsong Fan, Zhaoxiang Zhang, Tienniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *AAAI*, 2020.

[18] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[20] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[23] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *NeurIPS*, 2018.

[24] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018.

[25] 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.

[26] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In *ICCV*, 2021.

[27] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. In *ICLR*, 2021.

[28] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

[30] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[31] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 2011.

[32] 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.

[33] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 2019.

[34] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *CVPR*, 2021.

[35] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *CVPR*, 2021.

[36] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference

network. In *CVPR*, 2018.
[37] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In *AAAI*, 2021.
[38] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters inweakly-supervised semantic segmentation. In *ICCV*, 2021.
[39] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *CVPR*, 2022.
[40] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.
[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
[42] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.
[43] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.
[44] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *CVPR*, 2021.
[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
[46] Fatemehsadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*, 2016.
[47] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *ICCV*, 2019.
[48] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *ICCV*, 2019.
[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
[50] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016.
[51] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *CVPR*, 2019.
[52] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *ECCV*, 2020.
[53] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020.
[54] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*, 2017.
[55] Binglu Wang, Yongqiang Zhao, and Xuelong Li. Multiple instance graph learning for weakly supervised remote sensing object detection. *IEEE TGRS*, 60:1–12, 2021.
[56] Wenguan Wang, Tianfei Zhou, Fatih Porikli, David Crandall, and Luc Van Gool. A survey on deep learning technique for video segmentation. *arXiv preprint arXiv:2107.01153*, 2021.
[57] Wenguan Wang, Tianfei Zhou, Siyuan Qi, Jianbing Shen, and Song-Chun Zhu. Hierarchical human semantic parsing with comprehensive part-relation modeling. *IEEE TPAMI*, 2021.
[58] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021.
[59] Xiang Wang, Sifei Liu, Huimin Ma, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation by iterative affinity learning. *IJCV*, pages 1–14, 2020.
[60] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018.
[61] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. Cross-batch memory for embedding learning. In *CVPR*, 2020.
[62] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.
[63] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020.
[64] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017.
[65] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 39(11):2314–2320, 2016.
[66] 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.
[67] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2021.
[68] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.
[69] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.
[70] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021.
[71] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *ICCV*, 2021.
[72] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-

salient region object mining for weakly supervised semantic segmentation. In *CVPR*, 2021.
[73] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020.
[74] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, 2019.
[75] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020.
[76] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *ICCV*, 2021.
[77] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.
[78] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018.
[79] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018.
[80] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *ECCV*, 2020.
[81] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.
[82] Tianfei Zhou, Jianwu Li, Xueyi Li, and Ling Shao. Target-aware object discovery and association for unsupervised video multi-object segmentation. In *CVPR*, 2021.
[83] Tianfei Zhou, Liulei Li, Xueyi Li, Chun-Mei Feng, Jianwu Li, and Ling Shao. Group-wise learning for weakly supervised semantic segmentation. *IEEE TIP*, 31:799–811, 2021.
[84] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, 2020.
[85] Tianfei Zhou, Wenguan Wang, Si Liu, Yi Yang, and Luc Van Gool. Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In *CVPR*, 2021.