Hierarchical Semantic Contrast for Weakly Supervised Semantic Segmentation

Yuanchen Wu, **Xiaoqiang Li***, **Songmin Dai**, **Jide Li**, **Tong Liu** and **Shaorong Xie**
School of Computer Engineering and Science, Shanghai University, Shanghai, China
{yuanchenwu, xqli, laodar, iavtvai, tong_liu, srxie}@shu.edu.com

Abstract

Weakly supervised semantic segmentation (WSSS) with image-level annotations has achieved great processes through class activation map (CAM). Since vanilla CAMs are hardly served as guidance to bridge the gap between full and weak supervision, recent studies explore semantic representations to make CAM fit for WSSS better and demonstrate encouraging results. However, they generally exploit single-level semantics, which may hamper the model to learn a comprehensive semantic structure. Motivated by the prior that each image has multiple levels of semantics, we propose hierarchical semantic contrast (HSC) to ameliorate the above problem. It conducts semantic contrast from coarse-grained to fine-grained perspective, including ROI level, class level, and pixel level, making the model learn a better object pattern understanding. To further improve CAM quality, building upon HSC, we explore consistency regularization of cross supervision and develop momentum prototype learning to utilize abundant semantics across different images. Extensive studies manifest that our plug-and-play learning paradigm, HSC, can significantly boost CAM quality on both non-saliency-guided and saliency-guided baselines, and establish new state-of-the-art WSSS performance on PASCAL VOC 2012 dataset. Code is available at https://github.com/Wu0409/HSC_WSSS.

1 Introduction

Semantic segmentation is a fundamental task in computer vision, aiming at delineating target objects on the pixel level. With the recent advance of Deep Learning, fully supervised semantic segmentation (FSSS) models have made significant progress [Chen *et al.*, 2017; Yuan *et al.*, 2020; Xie *et al.*, 2021]. However, compared to other vision tasks, such as classification and object detection, acquiring sufficient pixel-level annotations for segmentation is exceedingly costly and time-consuming. Recently, many efforts have been devoted to weakly supervised semantic segmentation (WSSS) to lower the reliance on pixel-level annotations by utilizing "weak" labels (e.g., image-level labels [Wang *et al.*, 2020; Lee *et al.*, 2021c; Du *et al.*, 2022], scribbles [Lin *et al.*, 2016; Liang *et al.*, 2022; Vernaza and Chandraker, 2017], and bounding boxes [Lee *et al.*, 2021b; Oh *et al.*, 2021]). Thereinto, since image-level annotations can be obtained effortlessly, image-level WSSS is currently the popular direction, and we also adopt this form for WSSS.

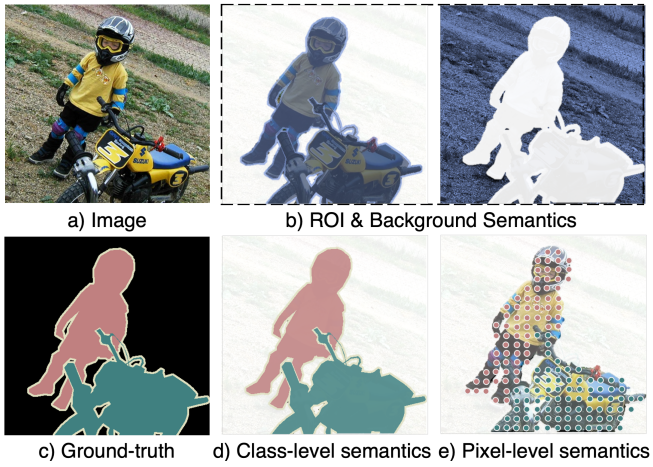

Figure 1: Motivating example of images with **multiple semantic hierarchies**. (a) Image; (b) Coarse-grained semantic hierarchy: ROI and background semantics; (c) Ground-truth; (d) Medium-grained semantic hierarchy: class-level semantics; (e) Fine-grained semantic hierarchy: pixel-level semantics.

In absence of pixel-wise annotations, the supervision signal only involves the existence of object classes without any locations and contours, leaving a huge supervision gap between WSSS and FSSS. One main solution is using Class Activation Map (CAM) [Zhou *et al.*, 2016], which reveals object regions through the internal activations of the classifier to generate pseudo pixel-level supervision. However, the CAMs are prone to cover the most discriminative regions of objects (under-activation) or activate irrelevant regions (over-

*Corresponding Author. This work was supported in part by Shanghai Science and Technology Committee under grant No. 21511100600 and No. 22511106000. We appreciate Shanghai Engineering Research Center of Intelligent Computing System for providing the computing resources.

activation), affecting the segmentation performance. Previous studies design various methods to improve CAM in different ways, e.g., region growing [Kolesnikov and Lampert, 2016; Huang *et al.*, 2018; Wang *et al.*, 2018], adversarial erasing [Kumar Singh and Jae Lee, 2017; Wei *et al.*, 2017; Hou *et al.*, 2018], and auxiliary saliency supervision [Zeng *et al.*, 2019; Oh *et al.*, 2021; Wu *et al.*, 2021].

Recently, driven by potent contrastive learning paradigms, several representation-based WSSS algorithms have been proposed [Du *et al.*, 2022; Zhou *et al.*, 2022]. They design different learning tasks on a single semantic level, such as pixel and region level, and manifest noticeable performance improvement. However, one prior knowledge is that **complex scenes are composed of multiple hierarchies of semantics** [Lu *et al.*, 2021]. For instance, from the coarse-grained level, Figure 1a can be separated into region-of-interests (ROI) and background (Figure 1b). Then, ROI can be decomposed to "*person*" and "*motorbike*" (Figure 1d), which further consist of more fine-grained components like "*face*" and "*body*", and "*handlebar*" and "*wheel*" (Figure 1e). For precise CAMs, the following semantics should be distinguishable from each other: 1) ROI and background (ROI level); 2) the objects of each class (class level); and 3) the disjoint pixels of objects across classes (pixel level). Only implementing single-level contrastive learning may hinder the model from learning a more comprehensive semantic structure where the semantics are self-consistent simultaneously on the ROI, class, and pixel levels. To mitigate the above problem, we propose **hierarchical semantic contrast (HSC)** in lieu of single-level semantic contrastive learning to exploit different levels of semantic relations, aiming at making the model learn better object pattern understanding and more precise CAM inference.

Specifically, in the training stage, HSC infers coarse CAMs to generate pixel-wise pseudo-labels. Since pixel-level semantic contrast is susceptible to noises (i.e., false pseudo-labels), we apply dense Conditional Random Field (CRF) to refine these pseudo-labels. Then, based on the activations of the CAMs and refined pseudo-labels, HSC implements semantic contrast in the embedding space from the prospects of three semantic hierarchies, i.e., ROI level, class level, and pixel level. Considering the semantic consistency under different views (transformations) of each image, HSC establishes a siamese network structure that adopts cross supervision on hierarchical semantic contrast, where one view of representations serves as the additional semantic supervision for the other view. Moreover, to utilize abundant semantics in the same hierarchy (e.g., class-level representations of *person* with various characteristics across different images), we establish momentum prototypes to complement more holistic and accurate representations for class-level and pixel-level semantic contrast. As the training goes, these prototypes merge representations that belong to the same hierarchy but across different images, and participate in contrastive learning as additional supervision. The results of extensive experiments demonstrate the superiority of our proposed approach.

Collectively, the main contributions of this paper can be summarized as follows:

- We propose hierarchical semantic contrast (HSC) for WSSS, which is motivated by the prior knowledge of semantic hierarchies harbored in each image. It involves three levels of semantic contrast and aims to learn more abundant semantic relations, enabling the model to earn better object pattern understanding and generate more precise CAMs. Compared to the prior art with single-level contrast, PPC [Du *et al.*, 2022], the CAMs of HSC have more complete objects and precise boundaries without any saliency supervision.
- The siamese architecture with cross supervision and momentum prototype learning effectively couples hierarchical semantic contrast and further improves the quality of CAM inference. Moreover, we prove that CRF in the training stage can refine pseudo labels from coarse CAMs for better representation learning, especially in the noise-sensitive pixel-level semantic contrast.
- Our approach supports plug-and-play in existing WSSS models. We demonstrate its effectiveness on both non-saliency-guided and saliency-guided baseline (SEAM [Wang *et al.*, 2020] and EPS [Lee *et al.*, 2021c]). The performance of their variants surpasses corresponding vanilla versions by large margins. With EPS, we record new state-of-the-art on PASCAL VOC 2012.

2 Related Work

2.1 WSSS with Image-level Labels

WSSS with image-label is a promising task that significantly alleviates the reliance on large numbers of pixel-wise annotations for training segmentation models. The convention of WSSS can be divided into two steps. Step 1 is to train a classification model and use its CAMs to generate pseudo-labels. Step 2 is to use these labels to train a segmentation model. Since CAM can only highlight the discriminative regions and thus hardly cover complete object regions, SEC [Kolesnikov and Lampert, 2016] proposes three principles to refine CAM, i.e., seed, expand, and constrain, which are followed by many subsequent works. They can be summarized as region growing [Kolesnikov and Lampert, 2016; Huang *et al.*, 2018; Wang *et al.*, 2018], incorporating self-supervised learning [Shimoda and Yanai, 2019; Wang *et al.*, 2020], adversarial erasing [Kumar Singh and Jae Lee, 2017; Wei *et al.*, 2017; Hou *et al.*, 2018], and auxiliary supervision [Zeng *et al.*, 2019; Oh *et al.*, 2021; Wu *et al.*, 2021]. These past efforts mainly focus on images individually, ignoring the abundant semantics in the images.

2.2 WSSS with Semantic-level Supervision

To utilize the semantics in the images, early works dedicate to exploit the semantic affinity to refine the pseudo-labels [Ahn and Kwak, 2018; Ahn *et al.*, 2019] or compute the semantic co-attention between pairs of images to obtain more consistent and integral CAM regions [Fan *et al.*, 2020; Sun *et al.*, 2020]. With the advance in contrastive learning, several current studies attempt to perform dense contrastive learning to improve object localization ability and obtain better CAMs. RCA [Zhou *et al.*, 2022] establishes memory banks for each class and applies class-level semantic contrast and aggregation on the dataset level. PPC [Du *et al.*, 2022]

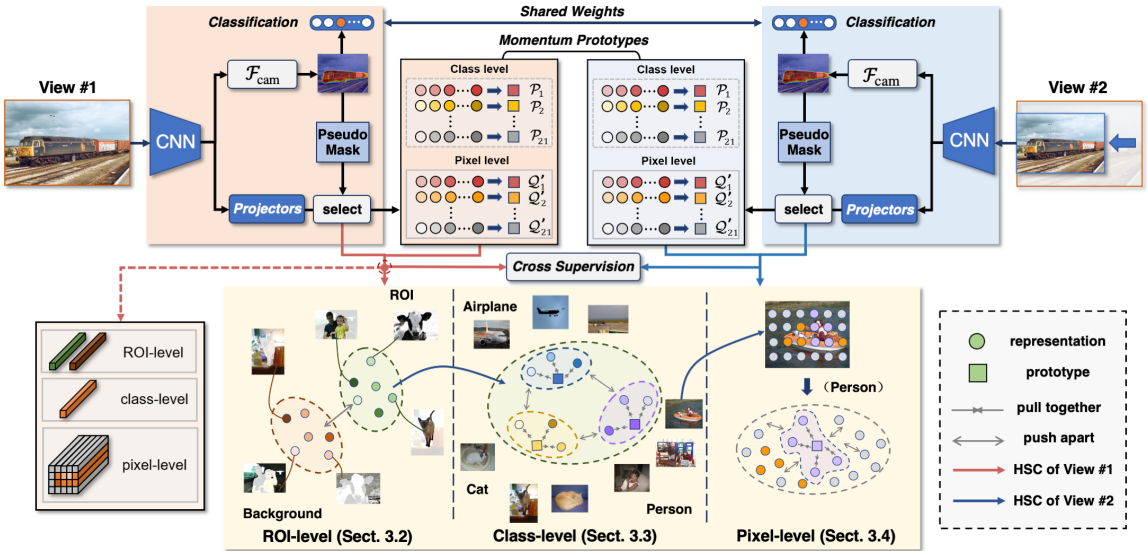

Figure 2: The **overview** of our proposed Hierarchical Semantic Contrast framework for WSSS. It adopts a siamese architecture with two views of inputs. HSC is implemented on **both** views of semantics. The hierarchical semantics and momentum prototypes of two views are applied cross supervision, where one view's hierarchical semantics and momentum prototypes serve as auxiliary supervision signals for another view.

introduces consistency regularization and proposes pixel-to-prototype learning based on [Wang *et al.*, 2021], constraining intra(inter)-class compactness(dispersion) in the feature space. In contrast, our work proposes hierarchical semantic contrast paradigm to build semantic relations simultaneously on ROI, class, and pixel levels, learning a comprehensive semantic structure for CAM inference.

3 Methodology

Our proposed HSC is implemented in Step 1 of WSSS to make the model generate more accurate CAMs. Building upon classification task ($\mathcal{L}_{\text{cls}}$), HSC can be interpreted as an auxiliary learning task, whose overall loss function for training is the linear combination of the ROI-level contrastive loss $\mathcal{L}_{\text{rsc}}$, class-level contrastive loss $\mathcal{L}_{\text{csc}}$, and pixel-level contrastive loss $\mathcal{L}_{\text{psc}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \gamma\left(\mathcal{L}_{\text{rsc}} + \mathcal{L}_{\text{csc}} + \mathcal{L}_{\text{psc}}\right), \tag{1}$$

where γ is a positive weight to rescale the loss of HSC. After generating our pseudo-masks, Step 2 is to use them to train DeepLab-ASPP [Chen *et al.*, 2017] segmentation model following common practice in WSSS. The framework overview of HSC is illustrated in Figure 2.

3.1 Preliminary

Class Activation Map

To generate CAM, the first step is to train a multi-label classification CNN (e.g., ResNet-38 [Wu *et al.*, 2019]). Given a batch of N images as input, its output of the last convolutional layer is $\boldsymbol{f} \in \mathbb{R}^{N\times D\times HW}$, where D is the channel dimension, and HW is the spatial size of $\boldsymbol{f}$. Then, $\boldsymbol{f}$ is aggregated to C channels by a 1×1 convolutional layer $\mathcal{F}_{\text{cam}}$:

$$\boldsymbol{f}' = \mathcal{F}_{\text{cam}}(\boldsymbol{f}) \in \mathbb{R}^{N\times C\times HW}. \tag{2}$$

Here, C is the number of total classes. Finally, global average pooling (GAP) is connected after $\mathcal{F}_{\text{cam}}$ to retrieve the final classification scores. In the above process, $\boldsymbol{f}'$ followed by ReLU function is theoretically equivalent to the CAMs of C classes [Zhang *et al.*, 2018]. As it is a more efficient way to compute CAM during forward propagation, we follow this manner in our paper.

Feature Representations

We set projectors to map original $\boldsymbol{f}$ to the embedding space for semantic contrast. Different from pixel-level semantics, ROI and class levels are more abstract semantic hierarchies. Therefore, we set a projector $\mathcal{F}_{\text{proj1}}$ for ROI-level and class-level semantic contrast and a projector $\mathcal{F}_{\text{proj2}}$ for pixel-level semantic contrast, respectively. $\mathcal{F}_{\text{proj1}}$ comprises two projection layers with convolution operation and activation to map $\boldsymbol{f}$ to representation $\boldsymbol{u}_1$, and $\mathcal{F}_{\text{proj2}}$ comprises one simple projection layer, mapping $\boldsymbol{f}$ to representation $\boldsymbol{u}_2$. It is worth noticing that the channel dimension of $\boldsymbol{u}_1$ is larger than $\boldsymbol{u}_2$, aiming at accommodating more ROI-level and class-level representation information. The details of this part can be viewed in Section 4.2.

Momentum Prototype Learning

Considering the semantic information of the same hierarchy is unidentical in different images (e.g., *cats* with diversified characteristics), we set momentum prototype learning to complement the holistic semantics for class-level and pixel-level contrast inspired by [Zhang *et al.*, 2021]. These momentum prototypes update via Exponential Moving Average (EMA) as training goes. For both class-level and pixel-level hierarchies, we set one momentum prototype for each class, which is updated by representations with high activations. Their momentum prototype learning and updating criteria will be detailed in Section 3.3 and Section 3.4, respectively.

Cross Supervision

Considering that the semantics of an image under different views (transformations) should be consistent [Wang *et al.*, 2020], we impose semantic consistency on the two views in our siamese architecture, where the representations of each hierarchy from one view can act as a supervisory signal for the other view, and vice versa. This design implicitly achieves consistency regularization to improve the robustness of semantic structure against various image scales. In HSC, each hierarchy's representations and momentum prototypes are involved in cross supervision. For simplicity, Section 3.2-3.4 only describe HSC from one view.

3.2 ROI-level Semantic Contrast (RSC)

Given an image, if one CAM can cover a more complete and precise area of ROI, its ROI and background representations should contain sufficiently different semantics, and thus having a large distance in the embedding space. Further, in a batch of images, their pairs of ROI and background representations should also be distinct. Based on this insight, we propose to push apart the ROI and background representations across images. Note that in the current hierarchy, we do not encourage the model to pull representations with the same property together (as adopted in many self-supervised contrastive learning paradigms), because the semantics of ROI and background in each image are unique in most cases. The pulling behavior will cause the collapse of the embedding space, deteriorating the quality of the CAM.

Representation of ROI and Background

For the i-th image I in one batch and its representation $\boldsymbol{u}_1^i$, we first conduct pixel-wise argmax function on its CAM to assign the category for each pixel, and generate the pseudo binary mask of its ROI and background. Thereinto, if one pixel p belongs to any non-background class, it will be marked in its ROI mask $\boldsymbol{m}_o^i \in \mathbb{R}^{HW}$, otherwise it will be marked in its background mask $\boldsymbol{m}_b^i \in \mathbb{R}^{HW}$. Guided by $\boldsymbol{m}_o^i$ and $\boldsymbol{m}_b^i$, we select their corresponding dense representations from $\boldsymbol{u}_1^i$, and then aggregate them into a pair of ROI and background representations as follows:

$$\boldsymbol{r}_w^i = \frac{\sum_{p\in I} \boldsymbol{m}_w^i(p)\boldsymbol{u}_1^i(p)}{\sum_{p\in I} \boldsymbol{m}_w^i(p)},\ w \in \{o, b\}, \tag{3}$$

where $\boldsymbol{r}_o^i$ and $\boldsymbol{r}_b^i$ is the i-th image's representation of the ROI $\boldsymbol{r}_o$ and that of background $\boldsymbol{r}_b$, respectively.

ROI-level Semantic Contrast

Given a batch of N images, the ROI-level semantic contrast applies the following formulation:

$$\mathcal{F}_{\text{rsc}}\left(\boldsymbol{r}_o, \boldsymbol{r}_b\right) = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log\left(1-\text{sim}(\boldsymbol{r}_o^i, \boldsymbol{r}_b^j)\right), \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity between pair of representations. Finally, the ROI-level semantic contrast with cross supervision is calculated by:

$$\mathcal{L}_{\text{rsc}} = \mathcal{F}_{\text{rsc}}\left(\boldsymbol{r}_o, \boldsymbol{r}_b\right) + \mathcal{F}_{\text{rsc}}\left(\boldsymbol{r}_o, \tilde{\boldsymbol{r}}_b\right), \tag{5}$$

where $\tilde{r}_b$ is the background representations of the other view.

3.3 Class-level Semantic Contrast (CSC)

Due to lacking pixel-wise annotations for the model, the CAMs of some classes probably suffer from over-activation or under-activation. In this case, some class-level representations are ambiguous because they are coupled with the representations of other classes. Therefore, the goal of class-level semantic contrast is to mine high-quality representations of each class from the entire dataset, utilizing them to calibrate the model to learn more precise representations of each class while pushing them apart from other classes' representations.

Class-level Prototype Mining

The class-level momentum prototype will be updated by the representations of pixels with high CAM values. Given a batch of N images, we first aggregate the representations of each class. Suppose $\boldsymbol{M}_c$ is the collection of the indices of the images with class c in the current batch, then the class-level representation of c will be:

$$\boldsymbol{r}_c = \frac{1}{|\boldsymbol{M}_c|}\sum_{i\in\boldsymbol{M}_c}\frac{\sum_{p\in\boldsymbol{A}_c^i}\boldsymbol{w}^i(p)\boldsymbol{u}_1^i(p)}{\sum_{p\in\boldsymbol{A}_c^i}\boldsymbol{w}^i(p)}, \tag{6}$$

where $|\cdot|$ represents the cardinality, $\boldsymbol{A}_c^i$ is the collection of pixels with top k_c activation values of class c at the i-th image, and each pixel p in the i-th image has its activation value of $\boldsymbol{w}^i(p)$ and representation $\boldsymbol{u}_1^i(p)$. Then, if c appears in the current batch, its momentum prototype $\mathcal{P}_c$ will be updated by:

$$\mathcal{P}_c \leftarrow \lambda\cdot\boldsymbol{r}_c + (1-\lambda)\cdot\mathcal{P}_c, \quad \text{if } c\in\mathcal{C}, \tag{7}$$

where λ is the momentum for the prototype update and $\mathcal{C}$ is the collection of the classes involved in the current batch.

Class-level Semantic Contrast

Given class c involved in the current batch and the prototpyes $\mathcal{P} = \{\mathcal{P}_c\}_{c=1}^{C}$, we pull $\boldsymbol{r}_c$ and its corresponding momentum prototype together, and meanwhile push apart other momentum prototypes via Info-NCE loss [Oord *et al.*, 2018]:

$$\mathcal{F}_{\text{csc}}\left(\boldsymbol{r}_c, \mathcal{P}\right) = \\ -\log\cdot\frac{e^{\text{sim}(\boldsymbol{r}_c,\mathcal{P}_c)/\tau}}{e^{\text{sim}(\boldsymbol{r}_c,\mathcal{P}_c)/\tau}+\sum_{\mathcal{P}_l\in\mathcal{P}\setminus\mathcal{P}_c}e^{\text{sim}(\boldsymbol{r}_c,\mathcal{P}_l)/\tau}}, \tag{8}$$

where $\tau = 0.1$ is a sharpening temperature used to reconcile the distribution of distances. Finally, the class-level semantic contrast is calculated with cross supervision as:

$$\mathcal{L}_{\text{csc}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\left[\mathcal{F}_{\text{csc}}\left(\boldsymbol{r}_c, \mathcal{P}\right) + \mathcal{F}_{\text{csc}}\left(\boldsymbol{r}_c, \widetilde{\mathcal{P}}\right)\right], \tag{9}$$

where $\widetilde{\mathcal{P}}$ denotes the class-level momentum prototpyes of the other view.

3.4 Pixel-level Semantic Contrast (PSC)

A recent study, PPC [Du *et al.*, 2022], has proposed pixel-to-prototype contrast for WSSS. Its training objective is to build pixel-level supervision from the reliable prototype of each class to narrow the gap between WSSS and FSSS. Based on its three contrast learning paradigms (i.e., *cross-view* contrast,

cross-CAM contrast, and *intra-view* contrast), we further improve this way by adding CRF refinement and Momentum prototype learning as conducted in CSC. The brief introduction of the above contrast learning paradigms and their corresponding improvements are described as follows.

CRF Refinement

Since each pixel will be involved in pixel-level semantic contrast, the validity of pseudo-labels generated by CAMs is important. However, we observe that the quality of pseudo-labels is not ideal in many cases, such as complex scenes and overlapping objects. If one pixel's pseudo-label is misclassified, its corresponding representation will be conducted PSC with undesirable prototypes, which degrades the quality of PSC. Consequently, before implementing PSC, we apply CRF to refine generated pseudo-labels: $y' = \mathrm{CRF}(y, I)$, where I denotes the whole image and y denotes its pseudo-label. The pseudo-labels and corresponding pixel representations are down-sampled before PSC, and thus CRF does not bring much computation load.

Pixel-level Prototype Mining

For each class c, its pixel-level prototype $\mathcal{Q}_c$ is produced from the representations of highly activated pixels that are assigned to c in the current batch:

$$\mathcal{Q}_c = \frac{\sum_{p \in \boldsymbol{B}_c} \boldsymbol{w}(p)\boldsymbol{u}_2(p)}{\sum_{p \in \boldsymbol{B}_c} \boldsymbol{w}(p)}, \tag{10}$$

where $\boldsymbol{B}_c$ is the collection of top k_p activated pixels in class c for the current batch, and each pixel p has activation value of $\boldsymbol{w}(p)$ and representation $\boldsymbol{u}_2(p)$. Then, class c's momentum prototype $\mathcal{Q}_c^{'}$ will be:

$$\mathcal{Q}'_c \leftarrow \lambda \cdot \mathcal{Q}_c + (1-\lambda) \cdot \mathcal{Q}'_c. \tag{11}$$

Pixel-level Semantic Contrast

Followed by PPC, given a pixel p and the prototypes $\mathcal{Q} = \{\mathcal{Q}_c\}_{c=1}^{C}$, the pixel-level semantic contrast $\mathcal{F}_{\mathrm{psc}}(\cdot)$ holds the following formulation:

$$\mathcal{F}_{\mathrm{psc}}\left(p, \boldsymbol{u}_2, \boldsymbol{y}', \mathcal{Q}\right) = -\log \frac{\exp\left(\boldsymbol{u}_2(p) \cdot \mathcal{Q}_{\boldsymbol{y}'(p)}/\tau\right)}{\sum_{\mathcal{Q}_c \in \mathcal{Q}} \exp\left(\boldsymbol{u}_2(p) \cdot \mathcal{Q}_c/\tau\right)}, \tag{12}$$

where $\boldsymbol{y}'(p)$ denotes the refined pseudo-label of p and $\mathcal{Q}_{\boldsymbol{y}'(p)}$ is its corresponding prototype.

For the *intra-view* contrast, it is not involved supervision signals from the other view, where pseudo-labels $\boldsymbol{y}'$, representation $\boldsymbol{u}_2$, and prototypes $\mathcal{Q}$ are derived from the same view. For the *cross-view* contrast, the prototypes of the current view are replaced with the prototypes $\tilde{\mathcal{Q}}$ from the other view. For the *cross-CAM* contrast, the pseudo-labels of the current view are replaced with the pseudo-labels $\tilde{\boldsymbol{y}}'$ from the other view. For our proposed *momentum prototype learning*, we replace the original prototypes $\mathcal{Q}$ with the momentum prototypes $\mathcal{Q}'$ to implement intra-view contrast. Finally, given one whole image I, the pixel-level semantic contrast is calculated as:

$$\begin{aligned}\mathcal{L}_{psc} = & \frac{1}{|I|} \sum_{p \in I} \big[\mathcal{F}_{\mathrm{psc}}(p, \boldsymbol{u}_2, \boldsymbol{y}', \mathcal{Q}) + \mathcal{F}_{\mathrm{psc}}(p, \boldsymbol{u}_2, \boldsymbol{y}', \tilde{\mathcal{Q}}) \\ & + \mathcal{F}_{\mathrm{psc}}(p, \boldsymbol{u}_2, \tilde{\boldsymbol{y}}', \mathcal{Q}) + \mathcal{F}_{\mathrm{psc}}(p, \boldsymbol{u}_2, \boldsymbol{y}', \mathcal{Q}')\big].\end{aligned} \tag{13}$$

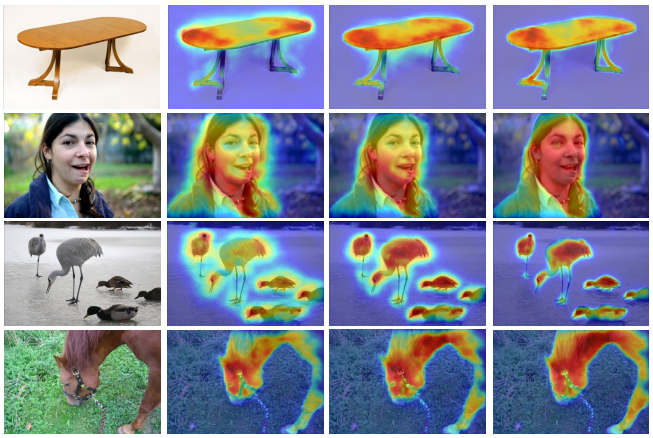

Figure 3: The visualization of **original CAMs**. (a) Images; (b) CAMs produced by SEAM; (c) CAMs produced by PPC (with pixel-to-prototype semantic contrast); (c) CAMs produced by our proposed HSC (with hierarchical semantic contrast). Our approach performs better in terms of object completeness and boundary precision.

4 Experiment

4.1 Dataset and Evaluation Metric

We evaluate our approach on the standard WSSS benchmark, PASCAL VOC 2012 (20 object classes and one background class). Following the convention in semantic segmentation, we adopted its augmented training set (SBD) [Hariharan *et al.*, 2011] that consists of 10582 images to train our classification model in step 1 and the segmentation model in step 2. Mean intersection over union (mIoU) is used to evaluate initial seeds, pseudo-labels, and final segmentation performance. The mIoU of segmentation performance on the *test* set is evaluated from the official evaluation server.

4.2 Implementation Details

Following the baseline, SEAM and EPS, ResNet38 [Wu *et al.*, 2019] with $\mathrm{output\_stride} = 8$ is adopted as the backbone network. For our siamese network architecture, the backbone and the projectors share weights between the two views. The projector $\mathcal{F}_{\mathrm{proj1}}$ is equipped with two $\mathrm{Conv\text{-}ReLU}$ blocks, mapping 4096-dimensional features to 256-dimensional embedding space, and the projector $\mathcal{F}_{\mathrm{proj2}}$ is equipped with one $\mathrm{Conv\text{-}ReLU}$ block, mapping 4096-dimensional features to 128-dimensional embedding space. For the training images of View #1, they are first randomly rescaled with the range of $[448,\ 768]$ by the longest edge and then randomly cropped by 448×448. Based on the images of View #1, they are downsampled to 128×128 for View #2. When integrating our approach to SEAM and EPS, we follow the same training and inference protocol in SEAM and EPS, including the training epoch, optimizer, learning rate, learning rate decay strategy, and weight decay. At the training stage, we set $\gamma = 0.1$ to balance the loss of $\mathcal{L}_{\mathrm{hsc}}$ and their supervision loss. At the inference stage, we adopt multi-scale inference with flipping transformation as conducted in previous works.

Method	Seed	+CRF	Masks
PSA [CVPR' 18] [Ahn and Kwak, 2018]	48.0	-	61.0
CONTA [NIPS' 20] [Zhang *et al.*, 2020]	56.2	65.4	66.1
EDAM [CVPR' 21] [Wu *et al.*, 2021]	52.8	58.2	68.1
AdvCAM [CVPR' 21] [Lee *et al.*, 2021a]	55.6	62.1	68.0
OC-CSE [ICCV' 21] [Kweon *et al.*, 2021]	56.0	62.8	66.9
ECS-Net [ICCV' 21] [Sun *et al.*, 2021]	56.6	58.6	-
Ru et al. [IJCAI' 21] [Ru *et al.*, 2021]	52.9	52.0	67.7†
PPC (w/ SEAM) [CVPR' 22] [Du *et al.*, 2022]	61.5	64.0	69.2
RCA (w/ EPS) [CVPR 22] [Zhou *et al.*, 2022]	-	74.1	-
PPC (w/ EPS) [CVPR' 22] [Du *et al.*, 2022]	70.5	73.3	73.3
SEAM (w/o saliency) [CVPR' 20]	55.4	56.8	63.6
Ours (w/ SEAM)	**64.3**	**66.5**	**69.5**
EPS (w/ saliency) [CVPR' 21]	69.5	71.4	71.6
Ours (w/ EPS)	**71.8**	**74.6**	74.6*

Table 1: Evaluation (mIoU (%)) of the initial seed (Seed), the seed after CRF (+ CRF), and the final pseudo-labels (Masks) refined by PSA on PASCAL VOC 2012 $train$ set. † means the pseudo-labels are refined by IRN [Ahn *et al.*, 2019]. * means the pseudo-labels are adopted the seeds with CRF without any refinement networks.

Method	BB.	Sup.	val	test
PSA [CVPR' 18] [Ahn and Kwak, 2018]	R38	$\mathcal{I}$	61.7	63.2
IRNet [CVPR' 19] [Ahn *et al.*, 2019]	R50	$\mathcal{I}$	63.5	64.8
SEAM [CVPR' 20] [Wang *et al.*, 2020]	R38	$\mathcal{I}$	64.5	65.7
CONTA [NIPS' 20] [Zhang *et al.*, 2020]	R101	$\mathcal{I}$	66.1	66.7
Ru et al. [IJCAI' 21] [Ru *et al.*, 2021]	R101	$\mathcal{I}$	67.2	67.3
EDAM [CVPR' 21] [Wu *et al.*, 2021]	R101	$\mathcal{I}+\mathcal{S}$	70.9	70.6
AdvCAM [CVPR' 21] [Lee *et al.*, 2021a]	R101	$\mathcal{I}$	68.1	68.0
URN [AAAI' 22] [Li *et al.*, 2022]	R101	$\mathcal{I}$	69.5	69.7
AMR [AAAI' 22][Qin *et al.*, 2022]	R101	$\mathcal{I}$	68.8	69.1
ReCAM [CVPR' 22] [Chen *et al.*, 2022]	R101	$\mathcal{I}$	68.5	68.4
MCTformer [CVPR' 22] [Xu *et al.*, 2022]	R38	$\mathcal{I}$	71.9	71.6
RCA [CVPR' 22] [Zhou *et al.*, 2022]	R101	$\mathcal{I}+\mathcal{S}$	72.2	72.8
PPC [CVPR' 22] [Du *et al.*, 2022]	R101	$\mathcal{I}+\mathcal{S}$	72.6	73.6
EPS [Lee *et al.*, 2021c]	R101	$\mathcal{I}+\mathcal{S}$	70.9	70.8
Ours (w/ EPS)	R101	$\mathcal{I}+\mathcal{S}$	**73.6**	**74.5**†

Table 2: Evaluation (mIoU (%)) of WSSS methods on PASCAL VOC 2012 val and $test$. "BB.": Backbone (R-ResNet); "Sup.": Supervision; "$\mathcal{I}$": Image-level class labels; "$\mathcal{S}$": Saliency supervision. †: http://host.robots.ox.ac.uk/anonymous/11DHLZ.html

CRF is implemented on the initial seeds as a post-processing procedure to refine the pseudo-labels. For the segmentation network, we train DeepLab-ASPP [Chen *et al.*, 2017] with ResNet101 backbone using the pseudo-labels generated from our approach with EPS. More detailed settings are available in our Appendix.

4.3 Initial Seed and Pseudo Label Evaluation

Table 1 reports the segmentation performance of initial seeds, seeds after CRF, and final pseudo-labels on PASCAL VOC 2012 *train* set. Following SEAM [Wang *et al.*, 2020], our initial seeds are generated by setting a range of thresholds to separate the objects and backgrounds in the original CAM inferences. Pseudo-labels are refined by PSA [Ahn and Kwak, 2018] from seeds after CRF. As can be seen, for SEAM with non-saliency guidance, HSC dramatically enhances its performance, resulting in an 8.9% and 9.7% increase in initial seeds and seeds after CRF, respectively. Furthermore, on EPS with saliency guidance, HSC can also boost its performance on initial seeds (+2.3%) and seeds after CRF (+3.2%), achieving new state-of-the-art performance. In comparison to the methods equipped with semantic contrast paradigms (i.e., PPC and RCA), our approach achieves noticeable improvement when integrating with EPS. In terms of our refined pseudo-labels, we note that our approach with SEAM obtain slight improvement compared to seeds after CRF. Considering previous studies [Lee *et al.*, 2021c; Du *et al.*, 2022] have manifested that PSA is limited to yield benefits when seeds after CRF are already exquisite enough, we reckon that this is mainly caused by the performance bottleneck of PSA, and thus we directly use seeds after CRF as pseudo-labels when equipping HSC to stronger baseline EPS.

To investigate how HSC improves the quality of initial seeds and seeds after CRF, we present the qualitative results of CAMs generated by SEAM, PPC, and HSC, respectively. As shown in Figure 3, we observe that our approach achieves more superior performance than SEAM and PPC in terms of object completeness and boundary precision. Especially in the boundaries, the CAMs of our method cover more accurately without any saliency supervision compared to PPC with single-level semantic contrast.

4.4 Final Segmentation Performance

Table 2 provides the comparison of HSC against representative methods in terms of final segmentation results on PASCAL VOC 2012 *val* and *test* set. As seen, building upon DeepLab-ASPP, our proposed approach brings significant gains over EPS. With equipping HSC, we increase the segmentation mIoU of EPS by 2.7% on *val* set and 3.7% on *test set*, setting a new state-of-the-art on WSSS. Figure 4 illustrates some qualitative segmentation results of HSC, from which we can observe that our method works well in both simple and complex scenes.

Furthermore, compared to PPC, the prior art with single-level semantic contrast, we explore which categories have performance improvement brought by hierarchical semantic contrast. Table 3 tabulates the categorical mIoU performance comparison on *test* set. We can find that our proposed HSC achieves significant performance improvements ($1.45\% \sim 4.54\%$) in eight categories. The objects in most of these categories are relatively challenging to segment, as they often appear in complex scenes (such as multiple instances, occlusions, and objects with different sizes or various characteristics). This result demonstrates that hierarchical semantic contrast can better guide the segmentation model to handle these cases.

4.5 Ablation Study

To investigate how each component in our proposed approach contributes to WSSS, we conduct extensive ablation studies

Methods	bird	boat	bottle	chair	table	dog	plant	sheep	sofa	train	tv	mIoU
PPC	89.77	65.35	69.89	31.42	48.23	85.12	59.55	87.66	52.89	80.31	46.39	73.6
Ours	89.09 (-0.68)	63.62 (-1.73)	72.95 (+3.06)	34.15 (+2.73)	49.68 (+1.45)	87.59 (+2.47)	64.09 (+4.54)	87.14 (-0.52)	55.03 (+2.14)	82.97 (+2.66)	48.39 (+2.00)	74.5

Table 3: Catagorial performance comparison (mIoU (%)) on PASCAL VOC 2012 *test* set. Both methods use DeepLab-ASPP as the final segmentation model. The categories with performance differences greater than 0.5% are listed.

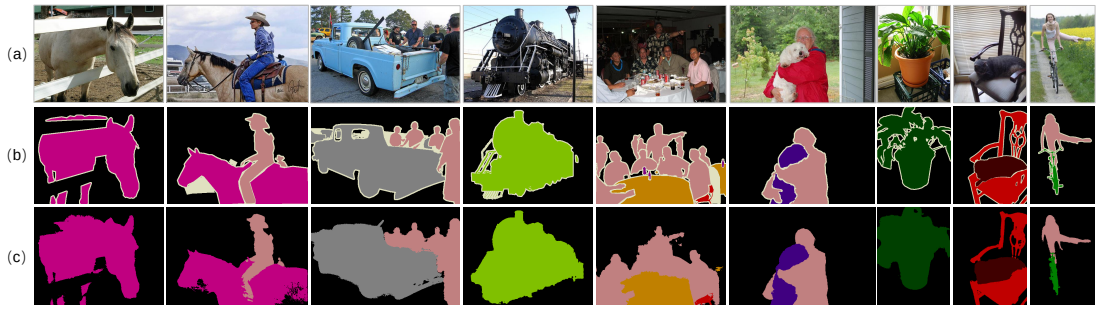

Figure 4: Qualitative segmentation results of HSC on PASCAL VOC 2012 *val* set. (a) Images; (b) Ground-truth; (c) Segmentation results predicted by DeepLab-ASPP model retrained on our produced pseudo-labels.

Baseline	PSC	CSC	RSC	CS	MPL	CRF †	CRF	Seed
✓								55.4
✓	✓			✓				59.3
✓	✓	✓		✓				60.5
✓	✓	✓	✓	✓				62.6
✓	✓	✓	✓					60.7
✓	✓	✓	✓	✓	✓			63.5
✓	✓	✓	✓	✓	✓	✓		64.2
✓	✓	✓	✓	✓	✓		✓	**64.3 (+8.9)**

Table 4: The ablation study (mIoU (%)) for each part of HSC. **PSC**: Pixel-level semantic contrast; **CSC**: Class-level semantic contrast; **RSC**: ROI-level semantic contrast; **CS**: Cross Supervision; **MPL**: Momentum Prototype Learning in PSC; **CRF**†: CRF Refinement in PSC. **CRF**: CRF Refinement in complete HSC.

in this section. Here, all ablation analyses are implemented with SEAM baseline on PASCAL VOC 2012 *train* set. The results are presented in Table 4.

Hierarchical Semantic Contrast. We investigate the necessity of learning hierarchical semantic contrast for WSSS. Based on cross supervision, each level's semantic contrast paradigm provides a performance boost, improving the baseline performance from 55.4% to 62.6% (**+7.2%**). This proves that our approach fulfills the goal of obtaining a better object pattern understanding by learning a holistic semantic structure across different semantic hierarchies.

Cross Supervision. It is an essential component in semantic contrast. When integrating it into HSC, we witness a noticeable performance increment from 60.7% to 62.6% (**+1.9%**). This indicates that the implicit consistency regularization can allow the model to learn the semantic consistency across different views, which effectively bridges the supervision gap between pixel-level and image-level annotations.

CRF Refinement. We note that CRF refinement on pseudo-labels works well with PSC, increasing from 63.5% to 64.2% (**+0.7%**). By contrast, it shows marginal improvement, from 64.2% to 64.3% (**+0.1%**), for CSC and RSC. This is in line with our assumption that PSC is more susceptible to false pseudo-labels. After CRF refinement, HSC can obtain better pseudo-labels to perform semantic contrast. In comparison to PSC, the representations of RSC and CSC are estimated from numerous representations, which are more robust against false pseudo-labels.

Momentum Prototype Learning. In CSC, we use momentum prototypes to calibrate the model to learn high-quality representations at class level. Further, we also introduce this paradigm to improve PSC. When MPL is integrated into PSC, the initial seeds become more precise, from 62.6% to 63.5% (**+0.9%**), indicating that the model learns holistic semantics across different images from momentum prototypes. We empirically choose the momentum λ and the prototype number k_c and k_p to implement this paradigm. In this paper, we set $\lambda = 0.9$, $k_c = 16$, and $k_p = 32$ for the optimal performance.

5 Conclusion

In this work, we propose a novel approach, HSC, to conduct semantic contrast at different levels as the auxiliary learning task to learn semantic segmentation using image-level supervision only. HSC conducts semantic contrast from fine-grained hierarchy to coarse-grained hierarchy, building a more comprehensive semantic structure in the embedding space. In comparison to representative approaches, this enables the model to have significant improvements in the object completeness and boundary precision of CAMs, which further boosts the final segmentation performance. Extensive experiments validate the effectiveness of HSC and manifest its leading performance of WSSS on PSCAL VOC 2012.

References

[Ahn and Kwak, 2018] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018.

[Ahn *et al.*, 2019] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019.

[Chen *et al.*, 2017] 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 transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[Chen *et al.*, 2022] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022.

[Du *et al.*, 2022] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022.

[Fan *et al.*, 2020] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10762–10769, 2020.

[Hariharan *et al.*, 2011] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011.

[Hou *et al.*, 2018] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *Advances in Neural Information Processing Systems*, 31, 2018.

[Huang *et al.*, 2018] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7014–7023, 2018.

[Kolesnikov and Lampert, 2016] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016.

[Kumar Singh and Jae Lee, 2017] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3524–3533, 2017.

[Kweon *et al.*, 2021] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6994–7003, 2021.

[Lee *et al.*, 2021a] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021.

[Lee *et al.*, 2021b] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2643–2652, 2021.

[Lee *et al.*, 2021c] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021.

[Li *et al.*, 2022] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1447–1455, 2022.

[Liang *et al.*, 2022] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16907–16916, 2022.

[Lin *et al.*, 2016] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

[Lu *et al.*, 2021] Xiankai Lu, Wenguan Wang, Jianbing Shen, David J Crandall, and Luc Van Gool. Segmenting objects from relational visual data. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7885–7897, 2021.

[Oh *et al.*, 2021] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Qin *et al.*, 2022] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang. Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2117–2125, 2022.

[Ru *et al.*, 2021] Lixiang Ru, Bo Du, and Chen Wu. Learning visual words for weakly-supervised semantic segmentation. In *IJCAI*, volume 5, page 6, 2021.

[Shimoda and Yanai, 2019] Wataru Shimoda and Keiji Yanai. Self-supervised difference detection for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5208–5217, 2019.

[Sun *et al.*, 2020] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *European conference on computer vision*, pages 347–365. Springer, 2020.

[Sun *et al.*, 2021] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7283–7292, 2021.

[Vernaza and Chandraker, 2017] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017.

[Wang *et al.*, 2018] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1354–1362, 2018.

[Wang *et al.*, 2020] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[Wang *et al.*, 2021] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021.

[Wei *et al.*, 2017] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

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

[Wu *et al.*, 2021] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16765–16774, 2021.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[Xu *et al.*, 2022] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022.

[Yuan *et al.*, 2020] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020.

[Zeng *et al.*, 2019] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7223–7233, 2019.

[Zhang *et al.*, 2018] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018.

[Zhang *et al.*, 2020] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

[Zhang *et al.*, 2021] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[Zhou *et al.*, 2022] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309, 2022.