Learning Weakly Supervised Semantic Segmentation Through Cross-Supervision and Contrasting of Pixel-Level Pseudo-Labels

Lucas David[a], Helio Pedrini[b] and Zanoni Dias[c]
Institute of Computing, University of Campinas, Campinas, Brazil
{*lucas.david, helio, zanoni*}@*ic.unicamp.br*

Keywords: Machine Learning, Computer Vision, Semantic Segmentation, Weak Supervision, Mutual Promotion, Contrastive Learning, Noise Mitigation.

Abstract: The quality of the pseudo-labels employed in training is paramount for many Weakly Supervised Semantic Segmentation techniques, which are often limited by their associated uncertainty. A common strategy found in the literature is to employ confidence thresholds to filter unreliable pixel labels, improving the overall quality of label information, but discarding a considerable amount of data. In this paper, we investigate the effectiveness of cross-supervision and contrastive learning of pixel-level pseudo-annotations in weakly supervised tasks, where only image-level annotations are available. We propose CSRM: a multi-branch deep convolutional network that leverages reliable pseudo-labels to learn to classify and segment a task in a mutual promotion scheme, while employing both reliable and unreliable pixel-level pseudo-labels to learn representations in a contrastive learning scheme. Our solution achieves 75.0% mIoU in Pascal VOC 2012 testing and 50.4% MS COCO 2014 validation datasets, respectively. Code available at github.com/lucasdavid/wsss-csrm.

1 INTRODUCTION

Semantic segmentation is a prominent task in Computer Vision, considering its multiple real-world applications (Mo et al., 2022). Nowadays, Deep Convolutional Networks have become the standard approach to semantic segmentation, yielding great effectiveness across different problems and domains (Chen et al., 2020). However, this success comes at the expense of large amounts of annotated data, resulting in laborious and extensive annotation work from human specialists to produce Fully Supervised Semantic Segmentation (FSSS) datasets.

In recent years, researchers have development strategies to mitigate this prominent annotation dependency, such as Semi-Supervised Semantic Segmentation (SSSS) (Zhang et al., 2020b) and Weakly Supervised Semantic Segmentation (WSSS) (Shen et al., 2023). In the former, a small subset of samples is (fully) annotated at a pixel-level, while the remaining samples are kept unlabeled. In the latter, all samples are (weakly) annotated with a degenerated form of the supervised information, such as image-level labels, bounding boxes, scribes, or points.

Approaches to SSSS problems often leverage the (otherwise wasted) unlabeled sample set in the training process by adopting pixel-level pseudo-labels, devised from visual patterns and similarity with the existing supervised set. Approaches to WSSS problems, on the other hand, create pixel-level pseudo-labels from any supervised information available, such as localization cues extracted with explaining methods (Samek et al., 2021).

The lack of quality in the pixel-level pseudo-labels is detrimental to the effectiveness in SSSS and WSSS tasks, and it is often mitigated by filtering pixels in the pseudo-label masks according to a *confidence* hyperparameter $\delta_{fg} \in [0,1]$. Pixels whose confidence is high (according to δ_{fg}) are used in training, while the remaining *unreliable* pixels are discarded.

While promising strategies for utilizing unreliable pixel-level pseudo-labels, through the contrastive learning of pixel representations, have been devised to improve the effectiveness of fully-supervised (Liu et al., 2022b) and semi-supervised (Wang et al., 2022) solutions, they still require (at least) a few fully human-annotated samples, implying in laborious human intervention. To the best of our knowledge, no work proposed thus far has studied the employment of unreliable pixel-level labels in WSSS tasks.

[a] https://orcid.org/0000-0002-8793-7300
[b] https://orcid.org/0000-0003-0125-630X
[c] https://orcid.org/0000-0003-3333-6822

David, L., Pedrini, H. and Dias, Z.
Learning Weakly Supervised Semantic Segmentation Through Cross-Supervision and Contrasting of Pixel-Level Pseudo-Labels.
DOI: 10.5220/0013238400003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 3: VISAPP*, pages 154-165
ISBN: 978-989-758-728-3; ISSN: 2184-4321

In this work, we investigate the effect of using unreliable pixel-level pseudo-labels for the learning of weakly supervised semantic segmentation problems. Our solution, namely Cross-Supervision and Relational Model (CSRM), comprises three branches: (i) the classification branch, responsible for learning the associated classification task, and produce CAMs that can be refined into coarse segmentation priors; (ii) the segmentation branch, responsible for learning the segmentation task from reliable pixel-level pseudo-labels, while regularizing the classification branch to produce better CAMs; and (iii) the representation branch, responsible for contrasting pixel-level feature representations of both reliable and unreliable regions in a metric space. We define reliability of regions based on their prediction confidence, entropy, and the image-level label information available, and extract hard queries that are compared with positive and negative anchors in a contrastive learning scheme, where unreliable regions (associated with high entropy) are pushed close together from their likely class prototype and farther away from negative (visually similar) anchors.

Our approach is inspired by recent advances in both FSSS and SSSS, but differs in key aspects from the works previously proposed (He et al., 2020; Liu et al., 2022b; Wang et al., 2022): (i) it does not rely on (however small) human-made annotation sets for building reliable class-specific representations, and (ii) it relies solely on image-level annotation information to extract pixel-level, class-specific anchors used in the contrastive learning task.

The remaining of this work is organized as follows. Section 2 introduces important concepts used in this work, as well as pertinent literature. Section 3 describes our approach in detail, while Section 4 details the training procedure, and hyperparameters employed. We present the main results in Section 5, and conclude the paper in Section 6.

2 RELATED WORK

In this section, we introduce concepts and strategies in the literature that important for the understanding of our work.

2.1 Weakly Supervised Semantic Segmentation (WSSS)

Various approaches to semantic segmentation using only image-level labels were proposed in the last decade. These are subdivided into two categories: single-stage, in which an end-to-end model capable of segmenting samples by itself is trained (Bircanoglu and Arica, 2022; Zhu et al., 2023), and multi-stage strategies (Kweon et al., 2021; Jo and Yu, 2021; David et al., 2024), involving multiple steps to train the segmentation model, often by deriving pseudo semantic segmentation labels from localization cues advent from AI explaining methods (such as CAM (Simonyan et al., 2013) and Grad-CAM (Selvaraju et al., 2017)) applied over models trained with complex regularization strategies that promote the emergence of useful properties for the segmentation task.

Mutual promotion is an example of such regularization method, and it has become an important aspect for single-stage WSSS approaches (Bircanoglu and Arica, 2022; Zhu et al., 2023). It consists of a WSSS model that has multiple branches responsible for performing concomitant tasks, leveraging the mutual information between tasks to improve the effectiveness of the associated processes. For example, a classification and a segmentation branches, in which the former provides localization cues to the latter, which, in turn, is used to regularize the first.

Though such strategies are successful, to some extent, most are strongly affected by the quality of localization cues. To alleviate this problem, WSSS strategies commonly adopt confidence thresholds δ_{fg}, retaining regions associated with “high” confidence and “low” noise, at the cost of discarding the remaining (potentially useful) data.

2.2 Self-Supervised Learning

Self-supervised learning techniques have shown great promise in modern machine learning. By learning to contrast the visual patterns of unsupervised samples, they explore consistency regularization and entropy minimization to mitigate prediction noise, thus improving model effectiveness. In semantic segmentation, various techniques have been proposed to leverage the unsupervised set in a contrastive learning setup, such as Pixel-Level Contrastive Learning (Liu et al., 2022b), Adaptive-Equalization Learning (Hu et al., 2021) and Contrastive Learning of Unreliable Pseudo Labels (Wang et al., 2022).

While our approach is similar to the aforementioned solutions, it differs by relying solely on the weakly supervised information available (image-level labels) to create pixel-level feature representations from image regions associated with both reliable and unreliable pseudo-labels, allowing for the application of self-supervised learning over WSSS tasks, in which no human-made pixel-level annotations are available.

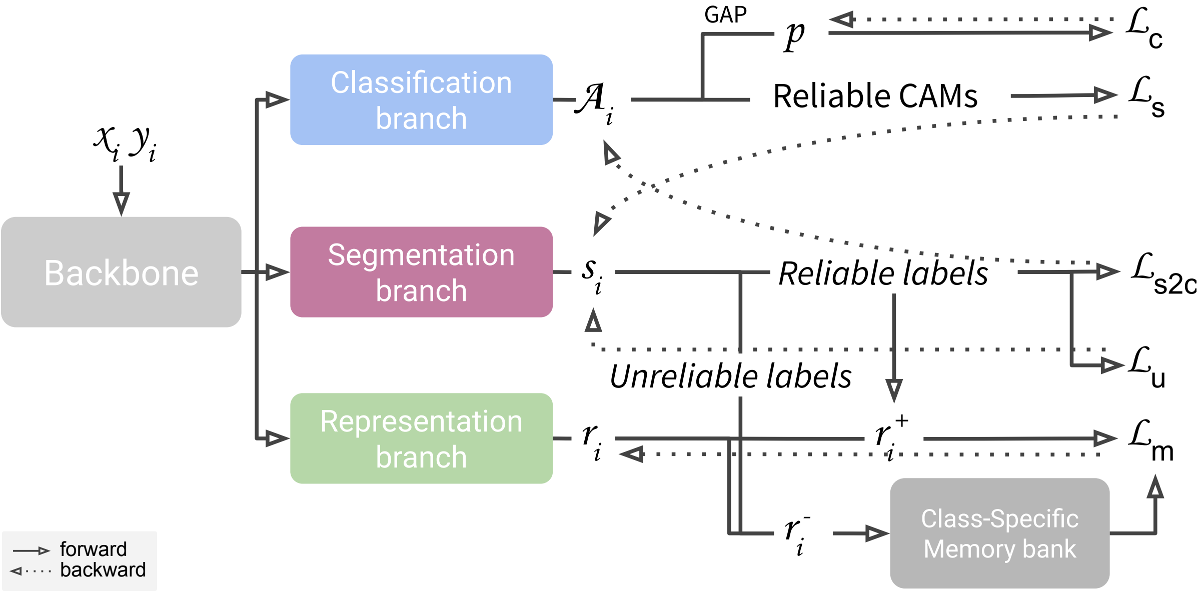

Figure 1: Overview of our CSRM approach: a single model that employs cross-supervision of the classification and semantic segmentation tasks to mutually promote their effectiveness, while learning pixel-level representations for both reliable and unreliable pixels in an unsupervised contrastive learning setup.

3 CROSS-SUPERVISION AND RELATIONAL MODEL

We propose Cross-Supervision and Relational Model (CSRM): the employment of Cross-Supervision between classification, segmentation and Contrastive Representation tasks to approach WSSS problems. Our strategy, illustrated in Figure 1, is inspired by recent success of contrastive learning approaches applied to learning better concept representations for Fully-Supervised and Semi-Supervised Semantic Segmentation problems (He et al., 2020; Liu et al., 2022b; Wang et al., 2022).

3.1 Architecture

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ be a training dataset, where $\mathbf{x}_i$ is the i-th sample image in the set, $\mathbf{y}_i = [y_i^1, \dots, y_i^{|C|}]$ is the one-hot class label vector indicating which classes are present in image $\mathbf{x}_i$, and C is the set of all classes.

Our approach consists of a single-stage network, comprising a feature extractor $\mathcal{F} : \mathbb{R}^{HW\times 3} \to \mathbb{R}^{hwk}$, where $(h,w) \ll (H,W)$; a classification branch $\mathcal{C} : \mathbb{R}^{hwk} \to \mathbb{R}^{hw\times|C|}$; a segmentation branch $\mathcal{S} : \mathbb{R}^{hwk} \to \mathbb{R}^{HW\times|C|}$; and a representation branch $\mathcal{R} : \mathbb{R}^{hwk} \to \mathbb{R}^{HW\times 256}$.

The **classification branch** consists of a 1×1 convolution layer with $|C|$ output channels, followed by a Global Average Pooling layer (GAP), and the *sigmoid* activation function. That is, the classification prediction for an image $\mathbf{x}_i$ is defined as:

$$\begin{gathered} \mathbf{A}_i \in \mathbb{R}^{hw\times|C|} \mid \mathbf{A}_{ic} = \mathcal{C}_c(\mathcal{F}(\mathbf{x}_i)) \\ \mathbf{p}_i \in [0,1]^{|C|} \mid \mathbf{p}_{ic} = \sigma(\mathrm{GAP}(\mathbf{A}_{ic})) \end{gathered} \tag{1}$$

The Class-specific Activation Map (CAM) (Selvaraju et al., 2017) of a class $c \in C$, represented by $\mathbf{A}_{ic}$, can be obtained by simply forwarding sample $\mathbf{x}_i$ onto the classification branch, and collecting the positional signal before the GAP layer.

The signal is upscaled to match the original sizes of the input, and a normalization function $\psi : \mathbb{R} \to [0,1]$ is further adopted to transform the CAM into a probability map:

$$\psi(\mathbf{A}_{ic}) = \frac{\mathrm{ReLU}(\mathrm{upscale}(\mathbf{A}_{ic}))}{\max_{ab\in HW} \mathrm{ReLU}(\mathrm{upscale}(\mathbf{A}_{iabc}))} \tag{2}$$

We employ the DeepLabV3+ (Chen et al., 2018) decoder as the **segmentation branch**, considering its well-established organization and extensively asserted effectiveness over multiple fully-supervised semantic segmentation tasks. This branch predicts segmentation maps for $|C|+1$ classes (the original C classes and the *background*), which is missing in the classification task.

We define the segmentation prediction for a samples $\mathbf{x}_i$ as:

$$\begin{gathered} \mathbf{S}_i \in \mathbb{R}^{HW\times|C|+1} \mid \mathbf{S}_i = \mathcal{S}(\mathcal{F}(\mathbf{x}_i)) \\ \mathbf{s}_i \in [0,1]^{HW\times|C|+1} \mid \mathbf{s}_{ic} = \mathrm{softmax}_c(\mathbf{S}_i) \end{gathered} \tag{3}$$

Finally, the **representation branch** is similar to the segmentation one, except for the last convolution layer, which maps each pixel to a vector representation in a 256-dimensional feature space:

$$\mathbf{R}_i \in \mathbb{R}^{HW\times 256} \mid \mathbf{R}_i = \mathcal{R}(\mathcal{F}(\mathbf{x}_i)) \tag{4}$$

3.2 Training and Objective Functions

The objective functions employed when training each component of CSRM are detailed as follows.

3.2.1 Classification Loss

The classification branch is refined (at a lower learning rate) to perform a multi-label classification task with the *multi-label soft-margin* loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{c}}(\mathbf{p}_i^s, \mathbf{y}_i) = -\frac{1}{C}\sum_c y_{ic}\log((1+e^{-p^s_{ic}})^{-1})\\ + (1-y_{ic})\log(e^{-p^s_{ic}}/(1+e^{-p^s_{ic}}))\end{aligned} \tag{5}$$

3.2.2 Segmentation Loss

Pseudo semantic segmentation labels $\mathbf{l}_i^t$ are devised from CAMs (extracted from the classification branch of the teacher model), and refined with dCRF (Krähenbühl and Koltun, 2011). Confident foreground and background regions are extracted considering thresholds δ_{fg} and δ_{bg}, respectively. The remaining pixels, whose intensity fall between δ_{bg} and δ_{fg}, are marked as "uncertain" and ignored:

$$\begin{aligned}\mathbf{l}_i^t &= \mathrm{dCRF}(\psi(\mathbf{A}_i^t))\\ \mathbf{M}^p(\mathbf{l}_i^t) &= \mathbb{1}\left[\max_{c\in C}\mathbf{l}_i^{tc} \notin (\delta_{\mathrm{bg}}, \delta_{\mathrm{fg}}]\right]\end{aligned} \tag{6}$$

where $(\delta_{\mathrm{bg}}, \delta_{\mathrm{fg}}]$ is the interval of "uncertainty", and $\mathbf{M}^p(\mathbf{l}_i^t)$ is the binary mask matrix with dimensions (H,W), indicating the reliability of every pixel in the pseudo-labels map $\mathbf{l}_i^t$.

The *categorical cross-entropy* loss function is used to train the student network to match its segmentation output signal to the pseudo-masks:

$$\mathcal{L}_{\mathrm{s}}(\mathbf{s}_i^s, \mathbf{l}_i^t) = -\frac{1}{|\mathbf{M}^p(\mathbf{l}_i^t)|}\sum_{hw}^{HW}\mathbf{M}^p_{hw}(\mathbf{l}_i^t)\sum_c^C \mathbf{l}^t_{ihwc}\log\mathbf{s}^s_{ihwc} \tag{7}$$

3.2.3 Activation Consistency Loss

To complete the mutual promotion scheme, we regularize predictions from the classification branch of the student with supervision from the segmentation branch $\mathcal{S}^t$ of the teacher, which guides the model to produce more organized activation signals, ultimately culminating in better CAMs.

Firstly, we define a threshold hyperparameter σ_{s2c} to create a binary mask $\mathbf{M}^s(\mathbf{s}_i^t)$, indicating which pixels in the teacher's segmentation map were predicted with high confidence (and low entropy):

$$\mathbf{M}^s(\mathbf{s}_i^t) = \mathbb{1}\left[\max_{c\in C}\mathbf{s}^t_{ic} > \sigma_{\mathrm{s2c}}\right] \tag{8}$$

We align the activation signal $\mathbf{A}_i^s$ with $\mathbf{s}_i^t$ by resizing it to the original input size, and concatenating it to the segmentation prediction for the *background* class ($\mathbf{S}^s_{i,\mathrm{bg}}$), as this class is not regarded in the classification task, nor is it represented within $\mathbf{A}_i$. $\mathcal{S}^s_{i,\mathrm{bg}}$ is treated as a constant in this step, not affecting the gradient propagation process.

The classification branch is then regularized to predict better organized activation maps, associated with lower entropy, through the employment of the *sparse categorical cross-entropy* loss function.

$$\begin{aligned}\mathbf{Q}_i^s &= \mathrm{softmax}\left(\left[\mathbf{S}^s_{i,\mathrm{bg}} \mid \mathrm{upscale}(\mathbf{A}_i^s)\right]\right)\\ \mathcal{L}_{\mathrm{s2c}}(\mathbf{A}_i^s, \mathbf{s}_i^t) &= -\frac{1}{|\mathbf{M}^s(\mathbf{s}_i^t)|}\sum_{hw}^{HW}\mathbf{M}^s_{hw}(\mathbf{s}_i^t)\log\mathbf{Q}^s_{ihwc_i^\star}\end{aligned} \tag{9}$$

where $c_i^\star = \arg\max_c \mathbf{s}^t_{ihwc}$.

3.2.4 Segmentation Consistency Loss

To reinforce prediction consistency in the segmentation branch, we employ an unsupervised weak-to-strong consistency (Yang et al., 2023) optimization objective $\mathcal{L}_u$, which reinforces the student model to output segmentation proposals similar to the teacher's, when the former is presented with a strongly augmented version of sample $\mathbf{x}_i$, and an un-augmented version is shown to the latter.

Firstly, a batch $\mathcal{B}_i = \{\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+b-1}\}$ is randomly drawn from the training set, and it is augmented with a strong augmentation technique (e.g., ClassMix (Olsson et al., 2021)), resulting in the augmented samples $\tilde{\mathcal{B}}_i = \{\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i+1}, \ldots, \tilde{\mathbf{x}}_{i+b-1}\}$.

The augmented and un-augmented samples are forward onto the student and teacher models, and the pseudo-labels produced by the teacher model are mixed with the same combinations used to build $\tilde{\mathcal{B}}_i$.

Similarly to previous works (Liu et al., 2022a; Wang et al., 2022; Yang et al., 2023), we also account for unreliable predictions from the teacher (specially in early stages) by first defining the expected proportion of reliable predictions $\alpha_t \in [0,1]$ at a train step t, and by computing the entropy associated to the posterior probability predicted by the teacher given the augmented samples:

$$\mathcal{H}_{hw}(\tilde{\mathbf{s}}_i^t) = -\sum_{c}^{|C|} \tilde{s}_{ihwc}^t \log \tilde{s}_{ihwc}^t \tag{10}$$

Pixels are sorted by their entropy, and those whose associated entropy are lower than the α_t-quantile are considered "reliable", while the highest $(100\% - \alpha_t)$ portion of the labels, associated with the highest entropy, is deemed "unreliable" and discarded:

$$\begin{aligned}\gamma_t^{\tilde{\mathcal{B}}_i} &= \text{quantile}\left(\mathcal{H}\left(\left[\tilde{\mathbf{s}}_i^t \mid \tilde{\mathbf{s}}_{i+1}^t \mid \ldots \mid \tilde{\mathbf{s}}_{i+b-1}^t\right]\right), 1 - \alpha_t\right) \\ \mathbf{M}^{\tilde{s}}(\tilde{\mathbf{s}}_i^t) &= \mathbb{1}\left[\mathcal{H}(\tilde{\mathbf{s}}_i^t) \leq \gamma_t^{\tilde{\mathcal{B}}_i}\right]\end{aligned} \tag{11}$$

Finally, the *sparse categorical cross-entropy* loss function is used to match the segmentation proposals of the student model to the reliable labels:

$$\mathcal{L}_{\text{u}}(\mathbf{s}_i^s, \mathbf{s}_i^t) = -\frac{1}{|\mathbf{M}^{\tilde{s}}(\tilde{\mathbf{s}}_i^t)|} \sum_{hw}^{HW} \mathbf{M}_{hw}^{\tilde{s}}(\tilde{\mathbf{s}}_i^t) \log \mathbf{s}_{ihwc_i^\star}^s \tag{12}$$

where $c_i^\star = \arg\max_c \mathbf{s}_{ihwc}^t$.

3.2.5 Contrastive Learning of Unreliable Pixel Labels

Similar to previous fully-supervised and semi-supervised semantic segmentation works (Liu et al., 2022b; Wang et al., 2022), we employ the InfoNCE (Oord et al., 2018) loss function to learn pixel-level representations:

$$\begin{aligned}\mathcal{L}_m = -\frac{1}{|C| \times P} \sum_{c}^{|C|} \sum_{p}^{P} \\ \log\left[\frac{e^{\langle \mathbf{r}_{pc}, \mathbf{r}_{pc}^+ \rangle} / \tau}{e^{\langle \mathbf{r}_{pc}, \mathbf{r}_{pc}^+ \rangle} / \tau + \sum_j^N e^{\langle \mathbf{r}_{pjc}, \mathbf{r}_{pjc}^- \rangle} / \tau}\right]\end{aligned} \tag{13}$$

where $\mathbf{r}_{pc}$ contains the p-th representation vector (anchor) for class c, $\mathbf{r}_{pc}^+$ a positive anchor for c, and $\mathbf{r}_{pjc}^-$ the j-th negative anchor for pixel p and class c. Pairs are compared with the cosine similarity function $\langle \cdot, \cdot \rangle$.

Representations are split based on their associated segmentation prediction entropy: the first *low entropy* group contains "reliable" representations, associated with entropy lower than the α_t-quantile, while the second *high entropy* group contains "unreliable" representations, associated with entropy higher than the $(100\% - \alpha_t)$-quantile.

We create the group of query candidates: for every class c in the set, we sample pixel representations in $\{\mathbf{R}_k \mid \mathbf{x}_k \in \mathcal{B}_i\}$ such that its label information is confidently known. That is, it belongs to the *low entropy* group, and it was inferred from the pseudo-labels devised from CAMs with confidence higher than δ_{fg} and segmented with confidence higher than 30%.

Positive class-specific anchors are formed from averaging representation vectors (from the teacher) with *low entropy*, confident predictions.

Finally, negative anchors are formed from representations associated to pixels segmented into class c with confidence lower than 1%, and belonging to the *high entropy* group. These are pushed into a FIFO memory bank of negative anchors for class c if: (a) the segmentation prediction rank of class c is less or equal than γ_l and class c does not occur in the label set $\mathbf{y}_i$ (the class does not appear in the image-level annotation); or (b) the segmentation prediction rank of class c is greater than γ_l and lower than γ_h.

The hyperparameters γ_l and γ_h are defined as small values (e.g., 3 and 6, respectively), indicating our preference for hard examples, associated to pixels easily mistaken for class c.

During forward, P queries are randomly drawn (with replacement) from the query candidates, for each class c, and N negative anchors are drawn (with replacement) for each query from the memory bank. InfoNCE is employed to project queries and positive anchors close together, while pushing negative anchors apart in the contrastive space. The model is hence reinforced to refine its shared feature space to account for these differences, hence providing additional regularization.

3.2.6 Pre-Training and Self-Supervision

We start with pretrained weights for the feature extractor and classification branch, recovered from pre-existing classification or WSSS solution strategies, such as Puzzle-CAM (Jo and Yu, 2021) or P-NOC (David et al., 2024), allowing us to bootstrap and accelerate training with "reasonable" segmentation pseudo-labels from the start. The parameters in the segmentation and representation branches are randomly drawn from a Kaiming normal distribution.

Following the typical self-supervised setup, we employ the student-teacher training framework to obtain more stable (and higher quality) pseudo-labels. In this setup, the "teacher" model $\theta^t = \{\mathcal{F}^t, \mathcal{C}^t, \mathcal{S}^t, \mathcal{R}^t\}$ shares the architecture and initial parameters of the student $S = \{\mathcal{F}^s, \mathcal{C}^s, \mathcal{S}^s, \mathcal{R}^s\}$, while having its parameters θ^t updated with the exponential moving average (EMA) of the student's weights θ^s.

Finally, the training of the student model is conducted with the following optimization objectives:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_s + \lambda_{\text{s2c}} \mathcal{L}_{\text{s2c}} + \lambda_u \mathcal{L}_u + \lambda_m \mathcal{L}_m \tag{14}$$

where λ_{s2c}, λ_u, and λ_m are importance coefficients useful for balancing the different tasks, when they have noticeably different loss values. For simplicity, we set λ_u and λ_m to 1 in our studies over the Pascal VOC 2012 and the MS COCO 2014 sets.

4 EXPERIMENTAL SETUP

In this section, we detail the training and evaluation procedures employed in this work.

Following previous work (Ahn and Kwak, 2018; Kweon et al., 2021; Jo and Yu, 2021), we train CSRM over the Pascal VOC 2012 dataset (Everingham et al., 2015) using Stochastic Gradient Descent (SGD) for 15 epochs with linearly decaying learning rates of 0.07 and 0.007 (for randomly initialized weights and pretrained weights, respectively), 1e-4 weight decay, and a batch size of 32. For the MS COCO 2014 dataset (Lin et al., 2014), we employ the learning rates 0.04 and 0.004 for randomly initialized weights and pretrained weights, respectively.

For Pascal VOC 2012, image samples are resized to a common resolution of 512 per 512 pixels² and augmented with random flip. For the MS COCO 2014 dataset, models are trained with images of 640 per 640 pixels². We employ ClassMix (Olsson et al., 2021) as strong augmentation technique, having 50% probability of being applied to images before they are fed to the student model.

We set δ_{bg} and δ_{fg} to the constants 0.05 and 0.35 throughout training, respectively. σ_{s2c} is initially set to 10%, and increases linearly up to 100%. For simplicity, we set the remaining coefficients in the $\mathcal{L}$ (λ_u and λ_m) to 1. We also consider a *warm-up* period in which the objectives $\mathcal{L}_{\text{s2c}}$, $\mathcal{L}_{\text{u}}$ and $\mathcal{L}_{\text{m}}$ are ignored (λ_{s2c}, λ_{u}, and λ_{m} are set to 0), in order to prevent noisy predictions (from the randomly initialized segmentation branch) to degenerate the activation signal of the classification and representation branches during early stages. We find these values to be reasonable assumptions, commonly adopted by many studies in the literature, and leave the quantitative analysis of their effect as future work.

Following (Wang et al., 2022), we use the following hyperparameters for the contrastive objective: $P = 50$, $N = 256$ and $\tau = 0.5$. Moreover, α_t is set to 20%, and linearly decreases to 0% as training progresses. The memory bank consists of $|C|$ class-specific lists, each of which containing N negative "hard" anchors for its respective class. New anchors are inserted into the bank in a FIFO order.

We employ Test-Time Augmentation (TTA) during inference to improve prediction stability, forwarding the images in multiple scales and combining the segmentation proposals. The results are then refined with Densely Connected Conditional Random Fields (dCRF) (Krähenbühl and Koltun, 2011) and/or SAM Enhanced Pseudo-Labels (SEPL) (Chen et al., 2023).

Pseudo-labels are evaluated with respect to their fidelity to human-annotated semantic segmentation masks, measured through the *mean Intersection over Union* (mIoU) metric. We report scores over both training and validation sets — a common practice in WSSS, as all ground-truth segmentation maps from both sets were unused during training.

To provide a fair comparison with literature, we execute a verification step, in which the pseudo-labels are used to train a fully-supervised semantic segmentation model. DeepLabV2 (Chen et al., 2017) and DeepLabV3+ (Chen et al., 2018) architectures are employed as semantic segmentation networks, and keep their original training procedures unchanged.

For DeepLabV2, we employ the RN-101 architecture as backbone, with weights pretrained over the ImageNet dataset. We set a batch size of 10 and crop each training image to the size of 321 per 321 pixels². We train the model for 10,000 iterations over the Pascal VOC 2012 dataset, using an initial learning rate of 2.5e-4. When training the segmentation model over the MS COCO 2014 dataset, we employ 20,000 training iterations over patches of 481 per 481 pixels², using an initial learning rate of 2e-4.

For DeepLabV3+, we employ the RS-269 as backbone, and train the model for 50 epochs. The batch size is set to 32, and the initial learning rates of 0.007 and 0.004 are used for Pascal VOC 2012 and MS COCO 2014, respectively.

5 RESULTS

We provide an ablation study over the impact (measured in mIoU) of each objective function used by CSRM in Table 1. The regularization of the classification signal with segmentation predictions ($\mathcal{L}_{\text{s2c}}$) inadvertently deteriorates the efficacy, indicating that an unregularized segmentation signal can be detrimental to mutual promotion. On the other hand, we observe a noticeable improvement in scores when considering

Table 1: Ablation study for each objective function in CSRM, measured in mIoU (%) over Pascal VOC 2012 *train* and *val* sets.

#	$\mathcal{L}_{\text{s}}$	$\mathcal{L}_{\text{s2c}}$	$\mathcal{L}_{\text{u}}$	$\mathcal{L}_{\text{m}}$	VOC12	
					Train	***Val***
1	✓				69.9 (baseline)	67.7 (baseline)
2	✓	✓			68.3 (↓ 02.3%)	66.6 (↓ 01.6%)
3	✓	✓	✓		70.8 (↑ 03.7%)	68.7 (↑ 03.2%)
4	✓	✓	✓	✓	**71.8** (↑ 01.4%)	**69.8** (↑ 01.6%)

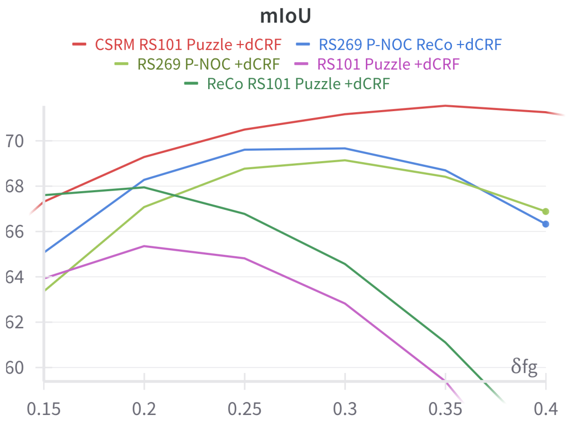

Figure 2: Score (mIoU) measured over Pascal VOC 2012 training set, for different choices of δ_{fg}.

prediction conformity ($\mathcal{L}_u$), suggesting that a strongly regularized segmentation branch is essential for mutual promotion. Lastly, learning contrasting representations ($\mathcal{L}_m$) produces the best results.

Segmentation proposals by CSRM can be further improved by its combination with other refinement methods (a strategy commonly employed by multi-stage WSSS solutions to refine its segmentation priors). Table 2 describes the effect of each refinement step on the quality of the pseudo-labels produced. The first row contains the score of the pretrained baseline model, ResNeSt-101 Puzzle-CAM. Training CSRM improves mIoU measured over the training and validation sets by 19.9% and 16.7%, respectively, while refining priors with dCRF results in a more modest improvement (9.0%). In turn, SEPL also produces a considerable improvement in mIoU (19.5% and 20.1% over training and validation sets, respectively). Finally, combining the three methods results in the highest mIoU scores observed, demonstrating that these techniques are complementary.

Figure 2 displays the segmentation score (measured in mIoU) of the semantic segmentation pseudo-labels devised by CSRM, considering different choices of δ_{fg}. While other WSSS techniques produce semantic segmentation pseudo-labels that are strongly affected by the choice of the threshold, the segmentation proposals obtained from CSRM are more robust against it, achieving a higher and less varying score for almost all choices of the threshold.

Table 2: Impact of each refinement procedure on mIoU (%), measured over Pascal VOC 2012 *train* and *val* sets.

#	CSRM	dCRF	SEPL	VOC12	
	(ours)			*Train*	*Val*
1				59.9 (baseline)	59.8 (baseline)
2	✓			71.8 (↑ 19.9%)	69.8 (↑ 16.7%)
3		✓		65.3 (↑ 09.0%)	65.2 (↑ 09.0%)
4			✓	71.6 (↑ 19.5%)	71.8 (↑ 20.1%)
4	✓	✓		72.6 (↑ 21.2%)	70.2 (↑ 17.4%)
5	✓	✓	✓	**77.6** (↑ 29.5%)	**76.2** (↑ 27.4%)

Table 3: Comparison with SOTA methods on Pascal VOC 2012, measured in mIoU (%)[1]. $\mathcal{F}$: fully-supervised; $\mathcal{I}$: image-level; $\mathcal{S}$: saliency. †Imagenet-21k pretrained; ‡MS COCO pretrained.

Method	Sup.	B.bone	Seg.	*Val*	*Test*
DeepLabV1 (Chen et al., 2017)	$\mathcal{F}$	RN-38	V1	78.1	78.2
AffinityNet (Ahn and Kwak, 2018)	$\mathcal{I}$	RN-38	V1	61.7	63.7
SEAM (Wang et al., 2020)	$\mathcal{I}$	RN-38	V1	64.5	65.7
CONTA (Zhang et al., 2020a)	$\mathcal{I}$	RN-38	V1	66.1	66.7
OC-CSE (Kweon et al., 2021)	$\mathcal{I}$	RN-38	V1	68.4	68.2
ADELE (Liu et al., 2022a)	$\mathcal{I}$	RN-38	V1	69.3	68.8
MCT-Former (Xu et al., 2022)	$\mathcal{I}$	RN-38	V1	71.9	71.6
ACR (Kweon et al., 2023)	$\mathcal{I}$	RN-38	V1	72.4	**72.4**
ICD (Fan et al., 2020)	$\mathcal{I}$	RN-101	V1	64.1	64.3
ICD (Fan et al., 2020)	$\mathcal{I}+\mathcal{S}$	RN-101	V1	67.8	68.0
EPS (Lee et al., 2021c)	$\mathcal{I}+\mathcal{S}$	RN-101	V1	71.0	**71.8**
ToCo (Ru et al., 2023)	$\mathcal{I}$	ViT-B	-	69.8	70.5
ToCo† (Ru et al., 2023)	$\mathcal{I}$	ViT-B	-	71.1	72.2
BECO (Rong et al., 2023)	$\mathcal{I}$	MiT-B2	-	73.7	73.5
SeCo (Yang et al., 2024)	$\mathcal{I}$	ViT-B	-	74.0	**73.8**
DeepLabV2 (Chen et al., 2017)	$\mathcal{F}$	RN-101	V2	76.8	76.2
IRNet (Ahn et al., 2019)	$\mathcal{I}$	RN-50	V2	63.5	64.8
AdvCAM (Lee et al., 2021b)	$\mathcal{I}$	RN-101	V2	68.1	68.0
RIB (Lee et al., 2021a)	$\mathcal{I}$	RN-101	V2	68.3	68.6
AMR (Qin et al., 2022)	$\mathcal{I}$	RN-101	V2	68.8	69.1
AMN (Lee et al., 2022)	$\mathcal{I}$	RN-101	V2	69.5	69.6
SIPE (Chen et al., 2022)	$\mathcal{I}$	RN-101	V2	68.8	69.7
URN (Li et al., 2022)	$\mathcal{I}$	RN-101	V2	69.5	69.7
RIB (Lee et al., 2021a)	$\mathcal{I}+\mathcal{S}$	RN-101	V2	70.2	70.0
SGWS (Yi et al., 2022)	$\mathcal{I}$	RN-101	V2	70.5	70.5
AMN‡ (Lee et al., 2022)	$\mathcal{I}$	RN-101	V2	70.7	70.6
EPS (Lee et al., 2021c)	$\mathcal{I}+\mathcal{S}$	RN-101	V2	**70.9**	70.8
ViT-PCM (Rossetti et al., 2022)	$\mathcal{I}$	RN-101	V2	70.3	**70.9**
P-NOC (David et al., 2024)	$\mathcal{I}$	RN-101	V2	70.3	**70.9**
CSRM (ours)[a]	$\mathcal{I}$	RN-101	V2	69.8	70.4
DeepLabV3+	$\mathcal{F}$	RS-269	V3+	80.6	81.0
Puzzle-CAM (Jo and Yu, 2021)	$\mathcal{I}$	RS-269	V3+	71.9	72.2
P-NOC (David et al., 2024)	$\mathcal{I}$	RS-269	V3+	73.8	73.6
CSRM (ours)[b]	$\mathcal{I}$	RS-269	V3+	**74.5**	**75.0**

Official evaluations: [a] BKRNCI and [b] TANYYC.

Table 3 shows the effectiveness of fully-supervised semantic segmentation models trained with pseudo-labels generated by CSRM, while comparing it with current state-of-the-art over the Pascal VOC 2012 dataset. A DeepLabV2 model trained with CSRM pseudo-labels achieves the competitive result of 70.4% test mIoU; while a DeepLabV3+ model trained over the same set of pseudo-labels achieve a mIoU score of 75.0%.

Table 4 illustrates the effectiveness of our approach over the MS COCO 2014 dataset. Once again, DeepLabV2 and DeepLabV3+ achieve competitive

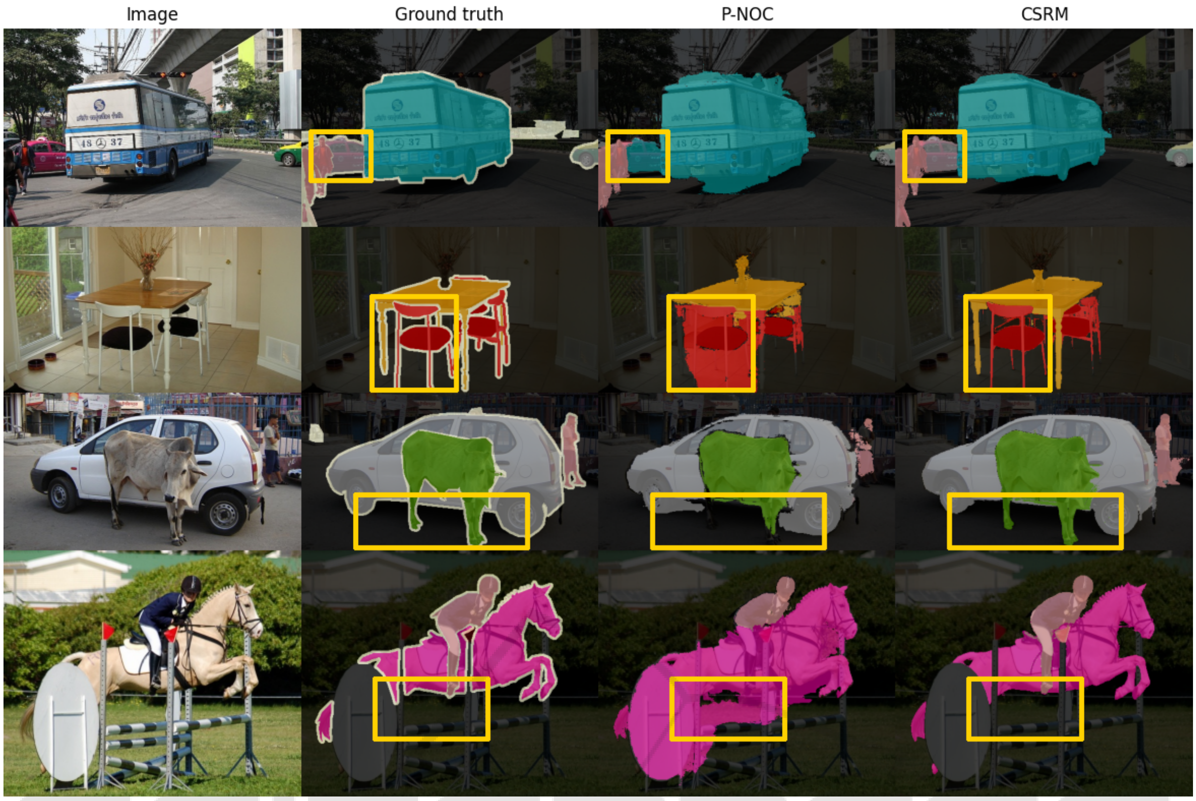

Figure 3: Qualitative comparison of pseudo-labels obtained from images in Pascal VOC 2012 *val* dataset. From left to right: (a) input image, (b) ground-truth; (c) P-NOC (David et al., 2024), and (d) CSRM (ours).

Table 4: Comparison with SOTA methods on MS COCO 2014 dataset, measured in mIoU (%). $\mathcal{F}$: fully-supervised; $\mathcal{I}$: image-level; $\mathcal{S}$: saliency. †Imagenet-21k pretrained.

Method	**Sup.**	**B.bone**	**Seg.**	***Val***
OC-CSE (Kweon et al., 2021)	$\mathcal{I}$	RN-38	V1	36.4
MCT-Former (Xu et al., 2022)	$\mathcal{I}$	RN-38	V1	42.0
ACR (Kweon et al., 2023)	$\mathcal{I}$	RN-38	V1	45.3
ToCo (Ru et al., 2023)	$\mathcal{I}$	ViT-B	-	41.3
ToCo† (Ru et al., 2023)	$\mathcal{I}$	ViT-B	-	42.3
SeCo (Yang et al., 2024)	$\mathcal{I}$	ViT-B	-	**46.7**
IRNet (Ahn et al., 2019)	$\mathcal{I}$	RN-50	V2	32.6
IRN+CONTA (Zhang et al., 2020a)	$\mathcal{I}$	RN-50	V2	33.4
EPS (Lee et al., 2021c)	$\mathcal{I}+\mathcal{S}$	VGG16	V2	35.7
PPM (Li et al., 2021)	$\mathcal{I}$	ScaleNet	V2	40.2
SIPE (Chen et al., 2022)	$\mathcal{I}$	RN-101	V2	40.6
URN (Li et al., 2022)	$\mathcal{I}$	RN-101	V2	40.7
P-NOC (David et al., 2024)	$\mathcal{I}$	RN-101	V2	42.9
RIB (Lee et al., 2021a)	$\mathcal{I}$	RN-101	V2	43.8
AMN (Lee et al., 2022)	$\mathcal{I}$	RN-101	V2	44.7
CSRM (ours)	$\mathcal{I}$	RN-101	V2	**46.2**
P-NOC (David et al., 2024)	$\mathcal{I}$	RS-269	V3+	44.6
CSRM (ours)	$\mathcal{I}$	RS-269	V3+	**50.5**

scores, compared to the current state-of-the-art, when trained over pseudo-labels devised by CSRM.

Table 5 displays the effectiveness (measured in IoU) of our approach for each individual class. CSRM achieves the best IoU scores for most classes.

A qualitative comparison between pseudo-labels devised by CSRM and similar WSSS techniques is provided in Figure 3, while Figures 4 and 5 show examples of proposals made by semantic segmentation models (from the verification step) trained with pseudo-labels devised by CSRM over Pascal VOC 2012 and MS COCO 2014 datasets, respectively.

We further remark that many strategies in literature achieve SOTA by relying on other forms of fully supervised information (such as pretrained saliency detectors (Fan et al., 2020; Lee et al., 2021a; Lee et al., 2021c)), or multiple training stages (such as semantic segmentation pseudo-label refinement with pixel-level affinity (Ahn and Kwak, 2018; Ahn et al., 2019; Jo and Yu, 2021; David et al., 2024)). Conversely, CSRM comprises a single training stage, entailing lower training time and a simpler inference.

Training CSRM over VOC12 with 4 NVIDIA P100 GPUs took approximately 15.7 minutes per epoch and 3.9 hours in total, achieving 99% of the best mIoU score observed in the fist 1.6 hours. Similarly, training over COCO14 took approximately 4.4 hours/epoch, achieving 99% of the best mIoU score observed after only the first epoch, which can be ex-

Table 5: Comparison of Intersection over Union (IoU %) scored by various WSSS techniques, measured over Pascal VOC 2012 test set. Evaluated classes are, from left to right: *background*, *airplane*, *bicycle*, *bird*, *boat*, *bottle*, *bus*, *car*, *cat*, *chair*, *cow*, *dining table*, *dog*, *horse*, *motorbike*, *person*, *potted plant*, *sheep*, *sofa*, *train*, and *tv monitor*.

Method	bg	air	bik	bir	boa	bot	bus	car	cat	cha	cow	din	dog	hor	mot	per	pot	she	sof	tra	tvm	mIoU
AffinityNet	89.1	70.6	31.6	77.2	42.2	68.9	79.1	66.5	74.9	29.6	68.7	56.1	82.1	64.8	78.6	73.5	50.8	70.7	47.7	63.9	51.1	63.7
OC-CSE	90.2	82.9	35.1	86.8	59.4	70.6	82.5	78.1	87.4	30.1	79.4	45.9	83.1	83.4	75.7	73.4	48.1	89.3	42.7	60.4	52.3	68.4
MCT-Former	92.3	84.4	37.2	82.8	60.0	72.8	78.0	79.0	89.4	31.7	84.5	59.1	85.3	83.8	79.2	81.0	53.9	85.3	**60.5**	65.7	**57.7**	71.6
EPS	91.9	89.0	39.3	88.2	58.9	69.6	86.3	83.1	85.8	35.0	83.6	44.1	82.4	86.5	81.2	80.8	56.8	85.2	50.5	**81.2**	48.4	71.8
AMN	90.7	82.8	32.4	84.8	59.4	70.0	86.7	83.0	86.9	30.1	79.2	56.6	83.0	81.9	78.3	72.7	52.9	81.4	59.8	53.1	56.4	69.6
ViT-PCM	91.1	88.9	39.0	87.0	58.8	69.4	89.4	85.4	89.9	30.7	82.6	62.2	85.7	83.6	79.7	**81.6**	52.1	82.0	26.5	80.3	42.4	70.9
Puzzle-CAM	91.1	87.2	37.4	86.8	61.5	71.3	92.2	86.3	91.8	28.6	85.1	**64.2**	91.9	82.1	82.6	70.7	69.4	87.7	45.5	67.0	37.8	72.3
P-NOC	91.7	89.1	38.3	80.9	65.4	70.1	**93.8**	85.5	**93.4**	37.3	83.6	61.3	**92.8**	84.1	83.8	80.7	63.6	82.0	53.3	76.7	36.8	73.6
CSRM (ours)	**92.3**	**92.0**	**43.9**	**90.1**	**66.4**	**75.0**	93.3	**87.1**	86.8	**41.4**	**89.7**	49.6	88.7	**87.9**	**85.1**	77.9	**72.2**	**91.5**	46.5	70.1	47.8	**75.0**

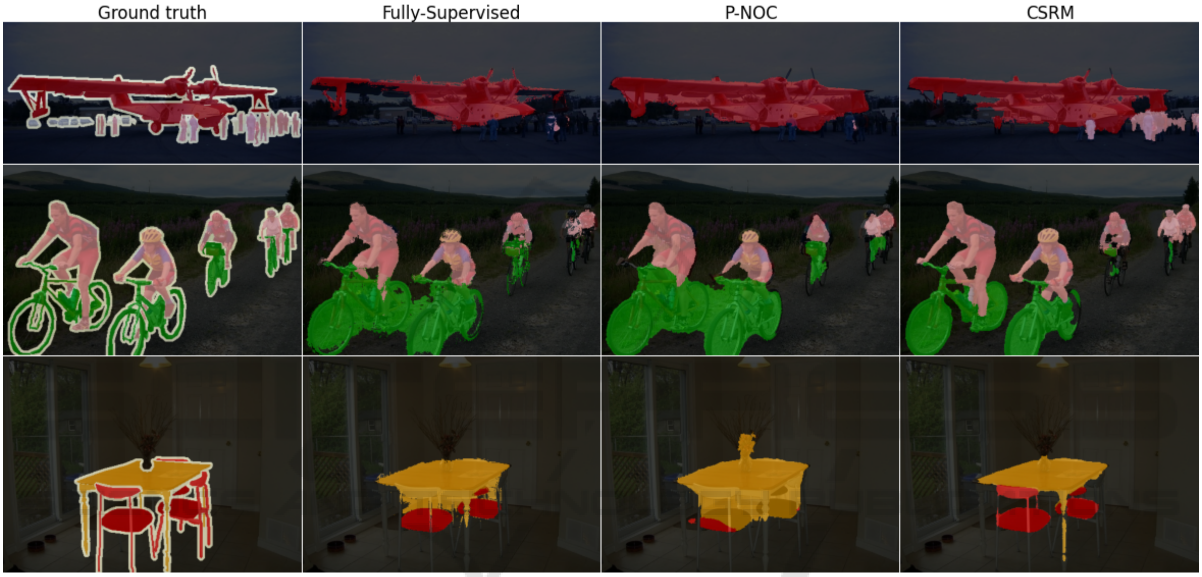

Figure 4: Segmentation proposals made by models trained over pseudo-labels devised from WSSS strategies, when fed from images in Pascal VOC 2012 *val* dataset. From left to right: (a) ground-truth; (b) Fully-Supervised; (c) P-NOC (David et al., 2024), and (d) CSRM (ours).

plained by the large number of iterations required to perform one pass over each sample in its training set.

6 CONCLUSIONS

In this work, we proposed an approach to determine and utilize (the otherwise wasted) “unreliable” regions in the pseudo-labels devised from CAMs in a weakly supervised semantic segmentation scheme, where only image-level annotations are available.

Empirical results suggest our approach can provide substantial improvement in segmentation effectiveness of weakly supervised models, achieving competitive mIoU in both Pascal VOC 2012 and MS COCO 2014 datasets, while requiring fewer training stages and lower computation footprint than other (multi-stage) techniques in the literature.

For future work, we will investigate the effect of our approach in functional segmentation and biological-related tasks, in which visual patterns are convoluted or unclear, and well as attempt to further reduce computational footprint.

ACKNOWLEDGEMENTS

The authors would like to thank CNPq (grants 140929/2021-5, 309330/2018-1, 302530/2022-3 and 304836/2022-2) and LNCC/MCTI for providing HPC resources of the SDumont supercomputer.

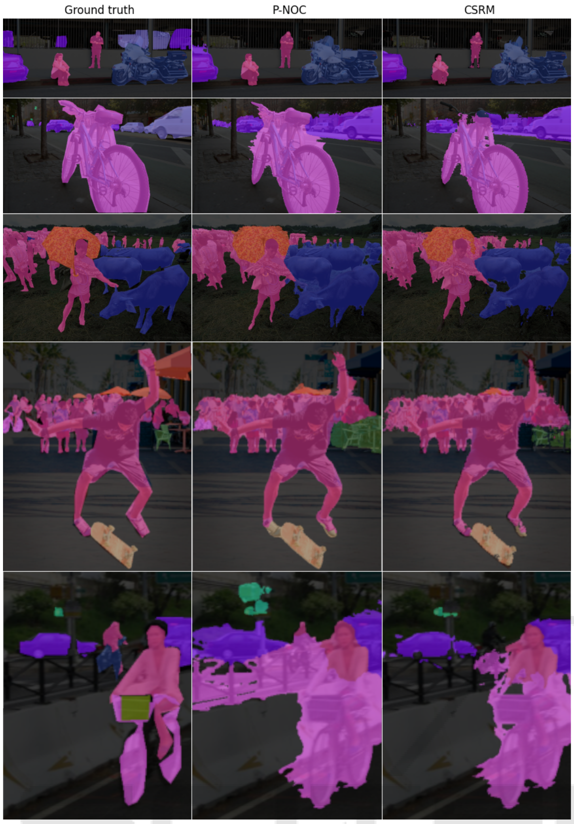

Figure 5: Segmentation proposals made by models trained over pseudo-labels of MS COCO 2014 *val* dataset. From left to right: (a) ground-truth; (b) P-NOC (David et al., 2024), and (c) CSRM (ours).

REFERENCES

Ahn, J., Cho, S., and Kwak, S. (2019). Weakly supervised learning of instance segmentation with inter-pixel relations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2209–2218.

Ahn, J. and Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4981–4990.

Bircanoglu, C. and Arica, N. (2022). ISIM: Iterative self-improved model for weakly supervised segmentation. *arXiv preprint arXiv:2211.12455*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818.

Chen, Q., Yang, L., Lai, J.-H., and Xie, X. (2022). Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4288–4298.

Chen, T., Mai, Z., Li, R., and Chao, W.-l. (2023). Segment Anything Model (SAM) Enhanced Pseudo Labels for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2305.05803*.

Chen, Y., Tao, J., Liu, L., Xiong, J., Xia, R., Xie, J., Zhang, Q., and Yang, K. (2020). Research of improving semantic image segmentation based on a feature fusion model. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13.

David, L., Pedrini, H., and Dias, Z. (2024). P-NOC: Adversarial training of cam generating networks for robust weakly supervised semantic segmentation priors. *Journal of Visual Communication and Image Representation*, 102:104187.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The PASCAL Visual Object Classes Challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

Fan, J., Zhang, Z., Song, C., and Tan, T. (2020). Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4283–4292.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., and Wang, L. (2021). Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118.

Jo, S. and Yu, I.-J. (2021). Puzzle-cam: Improved localization via matching partial and full features. In *IEEE International Conference on Image Processing (ICIP)*, pages 639–643.

Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected CRFs with Gaussian edge potentials. *Advances in Neural Information Processing Systems (NeurIPS)*, 24:109–117.

Kweon, H., Yoon, S.-H., Kim, H., Park, D., and Yoon, K.-J. (2021). Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6974–6983.

Kweon, H., Yoon, S.-H., and Yoon, K.-J. (2023). Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11329–11339.

Lee, J., Choi, J., Mok, J., and Yoon, S. (2021a). Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:27408–27421.

Lee, J., Kim, E., and Yoon, S. (2021b). Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4071–4080.

Lee, M., Kim, D., and Shim, H. (2022). Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4330–4339.

Lee, S., Lee, M., Lee, J., and Shim, H. (2021c). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5495–5505.

Li, Y., Duan, Y., Kuang, Z., Chen, Y., Zhang, W., and Li, X. (2022). Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 1447–1455.

Li, Y., Kuang, Z., Liu, L., Chen, Y., and Zhang, W. (2021). Pseudo-mask matters in weakly-supervised semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 6964–6973.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECVV)*, pages 740–755. Springer.

Liu, S., Liu, K., Zhu, W., Shen, Y., and Fernandez-Granda, C. (2022a). Adaptive early-learning correction for segmentation from noisy annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2606–2616.

Liu, S., Zhi, S., Johns, E., and Davison, A. J. (2022b). Bootstrapping semantic segmentation with regional contrast. In *International Conference on Learning Representations (ICLR)*.

Mo, Y., Wu, Y., Yang, X., Liu, F., and Liao, Y. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646.

Olsson, V., Tranheden, W., Pinto, J., and Svensson, L. (2021). Classmix: Segmentation-based data augmentation for semi-supervised learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1369–1378.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qin, J., Wu, J., Xiao, X., Li, L., and Wang, X. (2022). Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 2117–2125.

Rong, S., Tu, B., Wang, Z., and Li, J. (2023). Boundary-enhanced co-training for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19574–19584.

Rossetti, S., Zappia, D., Sanzari, M., Schaerf, M., and Pirri, F. (2022). Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 446–463. Springer.

Ru, L., Zheng, H., Zhan, Y., and Du, B. (2023). Token contrast for weakly-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3102.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626.

Shen, W., Peng, Z., Wang, X., Wang, H., Cen, J., Jiang, D., Xie, L., Yang, X., and Tian, Q. (2023). A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9284–9305.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., and Le, X. (2022). Semi-supervised semantic segmentation using unreliable pseudo-labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4248–4257.

Wang, Y., Zhang, J., Kan, M., Shan, S., and Chen, X. (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12275–12284.

Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., and Xu, D. (2022). Multi-class token transformer for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4300–4309.

Yang, L., Qi, L., Feng, L., Zhang, W., and Shi, Y. (2023). Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7236–7246.

Yang, Z., Fu, K., Duan, M., Qu, L., Wang, S., and Song, Z. (2024). Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3615.

Yi, S., Ma, H., Wang, X., Hu, T., Li, X., and Wang, Y. (2022). Weakly-supervised semantic segmentation with superpixel guided local and global consistency. *Pattern Recognition*, 124:108504.

Zhang, D., Zhang, H., Tang, J., Hua, X.-S., and Sun, Q. (2020a). Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666.

Zhang, M., Zhou, Y., Zhao, J., Man, Y., Liu, B., and Yao, R. (2020b). A survey of semi-and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, 53:4259–4288.

Zhu, L., He, H., Zhang, X., Chen, Q., Zeng, S., Ren, Q., and Lu, Y. (2023). Branches mutual promotion for end-to-end weakly supervised semantic segmentation. *arXiv preprint arXiv:2308.04949*.