REPRESENTATION MUTUAL LEARNING FOR END-TO-END WEAKLY-SUPERVISED SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, end-to-end solutions for Weakly Supervised Semantic Segmentation (WSSS) with image-level labels have been developed rapidly. Previous end-to-end methods usually rely on segmentation branches or decoders to predict segmentation masks, bringing additional parameter numbers and consumption time. In this paper, we propose a decoder-free Representation Mutual Learning (RML) framework to directly predict segmentation masks, which leverages collaborative learning and mutual teaching among multi-level feature representations to improve segmentation performance. Our RML is a straightforward and efficient end-to-end WSSS framework, which incorporates the instance-level, feature-level and pixel-level representation mutual learning strategies to improve segmentation quality. To enhance the Class Activation Map (CAM) representations, we propose a CAM-driven Instance-leave Mutual Learning strategy that preserves the equivariance of CAMs and expands the distance between different classes of semantic prototypes. Besides, we design a Multi-scale Feature-leave Mutual Learning strategy, which can align aggregated contextual representations and facilitate the representation capability of contextual representations. Furthermore, we also provide an Affinity-aware Pixel-level Mutual Learning strategy to learn semantic affinity representations. Experiments validate that our RML yields a significant performance improvement over recent end-to-end methods on the Pascal VOC 2012 dataset and the MS COCO 2014 dataset. The release code is available at supplementary material.

1 INTRODUCTION

Fully supervised semantic segmentation models require a large number of labor-intensive pixel-level labels to annotate images Chen et al. (2017). To alleviate the dependence on pixel-level labels, weakly supervised semantic segmentation (WSSS) methods use weak labels that are cheaper to acquire for semantic segmentation, usually taking the form of image-level labels Lee et al. (2021b); Ru et al. (2022), scribbles Lin et al. (2016), or bounding boxes Dai et al. (2015); Lee et al. (2021c). This paper focuses on using only image-level labels, as they are the cheapest and most challenging option for weakly supervised semantic segmentation.

Most WSSS methods with image-level labels are usually done in a multi-stage process. These multi-stage methods require at least three stages that are complex and time-consuming. To solve the problem of complex training pipeline and time-consuming computation of multi-stage methods, some end-to-end methods Zhang et al. (2020a; 2021b); Araslanov & Roth (2020); Ru et al. (2022) have been proposed recently. These end-to-end methods first directly generate Class Activation Maps (CAM) Zhou et al. (2016) as initial pseudo-labels. The refinement module then refines the initial pseudo-labels concurrently during training. Finally, current end-to-end methods typically use fine-grained pseudo-labels to supervise the segmentation branch or decoder for segmentation mask prediction. For example, Araslanov & Roth (2020) uses refined pseudo-labels as the supervision for the semantic segmentation branch. Araslanov & Roth (2020) and Ru et al. (2022) employ pixel-adaptive refinement modules PAMR and PAR respectively to improve pseudo-labels, and finally use a decoder to predict segmentation results, as shown in Fig. 1 (a). However, we found that the seg-

mentation branch or decoder is not indispensable for WSSS, we can directly predict the segmentation masks and achieve high segmentation accuracy via mutual learning between feature representations.

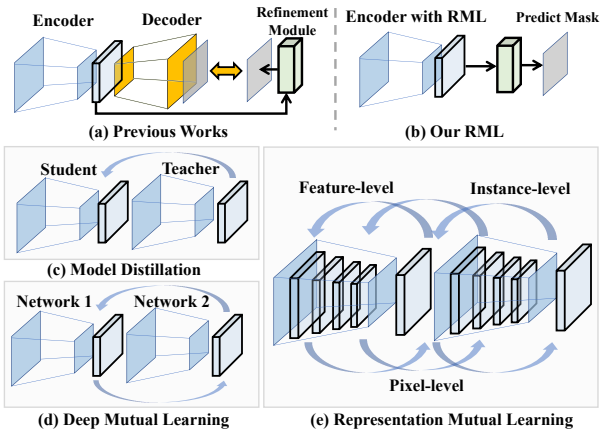

Figure 1: Previous Works vs. Representation Mutual Learning, and illustration of our multi-level RML.

For WSSS, the generation quality of CAM, global context information, and semantic affinity have been verified to be critical for segmentation performance Araslanov & Roth (2020); Ru et al. (2022). Therefore, we design a decoder-free Representation Mutual Learning (RML) framework includes the instance-level, feature-level and pixel-level mutual learning strategies, which improve the segmentation performance of the network by enhancing CAM representations, contextual representations and semantic affinity representations in a self-supervised manner, respectively. Unlike model distillation Hinton et al. (2015); Adriana et al. (2015), which usually allows the student network to learn the class probability of the teacher network, deep mutual learning Zhang et al. (2018b) does not distinguish the network as a teacher or student network, and allows the outputs of different networks to learn from each other, and achieves excellent performance, as shown in Fig. 1 (d). Inspired by this, we propose a Representation Mutual Learning (RML) framework, which aims to enable multi-level representations to learn from each other in a self-supervised manner, as shown in Fig. 1 (e). Our RML can enhance the feature representation capability of the network to improve segmentation accuracy, not just through mutual learning of class probabilities. For instance-level mutual learning, we present a CAM-driven Instance-leave Mutual Learning strategy to improve the generation quality of CAM representations by maintaining the equivariance of CAMs and expanding the distance between semantic prototypes (obtained by CAM aggregation) of different categories. As contextual information is crucial for the performance improvement of semantic segmentation models, we design the Multi-scale Feature-leave Mutual Learning strategy to facilitate contextual information learning by aligning aggregated contextual representations. Furthermore, inspired by Ahn & Kwak (2018) and Ru et al. (2022) predicting semantic affinity to facilitate segmentation performance, we provide the Affinity-aware Pixel-level Mutual Learning strategy for pixel-level mutual learning of semantic affinity representations by introducing mutual information and contrastive learning.

Specifically, our contributions are summarized as follows:

- We propose an efficient decoder-free Representation Mutual Learning (RML) framework that exploits the mutual promotion between feature representations via multi-level strategies to improve network performance, achieving SOTA performance on WSSS tasks.
- We present a CAM-driven Instance-leave Mutual Learning strategy to improve segmentation performance by improving the generation quality of of instance-leave CAM representations.
- We design a Multi-scale Feature-leave Mutual Learning strategy to facilitate the learning of feature-leave contextual representations.
- We give an Affinity-aware Pixel-level Mutual Learning strategy for mutual learning of pixel-level semantic affinity representations to improve segmentation accuracy.

2 Related Work

2.1 Weakly-Supervised Semantic Segmentation

Multi-stage methods. Prevailing Weakly Supervised Semantic Segmentation (WSSS) methods with image-level labels usually employ a multi-stage framework. To generate the initial seed masks,

multi-stage methods first use the trained classification model to extract the class activation map (CAM) Zhou et al. (2016). The initial seed masks are then refined by some refinement strategies. Lovasz et al. Lovász (1993) employ random walks to propagate seed regions of object classes. The boundary refinement strategies PSA Ahn & Kwak (2018) and IRN Ahn et al. (2019) refine the obtained initial seeds by learning semantic affinity. Erasure strategies Hou et al. (2018); Zhang et al. (2018a) prevent the classifier from focusing only on discriminative parts by erasing the most discriminative regions. Wu et al. proposed EDAM Wu et al. (2021) to integrate activation map generation into classification models for refinement. Some recent methods improve segmentation performance from the perspective of generating CAM, such as ReCAM Chen et al. (2022) inserting softmax cross-entropy loss into BCE-based model to reactivate CAM. Lee et al. Lee et al. (2021a) propose a method to reduce the information bottleneck in weakly supervised semantic segmentation. Finally, a fully supervised semantic segmentation model is trained using the refined pseudo-labels.

End-to-End methods. Unlike multi-stage methods that require training multiple models, end-to-end methods simplify the training pipeline. Pinheiro et al. Pinheiro & Collobert (2015) treat WSSS as a multi-instance learning problem and propose a CNN-based model to segment objects with only weak supervision. 1Stage Araslanov & Roth (2020) employs a normalized global weighted pooling, a local pixel-adaptive mask refinement module, and a stochastic gate to improve segmentation accuracy. Zhang et al. Zhang et al. (2020a) proposed Reliable Region Mining (RRM) uses CRF Krähenbühl & Koltun (2011) to refine the initial pseudo-labels as segmentation supervision. Ru et al. Ru et al. (2022) propose an Affinity from Attention (AFA) module to learn semantic affinity from transformers, and design a pixel-adaptive module to refine pseudo-labels. The above methods either use the refined pseudo-labels as supervision for the semantic segmentation branch or as supervision for the segmentation decoder. In contrast, our RML directly learns segmentation masks using a representation mutual learning framework.

2.2 DEEP MUTUAL LEARNING

Unlike collaborative learning He et al. (2016) where different models learn different tasks, or cooperative learning Batra & Parikh (2017) where multiple models are learned in different domains of the same task, in recent years, mutual learning Zhang et al. (2018b) has been proposed where all models deal with the same task and domain. Zhang et al. Zhang et al. (2021a) proposed robust mutual learning to effectively suppress noise in pseudo-labels in semi-supervised semantic segmentation tasks. Zhou et al. Zhou et al. (2021) proposed binocular mutual learning to improve few-shot classification through intra-view and cross-view modeling. Different from the mutual learning methods above which utilize the output of each sub-network for supervision, our representational mutual learning utilizes the feature representations within each sub-network for mutual supervision.

3 METHOD

Our model consists of a multi-level representation mutual learning framework. Specifically, we propose instance-level, feature-level, and pixel-level mutual learning strategies for representation enhancement in WSSS. Our overall network architecture is shown in Fig. 2.

3.1 PRELIMINARY

First we review the generation of class activation maps (CAMs) Zhou et al. (2016). Class scores are usually computed by global average pooling (GAP), and we denote the weight matrix in the classification layer by W. For a given specific class c and a given feature map $f \in \mathbb{R}^{H\times W\times D}$, an activation map M_c is generated by weighting the feature maps and their contributions to class c:

$$M_c = Relu(\sum_{i}^{D} W_{c,i} f_{i,:}). \tag{1}$$

where we scale M_c to [0, 1] using min-max normalization.

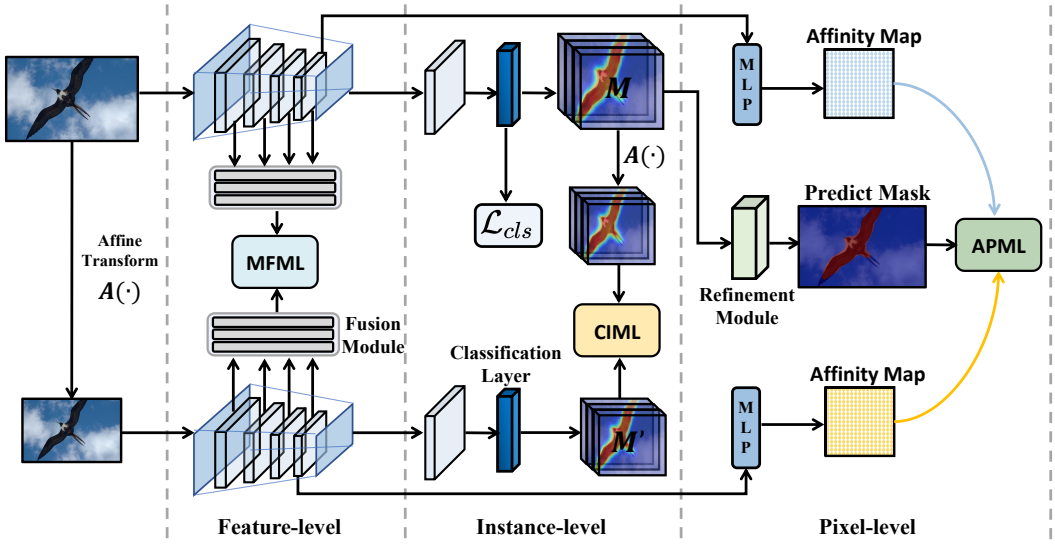

Figure 2: Schematic of Representation Mutual Learning (RML) for end-to-end WSSS. Our RML framework contains instance-level, feature-level, and pixel-level mutual learning strategies for representation enhancement. CIML: CAM-driven Instance-leave Mutual Learning strategy; MFML: Multi-scale Feature-leave Mutual Learning strategy; APML: Affinity-aware Pixel-level Mutual Learning strategy. We adopt PAR Ru et al. (2022) as the refinement module.

3.2 CAM-DRIVEN INSTANCE-LEAVE MUTUAL LEARNING

To enhance the feature representation capability of the network by constructing a representation mutual learning framework, we first apply an affine transformation to the original image, and input the original image and the transformed image into network N and network N', respectively. Here both the segmentation networks share the same parameters.

For instance-level representation mutual learning, we propose a CAM-driven Instance-leave Mutual Learning (CIML) strategy. CIML aims to enhance the representation capability of CAM representations by maintaining the equivariance of extracted CAMs and extending the distance between semantic prototypes of different categories. Specifically, given a affine transformation A and an input image I, the segmentation network N prefers to extract features that are equivariant, that is, $N(A(I)) = A(N(I))$. In this study, we constrain the CAM features (M, M') extracted by the two shared parameter segmentation networks (N, N') to be as equivariant as possible, defining the equivariance distance with $\mathcal{D}_{equ}$:

$$\mathcal{D}_{equ} = \|A(M) - M'\|_1 = \|A(N(I)) - N'(A(I))\|_1 . \tag{2}$$

Besides, as CAM representations are sparsely distributed in high-dimensional spaces, we apply an average pooling layer to aggregate the CAM representations to obtain semantic prototypes. We define the semantic prototype as the representative embedding of a class, which is estimated by the CAM representation. Here we obtain two semantic prototypes sets (P, P') by aggregating (M, M'). For the same set of semantic prototypes, the difference between semantic prototypes of different categories should be large, while their similarity is low. We use the cosine distance to measure the similarity of semantic prototypes of different classes:

$$\mathcal{D}(p_i, p_j) = \frac{p_i \cdot p_j}{\|p_i\| \times \|p_j\|}, \mathcal{D}(p'_i, p'_j) = \frac{p'_i \cdot p'_j}{\|p'_i\| \times \|p'_j\|} \tag{3}$$

where p_i and p_j represent the semantic prototype vectors belonging to the i-th class and the j-th class in $P = \{p_1, p_2, \dots p_C\}$, respectively. p'_i and p'_j represent the semantic prototype vectors belonging to the i-th class and the j-th class in $P' = \{p'_1, p'_2, \dots p'_C\}$, respectively. C means number of categories.

In our CIML strategy, we keep the equivariance of the extracted CAMs and push aside those semantic prototypes that belong to different categories to improve the generation quality of CAM representations. Specifically, we take the weighted sum of the equivariant distance and semantic prototype similarity defined in Eq. 2 and Eq. 3. The CAM-driven Instance-leave Mutual Learning loss is defined as:

$$\mathcal{L}_{CIML} = \alpha(\frac{1}{C^2}\sum_{i=1}^{C}\sum_{j=1}^{C} D(p_i, p_j) + \frac{1}{C^2}\sum_{i=1}^{C}\sum_{j=1}^{C} D(p'_i, p'_j)) + \mathcal{D}_{equ}. \tag{4}$$

where α balances different distance metrics.

3.3 Multi-scale Feature-leave Mutual Learning

The Multi-scale Feature-leave Mutual Learning (MFML) strategy aims to facilitate the learning of contextual information in the network, which includes a multi-scale context fusion module and feature-leave mutual learning loss. The multi-scale context fusion module extracts multi-scale context representations, and the feature-leave mutual learning loss is given to align the aggregated context representations.

Multi-scale context fusion module. The multi-scale context fusion module is shown in Fig. 2. To extract the multi-scale context representation, we linearly interpolate the outputs of each layer of the backbone to the same scale to obtain (f_1, f_2, f_3, f_4), and then concatenate them together. Because high-resolution low-level features provide valuable low-level concepts for semantic segmentation, high-level features provide deep semantic structure information, which are both important contextual information. For cascaded multi-scale features, we perform channel adjustment via 1×1 convolutions to obtain aggregated contextual representations R_c:

$$R_c = Conv(Concat(f_1, f_2, f_3, f_4)). \tag{5}$$

Feature-leave mutual learning loss. We apply a multi-scale context fusion module to network N and network N' to obtain context representations R_c and R'_c, respectively. The feature-leave mutual learning loss aims to align the aggregated context representations (R_c, R'_c), and we use the L1 distance to define the absolute distance between R_c and R'_c:

$$\mathcal{L}_{align} = \|R_c - R'_c\|_1 . \tag{6}$$

However, such strict alignment may lead to the loss of specific information in the two complementary contextual representations. Therefore we introduce mutual information Hjelm et al. (2018) to preserve the specific information of the representation. Mutual information is a measure of the amount of information shared between two random variables. Here, we measure the mutual information between two representations (R_c, R'_c) as:

$$I(R_c; R'_c) = E_{p(R_c, R'_c)}[log\frac{p(R_c, R'_c)}{p(R_c)p(R'_c)}] \tag{7}$$

where $p(R_c)$ and $p(R'_c)$ are the marginals of R_c and R'_c, and $p(R_c, R'_c)$ is the joint probability distribution between them.

To preserve the specificity information of the representations to some extent, we minimize the mutual information of the two representations while aligning the two contextual representations. Therefore we define the feature-leave mutual learning loss as:

$$\mathcal{L}_{MFML} = \mathcal{L}_{align} + \beta_1 \cdot I(R_c; R'_c) \tag{8}$$

where β_1 balances different losses.

3.4 Affinity-aware Pixel-level Mutual Learning

PSA Ahn & Kwak (2018) achieved semantic propagation by predicting the semantic affinity between a pair of adjacent image coordinates. AFA Ru et al. (2022) learns semantic affinity from

self-attention. This suggests that extracting effective semantic affinity representations can boost segmentation accuracy. Based on this observation, we propose an Affinity-aware Pixel-level Mutual Learning (APML) strategy to learn efficient semantic affinity representations. Specifically, our pixel-level mutual learning strategy includes pixel-level mutual information loss and affinity-aware contrastive learning loss.

Pixel-level mutual information loss. In this study, we use the mutual information Hjelm et al. (2018) to measure the semantic affinity between the network's self-attention representations and pseudo-labels. First, we cascade the last two layers of self-attention maps of the two segmentation networks (N, and N') together to construct initial semantic affinity representations R_a and R'_a. Then we use the segmentation mask refined by the pixel-adaptive module as the task-related pseudo-label y. To learn the efficient semantic affinity representation R_a, our main objective can be formulated as:

$$maxI(y; R_a|R'_a) \tag{9}$$

where y denotes pseudo-labels, and $I(y; R_a|R'_a)$ represents the amount of information related to the semantic segmentation task in R_a, excluding the information from the representation R'_a. And $I(y; R_a|R'_a)$ can be decomposed into:

$$I(y; R_a|R'_a) = I(y; R_a) - I(R_a; R'_a) + I(R_a; R'_a|y) \approx I(y; R_a) - I(y; R'_a) \tag{10}$$

where $I(R_a; R'_a|y)$ measures task-irrelevant information in representations R_a and R'_a, $I(R_a; R'_a)$ indicates the relevance between the two representations, and $I(y; R_a)$ represents the dependence between the pseudo-labels y and representation R_a. Due to the difficulty of conditional mutual information computation in neural networks Tian et al. (2021), we simplify it further. Intuitively, we can assume that task-related information will have an overwhelming presence over task-irrelevant information in training. Therefore, $I(R_a; R'_a|y)$ is negligible in efficient training.

In parallel, the goal of learning an efficient semantic affinity representation R'_a can also be simplified as:

$$I(y; R'_a|R_a) = I(y; R'_a) - I(R'_a; R_a) + I(R'_a; R_a|y) \approx I(y; R'_a) - I(y; R_a). \tag{11}$$

Therefore, our pixel-level mutual information loss can be formulated as:

$$\mathcal{L}_{mi} = I(y; R'_a) - I(y; R_a) - I(y; R'_a) + I(y; R_a). \tag{12}$$

Affinity-aware contrastive learning loss. Furthermore, we also enhance the representation of affinity representations by exploiting the intrinsic consistency of self-attention and semantic affinity in the network Ru et al. (2022). The affinity-aware contrastive learning loss utilizes the pseudo-affinity label generated by the pseudo-label y to supervise affinity maps generated by self-attention maps. The affinity loss term $\mathcal{L}_{aff}$ is constructed as:

$$\mathcal{L}_{aff} = \frac{1}{N^-}\sum_{S^-}(1 - \frac{1}{1+e^{-R_a}}) + \frac{1}{N^+}\sum_{S^+}(\frac{1}{1+e^{-R_a}}) \tag{13}$$

where S^- and S^+ denote the set of negative and positive samples in the pseudo-affinity label, and N^- and N^+ represent the number of S^- and S^- respectively. For the learning of semantic affinity representation, the Affinity-aware Pixel-level Mutual Learning loss is formulated as:

$$\mathcal{L}_{APML} = \mathcal{L}_{aff} + \beta_2 \cdot \mathcal{L}_{mi} \tag{14}$$

where β_2 balances different losses.

3.5 Integrated Objective

As shown in Fig. 2, our framework consists of four loss terms: CAM-driven Instance-leave Mutual Learning loss $\mathcal{L}_{CIML}$, Multi-scale Feature-leave Mutual Learning loss $\mathcal{L}_{MFML}$, Affinity-aware Pixel-level Mutual Learning loss $\mathcal{L}_{APML}$ and classification loss $\mathcal{L}_{cls}$.

For the classification loss, we follow the common practice and adopt the multi-label soft-margin loss as the classification loss function:

$$\mathcal{L}_{cls} = \frac{1}{C}\sum_{c=1}^{C}(z^c log(p^c) + (1-z^c)log(1-p^c)) \tag{15}$$

where p^c is the class probability vector for the classification layer, z is the ground truth image-level label, and C is the total number of classes.

The overall loss is the weighted sum of $\mathcal{L}_{CIML}$, $\mathcal{L}_{MFML}$, $\mathcal{L}_{APML}$ and $\mathcal{L}_{cls}$, which is formulated as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{CIML} + \lambda_2\mathcal{L}_{MFML} + \lambda_3\mathcal{L}_{APML} + \mathcal{L}_{cls} \tag{16}$$

where λ_1, λ_2 and λ_3 balance the contributions of different losses. In this paper, they are all set to 0.1.

3.6 Network Architecture Configuration

For the network architecture of the backbone, we utilize the Mix Transformer (MiT) Xie et al. (2021) framework, in which we simplify it's self-attention to speed up computation, and use FFN with convolution instead of positional embedding. For the refinement module of the initial pseudo-labels, we adopt PAR Ru et al. (2022). Besides, we initialize the backbone parameters with ImageNet-1k Deng et al. (2009) pre-trained weights and randomly initialize other parameters.

4 Experiments

4.1 Dataset and Implementation Details

We evaluate our method on the PASCAL VOC 2012 dataset Everingham et al. (2010) and the MS COCO 2014 dataset Lin et al. (2014), which are standard benchmarks for WSSS. The PASCAL VOC 2012 dataset consists of 20 foreground object classes and 1 background class, which is typically augmented with the SBD dataset Hariharan et al. (2011) for the training set. The augmented dataset has 10,582 images for training, 1,449 for validation, and 1,464 for testing, respectively. The MS COCO 2014 dataset consists of 80 foreground object classes and 1 background class. The dataset includes 82,081 training set images and 40,137 validation set images. Both PASCAL VOC 2012 and MS COCO 2014 datasets only use image-level labels as supervision.

Our RML is implemented with PyTorch. For the input image, we employ random rescaling in the range $[0.5, 2.0]$, random cropping and random horizontal flipping for data augmentation. The AdamW optimizer Loshchilov & Hutter (2018) is adopted to train our network with a weight decay factor of 0.01. The initial learning rate of the backbone parameter is set to 6×10^{-5}, and the initial learning rate of other parameters is set to 6×10^{-4}. Following the common practice, we warm up the classification branch for 2,000 iterations on the Pascal VOC dataset and 5,000 iterations on the MS COCO dataset. The total number of iterations for the VOC dataset and COCO dataset are 18,000 and 75,000, respectively. The weight factors in Eq. 4, Eq. 8 and Eq. 14 are 0.1, 100, and 100, respectively. The consumption time of our RML on the VOC dataset on a single NVIDIA 3090 GPU is about 3 hours, which is 3/4 of AFA Ru et al. (2022) and 1/60 of AdvCAM Lee et al. (2021b). See the appendix for more comparisons of the consumption time.

Table 1: Ablation studies of our proposed RML on the Pascal VOC 2012 val set.

Methods	CIML	MFML	APML	CRF	*val*
Our Baseline					50.3
RML (Ours)	✔				58.8
	✔	✔			61.7
	✔	✔	✔		64.9
	✔	✔	✔	✔	**67.2**

Table 2: Ablation studies of components in representation mutual learning strategies.

Methods	*val*
(a) w/o CIML	61.9
(b) w/o Multi-scale context fusion module	63.2
(c) w/o Feature-leave mutual learning loss	62.6
(d) w/o Pixel-level mutual information loss	63.0
(e) w/o Affinity-aware contrastive learning loss	62.4
RML	**64.9**

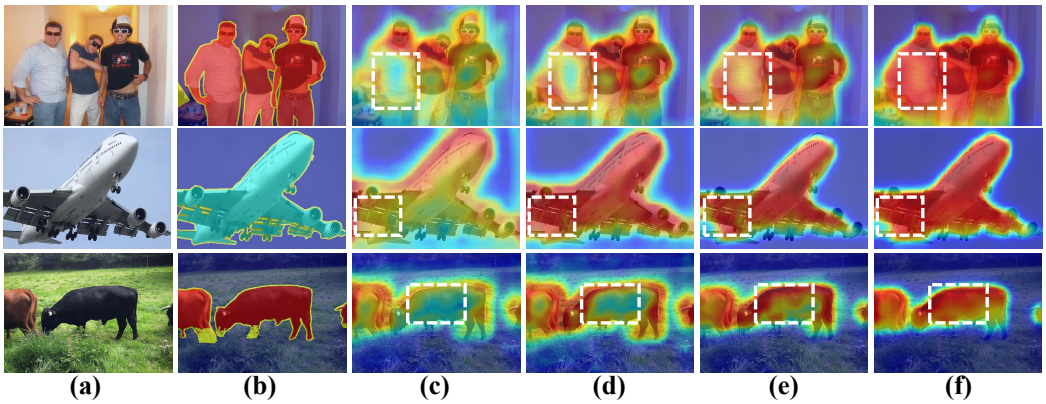

Figure 3: Visualization of CAMs. (a) Original images. (b) Ground truth. (c) CAMs generated by our baseline. (d) CAMs after applying our CIML. (e) CAMs after applying our CIML and MFML. (f) CAMs after applying our CIML, MFML and APML. The white box shows the difference.

4.2 Ablation Studies

We first evaluate the contribution of each strategy of RML to the overall performance in Tab. 1. The results show that the network using the proposed CIML improves by 16.9% compared to the baseline. The performance of the network is significantly improved by 4.9% after applying our MFML. Besides, our APML further improves the mIoU significantly to 64.9%. The final CRF Krähenbühl & Koltun (2011) post-processing improves the final performance to a mIoU of 67.2%. Briefly, the results in Tab. 1 demonstrate the effectiveness of our proposed instance-level, feature-level, and pixel-level mutual learning strategies. The generation quality of CAMs directly affects segmentation accuracy, so we also show the effect of mutual learning strategies on CAMs in Fig. 3, and it can be observed that our RML can generate more complete activation coverage.

Table 3: Performance comparisons on PASCAL VOC 2012 dataset. *F* means full supervision. *I* means image-level labels. *S* means saliency maps.

Method	*Sup*	Backbone	*val*	*test*
Fully-supervised methods				
DeepLab Chen et al. (2017)	*F*	R101	77.6	79.7
Segformer Xie et al. (2021)		MiT-B1	78.7	-
Multi-Stage weakly-supervised methods				
MCIS$_{ECCV'2020}$ Sun et al. (2020)	*I* + *S*	R101	66.2	66.9
AuxSegNet$_{ICCV'2021}$ Xu et al. (2021)		WR38	69.0	68.6
EPS$_{CVPR'2021}$ Lee et al. (2021d)		R101	**70.9**	**70.8**
SEAM$_{CVPR'2020}$ Wang et al. (2020)	*I*	WR38	64.5	65.7
SC-CAM$_{CVPR'2020}$ Chang et al. (2020)		R101	66.1	65.9
CDA$_{ICCV'2021}$ Su et al. (2021)		WR38	66.1	66.8
AdvCAM$_{CVPR'2021}$ Lee et al. (2021b)		R101	68.1	68.0
RIB$_{NeurIPS'2021}$ Lee et al. (2021a)		R101	**68.3**	**68.6**
End-to-End weakly-supervised methods				
EM$_{ICCV'2015}$ Papandreou et al. (2015)	*I*	VGG16	38.2	39.6
MIL$_{CVPR'2015}$ Pinheiro & Collobert (2015)		-	42.0	40.6
CRF-RNN$_{CVPR'2017}$ Roy & Todorovic (2017)		VGG16	52.8	53.7
RRM$_{AAAI'2020}$ Zhang et al. (2020a)		WR38	62.6	62.9
RRM+$_{AAAI'2020}$ Zhang et al. (2020a)		MiT-B1	63.5	-
1Stage$_{CVPR'2020}$ Araslanov & Roth (2020)		WR38	62.7	64.3
AA&LR$_{ACMMM'2021}$ Zhang et al. (2021b)		WR38	63.9	64.8
AFA$_{CVPR'2022}$ Ru et al. (2022)		MiT-B1	63.8	-
AFA+ CRF$_{CVPR'2022}$ Ru et al. (2022)		MiT-B1	66.0	66.3
RML (Ours)	*I*	MiT-B1	64.9	65.4
RML + CRF (Ours)		MiT-B1	**67.2**	**67.5**

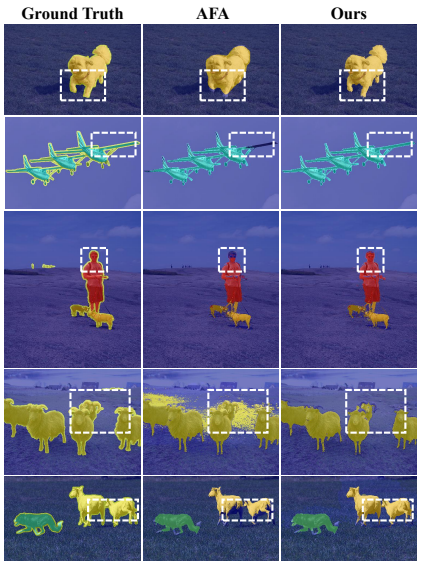

Figure 4: Qualitative segmentation results of AFA Ru et al. (2022) and our RML on PASCAL VOC benchmark.

To reveal the benefits of the components in each strategy, we further performed the ablation experiments in Tab. 2. Experiments (b) and (e) demonstrate the importance of learning contextual representations and affinity representations. Experiments (a), (c) and (d) demonstrate that instance-level, feature-level, and pixel-level representation mutual learning is crucial for segmentation network per-

formance. Removing any of these components significantly reduces the segmentation accuracy. In addition, we also provide the sensitivity analysis of α, β_1, β_2 and comparisons of consumption time in the appendix.

4.3 COMPARISON WITH STATE-OF-THE-ARTS

Segmentation performance on PASCAL VOC 2012. Tab. 3 provides a comparative overview of the current state-of-the-art on the PASCAL VOC 2012 val and test sets. In the image-level supervised setting, our method even achieves competitive performance with recent multi-stage methods. RIB Lee et al. (2021a) is trained on at least three stages and ends up only achieving 1.6% more mIoU than our method. Benefiting from our reinforcement and mutual learning of CAM representations, contextual representations and semantic affinity representations at multiple levels, the proposed RML significantly outperforms previous state-of-the-art end-to-end methods. Our method achieves a mIoU of 67.2% on VOC val set, which achieves 85.4% of the fully supervised counterpart Segformer Xie et al. (2021). The results in Tab. 3 illustrate that our method achieves state-of-the-art performance on the PASCAL VOC 2012 benchmark.

Table 4: Performance comparisons on MS COCO dataset.

Method	*Sup*	Backbone	*mIoU(%)*
Multi-Stage weakly-supervised methods			
AuxSegNet$_{ICCV'2021}$ Xu et al. (2021)	*I* + *S*	WR38	33.9
EPS$_{CVPR'2021}$ Lee et al. (2021d)		R101	**35.7**
SEAM$_{CVPR'2020}$ Wang et al. (2020)	*I*	WR38	31.9
CONTA$_{NeurIPS'2020}$ Zhang et al. (2020b)		WR38	32.8
CDA$_{ICCV'2021}$ Su et al. (2021)		WR38	31.7
CGNet$_{ICCV'2021}$ Kweon et al. (2021)		WR38	36.4
RIB$_{NeurIPS'2021}$ Lee et al. (2021a)		R101	**43.8**
End-to-End weakly-supervised methods			
AFA$_{CVPR'2022}$ Ru et al. (2022)	*I*	MiT-B1	38.0
AFA + CRF$_{CVPR'2022}$ Ru et al. (2022)		MiT-B1	38.9
RML (Ours)	*I*	MiT-B1	39.1
RML + CRF (Ours)		MiT-B1	**40.0**

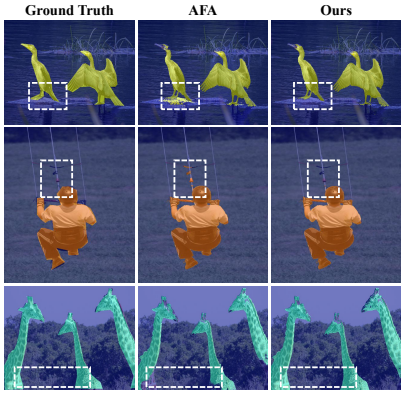

Figure 5: Qualitative segmentation results on MS COCO 2014 benchmark.

Segmentation performance on MS COCO 2014. In Tab. 4 we report the performance comparisons of our method with state-of-the-art methods on the MS COCO 2014 dataset. Our RML can achieve 40.0% mIoU on the COCO val set, significantly outperforming recent end-to-end methods and achieving competitive results among multi-stage methods. Our performance improvement does not come from a larger network structure, but mainly from the enhancement of multiple representations by our multi-level representation mutual learning strategies, which directly produces better segmentation masks for the WSSS task.

Qualitative analysis. We present qualitative results for PASCAL VOC and MS COCO in Fig. 4 and Fig. 5, respectively. On the PASCAL VOC dataset, we observe that our method outperforms AFA Ru et al. (2022), successfully segmenting fine-grained details with high fidelity. On the MS COCO dataset, our model produces segmentation masks that align well with object boundaries. See support materials for more qualitative results.

5 CONCLUSION

In this paper, we propose a direct and efficient Representation Mutual Learning (RML) framework that exploits the mutual promotion between multi-level feature representations to improve segmentation accuracy on WSSS tasks. Our RML does not require additional segmentation branches or decoders used by previous methods and directly predicts segmentation masks. RML mainly consists of a CAM-driven Instance-leave Mutual Learning strategy aimed at improving the quality of CAM representations, a Multi-scale Feature-leave Mutual Learning strategy to facilitate contextual representations, and an Affinity-aware Pixel-level Mutual Learning aimed at learning semantic affinity representations. Extensive experiments validate that our RML achieves significant improvements over previous state-of-the-art end-to-end techniques on the Pascal VOC dataset and the MS COCO dataset.

REFERENCES

Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and B Yoshua. Fitnets: Hints for thin deep nets. *Proc. ICLR*, 2, 2015.

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*, pp. 4981–4990, 2018.

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*, pp. 2209–2218, 2019.

Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4253–4262, 2020.

Tanmay Batra and Devi Parikh. Cooperative learning with visual attributes. *arXiv preprint arXiv:1705.05512*, 2017.

Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8991–9000, 2020.

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.

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*, pp. 969–978, 2022.

Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1635–1643, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

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

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Advances in neural information processing systems*, 29, 2016.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 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.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

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*, pp. 6994–7003, 2021.

Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 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*, pp. 4071–4080, 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*, pp. 2643–2652, 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*, pp. 5495–5505, 2021d.

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*, pp. 3159–3167, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

László Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4, 1993.

George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1742–1750, 2015.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1713–1721, 2015.

Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3529–3538, 2017.

Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7004–7014, 2021.

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*, pp. 347–365. Springer, 2020.

Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1522–1531, 2021.

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*, pp. 12275–12284, 2020.

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*, pp. 16765–16774, 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, 2021.

Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6984–6993, 2021.

Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12765–12772, 2020a.

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, 2020b.

Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, and Fang Wen. Robust mutual learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:2106.00609*, 2021a.

Xiangrong Zhang, Zelin Peng, Peng Zhu, Tianyang Zhang, Chen Li, Huiyu Zhou, and Licheng Jiao. Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5463–5472, 2021b.

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*, pp. 1325–1334, 2018a.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4320–4328, 2018b.

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*, pp. 2921–2929, 2016.

Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. Binocular mutual learning for improving few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8402–8411, 2021.