Pattern Recognition 162 (2025) 111348

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

M-SEE: A multi-scale encoder enhancement framework for end-to-end Weakly Supervised Semantic Segmentation

Ziqian Yang [a,b,1,2], Xinqiao Zhao [a,b,1,2], Chao Yao [c,2], Quan Zhang [a,2], Jimin Xiao [a,*,2]

[a] Xi'an Jiaotong-Liverpool University, Suzhou, China
[b] University of Liverpool, Liverpool, UK
[c] University of Science and Technology Beijing, Beijing, China

ARTICLE INFO

Keywords:
Weakly Supervised Semantic Segmentation
End-to-end framework
Knowledge distillation
Image-level labels

ABSTRACT

End-to-end image-level **W**eakly **S**upervised **S**emantic **S**egmentation (WSSS) has received increasing attention due to its simple but effective implementation. It helps to alleviate the laborious annotation costs required in semantic segmentation. In this work, we find that not all discriminative features can be extracted by a transformer encoder under image-level supervision. Thus, the decoder in end-to-end WSSS fails to predict a satisfying segmentation result. To solve this issue, we propose a **M**ulti-**S**cale **E**ncoder **E**nhancement (M-SEE) framework for enabling the encoder to extract comprehensive discriminative features and improve WSSS performance. Specifically, we first resize the original training image to various scales and calculate **C**lass **A**ctivation **M**ap (CAM) for each scale image. Then, reliable discriminative regions are mined based on the CAM and decoder segmentation result. Finally, a knowledge distillation loss is calculated among features of original scale and the scaled features of selected reliable discriminative regions. Experimental results show that our M-SEE framework achieves new state-of-the-art performances with 74.8% on PASCAL VOC 2012 test split and 45.8% on MS COCO 2014 validation split. Codes will be released.

1. Introduction

Due to the laborious pixel-wise annotation costs required in fully supervised semantic segmentation, image-level **W**eakly **S**upervised **S**emantic **S**egmentation (WSSS) is proposed to generate pixel-level predictions using only image-level annotation. Early WSSS approaches adopt a multi-stage framework to derive pseudo labels from **C**lass **A**ctivation **M**ap (CAM) based on the classification model, and then utilize the pseudo labels as the supervision for semantic segmentation model training. In order to simplify this tedious implementation process, [1,2] adopt an end-to-end framework, which encodes the image feature maps using one encoder, and then directly processes the encoder extracted feature maps to generate pseudo labels and final segmentation results parallelly.

In this work, we first find that the final segmentation performance of end-to-end WSSS method relies on the integrality of discriminative features in a feature map. However, the discriminative features are not thoroughly extracted from an input image by existing end-to-end WSSS methods with transformer encoders, degrading the final WSSS performances. This finding is depicted in Fig. 1 with an example of 'Boat'. The CAMs for the 'Boat' are all generated from the feature maps of **V**ision **T**ransformer (ViT) via a classifier following [3]. As can be found in Fig. 1(a), When the input is an original size image (*i.e.*, *size*×1), the 'Boat Mast' is not activated by any class classifier weights in CAM and its feature values are close to 0, which means the discriminative features of 'Boat Mast' are not extracted by the encoder. Under this circumstance, the 'Boat Mast' fails to be segmented by the decoder. This verifies that the discriminative feature encoding ability of a transformer encoder is not sufficient for the existing end-to-end WSSS methods and thus degrades WSSS performances.

On the other hand, we also find that certain discriminative features, which cannot be extracted from the original image, can be extracted from the image of a different scale under the end-to-end WSSS setting. As shown in Fig. 1(a), when the image of 'Boat' is down-sampled to half scale (*i.e.*, *size*×0.5), the 'Boat Mast' is activated by a classifier weight of 'Boat' in CAM and its feature values are significantly positive, which means the discriminative features of 'Boat Mast' are extracted by the encoder. The discriminative features of 'Boat Mast' can be extracted

* Corresponding author.
E-mail addresses: zqyang@liverpool.ac.uk (Z. Yang), xqz@liverpool.ac.uk (X. Zhao), yaochao1986@gmail.com (C. Yao), Quan.Zhang@xjtlu.edu.cn (Q. Zhang), jimin.xiao@xjtlu.edu.cn (J. Xiao).
[1] Co-first author.
[2] All the authors contributed equally to this research.

https://doi.org/10.1016/j.patcog.2025.111348
Received 30 July 2024; Received in revised form 26 November 2024; Accepted 6 January 2025
Available online 17 January 2025

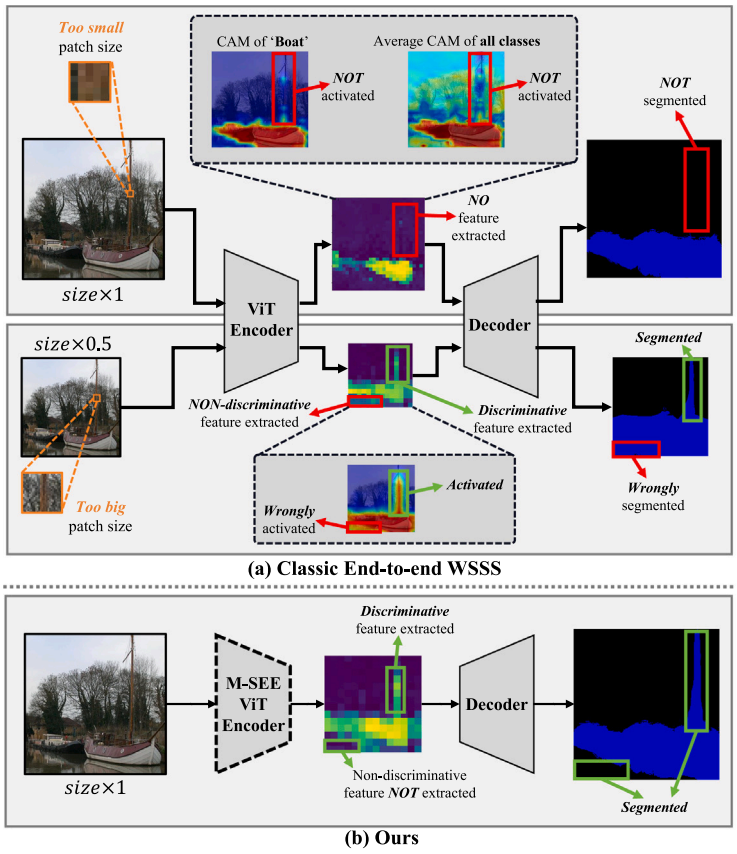

Fig. 1. How inadequate patch size of transformer encoder influences performances of conventional end-to-end WSSS methods and our method. From (a), it can be found that *too small* patch size (*e.g.*, for the image of size ×1) makes encoder cannot extract the discriminative features of 'Boat Mast', making the decoder cannot decode a complete segmentation result. However, an *adequate* patch size can make the encoder extract more discriminative features but also introduce more non-discriminative noisy features. Neither of the above cases can output a satisfying segmentation result. From (b), the encoder trained under our M-SEE method can extract complete discriminative features without introducing any non-discriminative noisy features, improving the final WSSS performance.

in the image with *size*×0.5 but cannot be extracted in the image with *size*×1. We think the reason behind it is that the image patch size of ViT encoder is too small and not suitable for the 'Boat Mast' in image with *size*×1. Thus, the encoder cannot distinguish it from other classes and regard it as a non-discriminative region with no discriminative feature extracted.

However, as shown in Fig. 1, the image patch size in the image with *size*×0.5 is too big. Though the discriminative features of 'Boat Mast' can be extracted, some noisy non-discriminative features are also extracted because one image patch covers both discriminative and non-discriminative regions. As a result, some non-discriminative regions of image with *size*×0.5 are wrongly segmented for 'Boat' class. Using a down-scaled image inevitably introduces noisy features because one patch feature of ViT now covers more pixels (including the noisy pixels), making the non-discriminative pixels also activated, harming the final semantic segmentation results. This motivates us to think if we can leverage the discriminative feature properties of multi-scale images described above to compel the encoder to extract intact discriminative features and meanwhile suppress the extraction of noisy non-discriminative features brought by multi-scale images, the final end-to-end WSSS performance can then be improved.

Inspired by the above findings, we propose a **M**uti-**S**cale **E**ncoder **E**nhancement (M-SEE) framework for end-to-end WSSS. It compels the encoder to extract comprehensive discriminative features from the original-scaled image. Through being supervised only by the reliable discriminative features, the encoder only learns useful discriminative features and does not learn the wrongly extracted non-discriminative noisy features of other scaled images, with an intact segmentation result predicted as shown in Fig. 1(b). In detail, we first resize the original training image to a smaller scale and a larger scale. Then, the CAMs are calculated for each scale image, and the reliable discriminative regions are selected based on the confidence threshold calculated CAM and decoder segmentation result. Finally, a knowledge distillation smooth ℓ1-loss is calculated among the features within the reliable discriminative regions.

Our main contributions are summarized as follows:

- We first point out that the integrality of discriminative features in feature map cannot be promised in image-level end-to-end WSSS and the decoder thus cannot decode an intact result, which degrades the final segmentation performance.
- Then, an M-SEE framework is proposed, with a multi-scale CAMs generation module, a reliable discriminative region selection module, and a multi-scaled encoder enhancement module, enhancing the discriminative feature extracting ability of a transformer encoder without bringing issues caused by noisy non-discriminative features.
- The experimental results show that our M-SEE framework achieves new state-of-the-art performances, with on PASCAL VOC 2012 validation split and test split and on MS COCO 2014 validation split.

2. Related work

2.1. Weakly supervised semantic segmentation

In general, image-level WSSS can be categorized into two branches: a multi-stage WSSS and an end-to-end single-stage one. Current multi-stage WSSS approaches [4–8], mostly follow the scheme that trains a

classification model at first to generate initial CAMs and then refines these initial CAMs as pseudo labels using methods like IRN [9] or PSA [10]. Next, the pseudo labels are used to train a semantic segmentation model for evaluating the final performance. A common drawback of CAM is that it can only focus on the most discriminative regions in the image. To solve it, several approaches provide new training schemes like erasing [11], cross-image semantic mining [12] and so on. A new trend of recent works is discovering better prototype representations and using them to encourage more completely activated object regions in CAM [13,14]. Some transformer-based works also consider the long-range modeling capability of Vision Transformer to generate more accurate CAM. MCTformer [15] designs multiple class tokens for each category to generate class-aware attention maps to refine the CAM. Other language-model-based approaches, such as CLIMS [16] and CLIP-ES [17], are proposed to incorporate extra text prompts during training, aiming at distinguishing the background parts in CAM.

The end-to-end WSSS approaches [3] perform the classification training, pseudo labels refinement, and semantic segmentation model training parallelly. RRM [1] proposes a new dense energy loss to optimize the training process, which achieves comparable performance with the multi-stage approaches. AFA [18] learns the reliable semantic affinity from multi-head self-attention in ViT, and exploits them to refine pseudo labels. Additionally, ToCo [3] addresses the issue of over-smoothing in ViT by deriving the auxiliary CAM from the intermediate layer to supervise the final patch tokens. Different from these approaches, which aim to improve pseudo labels based on the transformer architecture, our method focuses on enhancing the transformer encoder and achieving a more intact discriminative feature map for performance improvement.

2.2. Multi-scale training

In the fully supervised setting, multi-scale training has been widely applied in various tasks. For CNNs, a multi-scale data augmentation strategy [19] is first proposed for image recognition by randomly sampling training images of different scales. Then, various multi-scale training and testing paradigms are widely adopted for dense prediction tasks [20–23]. For transformers, a ResFormer [24] is built upon multi-resolution training for improved performance on a wide spectrum of testing resolutions. Other works [25,26] attempt to lay emphasis on the multi-scale spatial dimension of features instead on the input. However, as the supervision and prediction are different for image-level WSSS, the schemes used in fully supervised setting are no longer suitable. For solving this, Wang et al. propose a consistency regularization on the predicted CAMs from various transformed images with different scales to provide self-supervision for network learning [13]. Zhang et al. regard the sum of CAMs calculated from various scaled images as the pseudo label for guiding the training of end-to-end WSSS [1]. Additionally, both multi-stage and end-to-end WSSS methods employ the multi-scale operation on the CAM inference stage, and adding the suitable scaled CAMs can expand the activations and improve the CAM performance significantly. However, adding the excessive and unsuitable scaled CAMs inevitably leads to the over-activated issue and influence the CAM performance, this is because they directly combine all different scaled CAMs without any selection. Different from the above-mentioned works, we optimize the encoder through a knowledge distillation loss calculated within the reliable discriminative regions selected by our proposed reliable discriminative region selection module from multi-scaled images, achieving significant performance improvement for end-to-end WSSS.

2.3. ViT CAM generation

ViT has been widely used in WSSS CAM generation, with global feature interactions being better modeled [27]. However, unlike the well-studied CAM generated through Convolutional Neural Network (CNN) [28,29], whose one CAM value corresponds to one specific pixel region in image, the CAM generated through ViT is over-smoothing because one CAM value is related to different pixel patches in image due to the ViT self-attention mechanism. To deal with this issue, most existing methods [15,30] obtain CAM with the assistance of transformer features and corresponding self-attention maps in a two-stage manner. Recently, Ru et al. proposed a method named ToCo which supervises the final patch tokens with the pseudo token relations derived from intermediate layers and facilitates the representation consistency between uncertain local regions and global objects [3], achieving a more accurate CAM compared with previous methods. In this study, we focus on the CAM flaws caused by the fixed-size image patch separation in ViT. To achieve the best performance of our M-SEE framework for fair comparison, we adopt ViT CAM generated by ToCo [3] as our framework input CAM.

3. Methodology

To improve the integrality of discriminative features in a feature map, as explained in Section 1 for achieving better end-to-end WSSS performance, we adopt the ToCo [3] as our baseline model and its patch contrast loss is involved to address the over-smoothing issue in ViT, and upon it we further propose a **M**ulti-**S**caled **E**ncoder **E**nhan- cement (M-SEE) framework. Our framework (depicted in Fig. 2) involves multi-scale CAMs generation, reliable discriminative region selection, and multi-scale encoder enhancement. Each step of our M-SEE framework will be elaborated on in the following subsections.

3.1. Multi-scale CAMs generation

For enhancing the encoder to extract more discriminative features without introducing noisy non-discriminative features (depicted in Fig. 1), we design to calculate the CAMs for various-scale images first and then select reliable discriminative regions in them based on CAM activation values for further optimization.

With regard to the multi-scale CAMs calculation, for each original-scale training image $\mathbf{I}^o$, we down-sample and up-sample it to a small-scale image $\mathbf{I}^s$ and a large-scale image $\mathbf{I}^l$, respectively. Specifically, for $\mathbf{I}^s$, we down-sample $\mathbf{I}^o$ using bi-linear interpolation with a scale factor $SF = 0.5$; for $\mathbf{I}^l$, we up-sample $\mathbf{I}^o$ also using bi-linear interpolation but with $SF = 1.5$. In this way, three images with different sizes $\{\mathbf{I}^s, \mathbf{I}^o, \mathbf{I}^l\}$ are obtained.

Then, all images are input to a ViT encoder with a classifier for feature extraction and classification. The encoder and classifier are weakly supervised by image-level labeled data. For the input images $\{\mathbf{I}^s, \mathbf{I}^o, \mathbf{I}^l\}$, we calculate CAMs $\{\mathbf{M}^s, \mathbf{M}^o, \mathbf{M}^l\}$ for all images based on the classifier and feature maps $\{\mathbf{F}^s, \mathbf{F}^o, \mathbf{F}^l\}$ encoded by ViT encoder, following [3] through Eq. (1) as follows:

$$(\mathbf{M}^j)_c = \frac{\mathbf{ReLU}(\mathbf{W}_c\mathbf{F}^j)}{\mathbf{max}(\mathbf{ReLU}(\mathbf{W}_c\mathbf{F}^j))}, \tag{1}$$

where $j \in \{s, o, l\}$; $\mathbf{W}_c$ is the classifier weight of class c; $(\mathbf{M}^j)_c$ is the CAM of class c. Next, through the bi-linear interpolation, $\mathbf{M}^s$ is up-sampled to the size of $\mathbf{M}^o$ as $\mathbf{M}^{so}$ and $\mathbf{M}^l$ is down-sampled to the size of $\mathbf{M}^o$ as $\mathbf{M}^{lo}$. Finally, the calculated CAMs $\{\mathbf{M}^{so}, \mathbf{M}^o, \mathbf{M}^{lo}\}$ are regarded as the data foundation for reliable discriminative regions selection of our M-SEE framework.

3.2. Reliable discriminative region selection

3.2.1. Analysis of discriminative features

As analyzed in Section 1, the discriminative features, which cannot be extracted from the original-scale image, can be extracted from the image of a different scale in end-to-end WSSS. Thus, improving the consistency between the features extracted from different-scale images could be an option for compelling the encoder to learn to extract intact

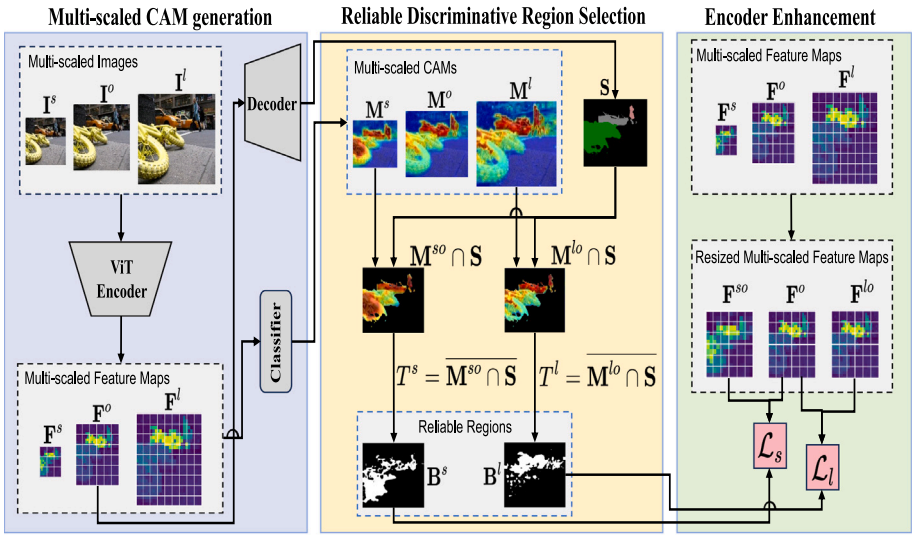

Fig. 2. An overview of our proposed framework, we first input three different-scale images $\{\mathbf{I}^s, \mathbf{I}^o, \mathbf{I}^l\}$ into the ViT encoder. The encoded three-scale feature maps $\{\mathbf{F}^s, \mathbf{F}^o, \mathbf{F}^l\}$ are then fed into the classifier for multi-scale CAMs (*i.e.*, $\{\mathbf{M}^s, \mathbf{M}^o, \mathbf{M}^l\}$) calculation, and the feature map of original-scale image $\mathbf{I}^o$ is fed into the decoder for semantic segmentation. Next, by calculating two CAM confidence thresholds T^s and T^l, we select two reliable discriminative regions $\mathbf{B}_s$ and $\mathbf{B}_l$. Based on $\mathbf{B}_s$ and $\mathbf{B}_l$, two knowledge-distillation losses $\mathcal{L}_l$ and $\mathcal{L}_s$ are calculated for enhancing the encoder to learn to extract an intact discriminative feature map without any noisy non-discriminative features. During inference, only the feature map of original-scale image is used for semantic segmentation result decoding.

discriminative features from any-scale images. Furthermore, when the encoded feature map has more discriminative features, the decoder could generate a more complete semantic segmentation result based on it. However, using a different-scale image inevitably introduces noisy non-discriminative features (depicted in Fig. 1(a)) as ViT feature of one image patch could cover both discriminative and non-discriminative pixels due to the change of image scale.

Considering this, particularly for the end-to-end WSSS structure, we introduce a reliable discriminative region selection procedure and ensure that feature consistency is only conducted in the reliable discriminative regions of different-scale images, avoiding the side effects brought by the noisy non-discriminative features and enhancing the integrality of discriminative features in feature map.

3.2.2. Thresholds generation

Specifically, given the fact that DeepLab decoder has the merit of filtering out the noisy features [31], we first regard the DeepLab decoder semantic segmentation result $\mathbf{S}$ in end-to-end WSSS as a reliable region seed. This seed region has filtered out most noisy non-discriminative features that are conflicted in pixel-level classification training [32].

Then, as a feature with a higher CAM activation value tends to be a discriminative feature, using a CAM threshold for determining if a feature is a discriminative one could be an option for completing the reliable discriminative region seed. However, as the model keeps being optimized and different images have different noise levels, the predicted CAM values of ground truths fluctuate [33]. Thus, it is hard to set a suitable fixed CAM threshold. Considering this, we choose to calculate an average CAM value within the reliable region seed as the threshold for discriminative feature determination. In this way, the threshold is calculated only based on the reliable discriminative regions of current model, without the interference of model optimization and image noises.

Specifically, we calculate two CAM confidence thresholds T^s and T^l for the small-scale feature map $\mathbf{F}^s$ and large-scale feature map $\mathbf{F}^l$ respectively as follows:

$$\begin{aligned} T^s &= \overline{\mathbf{M}^{so} \cap \mathbf{S}}, \\ T^l &= \overline{\mathbf{M}^{lo} \cap \mathbf{S}}, \end{aligned} \tag{2}$$

where $\mathbf{S}$ is the segmentation result of $\mathbf{I}^o$ output by DeepLab decoder. Here, we use $\mathbf{S}$ to mask out the initial reliable discriminative seed region first and then calculate the average CAM value of masked-out region as the CAM confidence threshold.

3.2.3. Binary masks generation

Next, for $\mathbf{I}^l$, we regard the patch region whose CAM value is higher than its corresponding CAM confidence threshold as a reliable discriminative region. Thus, we achieve a binary-masked reliable discriminative region $\mathbf{B}^l$ as follows:

$$\mathbf{B}^l_{i,j} = \begin{cases} 1, & \text{if} \quad \mathbf{M}^{lo}_{i,j} > T^l \\ 0, & otherwise, \end{cases} \tag{3}$$

where i and j denote the coordinate values. For $\mathbf{I}^s$, to prevent selecting the same reliable discriminative region, we stipulate that once a region is masked in $\mathbf{B}^l$, it will not be masked again for $\mathbf{I}^s$. As a result, we have the binary-masked reliable discriminative region for $\mathbf{I}^s$ as follows:

$$\mathbf{B}^s_{i,j} = \begin{cases} 1, & \text{if} \quad \mathbf{M}^{so}_{i,j} > T^s \quad and \quad \mathbf{B}^l_{i,j} = 0 \\ 0, & otherwise. \end{cases} \tag{4}$$

3.3. Multi-scale encoder enhancement

After having two reliable discriminative regions $\mathbf{B}^s$ and $\mathbf{B}^l$, we first down-sample $\mathbf{F}^l$ to $\mathbf{F}^{lo}$ and up-sample $\mathbf{F}^s$ to $\mathbf{F}^{so}$ through bi-linear interpolation to fit the size of $\mathbf{F}^o$. Then, we regard $\mathbf{F}^{lo}$ and $\mathbf{F}^{so}$ as two teacher feature maps and adopt knowledge distillation to transfer the reliable discriminative feature knowledge from the teacher feature maps to student feature map $\mathbf{F}^o$. The knowledge distillation is conducted through smooth $\ell 1$ loss [34,35] as follows:

$$\mathcal{L}^l_{i,j} = \begin{cases} \frac{\|\mathbf{F}^o_{i,j} - \mathbf{F}^{lo}_{i,j}\|_2^2}{2\beta}, & \text{if} \quad \|\mathbf{F}^o_{i,j} - \mathbf{F}^{lo}_{i,j}\|_2^1 \le \beta \\ \|\mathbf{F}^o_{i,j} - \mathbf{F}^{lo}_{i,j}\|_2^1 - \frac{\beta}{2}, & otherwise, \end{cases} \tag{5}$$

$$\mathcal{L}^s_{i,j} = \begin{cases} \frac{\|\mathbf{F}^o_{i,j} - \mathbf{F}^{so}_{i,j}\|_2^2}{2\beta}, & \text{if} \quad \|\mathbf{F}^o_{i,j} - \mathbf{F}^{so}_{i,j}\|_2^1 \le \beta \\ \|\mathbf{F}^o_{i,j} - \mathbf{F}^{so}_{i,j}\|_2^1 - \frac{\beta}{2}, & otherwise, \end{cases} \tag{6}$$

where β is set to 2 by default; i and j are the coordinate values. Then, the overall multi-scale feature enhancement loss $\mathcal{L}_{\mathbf{M-SEE}}$ is formulated as follows:

$$\mathcal{L}_{\mathbf{M-SEE}} = \sum_{i,j} \mathbf{B}^l_{i,j} \mathcal{L}^l_{i,j} + \sum_{i,j} \mathbf{B}^s_{i,j} \mathcal{L}^s_{i,j}. \tag{7}$$

With the help of $\mathbf{B}^l_{i,j}$ and $\mathbf{B}^s_{i,j}$, the knowledge distillation is only conducted within the reliable discriminative regions of up-scale image $\mathbf{I}^l$

Table 1
Performance comparison with other methods on PASCAL VOC 2012 validation and test split. 'Sup.' denotes the supervision type. '$\mathcal{I}$' denotes using image-level supervision. '$\mathcal{S}$' denotes using saliency map supervision. '**L**' denotes using language supervision. 'Net.' denotes the semantic segmentation model of multi-stage WSSS methods or the backbone of end-to-end WSSS methods. 'MiT' denotes mix transformer and 'ViT' denotes vision transformer.

Method	Sup.	Net.	Val	Test
Multi-stage WSSS methods				
RIB [36]	$\mathcal{I}+\mathcal{S}$	ResNet101	70.2	70.0
EDAM [37]	$\mathcal{I}+\mathcal{S}$	ResNet101	70.9	70.6
EPS [38]	$\mathcal{I}+\mathcal{S}$	ResNet101	71.0	71.8
L2G [39]	$\mathcal{I}+\mathcal{S}$	ResNet101	72.1	71.7
RCA [40]	$\mathcal{I}+\mathcal{S}$	ResNet38	72.2	72.8
HSC [41]	$\mathcal{I}+\mathcal{S}$	ResNet101	73.6	74.5
SEAM [13]	$\mathcal{I}$	ResNet38	64.5	65.7
ViT-PCM [42]	$\mathcal{I}$	ResNet101	70.3	70.9
ESOL [43]	$\mathcal{I}$	ResNet101	69.9	69.3
ReCAM [44]	$\mathcal{I}$	ResNet101	68.5	68.4
SIPE [14]	$\mathcal{I}$	ResNet101	68.8	69.7
AMN [45]	$\mathcal{I}$	ResNet101	70.7	70.6
MCTformer [15]	$\mathcal{I}$	ResNet38	71.9	71.6
SAS [46]	$\mathcal{I}$	ResNet101	69.5	70.1
OCR [47]	$\mathcal{I}$	ResNet38	72.7	72.0
BECO [48]	$\mathcal{I}$	MiT-B2	73.7	73.5
FPR [49]	$\mathcal{I}$	ResNet101	70.3	70.1
CLIMS [16]	$\mathcal{I}+\mathbf{L}$	ResNet101	70.4	70.0
CLIP-ES [17]	$\mathcal{I}+\mathbf{L}$	ResNet101	73.8	73.9
End-to-end WSSS methods				
RRM [1]	$\mathcal{I}$	ResNet38	62.6	62.9
1Stage [50]	$\mathcal{I}$	ResNet38	62.7	64.3
AFA [18]	$\mathcal{I}$	MiT-B1	66.0	66.3
TSCD [51]	$\mathcal{I}$	MiT-B1	67.3	67.5
ToCo [3]	$\mathcal{I}$	ViT-B	71.1	72.2
Ours	$\mathcal{I}$	ViT-B	**74.9**	**74.8**

and down-scale image $\mathbf{I}^s$. The encoder is thus compelled to learn to extract more reliable discriminative features from the original feature map $\mathbf{F}^o$, and the DeepLab decoder can then output a complete semantic segmentation result based on the intact feature map. Besides, as DeepLab decoder benefits from our M-SEE and outputs more accurate segmentation results, the reliable discriminative regions selected by M-SEE become more accurate. Then, these more accurate selected regions in turn improve the feature enhancement process, benefitting the decoder. This cycling training manner further boosts the performance improvement brought by our M-SEE.

4. Inference

For inference, considering there are noisy features in large- and small-scale images (*i.e.*, $\mathbf{I}^l$ and $\mathbf{I}^s$) and the feature map of original-scale image already contains intact discriminative features, we do not adopt the multi-scaled feature maps and only use the feature map of original-size image for decoding final semantic segmentation result.

5. Experiment

5.1. Experimental settings

Datasets and Evaluation Metrics. Our proposed method is evaluated on two benchmarks: PASCAL VOC 2012 with 21 classes and MS COCO dataset with 81 classes. Following common practice [1,3], PASCAL VOC 2012 dataset is augmented with the SBD dataset [52], including 10 582, 1449, and 1456 images for training, validation, and testing, respectively. The MS COCO 2014 dataset has 82 081 images for training and 40 137 images for validation. During the training phase, we only use image-level labels. **M**ean **I**ntersection **o**ver **U**nion (mIoU) is reported as the evaluation metric.

Table 2
Performance comparison with other methods on MS COCO 2014 validation split. 'RN38' denotes ResNet38 and 'RN101' denotes ResNet101.

Method	Sup.	Net.	Val
Multi-stage WSSS methods			
RIB [36]	$\mathcal{I}+\mathcal{S}$	RN101	43.8
EPS [38]	$\mathcal{I}+\mathcal{S}$	RN101	35.7
SIPE [14]	$\mathcal{I}$	RN101	40.6
AMN [45]	$\mathcal{I}$	RN101	44.7
MCTformer [15]	$\mathcal{I}$	RN38	42.0
ESOL [43]	$\mathcal{I}$	RN101	42.6
SAS [46]	$\mathcal{I}$	RN101	44.8
OCR [47]	$\mathcal{I}$	RN38	42.5
BECO [48]	$\mathcal{I}$	RN101	45.1
End-to-end WSSS methods			
AFA [18]	$\mathcal{I}$	MiT-B1	38.9
TSCD [51]	$\mathcal{I}$	MiT-B1	40.1
ToCo [3]	$\mathcal{I}$	ViT-B	42.3
Ours	$\mathcal{I}$	ViT-B	**45.8**

Implementation Details. We adopt the ViT-B [53] as the transformer encoder, which is pretrained on ImageNet. The decoder is a DeepLab-LargeFOV-based segmentation head following [54], which contains two 3×3 convolutional layers with a dilation rate of 5 and a 1×1 prediction layer. We optimized the entire network via AdamW [55] with default parameters. The learning rate is warmed up to $6e^{-5}$ at the first 1500 iterations and decayed with a rate 0.9 of the polynomial scheduler for the rest of iterations. The data augmentation is implemented following [3,18]: random resized cropping to 448^2, random horizontal flipping, and random color jittering. Besides, the network is trained for 20 000 iterations with a batch size equal to 4 in PASCAL VOC 2012 dataset, and trained for 80 000 iterations with a batch size equal to 8 in MS COCO 2014 dataset. During the testing stage, we use the additional dense CFR processing [56,57].

5.2. Comparison with state-of-the-art

PASCAL VOC 2012. In Table 1, we compare the segmentation results of our proposed **M**ulti-**S**caled **E**ncoder **E**nhancement (M-SEE) with other state-of-the-art multi-stage and end-to-end WSSS methods on PASCAL VOC 2012. It can be seen from the results that our M-SEE outperforms all previous end-to-end WSSS methods, obtaining 74.9% and 74.8% mIoU on the validation split and test split. Our M-SEE also surpasses the previous state-of-the-art end-to-end method ToCo [3] by margins of 3.8% and 2.6%, which is also the baseline method of our M-SEE. Besides, M-SEE still shows a competitive performance compared with multi-stage methods using only image-level labels, (*e.g.*, exceeding BECO [48] by 1.2% and 1.3%, and exceeding MCTformer [15] by 3.0% and 3.2%). Although some methods (*e.g.*, HSC [41] and L2G [39]) use additional saliency maps as the supervision to reduce the background noise, our method still shows a slightly better performance than theirs. Additionally, for the methods with language supervision (*e.g.*, CLIMS [16] and CLIP-ES [17]), which design the background text prompts to better differentiate the background regions and foreground objects, they exhibit superior performance than contemporaneous methods. Nevertheless, our M-SEE still achieves better results than them.

MS COCO 2014. In Table 2, we conduct comparisons in a more challenging benchmark MS COCO 2014. M-SEE achieves a performance of 45.8% on the validation set, which also outperforms end-to-end competitors ToCo [3] and AFA [18]. Moreover, we are slightly ahead of the state-of-the-art multi-stage WSSS method BECO [48] by 0.7%. All comparisons on both benchmarks verify the effectiveness of our method.

Table 3
Ablation study of our M-SEE losses (*i.e.*, $\mathcal{L}_l$ and $\mathcal{L}_s$) on PASCAL VOC 2012 validation split. 'CAM' denotes the mIoU (%) of CAM and 'Seg.' denotes the mIoU (%) of DeepLab output without the additional dense CRF processing.

#	$\mathbf{B}^l\mathcal{L}^l$	$\mathbf{B}^s\mathcal{L}^s$	CAM	Seg.
0			72.3	69.2
1	✓		74.1	71.6
2		✓	73.5	71.1
3	✓	✓	**75.1**	**72.8**

Table 4
Ablation study of our binary-masked reliable discriminative regions (*i.e.*, $\mathbf{B}^l$ and $\mathbf{B}^s$) in M-SEE losses (*i.e.*, $\mathcal{L}_l$ and $\mathcal{L}_s$) on PASCAL VOC 2012 validation split. 'w/o $\mathbf{B}^l$' or 'w/o $\mathbf{B}^s$' means discarding the binary-masked reliable discriminative region in Eq. (7)(*i.e.*, $\mathbf{B}^l$ or $\mathbf{B}^s$) and directly calculate $\mathcal{L}_l$ or $\mathcal{L}_s$ based on all features in images. 'Seg.' denotes the mIoU (%) of DeepLab decoder output without the additional dense CRF processing.

#	Scale Factor (SF)	$\mathcal{L}^l$	$\mathcal{L}^s$	Seg.
0	w/o M-SEE framework			69.2
1	1.5	w/o $\mathbf{B}^l$	w/o $\mathbf{B}^s$	70.7
2		w/ $\mathbf{B}^l$	w/o $\mathbf{B}^s$	71.6
3	0.5	w/o $\mathbf{B}^l$	w/o $\mathbf{B}^s$	70.5
4		w/o $\mathbf{B}^l$	w/ $\mathbf{B}^s$	71.1
5	1.5 & 0.5	w/o $\mathbf{B}^l$	w/o $\mathbf{B}^s$	70.2
6		w/ $\mathbf{B}^l$	w/ $\mathbf{B}^s$	72.8

5.3. Ablation studies of losses

In Table 3, we first verify the effectiveness of two losses $\mathcal{L}_l$ and $\mathcal{L}_s$ in our M-SEE framework (*i.e.*, Eqs. (5) and (6)). Setting #0 denotes the base framework without any losses (*i.e.*, without our M-SEE framework). On the one hand, by only introducing $\mathcal{L}_l$, we achieve 71.6% mIoU, 2.4% higher than base framework. On the other hand, by only introducing $\mathcal{L}_s$, we achieve 71.1% mIoU, 1.9% higher than base framework. Moreover, by adopting $\mathcal{L}_l$ and $\mathcal{L}_s$ simultaneously, we achieve the highest performance 72.8%. This verifies that either one of $\mathcal{L}_l$ and $\mathcal{L}_s$ can work well for enhancing the encoder, but adopting both can bring the most significant improvement.

5.4. Ablation studies of binary-masked reliable discriminative regions

In Table 4, we also verify the effectiveness of our binary-masked reliable discriminative regions (*i.e.*, $\mathbf{B}^l$ and $\mathbf{B}^s$) in $\mathcal{L}_l$ and $\mathcal{L}_s$. Setting #1 denotes that we implement $\mathcal{L}_l$ on the base framework without any reliable discriminative region selections, which means the feature knowledge is distilled from the whole $SF = 1.5$ feature map that inevitably distills noisy non-discriminative features. We achieve 70.7% in this setting #1, with 0.9% lower than the one of adding our binary-masked reliable discriminative region $\mathbf{B}^l$ in setting #2, which proves that we successfully select the reliable feature regions and eliminate the noisy feature regions in the feature map of images with $SF = 1.5$.

Besides, for verifying the effectiveness of another binary-masked reliable discriminative region $\mathbf{B}^s$, we conduct similar experiments in setting #3 and #4. The final semantic segmentation results still demonstrate the binary-masked reliable discriminative region $\mathbf{B}^s$ is able to effectively locate the reliable discriminative features and make the knowledge distillation losses more precise without learning any noisy non-discriminative features. Finally, in setting #5, we also conduct the feature knowledge distillation on the feature maps of images with $SF = 1.5$ and $SF = 0.5$, without adopting $\mathbf{B}^l$ and $\mathbf{B}^s$. Interestingly, the segmentation result only reaches 70.2%, lower than other single-scale settings (*i.e.*, setting #1 and setting #3). We think this is mainly because the noisy non-discriminative features are introduced too much for knowledge distillation without the help of $\mathbf{B}^l$ and $\mathbf{B}^s$, and the side-effect of noisy non-discriminative features is higher than the positive effects brought by multi-scale images (see Table 5).

Table 5
Ablation study of Scale Factor (SF) selection on PASCAL VOC 2012 validation split. 'w/o $\mathbf{B}^l$' or 'w/o $\mathbf{B}^s$' means discarding the binary-masked reliable discriminative region. 'Seg.' denotes the mIoU (%) of DeepLab decoder output without the additional dense CRF processing.

#	Scale Factor (SF)	$\mathcal{L}^l$	$\mathcal{L}^s$	Seg.
0	w/o M-SEE framework			69.2
1	0.5	w/o $\mathbf{B}^l$	w/ $\mathbf{B}^s$	71.1
2	0.75	w/o $\mathbf{B}^l$	w/ $\mathbf{B}^s$	70.5
3	1.25	w/ $\mathbf{B}^l$	w/o $\mathbf{B}^s$	70.9
4	1.5	w/ $\mathbf{B}^l$	w/o $\mathbf{B}^s$	71.6
5	2	w/ $\mathbf{B}^l$	w/o $\mathbf{B}^s$	71.6

Table 6
Ablation study of β selection in our M-SEE loss ($\mathcal{L}_{\mathbf{M-SEE}}$) on PASCAL VOC 2012 validation split. 'CAM' denotes the mIoU (%) of CAM and 'Seg.' denotes the mIoU (%) of DeepLab output without the additional dense CRF processing.

#	β	CAM	Seg.
0	1.0	73.2	71.9
1	2.0	75.1	72.8
2	3.0	72.9	70.4

5.5. Ablation studies of scale factors

In Table 7, we conduct the experiments for different Scale Factors (SF) with M-SEE losses ($\mathcal{L}_l$ or $\mathcal{L}_s$) on PASCAL VOC 2012 validation split. For the small Scale Factor settings (setting #1 and setting #2), $SF = 0.5$ achieves a higher performance with 71.1%, which surpasses the $SF = 0.75$ by a margin of 0.6%. For the large Scale Factor settings (setting #3, setting #4 and setting #5), both $SF = 1.5$ and $SF = 2$ reach 71.6%, higher than the $SF = 1.25$. Considering the $SF = 2$ images will be divided into more patches and cost more computations in the ViT encoder, $SF = 1.5$ should be a more suitable large Scale Factor for our M-SEE. Therefore, we select $SF = 0.5$ and $SF = 1.5$ as our small-scale image and large-scale image Scale Factor (SF), respectively.

5.6. Ablation studies of hyper-parameter β

In Table 6, we study the selection of β with regard to $\mathcal{L}_l$ and $\mathcal{L}_s$. The segmentation results demonstrate that the setting of $\beta = 2.0$ achieves the highest performance among $\beta = 1.0$, $\beta = 2.0$, and $\beta = 3.0$. So, we decide to set $\beta = 2.0$ as the default value.

5.7. Ablation studies of extreme scale variations

To further verify the impact of the scale number on our method, we also conduct the experiments by incorporating more scale factors. In Table 7, it shows when using 3 scales, M-SEE achieves 73.0% mIoU which is higher than the one of using 2 scales. We argue that it is because when introducing more scales, more comprehensive discriminative features can be learned by encoder through M-SEE. However, the gain brought by M-SEE will saturate when using more scales as the amount of discriminative features is limited. Meanwhile, adopting more scales inevitably brings more noises when the encoder patch size does not fit with discriminative region size of different scaled images. When the side-effect is higher than the positive effect brought by newly learned discriminative features, the final segmentation performance inevitably decreases, as reflected by the result of using 4 scales.

5.8. Visualization comparisons

To verify the effectiveness of our M-SEE qualitatively, we also visualize the CAMs and segmentation results derived from our method M-SEE and baseline method ToCo [3] in Fig. 3. As can be found in the results, compared with the baseline method ToCo [3], our M-SEE

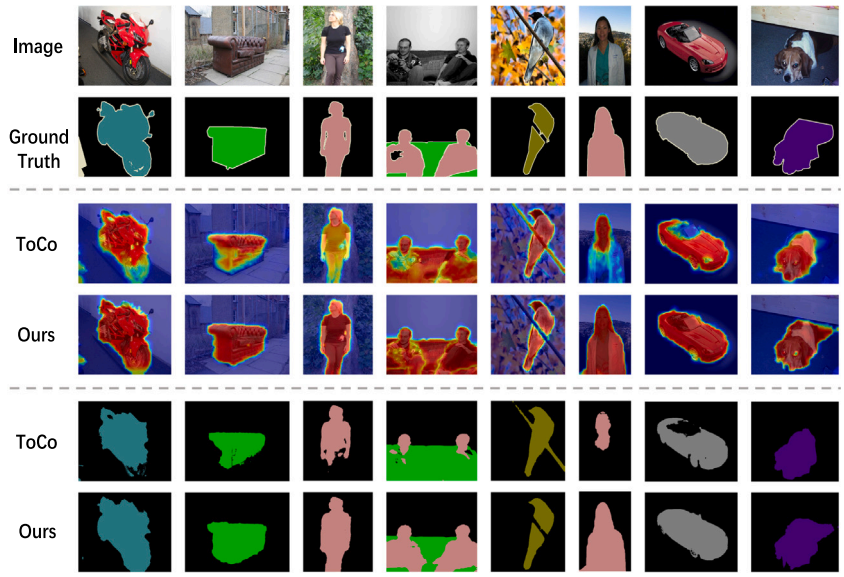

Fig. 3. Visualization comparison of CAMs and final semantic segmentation results between our baseline method ToCo and our M-SEE framework. The experimental dataset is PASCAL VOC 2012 validation split.

Table 7
Exp. of more Scale Factors (SF) on PASCAL VOC 2012 validation split. 'Seg.' denotes the mIoU (%) of decoder output.

Settings	Scale Factor (SF)	Decoder segmentation mIoU (%)
2 scales	1.5 & 0.5	72.8
3 scales	1.5 & 2.0 & 0.5	73.0
4 scales	1.5 & 2.0 & 0.5 & 0.75	72.6

generates more integral CAMs and higher quality semantic segmentation results. For instance, the front wheel of 'motorcycle' can only be mildly activated and incompletely segmented by ToCo while it can be strongly activated and fully segmented by our M-SEE. This proves the effectiveness of our method for enhancing the intact extraction of discriminative features from the feature encoder. Additionally, the 'bird' is able to be isolated from the background part 'branch' in M-SEE, instead of confusing with it in ToCo. This further confirms the usefulness of our reliable discriminative region selection procedure, which only conducts knowledge distillation among the selected reliable discriminative features and thus suppresses the noisy background (*non-discriminative*) feature extraction.

5.9. Visualizations of reliable discriminative regions

In this part, we verify the effectiveness of reliable discriminative regions (*i.e.*, $\mathbf{B}^l$ and $\mathbf{B}^s$) in M-SEE qualitatively.

In Fig. 4, for Scaling Factor $SF = 1.5$, we can see that the reliable discriminative regions (*i.e.*, B^l) are derived from the $SF = 1.5$ CAMs (*i.e.*, M^l) and the initial segmentation results (*i.e.*, S). The segmentation results are incomplete at the beginning (*i.e.*, S), they can be enhanced via our M-SEE losses as the reliable discriminative regions are effectively discovered. For example, the rearview mirror of the 'motorcycle' is calculated as the reliable discriminative regions in B_l. The features of these regions are then found and used to enhance the feature map in our method, thereby resulting in final complete segmentation results (*i.e.*, S^f).

Similarly, for Scaling Factor $SF = 0.5$, the reliable discriminative regions (B^s) and the changes of segmentation results are visualized in Fig. 5. It is obvious that many reliable discriminative regions can be detected by the $SF = 0.5$ CAMs. For example, the 'chair legs' and the 'dog tail' are able to be selected in B^s and used to enhance the feature map. Finally, the segmentation results are improved from S to S^f by our M-SEE.

5.10. Visualizations of more complex scenarios

We show visualizations for images containing multiple objects on PASCAL VOC and MS COCO in Fig. 6. The results indicate M-SEE extracts comprehensive discriminative features in multi-object images, generating intact CAMs in a variety of multi-object scenarios, which are consistent with our state-of-the-art quantitative results. It also shows our M-SEE does well on multi-class images. We argue this is because the discriminative feature extraction issue solved by M-SEE is uncorrelated with the class number in one image.

6. Discussions

6.1. Discussion of potential limitations

Despite the effectiveness of our method, there are still some limitations that need to be addressed in future work. One notable concern with multi-scale approaches is the increased computational overhead. In our implementation, two additional images are simultaneously fed into the ViT-based encoder during training, which inevitably increases both computational memory requirements and processing time. However, through our reliable discriminative region selection module, only the selected discriminative features from the two feature maps are utilized in the multi-scale feature enhancement loss, while irrelevant features are detached without additional computational resources. Unlike other methods, we remove the multi-scale processing in the inference stage, reducing both computational overhead and inference time when generating CAMs and segmentation results. The hyperparameters of our work are scale factors (SF) and β in smooth $\ell 1$ loss, different hyperparameter values cause obviously different outcomes. We will explore better ways to handle it, such as adopting adaptive scale factor selection strategies tailored to each class or each object with different regions, minimizing the impact of scale factor variations on the final results. For example, we report a representative M-SEE failure case in Fig. 7. When using $SF = 1.5\&0.5$, the horse legs are suppressed in CAMs **M** and segmentation result **S**, and the person legs are mis-classified to horses. We think this is because $SF = 1.5\&0.5$ is not perfectly suitable in this case and the transformer patch size does not fit well with the leg region size, this will be solved in our future work. Moreover, since our method relies on real-time information from the segmentation decoder, it cannot be immediately applied to multi-stage methods. Developing a suitable solution for multi-stage methods remains an open challenge that we aim to address in the future.

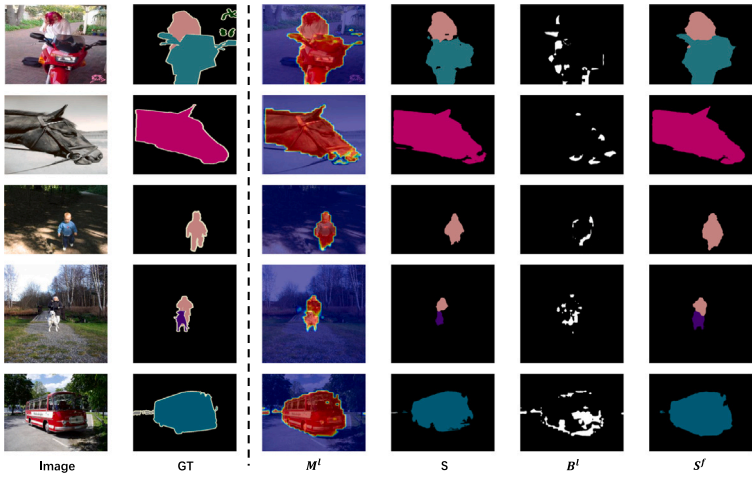

Fig. 4. Visualizations of semantic segmentation results and reliable discriminative regions in our M-SEE framework. The experimental dataset is PASCAL VOC 2012 validation split. M^l indicates $SF = 1.5$ CAMs; S indicates initial semantic segmentation results; B^l indicates reliable discriminative regions generated from M^l and S; S^f indicates the final semantic segmentation results.

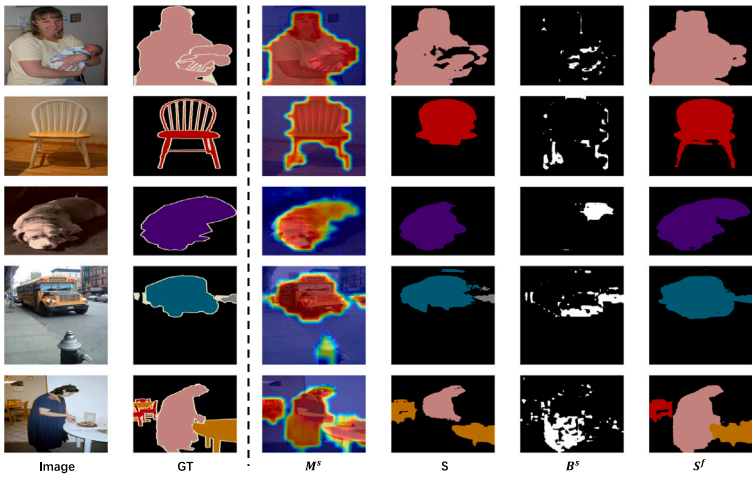

Fig. 5. Visualizations of semantic segmentation results and reliable discriminative regions in our M-SEE framework. The experimental dataset is PASCAL VOC 2012 validation split. M^s indicates $SF = 0.5$ CAMs; S indicates initial semantic segmentation results; B^s indicates reliable discriminative regions generated from M^s and S; S^f indicates the final semantic segmentation results.

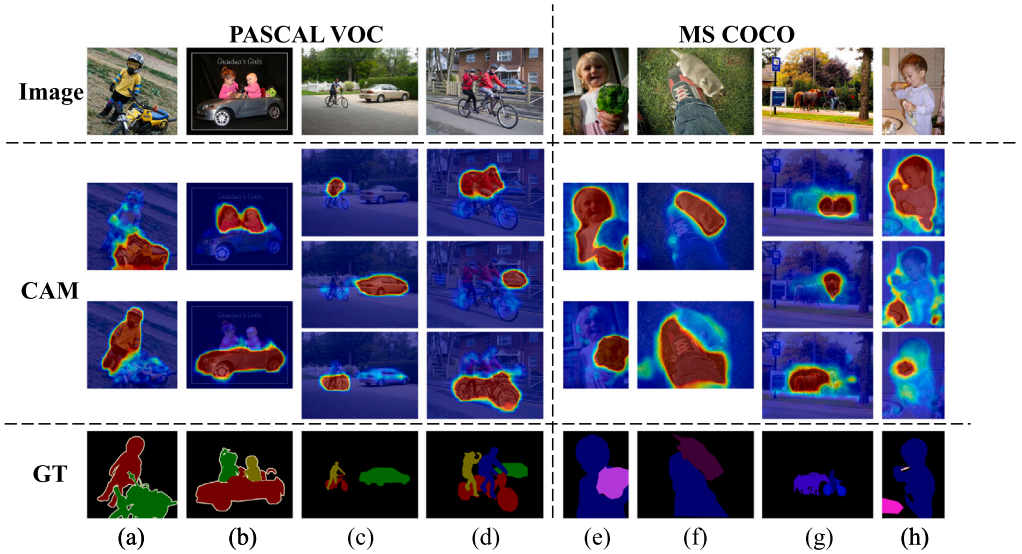

Fig. 6. (a)–(d) are from PASCAL VOC 2012; (e)–(h) are from MS COCO 2014; All images are multi-object and multi-class images.

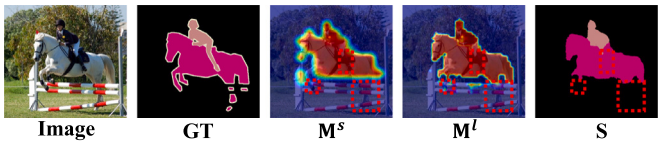

Fig. 7. M-SEE failure case with scale factor $SF = 1.5\&0.5$.

6.2. Discussion of findings in the broader field of WSSS

Under the WSSS task, we observe that some discriminative features fail to be effectively extracted by the encoder backbone, leading to under-activation on CAMs, ultimately degrading WSSS performance. We identify the underlying cause of this issue. In the ViT-based encoder, the input image is divided into patches. For certain object regions, when covered by regular-sized patches, some patches only cover limited semantic information, making it difficult for the ViT encoder to distinguish these patches, resulting in misclassification. In contrast, patches from small-scaled images can capture more global semantic information, making the patches easier to classify. This observation can be found in all ViT-based methods, including multi-stage methods.

7. Conclusion

In this paper, we demonstrate not all discriminative features can be encoded by the transformer encoder in end-to-end WSSS setting, causing the issue of incomplete semantic segmentation. For solving this, we propose a **M**ulti-**S**cale **E**ncoder **E**nhancement (M-SEE) framework for enabling the encoder to extract intact discriminative features without introducing noisy non-discriminative features, improving end-to-end WSSS performance. This finding and solution can also be adopted to other end-to-end works for improving the final segmentation performance. Moreover, our work still faces some potential limitations, particularly regarding the computational overhead and deployment challenges associated with multi-stage methods. We will overcome all these limitations and investigated other possible solutions in the future works.

CRediT authorship contribution statement

Ziqian Yang: Writing – original draft, Validation, Methodology. **Xinqiao Zhao:** Formal analysis, Data curation, Conceptualization. **Chao Yao:** Resources, Investigation, Conceptualization. **Quan Zhang:** Writing – review & editing, Supervision. **Jimin Xiao:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledge

This work was supported by the National Natural Science Foundation of China (No. 62471405, 62331003, 62301451), Jiangsu Basic Research Program Natural Science Foundation (SBK2024021981), Suzhou Basic Research Program (SYG202316) and XJTLU REF-22-01-010, XJTLU AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and SIP AI innovation platform (YZCXPT2022103).

Data availability

Data will be made available on request.

References

[1] B. Zhang, J. Xiao, Y. Wei, M. Sun, K. Huang, Reliability does matter: An end-to-end weakly supervised semantic segmentation approach, in: AAAI, 2020, pp. 12765–12772.
[2] G. Papandreou, L.-C. Chen, K.P. Murphy, A.L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: ICCV, 2015, pp. 1742–1750.
[3] L. Ru, H. Zheng, Y. Zhan, B. Du, Token contrast for weakly-supervised semantic segmentation, in: CVPR, 2023, pp. 3093–3102.
[4] H. Kweon, S.-H. Yoon, K.-J. Yoon, Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor, in: CVPR, 2023, pp. 11329–11339.
[5] S. Jo, I.-J. Yu, K. Kim, MARS: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation, 2023, arXiv preprint arXiv:2304.09913.
[6] X. Zhang, Z. Peng, P. Zhu, T. Zhang, C. Li, H. Zhou, L. Jiao, Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation, in: ACM MM, 2021.
[7] Z. Peng, G. Wang, L. Xie, D. Jiang, W. Shen, Q. Tian, Usage: A unified seed area generation paradigm for weakly supervised semantic segmentation, in: ICCV, 2023.
[8] G. Wang, X. Zhang, Z. Peng, T. Zhang, X. Tang, H. Zhou, L. Jiao, Negative deterministic information-based multiple instance learning for weakly supervised object detection and segmentation, IEEE TNNLS (2024).
[9] J. Ahn, S. Cho, S. Kwak, Weakly supervised learning of instance segmentation with inter-pixel relations, in: CVPR, 2019, pp. 2209–2218.
[10] J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in: CVPR, 2018, pp. 4981–4990.
[11] Q. Hou, P. Jiang, Y. Wei, M.-M. Cheng, Self-erasing network for integral object attention, in: NeurIPS, 2018.
[12] G. Sun, W. Wang, J. Dai, L. Van Gool, Mining cross-image semantics for weakly supervised semantic segmentation, in: ECCV, 2020, pp. 347–365.
[13] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: CVPR, 2020, pp. 12275–12284.
[14] Q. Chen, L. Yang, J.-H. Lai, X. Xie, Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation, in: CVPR, 2022, pp. 4288–4298.
[15] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, D. Xu, Multi-class token transformer for weakly supervised semantic segmentation, in: CVPR, 2022, pp. 4310–4319.
[16] J. Xie, X. Hou, K. Ye, L. Shen, Clims: Cross language image matching for weakly supervised semantic segmentation, in: CVPR, 2022, pp. 4483–4492.
[17] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, X. He, Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation, in: CVPR, 2023, pp. 15305–15314.
[18] L. Ru, Y. Zhan, B. Yu, B. Du, Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers, in: CVPR, 2022, pp. 16846–16855.
[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: ECCV, 2020, pp. 213–229.
[21] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: ECCV, 2018, pp. 734–750.
[22] B. Singh, L.S. Davis, An analysis of scale invariance in object detection snip, in: CVPR, 2018, pp. 3578–3587.
[23] W. Shen, Z. Peng, X. Wang, H. Wang, J. Cen, D. Jiang, L. Xie, X. Yang, Q. Tian, A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction, 2023, IEEE TPAMI.
[24] R. Tian, Z. Wu, Q. Dai, H. Hu, Y. Qiao, Y.-G. Jiang, Resformer: Scaling vits with multi-resolution training, in: CVPR, 2023, pp. 22721–22731.
[25] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, in: CVPR, 2021, pp. 6824–6835.
[26] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, D.Z. Pan, Multi-scale high-resolution vision transformer for semantic segmentation, in: CVPR, 2022, pp. 12094–12103.
[27] L. Zhu, Y. Li, J. Fang, Y. Liu, H. Xin, W. Liu, X. Wang, Weaktr: Exploring plain vision transformer for weakly-supervised semantic segmentation, 2023, arXiv preprint arXiv:2304.01184.
[28] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, M.-H. Yang, Weakly-supervised semantic segmentation via sub-category exploration, in: CVPR, 2020, pp. 8991–9000.
[29] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, Q. Sun, Causal intervention for weakly-supervised semantic segmentation, in: NeurIPS, 2020, pp. 655–666.
[30] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, Q. Ye, Ts-cam: Token semantic coupled attention map for weakly supervised object localization, in: ICCV, 2021, pp. 2886–2895.

[31] D. Kim, S. Lee, J. Choe, H. Shim, Weakly supervised semantic segmentation for driving scenes, 2023, arXiv preprint arXiv:2312.13646.
[32] X. Zhao, J. Xiao, S. Yu, H. Li, B. Zhang, Weight-guided class complementing for long-tailed image recognition, Pattern Recognit. (2023) 109374.
[33] J. Fan, Z. Zhang, T. Tan, Employing multi-estimations for weakly-supervised semantic segmentation, in: ECCV, 2020, pp. 332–348.
[34] R. Girshick, Fast r-cnn, in: ICCV, 2015, pp. 1440–1448.
[35] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, B. Guo, Contrastive learning rivals masked image modeling in fine-tuning via feature distillation, 2022, arXiv preprint arXiv:2205.14141.
[36] J. Lee, J. Choi, J. Mok, S. Yoon, Reducing information bottleneck for weakly supervised semantic segmentation, in: NeurIPS, 2021, pp. 27408–27421.
[37] T. Wu, J. Huang, G. Gao, X. Wei, X. Wei, X. Luo, C.H. Liu, Embedded discriminative attention mechanism for weakly supervised semantic segmentation, in: CVPR, 2021, pp. 16765–16774.
[38] S. Lee, M. Lee, J. Lee, H. Shim, Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation, in: CVPR, 2021, pp. 5495–5505.
[39] P.-T. Jiang, Y. Yang, Q. Hou, Y. Wei, L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation, in: CVPR, 2022, pp. 16886–16896.
[40] T. Zhou, M. Zhang, F. Zhao, J. Li, Regional semantic contrast and aggregation for weakly supervised semantic segmentation, in: CVPR, 2022, pp. 4299–4309.
[41] Y. Wu, X. Li, S. Dai, J. Li, T. Liu, S. Xie, Hierarchical semantic contrast for weakly supervised semantic segmentation, in: IJCAI, 2023, pp. 1542–1550.
[42] S. Rossetti, D. Zappia, M. Sanzari, M. Schaerf, F. Pirri, Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation, in: ECCV, 2022, pp. 446–463.
[43] J. Li, Z. Jie, X. Wang, X. Wei, L. Ma, Expansion and shrinkage of localization for weakly-supervised semantic segmentation, in: NeurIPS, 2022, pp. 16037–16051.
[44] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, Q. Sun, Class re-activation maps for weakly-supervised semantic segmentation, in: CVPR, 2022, pp. 969–978.
[45] M. Lee, D. Kim, H. Shim, Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds, in: CVPR, 2022, pp. 4330–4339.
[46] S. Kim, D. Park, B. Shim, Semantic-aware superpixel for weakly supervised semantic segmentation, in: AAAI, 2023, pp. 1142–1150.
[47] Z. Cheng, P. Qiao, K. Li, S. Li, P. Wei, X. Ji, L. Yuan, C. Liu, J. Chen, Out-of-candidate rectification for weakly supervised semantic segmentation, in: CVPR, 2023, pp. 23673–23684.
[48] S. Rong, B. Tu, Z. Wang, J. Li, Boundary-enhanced co-training for weakly supervised semantic segmentation, in: CVPR, 2023, pp. 19574–19584.
[49] L. Chen, C. Lei, R. Li, S. Li, Z. Zhang, L. Zhang, FPR: False positive rectification for weakly supervised semantic segmentation, in: ICCV, 2023, pp. 1108–1118.
[50] N. Araslanov, S. Roth, Single-stage semantic segmentation from image labels, in: CVPR, 2020, pp. 4253–4262.
[51] R. Xu, C. Wang, J. Sun, S. Xu, W. Meng, X. Zhang, Self correspondence distillation for end-to-end weakly-supervised semantic segmentation, 2023, arXiv preprint arXiv:2302.13765.
[52] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: ICCV, 2011, pp. 991–998.
[53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.
[54] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017, pp. 834–848, IEEE TPAMI.
[55] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv:1711.05101.
[56] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, 2014, arXiv preprint arXiv:1412.7062.
[57] J. Wang, S. Yu, B. Zhang, X. Zhao, Á.F. García-Fernández, E.G. Lim, J. Xiao, Cross-frame feature-saliency mutual reinforcing for weakly supervised video salient object detection, Pattern Recognit. (2024) 110302.