ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230480

Exploring Information Bottleneck for Weakly Supervised Semantic Segmentation

Jie Qin[1,2*], **Yueming Lyu**[1,2*] **and Xingang Wang**[2]

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences

Abstract. Image-level weakly supervised semantic segmentation (WSSS) has attracted much attention due to the easily acquired class labels. Most existing methods resort to utilizing Class Activation Maps (CAMs) obtained from the classification network to play as the initial pseudo labels. However, the classifiers only focus on the most discriminative regions of the target objects, which is referred to as the information bottleneck from the perspective of the information theory. To alleviate this information bottleneck limitation, we propose an Information Perturbation Module (IPM) to explicitly obtain the information difference maps, which provide the accurate direction and magnitude of the information compression in the classification network. After that, an information bottleneck breakthrough mechanism with three branches is proposed to overcome the information bottleneck in the classification network for segmentation. Additionally, a diversity regularization on the generated two information difference maps is proposed to improve the diversity of the output CAMs. Extensive experiments on PASCAL VOC2012 val and test sets demonstrate that the proposed method can effectively improve the weakly supervised semantic segmentation performance of the advanced approaches.

1 Introduction

Owing to the development of deep neural networks (DNNs), semantic segmentation [21, 5] has made outstanding progress in the recent few years. However, the fully supervised semantic segmentation requires time-consuming pixel-wise annotations for each image. To alleviate such expensive and unwieldy annotations, the semi-supervised segmentation [7, 22] and the image-level weakly supervised semantic segmentation (WSSS)[17, 32] has been widely studied recently. Image-level WSSS methods [17, 32, 8] usually rely on the initial seeds of Class Activation Maps(CAMs) [41] generated by the classification networks. However, the trained classifiers tend to highlight the most discriminative regions oriented to classification, while neglecting other non-discriminative but category-related regions of the target objects, making the pseudo labels obtained from such coarse CAMs unsuitable for segmentation. Some approaches[2, 32, 37] aim to refine the pseudo labels based on the initial coarse CAMs. AffinityNet [2] propagates the same semantic pixels by random walk strategy. SEAM [32] expands the activate regions by adding equivariance regularization on different augmented inputs.

∗ Equal contribution.

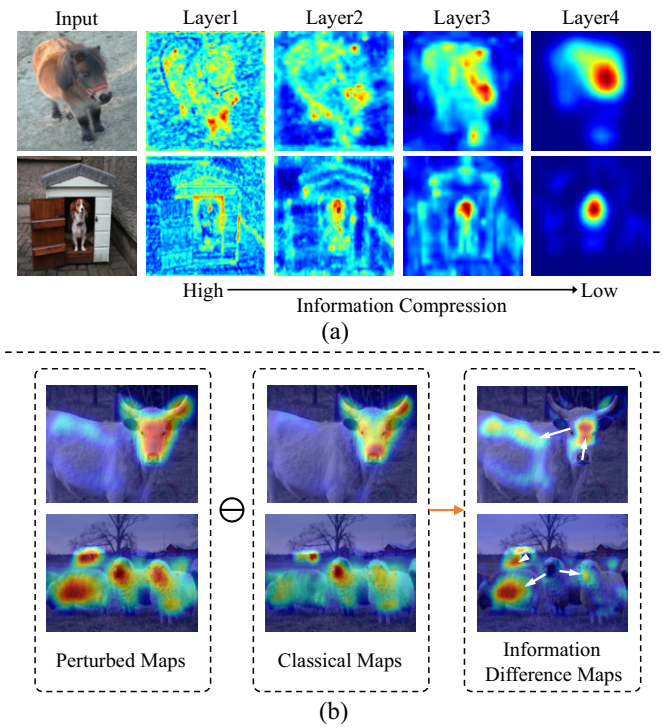

Figure 1. (a) Visualization of the information compression in the classification network ResNet-50 [11]. (b) The process of generating the information difference maps that illustrate the direction and magnitude of the information compression (white arrows).

Albeit these methods increase the response regions, they refine the initial CAMs through the results-driven approach and ignore the essential problem of the classification network that exists as the information bottleneck limitation for the segmentation task. Different from these approaches, we explicitly explore the underlying reason for such coarse CAMs from the perspective of information theory and locate the specific position of information compression in the classification network, and further reduce the information bottleneck limitation for segmentation.

We first visualize the features in each layer of the classification network in Figure 1 and observe that the amount of total activated information in each layer is gradually compressed as the layers go

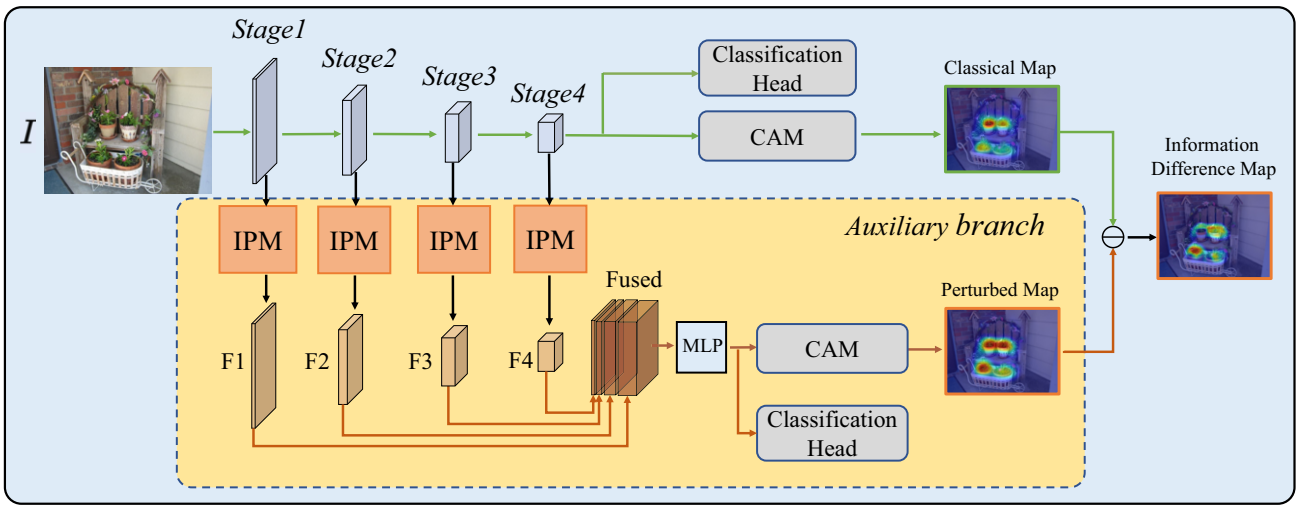

Figure 2. The pipeline to obtain the information difference maps, where the Information Perturbation Module (IPM) regulates the information flow by adding noise perturbation. The auxiliary branch is added to generate the comprehensive perturbed maps.

deeper, which is referred as the information bottleneck[30] in the classification network. To this end, the response regions of the features in the last layer only focus on the most important part of the target object since other task-irrelevant information is restricted by the information bottleneck.

In order to locate the position of the information compression, we introduce Information Perturbation Module (IPM) to the pre-trained classification model. The proposed IPM adds noise perturbation in the internal feature space to regulate the information flow process, so that the network can extract more information from the non-discriminative regions. After that, perturbed CAMs covering more response regions can be leveraged with the classical CAMs to generate the information difference maps. As shown in Figure 1, the information difference maps reflect the direction and the magnitude of the information compression.

With the explicit representation of the information bottleneck, we suggest to break through the information bottleneck limitation and provide complementary CAMs for WSSS. We propose a novel three-branch framework, which consists of a basic branch network and two auxiliary branches. The basic branch network generates the classical CAMs, which only highlight the classification-related regions. The other two auxiliary branches, equipped with Information Perturbation modules, are to compute the information difference maps with the classical CAMs. Additionally, we leverage the diversity regularization on the two difference maps to make the information compression direction be orthogonal. Therefore, the perturbed CAMs generated by two auxiliary branches are diversity. Our contributions can be summarized as follows:

- To alleviate the information bottleneck limitation, we propose a novel information perturbation module to locate the position and the direction of the information compression, named information difference maps.
- We design a three-branch framework to break through the information bottleneck limitation in the weakly supervised semantic segmentation task. The information difference maps are used to extend the classical CAMs. Diversity regularization is used to ensure the diversity of output CAMs.
- Extensive experiments demonstrate that the designed training framework can effectively improve the performance of weakly supervised semantic segmentation, *e.g.*, our method achieves 2.2% and 1.2% mIoU improvement on val and test set based on IRNet[2].

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

In recent years, there has been a significant advancement in image-level weakly supervised semantic segmentation (WSSS). Many state-of-the-art methods commonly utilize classification networks to generate Class Activation Maps (CAMs) [41] as localization maps. Most of these efforts[34, 32, 28, 18, 2, 14, 23] attempt to refine the pseudo labels obtained from the CAMs. [33] enforces CNN to focus on more regions by erasing the most discriminative regions continuously. [34] uses dilated convolution with different dilate rates to increase the response regions. AffinityNet [2] learns the relation of pixels and propagates the similar semantic pixels by a random walk algorithm. SEAM [32] captures different regions from transformed images via equivariance regularization in classification networks. OAA [14] fuses multi-attention maps in different training processes. [10] and [28] capture the information of cross-image semantic similarities and differences. [37] introduces a graph-based global reasoning unit to discover the objects in the non-salient regions. However, these approaches mentioned above resort to extract the initial CAM seeds by directly employing classification networks, which exist the information bottleneck limitation that is hard to localize the whole target region for segmentation.

2.2 Information Bottleneck

Neural networks are analyzed that tend to learn the minimal sufficient statistics of the input with respect to the output [30], which is illustrated as information bottleneck. [3] proposes a variational inference algorithm to estimate the bounds on mutual information. Recently, [40] employs the information bottleneck in classification

networks to visualize the perturbation maps that highlight the important regions for the classification task. The information bottleneck principle is exploited by DICE [24] to curtail the mutual information between the internal representation features and the inputs. When multiple models share information, the redundant information across features is also mitigated. [16] reduces the information bottleneck effects by inserting new classified layers to find more regions. [25] counts the amount of information by restricting the information flow. In this paper, we analyze the problem of the information bottleneck in classification for semantic segmentation, and propose to improve the performance of image-level weakly supervised semantic segmentation from the perspective of the information theory.

3 Methodology

3.1 Existing Information Bottleneck Problem

Most weakly supervised semantic segmentation approaches [17, 27, 32, 8] resort to obtain the CAM seeds as the initial pseudo labels from classification networks. However, with the classification loss function, the classifiers tend to focus on the most discriminative regions of the target object while neglecting other non-discriminative parts of the target object which are crucial for segmentation.

This phenomenon can be interpreted from the perspective of information theory. Since the trained classification network aims to extract maximally compressed information (the most discriminative information) from the input for accurate category distinction [25], it inevitably restricts other non-discriminative but class-relevant information flow to the output, which is referred as the information bottleneck[30]. The information bottleneck urges the model to find the compressed features by minimizing the mutual information of the input X and the features F, and maximizing the mutual information between F and the output $\widetilde{Y}$:

$$min(I(X,F) - \beta I(\widetilde{Y},F)), \tag{1}$$

where $I(\cdot,\cdot)$ denotes the mutual information, and $\beta \in [0,1]$ is a trade-off parameter between encouraging F to be predictive of $\widetilde{Y}$ and encouraging F to "forget" X. To this end, the compressed representations F are commonly the minimal sufficient features about the output and eliminate the irrelevant information which does not contribute to classification. As shown in Figure 1, the deeper layer of the network, the more severely the feature information is compressed. The last layer of features contains only the salient features of the most important part of the region, *e.g.*, head of dog.

3.2 Information Difference Maps

In order to solve the limitation of the information bottleneck problem and make the features of classification networks satisfy the requirements of the segmentation task, we first need to locate the position where the information is compressed. We propose a novel information perturbation module (IPM) to find the compression direction and region of feature information flow, which is called information difference map. As shown in Figure 2, the information perturbation module is employed in each layer to generate the intermediate perturbed features, which can reflect more context features rather than focusing too much on local region.

We incorporate the proposed IPM in each layer of the pre-trained classification network. Given the feature of i^{th} layer F_i, IPM introduces the noise perturbation into F_i and outputs more comprehensive features Z_i. After obtaining all the internal features, we fuse these multi-level features via concatenation:

$$Z_o = Cat(Z_1, Z_2, ..., Z_n), \tag{2}$$

where Z_o denotes the fused features and Cat represents the concatenate operation. Then a MLP module is used for the fused features Z_o to increase the information exchange between different layers. Finally, we leverage the classification head to complete the goal of the classification and the CAM module to obtain the perturbed maps M_p. M_p emphasizes more diverse context features than classical map, which is beneficial to locate the direction of the information compression according to the difference maps M_d.

We obtain the difference map by utilizing the new CAMs M_p and the classical CAMs M_c of the pre-trained network:

$$M_d = M_p - M_c, \tag{3}$$

where M_d is the information difference map. The obtained M_d indicates the direction and magnitude of the information compression,

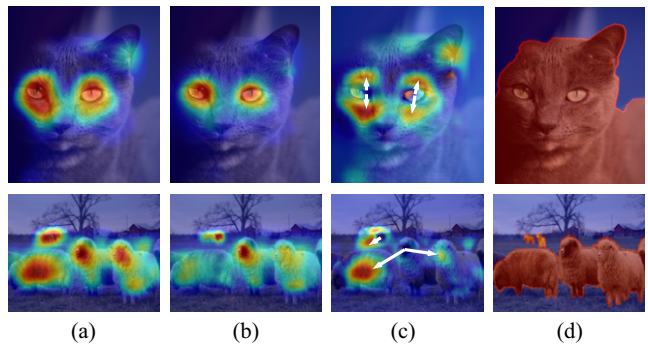

Figure 3. Illustration of different CAMs. (a) The perturbed CAMs generated by our proposed IPM module. (b) The original CAMs generated by the traditional classification networks. (c) The information difference maps where the white arrows represent the direction of the information bottleneck from the original most discriminative regions to the non-discriminative regions of the target objects. (d) The ground truth labels of the images.

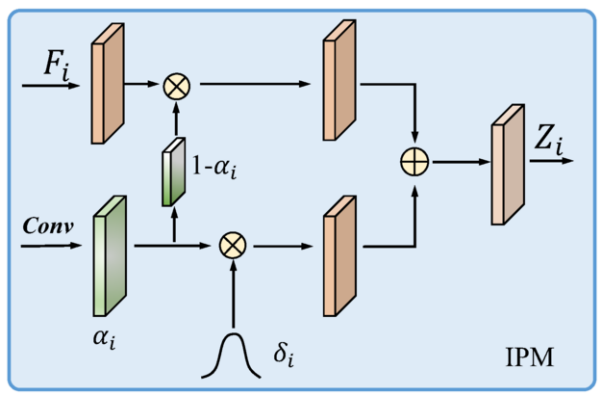

Figure 4. The architecture of the proposed information perturbation module. δ_i denotes the noise perturbation. α_i denotes the attentive map obtained from the multi-level features.

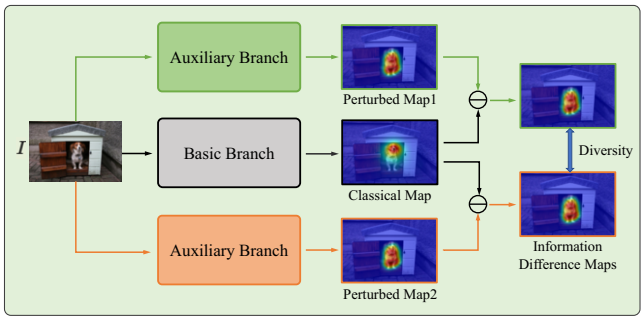

Figure 5. The framework of the proposed method for breakthrough the information bottleneck in weakly supervised semantic segmentation.

and also represents the existence of the information bottleneck. As shown in Figure 3, the visualization of the information difference maps help us to exploit the solutions of breaking through the limitation of information bottleneck for WSSS.

The proposed IPM is designed to perturb the internal feature F_i. It utilizes all internal features F_1, ..., F_n to calculate the attentive map α_i, which is defined as:

$$\alpha_i = H_i(F_1, ..., F_n), \tag{4}$$

where H_i is the hiden convolution layer. The attentive map α_i reflects the importance of each area to the target. The whole process is shown in Figure 4, the random noise is added to the current features according to α_i,

$$Z_i = (1 - \alpha_i)F_i + \alpha_i\delta_i, \tag{5}$$

where the noise δ_i is sampled from the gaussian distribution $N(\mu_i, \sigma_i^2)$ defined with the mean and variance obtained from the i-th features. F_i and Z_i are the input and output features of IPM. Since the attentive map α_i is obtained from features, the noise makes more perturbation on the important regions of the internal features. The final perturbed maps contain larger regions of target objects.

To add supervision for the information perturbation module, the mutual information between the perturbed features and the input features is formulated as the information loss:

$$L_{info} = \frac{1}{n}\sum_{i=1}^{n} I(Z_i, F_i), \tag{6}$$

where $I(Z_i, F_i)$ represents the mutual information between Z_i and F_i. In the word, the final loss function for the information difference map is summarized as:

$$L = L_{cls} + L_{info}, \tag{7}$$

3.3 The Overall Activation Map Framework

To break through the limitation of the information bottleneck in the classification networks, we design a novel architecture for generating comprehensive CAMs using the information difference maps, as shown in Figure 5. The overall framework consists of a basic classification network and two pairwise auxiliary branches. The auxiliary branches contain the information perturbation module proposed above, which aims to generate the information difference map M_d.

After obtaining two information difference maps, we employ a diversity regularization loss function L_{div} as,

$$L_{div} = cos(M_{d1}, M_{d2}), \tag{8}$$

where cos denotes the cosine distance between the two difference maps M_{d1} and M_{d2}. The diversity regularization aims to guarantee the diversity of information difference maps by increasing the cosine distance. To sum up, the whole loss function of training the framework is composed as:

$$L_{all} = L_{cls} + \lambda_1 L_{info} + \lambda_2 L_{div}, \tag{9}$$

where L_{all} denotes the final loss. λ_1 and λ_2 denote the coefficients of different losses. Finally, we leverage two information difference maps and the classical maps to expand the discriminative region for weakly supervised semantic segmentation. The final CAM M_f is obtained as:

$$M_f = M_c + \frac{1}{2}\gamma(M_{d1} + M_{d2}), \tag{10}$$

where M_c denotes the classical CAMs. γ denotes the weight of difference maps.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct all our experiments on the PASCAL VOC2012 dataset. It contains 20 foreground object classes and one background class. Following the common methods [32, 1], we use 10,582 images for training, 1,449 images for validation, and 1,456 for testing. During the whole training process, we only adopt the image-level class labels for supervision. We calculate the mean intersection over union (mIoU) of all classes to evaluate the performance of the experiments.

4.2 Implementation Details

To verify the effectiveness of our method, we conduct experiments on three basic methods, *i.e.*, IRNet[1], SEAM[32], and EPS[19]. The main classification backbones are ResNet50[11] and WideResNet38. The parameters of the backbones are pre-trained with the two basic methods and will be fixed while training our proposed framework. All the training settings are followed the two methods. Based on IRNet, we train the network for 4 epochs with a batch size of 16. The initial learning rate is set to 0.01 with a momentum of 0.9. For SEAM, we also train 4 epochs with a batch size of 4. The learning rate is set to 0.02. λ_1 and λ_2 are equal to 0.01 and 0.1. $\gamma = 0.8$ and the hardware setup is 4 NVIDIA V100 GPUs. Following the works [2, 1], we exploit the random walk algorithm on the expanded CAMs to refine the pseudo labels. After obtained the final pseudo labels for segmentation, we train the DeepLab-v2 [5] with the backbone of ResNet101 [11], which is pre-trained on the ImageNet.

4.3 Per-class Results

The comparison of segmentation results of all categories on VOC2012 validation set is elaborated, as visualized in Table 1. This provides a comprehensive analysis of the performance of each category. Our method is evaluated based on two different baseline approaches, IRNet and SEAM, demonstrating its generalization ability. The separate performance gain of applying our method on each

Table 1. Per-class segmentation results on PASCAL VOC2012 val set with DeepLab-v2 [5].

Method	*Bkg*	*Aero*	*Bike*	*Bird*	*Boat*	*bottle*	*Bus*	*Car*	*Cat*	*Chair*	*Cow*	*Table*	*Dog*	*Horse*	*Motor*	*Person*	*Plant*	*Sheep*	*Sofa*	*Train*	*Tv*	*mIoU*
MCOF [31]	87.0	78.4	29.4	68.0	44.0	67.3	80.3	74.1	82.2	21.1	70.7	28.2	73.2	71.5	67.2	53.0	47.7	74.5	32.4	71.0	45.8	60.3
FickleNet [17]	89.5	76.6	32.6	74.6	51.5	71.1	83.4	74.4	83.6	24.1	73.4	47.4	78.2	74.0	68.8	73.2	47.8	79.9	37.0	57.3	64.6	64.9
AffinityNet [2]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
SEAM [32]	88.8	68.5	33.3	85.7	40.4	67.3	78.9	76.3	81.9	29.1	75.5	48.1	79.9	73.8	71.4	75.2	48.9	79.8	40.9	58.2	53.0	64.5
IRNet[1]	87.6	70.2	30.7	76.4	47.5	63.9	75.4	61.6	82.3	33.6	74.8	68.9	75.3	71.7	56.3	63.0	50.3	69.8	44.2	67.4	62.5	63.5
IRNet + Ours	88.2	77.6	38.0	72.7	55.1	69.7	79.9	68.4	78.5	37.1	69.0	72.2	79.1	69.7	53.2	65.7	52.9	75.3	40.1	69.0	67.8	**65.7**
SEAM + Ours	89.1	71.6	37.6	87.2	44.2	63.8	73.3	80.1	83.0	34.9	76.4	52.5	74.3	76.9	72.1	78.3	51.4	70.3	43.3	61.5	56.7	65.6

Table 2. Quality results (mIoU) of pseudo labels on the VOC2012 train images. The "CAM" and "Pseudo" indicate the results of the CAMs and the refined pseudo labels with different methods.

Method	Backbone	Publication	CAM	Pseudo
AffinityNet [2]	Wide ResNet38	CVPR2018	48.0	59.7
Chang *et al*. [4]	Wide ResNet38	CVPR2020	50.9	63.4
CONTA [39]	ResNet50	NIPS2020	48.8	67.9
EDAM[35]	Wide ResNet38	CVPR2021	52.8	58.2
AdvCAM[18]	ResNet50	CVPR2021	55.6	69.9
CSE[29]	Wide ResNet38	ICCV2021	56.6	58.6
IRNet [1]	ResNet50	CVPR2019	48.3	66.5
IRNet + Ours	ResNet50	-	**52.5** +4.2	**68.1** +1.6
SEAM [32]	Wide ResNet38	CVPR2020	55.4	63.6
SEAM + Ours	Wide ResNet38	-	**56.8** +1.4	**64.8** +1.2
EPS [19]	Wide ResNet38	CVPR2021	69.4	71.6
EPS + Ours	Wide ResNet38	-	**70.0** +0.6	**71.9** +0.3

baseline approach further proves the effectiveness of our method itself. It is highlighted that our method primarily improves the performance of baseline categories with poor results, such as "bottle" of IRNet and "sofa" of SEAM. This indicates that our method excels in boosting the performance of long-tail categories. The final mIoU of 65.7% and 65.6% achieved by our method based on IRNet and SEAM on VOC2012 validation set are reported, providing quantitative results for analysis. This also provides evidence for the following qualitative discussion. The qualitative analysis in the end concludes that our method helps poorly performed categories overcome the information bottleneck, especially benefiting long-tail categories. This further proves the validity of our method.

4.4 Improvements of the Pseudo Labels

To verify the effectiveness of our proposed method in generating CAMs and pseudo labels, we conducted experiments on the VOC2012 training set and compared the initial CAMs with the final pseudo labels, as summarized in Table 2. It is worth noting that we applied our method to three state-of-the-art techniques, namely IRNet [1], SEAM [32], and EPS [19].

Our approach yield substantial performance gains over the baseline method IRNet, improving the CAM and pseudo label accuracy by 4.2% and 1.6%, respectively. We observed similar improvements in the pseudo label accuracies of the SEAM and EPS methods. These results demonstrate that our method is highly effective in alleviating the information bottleneck present in conventional CAM-based paradigms and can be seamlessly integrated with existing techniques to enhance their performance in generating pseudo labels.

Our extensive experiments demonstrate the superiority of our proposed method over existing CAM-based techniques. We believe that our method has the potential to significantly advance the field of weakly supervised learning and enable the development of more accurate and efficient models for a wide range of applications.

4.5 Comparison with State-of-the-art methods

To evaluate the effectiveness of our proposed approach in generating high-quality pseudo labels, we conducted fully supervised segmentation experiments on the PASCAL VOC2012 dataset using the DeepLab v2[5] network. In our experiments, we trained the network using the pseudo labels generated by our approach. We compared the segmentation results obtained using our approach with those obtained using state-of-the-art methods on both the validation and test sets of the PASCAL VOC2012 dataset, as shown in Table 3. Our results demonstrate that our approach outperforms the baseline method IRNet by 2.2% and 1.2% on the validation and test sets, respectively. Moreover, our approach consistently improves the performance of the SEAM baseline.

To further demonstrate the effectiveness of our approach, we evaluated it on the state-of-the-art EPS method, which uses image-level and saliency supervision. Our method produced a significant improvement in segmentation performance, achieving an accuracy of 71.4% and 72.0% on the validation and test sets, respectively. These results are a testament to the ability of our approach to extract more comprehensive features, thereby enhancing the performance of the segmentation task. The success of our approach can be attributed to its ability to generate high-quality pseudo labels that capture the true object boundaries and semantics. Our approach also leverages the power of deep neural networks to learn more features, which are crucial for accurate segmentation. Additionally, our approach is highly flexible and can be easily integrated with existing methods to improve their performance in generating pseudo labels.

In summary, our experiments demonstrate the efficacy of our proposed approach in generating high-quality pseudo labels for weakly supervised segmentation. Our approach outperforms state-of-the-art methods on the PASCAL VOC2012 dataset, highlighting its potential to advance the field of weakly supervised learning and enable the development of more accurate and efficient segmentation models.

4.6 Visualization of Segmentation Results

Figure 6 shows the segmentation results obtained using our proposed approach with the IRNet[1] and SEAM[32] methods on the validation set of the PASCAL VOC2012 dataset. As can be seen from the figure, the segmentation results obtained using IRNet and SEAM often suffer from the loss of fine details in the border regions and confusion between the regions where objects intersect. In contrast, our proposed approach successfully identifies the important regions of

Table 3. Comparison with the state-of-the-art methods on PASCAL VOC2012 val and test set. All results are evaluated in mIoU(%). $\mathcal{I}$ represents the image-level label and $\mathcal{S}$ indicates the salient label.

Methods	Backbone	Sup.	Val	Test	Methods	Backbone	Sup.	Val	Test
AffinityNet [2]	Wide ResNet38	$\mathcal{I}$	61.7	63.7	MCOF [31]	ResNet101	$\mathcal{I}+\mathcal{S}$	60.3	61.2
IRNet [1]	ResNet101	$\mathcal{I}$	63.5	64.8	SeeNet [12]	ResNet101	$\mathcal{I}+\mathcal{S}$	63.1	62.8
CIAN [10]	ResNet101	$\mathcal{I}$	64.3	65.3	DSRG [13]	ResNet101	$\mathcal{I}+\mathcal{S}$	61.4	63.2
SSDD [26]	ResNet101	$\mathcal{I}$	64.9	65.5	AuxSegNet [36]	ResNet101	$\mathcal{I}+\mathcal{S}$	69.0	68.6
OAA+ [14]	ResNet101	$\mathcal{I}$	65.2	66.9	FickleNet [17]	ResNet101	$\mathcal{I}+\mathcal{S}$	64.9	65.3
SEAM [32]	Wide ResNet38	$\mathcal{I}$	64.5	65.7	MCIS [28]	ResNet101	$\mathcal{I}+\mathcal{S}$	66.2	66.9
Chang *et al*. [4]	ResNet101	$\mathcal{I}$	66.1	65.9	ICD [9]	ResNet101	$\mathcal{I}+\mathcal{S}$	67.8	68.0
Zhang *et al*. [38]	ResNet101	$\mathcal{I}$	66.3	66.5	Yao *et al*. [37]	ResNet101	$\mathcal{I}+\mathcal{S}$	68.3	68.5
Chen *et al*. [6]	ResNet101	$\mathcal{I}$	65.7	66.7	EDAM [35]	ResNet101	$\mathcal{I}+\mathcal{S}$	70.9	70.6
CONTA [39]	ResNet101	$\mathcal{I}$	66.1	66.7	G-WSSS[20]	ResNet101	$\mathcal{I}+\mathcal{S}$	68.2	68.5
DRS [15]	ResNet101	$\mathcal{I}$	66.8	67.4	NSROM [37]	ResNet101	$\mathcal{I}+\mathcal{S}$	70.4	70.2
AdvCAM [18]	ResNet101	$\mathcal{I}$	68.1	68.0	EPS [19]	ResNet101	$\mathcal{I}+\mathcal{S}$	71.0	71.8
IRNet + Ours	ResNet101	$\mathcal{I}$	**65.7** +2.2	66.0 +1.2	EPS + Ours	ResNet101	$\mathcal{I}+\mathcal{S}$	**71.4** +0.4	**72.0** +0.2
SEAM + Ours	Wide ResNet38	$\mathcal{I}$	65.6 +1.1	**66.5** +0.8					

Figure 6. Qualitative comparison on the PASCAL VOC2012 validation set. (a) Input images. (b) The segmentation results of IRNet[1]. (c) The segmentation results of SEAM[32]. (d) The segmentation results of our method. (e) Ground truth labels.

the target objects and accurately segments them. The results obtained using our approach are more visually appealing and exhibit fewer errors in the border regions. This can be attributed to the fact that our approach generates high-quality pseudo labels that capture the true object boundaries and semantics. These pseudo labels are then used to train the network in a weakly supervised manner, enabling it to learn more discriminative features that are crucial for accurate segmentation.

4.7 Effect of main components

We conduct the ablation study to verify the effect of the proposed method. As shown in Table 4, with IPM, our method can improve the original CAM by 1.4% of mIoU. To increase the diversity of the expanded CAMs, the two auxiliary branch with IPM are introduced for consistent regularization. Our method achieves 51.9% performance. The proposed diversity regularization L_{div} is proposed to increase the diversity of the expanded CAMs, which bring out the perfor-

Table 4. Effectiveness of different designed modules. Baseline indicates the original CAM generation. IPM represents that only one auxiliary branch is employed. And IPM^2 represents the two auxiliary branches. L_{div} denotes the diversity regularization employed on the two information difference maps.

Baseline	IPM	IPM^2	L_{div}	mIoU (%)
√				48.3
√	√			49.7
√	√	√		51.9
√	√	√	√	52.5

mance improvement by 0.6%. By combining all these components, our full method performs significantly performance of CAMs of 52.5%. Additionally, the results show that our method break through information bottleneck in the pre-trained classification networks and releases the category information to improve the segmentation performance.

We conducted an ablation study to analyze the impact of each component. As summarized in Table 4, our method achieved an improvement of 1.4% in mIoU by incorporating IPM. This demonstrates the effectiveness of IPM in generating high-quality CAMs that accurately capture the object boundaries and semantics. To further increase the diversity of the expanded CAMs, we introduced two auxiliary branches with IPM for consistent regularization. This resulted in a further improvement of 0.6% in mIoU, bringing the overall performance of our approach to 51.9%. This highlights the importance of regularization in producing diverse and accurate CAMs. We also proposed a diversity regularization term L_{div} to encourage the network to produce more diverse CAMs. This component improved the performance of our approach by an additional 0.6% in mIoU, demonstrating the importance of diversity in producing accurate and robust CAMs. By combining all these components, our full method achieved a significant improvement in the performance of CAMs, achieving an mIoU of 52.5%. These results demonstrate the effectiveness of our proposed method in overcoming the information bottleneck in pre-trained classification networks and releasing the category information to improve the segmentation performance.

Overall, our ablation study provides a comprehensive analysis of the effectiveness of each component of our proposed approach. The results demonstrate that each component plays a crucial role in producing accurate and diverse CAMs.

5 Discussions and Limitations

We use publicly available training datasets with official authorization to avoid data privacy issues. WSSS methods may have negative societal consequences, such as in surveillance or facial recognition systems. It is vital to consider potential misuses and take ethical measures. Our method utilizes general scene images, avoiding bias against specific groups or regions.

Our framework has limitations in certain scenarios or data conditions, such as complexity of target objects. Our framework may struggle to capture complex visual patterns and structures of certain objects, particularly those with highly variable appearances or occlusions.

6 Conclusion

In this paper, we attempt to explore the information bottleneck problem for image-level weakly supervised semantic segmentation that the trained classifiers only focus on the most discriminative regions of the target objects. To alleviate this limitation, we propose an Information Perturbation Module (IPM) to locate the position of the information compression with information difference maps. With the advantages of the information difference maps, we then design a three-branch framework to overcome the information bottleneck limitation, making it more suitable for WSSS. The qualitative and quantitative experiments show that the proposed method can effectively improve the weakly supervised semantic segmentation performance of the advanced methods.

References

[1] 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).

[2] 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).

[3] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, 'Deep variational information bottleneck', *arXiv preprint arXiv:1612.00410*, (2016).

[4] 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).

[5] 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).

[6] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang, 'Weakly supervised semantic segmentation with boundary exploration', in *European Conference on Computer Vision*, pp. 347–362. Springer, (2020).

[7] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang, 'Semi-supervised semantic segmentation with cross pseudo supervision', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, (2021).

[8] Junsuk Choe, Seungho Lee, and Hyunjung Shim, 'Attention-based dropout layer for weakly supervised single object localization and semantic segmentation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2020).

[9] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan, 'Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4283–4292, (2020).

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

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[12] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng, 'Self-erasing network for integral object attention', *arXiv preprint arXiv:1810.09821*, (2018).

[13] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang, 'Weakly-supervised semantic segmentation network with deep seeded region growing', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7014–7023, (2018).

[14] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong, 'Integral object mining via online attention accumulation', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2070–2079, (2019).

[15] Beomyoung Kim, Sangeun Han, and Junmo Kim, 'Discriminative region suppression for weakly-supervised semantic segmentation', in

Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1754–1761, (2021).

[16] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon, 'Reducing information bottleneck for weakly supervised semantic segmentation', *Advances in Neural Information Processing Systems*, **34**, (2021).

[17] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon, 'Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5267–5276, (2019).

[18] 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, (2021).

[19] 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, (2021).

[20] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang, 'Group-wise semantic mining for weakly supervised semantic segmentation', *arXiv preprint arXiv:2012.05007*, (2020).

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell, 'Fully convolutional networks for semantic segmentation', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, (2015).

[22] Jie Qin, Jie Wu, Ming Li, Xuefeng Xiao, Min Zheng, and Xingang Wang, 'Multi-granularity distillation scheme towards lightweight semi-supervised semantic segmentation', in *European Conference on Computer Vision*, pp. 481–498. Springer, (2022).

[23] Jie Qin, Jie Wu, Xuefeng Xiao, Lujun Li, and Xingang Wang, 'Activation modulation and recalibration scheme for weakly supervised semantic segmentation', in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2117–2125, (2022).

[24] Alexandre Rame and Matthieu Cord, 'Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation', *arXiv preprint arXiv:2101.05544*, (2021).

[25] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf, 'Restricting the flow: Information bottlenecks for attribution', *arXiv preprint arXiv:2001.00396*, (2020).

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

[27] Krishna Kumar Singh and Yong Jae Lee, 'Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization', in *2017 IEEE international conference on computer vision (ICCV)*, pp. 3544–3553. IEEE, (2017).

[28] 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).

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

[30] Naftali Tishby and Noga Zaslavsky, 'Deep learning and the information bottleneck principle', in *2015 ieee information theory workshop (itw)*, pp. 1–5. IEEE, (2015).

[31] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma, 'Weakly-supervised semantic segmentation by iteratively mining common object features', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1354–1362, (2018).

[32] 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).

[33] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan, 'Object region mining with adversarial erasing: A simple classification to semantic segmentation approach', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1568–1576, (2017).

[34] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang, 'Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7268–7277, (2018).

[35] 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).

[36] 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).

[37] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang, 'Non-salient region object mining for weakly supervised semantic segmentation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2623–2632, (2021).

[38] 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*, pp. 12765–12772, (2020).

[39] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun, 'Causal intervention for weakly-supervised semantic segmentation', *arXiv preprint arXiv:2009.12547*, (2020).

[40] Andrey Zhmoginov, Ian Fischer, and Mark Sandler, 'Information-bottleneck approach to salient region discovery', in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 531–546. Springer, (2020).

[41] 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).