A Novel Evaluation Metric for Class Activation Mapping Methods in Weakly Supervised Learning

Qingdong Cai
School of Electronic and Electrical Engineering
University of Sheffield
Sheffield, UK
qcai7@sheffield.ac.uk

Charith Abhayaratne
School of Electronic and Electrical Engineering
University of Sheffield
Sheffield, UK
c.abhayaratne@sheffield.ac.uk

***Abstract*—Class activation mapping (CAM) methods are a fundamental component of weakly supervised learning tasks, and research on CAMs specifically tailored for these tasks has been steadily increasing. However, existing evaluation metrics primarily assess the faithfulness and interpretability of CAM algorithms. These metrics do not adequately assess whether an activation method can precisely activate the true target region and produce sufficiently strong activation values, both of which are crucial for downstream weakly supervised learning tasks. Furthermore, CAM performance is also evaluated using metrics derived from downstream tasks; however, these metrics rely on manually selected thresholds, which can result in substantially different, or even contradictory, evaluation outcomes across thresholds. This threshold sensitivity introduces the risk of implicit cherry-picking, thereby undermining the fairness, robustness, and reproducibility of CAM evaluation. To address the inappropriateness and threshold sensitivity for existing metrics, we propose a new metric, termed Mask-Based Intersection Energy (MBIE), which is threshold-insensitive and leverages object masks to more accurately evaluate the activation outcomes of CAM methods. The implementation is publicly available at https://github.com/caiqingdong111111/MBIE_New_CAM_Evaluation.**

***Index Terms*—Evaluation Metrics, Class Activation Mapping, Convolutional Neural Network, Weakly Supervised Learning,**

I. Introduction

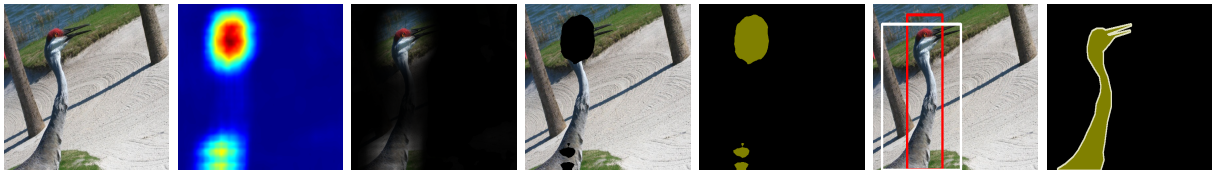
Class activation mapping (CAM) methods serve as a fundamental component in weakly supervised learning tasks [1]–[3]. By generating activation maps that highlight regions contributing most strongly to a model's prediction for a given class [4]–[6], CAM provides coarse object information that helps compensate for the lack of supervision in weakly supervised learning tasks [3], [4], [7]–[9]. However, early CAM methods often generate coarse and incomplete semantic cues for target objects. These deficiencies can be amplified during subsequent training stages, resulting in biased or partial object representations and, consequently, reduced performance in downstream weakly supervised tasks, including Weakly Supervised Semantic Segmentation (WSSS) and Weakly Supervised Object Localization (WSOL) [10]–[13].

Driven by the need for semantic completeness in weakly supervised learning, a wide range of CAM variants have been developed or adapted to support such tasks [1], [3], [14], [15]. However, despite broad acknowledgment of the importance of task-specific CAM design, suitable evaluation metrics for these activation algorithms remain largely underexplored. This limitation stems from the original role of CAM methods as visual explanation tools for convolutional neural networks (CNNs) [16], which has led existing metrics, such as Average Drop (AD), Increase in Confidence (IC) [17], [18], Deletion and Insertion curves [19], and Image Occlusion (IO) [3], [14], [20], to predominantly emphasizing faithfulness and effectiveness. These metrics mainly assess whether the regions highlighted by CAM algorithms contribute to model predictions, but they do not evaluate whether the activated regions accurately and completely cover the target objects, nor do they assess the corresponding activation intensity [3], [21]–[23].

Previous studies have employed task-specific metrics, such as mean Intersection over Union (mIoU) and Top-1 localization accuracy (Loc1) [3], [14], [24], to evaluate the ability of CAM algorithms to localize target objects. However, these evaluation metrics rely on an additional thresholding step that converts continuous activation maps into representations suitable for evaluation [3], [18], [20]. Due to the absence of a principled and consistent threshold selection criterion, this process introduces a significant risk of implicit cherry-picking [25], whereby evaluation results can be selectively reported under favorable threshold settings while disregarding less supportive outcomes, ultimately undermining the fairness and reproducibility of CAM evaluation.

To address the mismatch and inappropriateness for existing metrics, we propose a novel metric, termed Mask-Based Intersection Energy (MBIE), which is designed to be threshold-insensitive and leverages object masks to accurately evaluate the activation outcomes of CAM methods. MBIE comprises two complementary components MBIE_c and MBIE_b. The former quantifies the activation intensity of the target class within its corresponding semantic mask, while the latter measures the extent of false or unintended activation in background regions. By operating on class-specific semantic masks, MBIE enables accurate and class-aware evaluation of CAM methods tailored to downstream weakly supervised tasks.

1 MBIE operates on a per-class basis, enabling precise and class-specific evaluation of CAM methods. To the best of our knowledge, this work is the first to introduce evaluation metrics tailored to activation methods used in downstream weakly supervised tasks.



(a) Original image (b) Activation map (c) Explanation map (d) Occlusion image (e) Segmentation (f) Bounding box (g) Mask

Fig. 1: The original image and its corresponding activation map and explanation map. (a) The Original image. (b) The activation map is obtained by Grad-CAM++ method with Resnet50. The red highlighted points represent high activation values and strong activation. Conversely, the blue highlighted points represent low activation values and weak activation. (c) The explanation map. (d) The occlusion image. (e) The segmentation result is obtained by converting the activation map. (f) The red detection box is the estimated object bounding box. The white one is the ground-truth bounding box. (g) The semantic segmentation ground truth mask.

2 MBIE is a threshold-insensitive metric which eliminates implicit cherry-picking risk, providing a consistent and unbiased quantifies of CAM performance.
3 MBIE enables a comprehensive evaluation of CAM methods by quantifying both the activation strength within the target region and the activation leakage in background regions based on ground truth mask, ensuring a more accurate and complete assessment.

II. Related work

This section reviews CAM methods commonly used in weakly supervised learning and subsequently discusses the corresponding evaluation metrics

A. CAM Methods for Weakly Supervised Learning

The original CAM method, introduced by Zhou et al. [26], was developed to provide visual explanations for CNNs and has become widely utilized activation algorithms, spanning both visual explanation research and weakly supervised learning applications [27]–[29]. Subsequently, Grad-CAM and Grad-CAM++ were introduced to generalize the original CAM method by removing its dependence on a specific network architecture, allowing activation maps to be generated without requiring a global average pooling layer [17], [29]. As these methods have been increasingly adopted in weakly supervised learning [30], their inherent limitations have become more evident. In particular, the activation maps they produce tend to be spatially coarse and emphasize only the most discriminative regions, which is insufficient for weakly supervised tasks that require detailed object coverage. This limitation is especially pronounced in WSSS, where accurate pixel-level delineation of object regions is essential [31], [32]. In response to the demands of weakly supervised learning, a growing body of activation algorithms has been proposed [3], [14], [15], [30], [33]. These methods exploit intermediate network representations to generate more fine-grained activation maps, thereby alleviating the limitations of earlier approaches and giving rise to a distinct and rapidly evolving research direction.

B. Evaluation Metrics for CAM algorithms

Evaluation metrics for CAM methods were introduced alongside the development of the methods themselves. IO is among the earliest metrics developed to evaluate the quality of visualization algorithms [20]. This method systematically involves occluding specific regions of an image and analyzing the resulting changes in the model's predictions [14]. Similarly, AD and IC are also metrics designed to assess the quality and fidelity of visual explanation results. They were first introduced in Grad-CAM++ and have since become one of the standard evaluation criteria [29]. Subsequently, Deletion and Insertion curves [19], as well as the Energy-Based Pointing Game (EBPG) [18], were proposed, but these two metrics are rarely used in practice. Deleting and inserting curves involve the iterative addition or removal of pixels according to their activation values, while continuously monitoring the resulting changes in model confidence. Consequently, the inference process for a single image may require dozens or even hundreds of iterations, and when applied to large-scale datasets comprising tens of thousands of images, this procedure can impose a substantial computational burden. EBPG computes the ratio of total activation energy within the object's ground-truth bounding box to the total energy of the entire activation map. Since bounding boxes often include background regions with irregular shapes and sizes, the metric is influenced by both box size and object shape, causing background activations to be mistakenly counted as part of the object's energy. Therefore, an image is included for metric calculation only if the object occupies less than 50% of the total image area, in order to reduce the influence of excessive background regions [18]. This constraint substantially limits the applicability of the metric.

The other two evaluation metrics are mIoU and *Loc1*, which originate from the WSSS and WSOL domains, respectively [3], [14], [20], [29]. As noted earlier, both *Loc1* and mIoU depend on manually set thresholds, which introduces the risk of implicit cherry-picking [34], [35]. Consequently, existing metrics, whether derived from explanations studies or application-focused domains, are inadequate for evaluating CAM methods specifically designed for downstream tasks. This study presents the MBIE metric, which provides a consistent and unbiased evaluation of CAM performance.

III. Limitations of Existing Evaluation Metrics

This section aims to further demonstrate that existing evaluation methods are unsuitable for evaluating CAM methods specifically designed for downstream tasks through their computational methods. Furthermore, due to the numerous

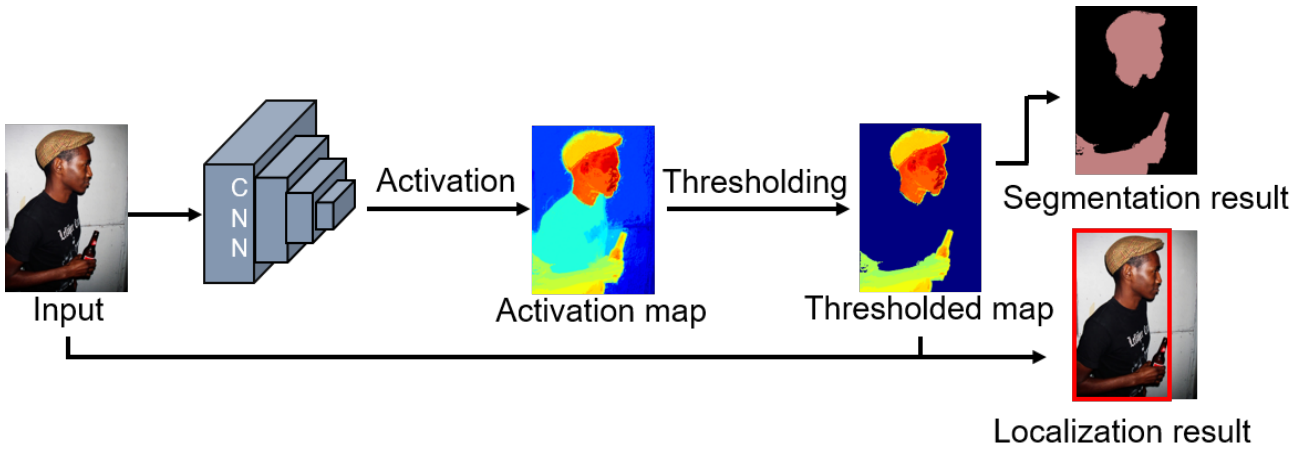


Fig. 2: The process of converting activation maps into localization and segmentation results

limitations and low frequency of use of EBPG and missing insertion curves, they will not be discussed further in this paper.

A. Lack of Class-Specific Evaluation

AD and IC are widely utilized indicators in visual explanations. AD measures the average percentage decrease in the model's confidence in the predicted class when the model is fed an explanation map and the original image. An example of the explanation map is shown in Fig. 1c. A lower AD value indicates that occluding non-essential regions results in only a marginal reduction in model confidence, suggesting that the CAM effectively identifies the most informative regions and that the retained input information remains sufficient to support the prediction. AD can be expressed as:

$$AD = \sum_{i=1}^{N} \frac{\max(0, Y_i^c - O_i^c)}{Y_i^c} \times \frac{100}{N}, \tag{1}$$

where Y_i^c is the classification predicated score for class c on an original image i and O_i^c is the predicated score when the explanation map is fed. N is the total number of images used to calculate AD. max function is applied to eliminate the case where $O_i^c > Y_i^c$. The IC metric measures the number of times when the network's confidence with the explanation map exceeds its confidence with the original image. A higher IC value implies that the activation maps have effectively identified regions that are critical to the model's decision-making. IC can be expressed as:

$$IC = \sum_{i=1}^{N} \frac{\text{Sign}(Y_i^c < O_i^c)}{N} \times 100, \tag{2}$$

where Sign presents an indicator function that returns 1 if input is True.

Equations (1) and (2) clearly demonstrate that AD and IC are class-independent metrics. They quantify changes in model confidence before and after occluding image regions and use these variations to assess the interpretability of activation algorithms. Although the equations include a predicted value Y_i^c for class c, this c corresponds to the network's top-1 predicted class rather than the image's true label [29], [36].

Another similar metric is IO, which assesses the interpretability of activation algorithms by measuring the model's Top-1 and Top-5 accuracy and confidence on occluded images. If the accuracy and confidence on occluded images in the network show a significant drop compared to when predicting on the original image, then the CAM method is considered to have strong interpretability. Unlike AD and IC, IO employs different occlusion strategies during its calculation [3], [14], [28], [36]. An example of the occluded image used in IO is shown in Fig. 1d. Although IO relies on classification prediction results, it computes the model's overall accuracy rather than class-specific accuracy. Therefore, these metrics are category-independent and cannot determine whether a CAM method fully activates a specific category region, or suppresses spurious activations in background areas. Consequently, they are not suitable for evaluating CAM methods designed for downstream tasks.

B. Threshold Sensitivity and Information Loss

mIoU and *Loc1* can, to some extent, be used to verify the CAM algorithm's ability to localize and activate specific targets. However, these metrics require converting continuous activation maps into discrete localization or segmentation outputs for metric calculation. Activation maps produced by CAM methods assign continuous values to pixels, reflecting the likelihood that each pixel belongs to the target class, yet they do not explicitly model the background category. Consequently, a thresholding operation must be manually introduced to distinguish foreground from background, assigning pixels with low activation values to the background class. The choice of this threshold directly influences the resulting localization or segmentation masks. The tables presented in subsequent chapters, along with prior studies [3], corroborate this claim. Fig. 2 illustrates the process of converting an activation map

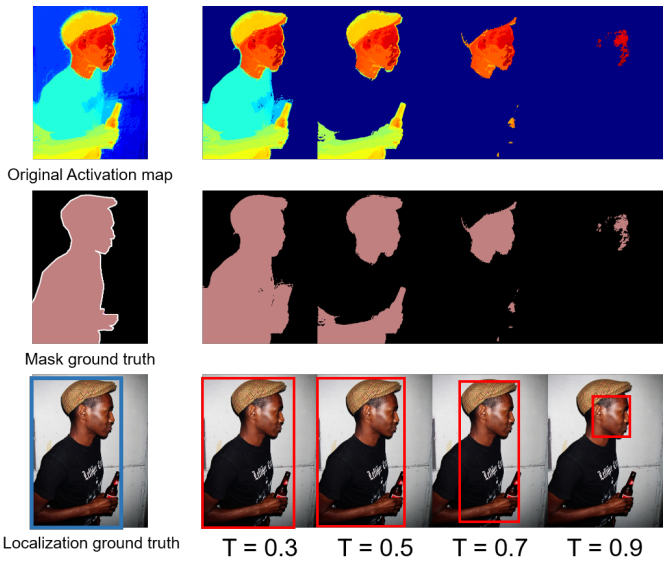


Fig. 3: The activation maps, segmentation and localization results vary with the threshold. From top to bottom: the thresholded activation map, the segmentation and localization results. The blue bounding box is the ground-truth localization labels, and T means threshold. The original activation map is obtained using the Region-CAM algorithm with ResNet-50 on the VOC validation dataset.

from a single image into a localization or segmentation result, while Fig. 3 also shows how the segmentation and localization results for the same activation map changes with different thresholds.

Manually setting thresholds can inadvertently introduce the risk of cherry-picking, as researchers may adjust thresholds in ways that favour certain outcomes. Such subjective decisions can cause reported performance to deviate from the intrinsic quality of the activation method and depend instead on evaluation choices made after the fact, undermining objectivity [36], [37]. More seriously, cherry-picking can exaggerate performance, conceal limitations, or make results appear more favorable than they actually are, distorting our understanding of model behavior and potentially fostering overconfidence in methods that perform poorly in practice [34]. Over time, these effects undermine trust in published results and hinder genuine progress in the field [38], because CAM methods were developed as tools to help researchers understand models and develop algorithms.

Moreover, continuous activation values provide fine-grained numerical information, reflecting the strength, importance, or probability of features and preserving subtle nuances. In contrast, converting these continuous values to discrete classification inevitably loses detail, amplitude, and subtlety, treating different activation intensities within the same category as equal. While discretization makes the output easier to understand and evaluate using standard metrics, the cost is the loss of rich spatial or intensity information. This means that

TABLE I: The AD and IC results obtained under a threshold of 0.5.

Model	Method	AD (%) ↓	IC (%) ↑
VGG-16	Grad-CAM	44.7	10.3
	Grad-CAM++	70.3	3.8
	Layer-CAM	43.5	14.8
	Region-CAM	32.7	17.0
ResNet-50	CAM	20.0	43.1
	Grad-CAM	21.0	41.3
	Grad-CAM++	20.7	41.4
	Layer-CAM	19.1	42.8
	Region-CAM	21.2	37.8

for tasks requiring precision, visualization, or interpretability, continuous values are more informative. To address the inapplicability of existing evaluation methods, as well as the cherry-picking problem and inherent information loss caused by thresholding, we propose MBIE to evaluate CAM more robustly and accurately without relying on arbitrary thresholds.

IV. Proposed Metric: MBIE

A. Class-specific Activation Intensity

The proposed MBIE consists of two complementary components: MBIE_c, measuring class-specific activation intensity, and MBIE_b, measuring background leakage intensity. Both are reported separately, allowing for a detailed evaluation of CAM performance, and are ultimately combined into a single score.

The computation of MBIE_c will be broken down into several steps and explained in detail. Suppose an image i contains category c. The intersection energy of category c in this image can be calculated as follows:

$$\text{MBIE}_i^c = \frac{\sum_{(h,w)\in mask^c} P^c_{(h,w)}}{N_i^c}, \tag{3}$$

where $P^c_{(h,w)}$ represents the activation value belong to class c activation map at the spatial point (h, w), and $mask^c$ means the semantic segmentation masks of target c. The N_i^c represents the number of pixels in $mask^c$. The green area in Fig. 1g is the semantic segmentation mask of the category 'bird'. The activation values produced by CAM methods are typically normalized to the range [0,1]. To establish a reference for comparison, we assign an activation value of 1 to pixels belonging to the target class in the ground truth mask. The activation rate of a pixel is computed by dividing its CAM activation value by the corresponding mask pixel value. This process yields a normalized measure that quantifies the intensity, or degree, of CAM activation for the target; when a pixel is fully activated, this rate equals 1. Therefore, in Eq.3, the denominator, which equals the total number of pixels of the target category, also represents the maximum activation amount of the target category. The activation intensity of category c across the entire dataset can be calculated as follows:

$$\text{MBIE}_c = \frac{1}{N_c}\left(\sum_{n=1}^{N_c} \frac{\sum_{(h,w)\in mask^c} P^c_{(h,w)}}{N_i^c}\right), \tag{4}$$

TABLE II: The Top-1 accuracy results of IO experiment with different activation methods at different thresholds, when $T = 1.0$, the reported value corresponds to the top-1 accuracy of the original VGG16 without occlusion.

Model	Method	Top-1 accuracy (%) under different occlusion thresholds ↓							
		$T = 0.3$	$T = 0.4$	$T = 0.5$	$T = 0.6$	$T = 0.7$	$T = 0.8$	$T = 0.9$	$T = 1.0$
VGG-16	Grad-CAM	5.9	13.9	25.5	37.9	49.5	58.9	65.9	70.0
	Grad-CAM++	8.3	17.6	29.8	42.5	53.5	62.1	67.4	70.0
	Layer-CAM	19.2	35.3	48.2	57.3	63.5	67.3	69.1	70.0
	Region-CAM	7.2	13.7	22.2	31.9	42.3	52.3	61.9	70.0
ResNet-50	CAM	46.9	56.9	64.7	70.1	74.0	77.0	79.0	80.9
	Grad-CAM	47.7	57.5	64.9	70.3	74.1	77.1	79.1	80.9
	Grad-CAM++	47.3	57.4	65.1	70.4	74.1	77.1	79.1	80.9
	Layer-CAM	11.4	23.7	39.4	54.4	65.5	72.4	76.9	80.9
	Region-CAM	9.5	15.2	23.0	33.2	45.3	57.4	69.0	80.9

TABLE III: The *Loc1* of each activation method changes with threshold on the ILSVRC2012 validation set.

Model	Method	*Loc1* varies with the threshold ↑						
		$T = 0.3$	$T = 0.4$	$T = 0.5$	$T = 0.6$	$T = 0.7$	$T = 0.8$	$T = 0.9$
VGG-16	Grad-CAM	41.3	40.2	35.5	26.8	15.9	7.0	1.6
	Grad-CAM++	38.9	39.2	37.6	32.9	24.8	14.6	4.8
	Layer-CAM	40.1	39.0	35.1	28.6	20.0	10.2	3.1
	Region-CAM	51.0	51.3	47.4	40.1	30.0	18.7	7.5
ResNet-50	CAM	54.1	41.6	27.1	14.2	5.8	1.7	0.3
	Grad-CAM	54.4	41.7	27.3	14.2	5.7	1.7	0.3
	Grad-CAM++	54.3	42.2	27.8	14.6	6.1	1.6	0.3
	Layer-CAM	54.3	59.4	60.3	53.7	37.1	15.5	2.3
	Region-CAM	50.8	55.1	58.2	58.7	54.8	44.1	26.6

where N_c represents the number of times class c appears in the entire dataset. This term was introduced to mitigate the long-tail effect, where a small number of classes dominate the dataset while the majority of classes have very few examples, thereby enabling a more balanced evaluation across all categories [39]–[41].

Based on MBIE_c, we further calculate the average activation level across all classes in the dataset as follows:

$$\text{mMBIE}_c = \frac{\sum_{c=1}^{C} \text{MBIE}_c}{C}, \tag{5}$$

where C is the total number of classes. The metric mMBIE_c represents the mean class-wise activation energy, providing a single scalar that reflects the overall alignment between CAM activation maps and ground-truth masks. mMBIE_c also makes evaluation more concise and interpretable.

B. Background Leakage Intensity

MBIE_c measures the activation value within the target mask, providing a quantitative assessment of a CAM algorithm's ability to correctly activate regions corresponding to a single target category. However, CAM algorithms often produce spurious activations in background regions that are unrelated to the target. Such incorrect activations may introduce noisy supervision, causing models to learn spurious associations that can lead to over-segmentation and reduced accuracy in downstream tasks. Prior studies have shown that noisy labels degrade segmentation performance and generalization, as deep networks may gradually fit incorrect labels, thereby distorting learned feature representations and decision boundaries [42]. In addition, label noise can hinder training stability and slow convergence, ultimately impairing the model's ability to generalize to clean test data [43].

To account for this issue, it is essential to compute the background leakage intensity for the same target class c, which quantifies the degree of false activation outside the ground-truth mask. By jointly considering the target activation MBIE_c and the corresponding background leakage MBIE_b, the proposed evaluation framework provides a more accurate and comprehensive measure of CAM performance, capturing both its precision in localizing target regions and its susceptibility to spurious activations. The MBIE_b can be calculated as follows:

$$\text{MBIE}_b = \frac{1}{\sum_{i=1}^{N} |\mathcal{C}_j|} \left(\sum_{i=1}^{N} \sum_{c_j \in \mathcal{C}_j} \frac{\sum_{(h,w) \in mask_b^{c_j}} P_{(h,w)}^{c_j}}{N_i^{c_j}} \right), \tag{6}$$

where $\mathcal{C}_j$ is the set of classes present in image i, $P_{(h,w)}^{c_j}$ denotes the activation value for class c_j in image i at the spatial point (h, w), $mask_b^{c_j}$ represents the background region in classes c_j activation map, $N_i^{c_j}$ means the number of pixels of $mask_b^{c_j}$. $\sum_{i=1}^{N} |\mathcal{C}_j|$ represents the total number of image–class pairs in the dataset, ensuring equal contribution from each class occurrence.

Furthermore, MBIE_b complements MBIE_c by addressing situations where the CAM algorithm arbitrarily activates all pixels to 1. In such cases, MBIE_c will reach 1, indicating complete activation of the target area, while MBIE_b also equals 1, reflecting the unintended activation in the background. Therefore, MBIE metric can calculate as:

$$\text{MBIE} = \max\big(\text{mMBIE}_c - \text{MBIE}_b,\, 0\big) \tag{7}$$

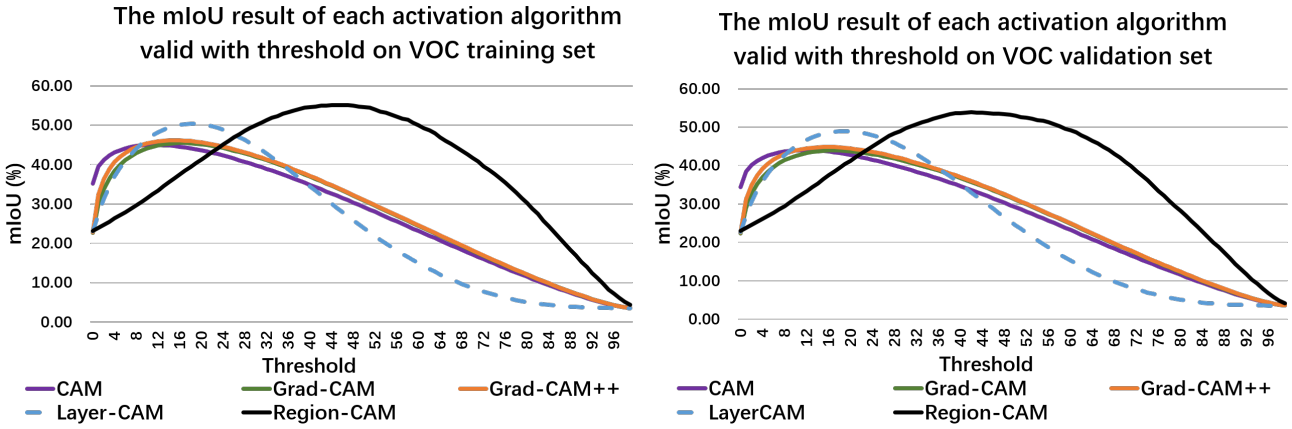


Fig. 4: The mIoU changes with threshold on VOC dataset. For visualization purposes, the threshold values in the graph have been scaled by a factor of 100.

TABLE IV: The evaluation results of each algorithm on MBIE (%) metric, including mMBIE_c (%) and MBIE_b (%).

		VGG-16				Resnet-50				
		Grad-CAM	Grad-CAM++	LayerCAM	Region-CAM	CAM	Grad-CAM	Grad-CAM++	LayerCAM	Region-CAM
VOC train.	$\text{mMBIE}_c \uparrow$	32.3	43.5	35	**69**	47.1	51.3	50.6	44.2	**71**
	$\text{MBIE}_b \downarrow$	**6.4**	15.7	7.4	21.1	**6.1**	9.5	8.4	8.4	19.9
	MBIE $\uparrow$	25.8	27.8	27.6	**47.9**	41	41.8	42.2	33.8	**51.1**
VOC val.	$\text{mMBIE}_c \uparrow$	31.7	42.9	34.3	**68.7**	45.6	50.9	50.1	43.7	**69.2**
	$\text{MBIE}_b \downarrow$	**6.9**	16.6	7.9	21.7	**6.1**	9.9	8.7	8.7	20
	MBIE $\uparrow$	24.8	26.3	26.4	**47**	39.5	41	41.4	35	**49.2**

TABLE V: MBIE_c (%) quantifies the activation intensity of each activation algorithm within each category on VOC validation set. The used model is ResNet-50.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mMBIE_c
CAM	60.2	47.5	53.1	56.7	49.3	38.4	48.4	36.1	37.5	44.3	32.0	39.3	42.1	48.8	46.7	50.5	49.4	37.6	42.5	50.9	45.6
Grad-CAM	64.0	51.8	57.4	61.5	55.6	46.8	54.5	39.8	46.1	50.2	37.6	43.1	48.6	54.6	51.8	54.2	55.9	43.1	47.6	53.2	50.9
Grad-CAM++	63.8	51.9	56.9	60.3	54.6	46.2	53.1	38.9	44.3	49.5	36.9	42.7	48.1	53.9	49.0	52.5	55.7	42.6	47.2	53.2	50.1
LayerCAM	54.5	49.0	46.8	49.4	48.2	41.1	48.4	33.4	41.8	41.1	33.5	34.8	41.1	48.0	41.9	45.7	46.3	38.4	45.3	46.3	43.7
Region-CAM	**82.3**	**70.6**	**72.2**	**74.4**	**64.9**	**70.5**	**66.2**	**65.7**	**63.4**	**67.8**	**60.9**	**64.6**	**69.3**	**75.4**	**65.8**	**66.8**	**69.5**	**68.0**	**73.4**	**71.1**	**69.2**

where max is maximum operation, which ensures that MBIE will not be less than 0.

V. Results and Discussion

In this section, multiple validation metrics will be calculated using two networks and two datasets to further support the claim that the current metrics for CAM methods are unsuitable for evaluating CAM methods specifically designed for downstream tasks. Several activation methods are considered in this study, including CAM [26], Grad-CAM [17], Grad-CAM++ [29], LayerCAM [14], and Region-CAM [3]. Although CAM, Grad-CAM and Grad-CAM++ were not designed for weakly supervised tasks, they are widely used in this context. LayerCAM [14] and Region-CAM [3], on the other hand, are specifically designed for weakly supervised tasks. LayerCAM is more widely used than Region-CAM, but Region-CAM is the latest algorithm.

A. Experimental Setup

During the experiment, we utilized two backbone networks: VGG-16 and Resnet-50, as well as the ILSVRC 2012 validation set [44] and Pascal VOC 2012 [45] dataset. We calculate AD, IC, IO, and *Loc1* on the ILSVRC validation set. When computing AD and IC, we selected only images correctly classified by the network from the ILSVRC 2012 validation set to improve reproducibility and reduce experimental randomness. Previous studies typically selected a fixed number of images randomly [3], [14], [18], [26], [28], [29], [36]. **Notably**, the original CAM method cannot be implemented by pre-trained VGG-16, because VGG-16 lacks a global averaging layer. The input images are resized to 244 × 244 and normalized to the [0, 1] range using a mean vector of [0.485, 0.456, 0.406] and a standard deviation vector of [0.229, 0.224, 0.225]. The weights used are from the PyTorch model zoo [46].

mIoU and MBIE are computed using the Pascal VOC 2012 dataset, which provides fine-grained target masks. Pascal

VOC was historically used to evaluate activation algorithms; however, as a multi-label dataset, it is not well suited for calculating AD and IC and is therefore no longer used [1], [29], [47]. The input images keep the original size and normalized to the [0, 1] range using a mean vector of [0.485, 0.456, 0.406] and a standard deviation vector of [0.229, 0.224, 0.225]. The weights are trained by ourselves.

B. Metric Behavior Analysis

Tab. I presents AD and IC scores for several CAM methods evaluated on VGG-16 and ResNet-50 under a threshold of 0.5. Region-CAM, a recently proposed method specifically designed for weakly supervised tasks [3], achieves the lowest AD and highest IC on VGG-16, indicating strong performance according to these metrics. However, on ResNet-50, all methods obtain similar AD and IC values, and Region-CAM does not consistently outperform other approaches. These observations suggest that AD and IC are sensitive to network architecture and may fail to fully capture the advantages of CAM methods designed for weakly supervised semantic segmentation and object localization. Task-specific methods such as Layer-CAM and Region-CAM produce spatially broader or smoother activation maps, which improve downstream localization but are not necessarily rewarded by classification-oriented metrics like AD and IC.

Compared with AD and IC, IO mIoU and *Loc1* is threshold sensitivity and carry a risk of introducing cherry-picking, which is supported by Tab. II, Tab. III and Fig. 4. Since the thresholds used in these metrics are not fixed and there are no standards to follow. Previous studies have adopted different threshold values [3], [14], [20]. In Tab. II, on VGG-16, when a low threshold is applied (e.g., T=0.3), Grad-CAM outperforming Region-CAM. However, this advantage diminishes as the threshold increases. Depending on the chosen threshold, two contradictory conclusions are reached: either Grad-CAM or Region-CAM is optimal. Tab. III and Fig. 4 also demonstrate the limitations of threshold-sensitive metrics. Region-CAM generally outperforms Layer-CAM; however, Layer-CAM can surpass it under certain threshold settings on ResNet-50. This indicates that selective threshold choices can artificially inflate the reported performance of a given algorithm or obscure performance differences between methods, potentially resulting in unreliable cross-paper comparisons and unfair algorithm evaluation.

In contrast, our proposed metric effectively avoids the aforementioned limitations and can well reflect the activation algorithm's ability to activate the correct target region. As shown in Tab. IV, methods that have been reported in prior studies to produce activation maps with broader spatial coverage and more complete object localization,such as Region-CAM, achieve higher mMBIE and overall MBIE scores. Conversely, methods that generate sparse or highly localized activations (CAM) receive substantially lower scores [3]. Moreover, the validity of the mMBIE and MBIE scores is further corroborated by the results shown in Fig. 4. Specifically, Region-CAM attains its peak mIoU at a threshold of approximately 48%, indicating that its segmentation results at this threshold most closely align with the ground-truth annotations. This observation suggests that Region-CAM correctly activates a large proportion of target regions with activation values exceeding 48%. Consistently, the MBIE metric accurately captures this behavior: when evaluated on ResNet-50, Region-CAM achieves an MBIE score of 51.1%. Therefore, the MBIE metric effectively evaluates and reflects the activation capabilities of CAM methods designed for weakly supervised tasks. It is threshold-insensitive, avoids cherry-picking risks, and is laser-specific, enabling verification of the activation algorithm's performance for each individual class. Activation results for each class are shown in Tab. V.

VI. Conclusions

This work proposes a novel metric, MBIE, which is designed to be threshold-insensitive and class-specific for accurately evaluating class activation map (CAM) methods in weakly supervised learning settings. MBIE consists of two complementary components, MBIE_c and MBIE_b, which jointly and effectively assess the activation strength of the target class and the extent of unintended activation in background regions. In particular, MBIE_c emphasizes the model's ability to correctly highlight semantically relevant regions, while MBIE_b penalizes excessive or noisy activations outside the target object, thereby encouraging precise localization. Our analysis demonstrates that MBIE correlates well with qualitative visual assessments and remains robust across different thresholds and datasets. As a result, MBIE offers a reliable and practical evaluation tool for benchmarking CAM methods and has the potential to facilitate more faithful interpretation and comparison of weakly supervised localization approaches.

References

[1] S.-A. Rebuffi, R. Fong, X. Ji, and A. Vedaldi, "There and back again: Revisiting backpropagation saliency methods," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8839–8848.

[2] Q. Li, A. Arnab, and P. H. Torr, "Weakly-and semi-supervised panoptic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 102–118.

[3] Q. Cai and C. Abhayaratne, "Region-cam: Toward accurate object regions in class activation maps for weakly supervised learning tasks," *IEEE Access*, vol. 13, pp. 208 361–208 375, 2025.

[4] J. Li, Z. Jie, X. Wang, Y. Zhou, L. Ma, and J. Jiang, "Weakly supervised semantic segmentation via self-supervised destruction learning," *Neurocomputing*, vol. 561, p. 126821, 2023.

[5] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3791–3800.

[6] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2209–2218.

[7] X. Wang, S. Liu, H. Ma, and M.-H. Yang, "Weakly-supervised semantic segmentation by iterative affinity learning," *International Journal of Computer Vision*, vol. 128, pp. 1736–1749, 2020.

[8] K. Kumar Singh and Y. Jae Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3524–3533.

[9] J. Xu, J. Hou, Y. Zhang, R. Feng, R.-W. Zhao, T. Zhang, X. Lu, and S. Gao, "Cream: Weakly supervised object localization via class re-activation mapping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9437–9446.

[10] V. C. Burkapalli and P. C. Patil, "Food image segmentation using edge adaptive based deep-cnns," *International Journal of Intelligent Unmanned Systems*, vol. 8, no. 4, pp. 243–252, 2020.

[11] L. Zhu, X. Wang, J. Feng, T. Cheng, Y. Li, B. Jiang, D. Zhang, and J. Han, "WeakCLIP: Adapting clip for weakly-supervised semantic segmentation," *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1085–1105, 2025.

[12] W. Wu, X. Qiu, S. Song, Z. Chen, X. Huang, F. Ma, and J. Xiao, "Prompt categories cluster for weakly supervised semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3198–3207.

[13] Q. Cai and C. Abhayaratne, "SSDB-Net: A single-step dual branch network for weakly supervised semantic segmentation of food images," in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2023, pp. 1–6.

[14] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

[15] X. Shi, S. Khademi, Y. Li, and J. van Gemert, "Zoom-CAM: Generating fine-grained pixel annotations from image labels," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 10 289–10 296.

[16] G. Ciocca, P. Napoletano, and R. Schettini, "Learning cnn-based features for retrieval of food images," in *New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers 19*. Springer, 2017, pp. 426–434.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[18] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[19] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models. arxiv 2018," *arXiv preprint arXiv:1806.07421*, vol. 6, 1806.

[20] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[21] R. R. Asaad and R. I. Ali, "Back propagation neural network (bpnn) and sigmoid activation function in multi-layer networks," *Academic Journal of Nawroz University*, vol. 8, no. 4, pp. 216–221, 2019.

[22] Y. Chen and G. Zhong, "Score-CAM++: Class discriminative localization with feature map selection," in *Journal of Physics: Conference Series*, vol. 2278, no. 1. IOP Publishing, 2022, p. 012018.

[23] X. Xu, P. Zhang, W. Huang, Y. Shen, H. Chen, J. Lin, W. Li, G. He, J. Xie, and S. Lin, "Weakly supervised semantic segmentation via progressive confidence region expansion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9829–9838.

[24] S. Yasuki and M. Taki, "CAM back again: Large kernel cnns from a weakly supervised object localization perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 341–351.

[25] D. Balduzzi, K. Tuyls, J. Perolat, and T. Graepel, "Re-evaluating evaluation," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[27] S. Syed, K. E. Anderssen, S. K. Stormo, and M. Kranz, "Weakly supervised semantic segmentation for mri: exploring the advantages and disadvantages of class activation maps for biological image segmentation with soft boundaries," *Scientific reports*, vol. 13, no. 1, p. 2574, 2023.

[28] H. G. Ramaswamy *et al.*, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 983–991.

[29] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[30] J. Lin, G. Han, X. Xu, C. Liang, T.-T. Wong, C. Chen, Z. Liu, and C. Han, "Broadcam: outcome-agnostic class activation mapping for small-scale weakly supervised applications," *arXiv preprint arXiv:2309.03509*, 2023.

[31] F. Meng, K. Luo, H. Li, Q. Wu, and X. Xu, "Weakly supervised semantic segmentation by a class-level multiple group cosegmentation and foreground fusion strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4823–4836, 2019.

[32] F. Wu, J. He, Y. Yin, Y. Hao, G. Huang, and L. Cheng, "Masked collaborative contrast for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 862–871.

[33] Z. Chen and Q. Sun, "Extracting class activation maps from non-discriminative features as well," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3135–3144.

[34] L. Roque, V. Cerqueira, C. Soares, and L. Torgo, "Cherry-picking in time series forecasting: How to select datasets to make your model shine," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 19, 2025, pp. 20 192–20 199.

[35] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3133–3142.

[36] I. Gkartzonika, N. Gkalelis, and V. Mezaris, "Learning visual explanations for dcnn-based image classifiers using an attention mechanism," in *European Conference on Computer Vision*. Springer, 2022, pp. 396–411.

[37] J. Hinns, S. Goethals, S. Van der Veeken, T. Evgeniou, and D. Martens, "On the definition and detection of cherry-picking in counterfactual explanations," *arXiv preprint arXiv:2601.04977*, 2026.

[38] G. Leech, J. J. Vazquez, N. Kupper, M. Yagudin, and L. Aitchison, "Questionable practices in machine learning," *arXiv preprint*, vol. abs/2407.12220, 2024.

[39] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 9, pp. 10 795–10 816, 2023.

[40] C. De Alvis and S. Seneviratne, "A survey of deep long-tail classification advancements," *arXiv preprint arXiv:2404.15593*, 2024.

[41] J. Wang, W. Zhang, Y. Zang, Y. Cao, J. Pang, T. Gong, K. Chen, Z. Liu, C. C. Loy, and D. Lin, "Seesaw loss for long-tailed instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9695–9704.

[42] Z. Huang, H. Zhang, A. Laine, E. Angelini, C. Hendon, and Y. Gan, "Co-seg: An image segmentation framework against label corruption," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 550–553.

[43] R. Yi, Y. Huang, Q. Guan, M. Pu, and R. Zhang, "Learning from pixel-level label noise: A new perspective for semi-supervised semantic segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 623–635, 2021.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[45] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8024–8035.

[47] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.