Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU34306

Mitigating the Black-Box Problem with CAM in Classification, Detection, and Segmentation

Hongkai Zeng

Manchester Metropolitan Joint Institute, Hubei University, Wuhan, China
21907766@stu.mmu.ac.uk

Abstract. Deep visual models have achieved strong results in image classification, object detection, and image segmentation. Yet the basis of their internal decisions is still hard to observe directly. Background interference, reliance on local textures, and limited object coverage remain common problems. Class Activation Mapping (CAM) and related methods use heat maps to show class-related spatial responses. These methods can verify discriminative evidence in classification, generate location cues in detection, and build pseudo labels in segmentation. This paper examines relevant methods and representative results in these three tasks. Existing studies have improved faithfulness, object coverage, box-level localisation, and segmentation accuracy. Yet fragmented responses, limited object coverage, and blurred boundaries still remain. CAM has moved from visual explanation to localisation and segmentation pipelines. It therefore has practical value for mitigating the black-box problem in visual models. This paper aims to explore the development and application of interpretability methods in visual tasks, revealing their potential to improve model transparency and practicality by reviewing relevant research. Studies show that, despite certain limitations, these methods are of significant value in driving visual models towards greater reliability and understandability.

Keywords: Class Activation Mapping (CAM), model interpretability, black-box problem, weakly supervised object localization, semantic segmentation

1. Introduction

Deep neural networks are widely used in image classification, object detection, and semantic segmentation. Their performance keeps rising. Yet the reason behind a specific prediction is still not fully transparent. Stable links are often missing among input regions, feature representations, and final outputs. Category scores or overall accuracy alone cannot show whether the model truly uses object-level information. They also cannot tell whether an error comes from background distraction, texture bias, or weak spatial localisation. The black-box problem is no longer a minor issue in interpretation. It also affects model evaluation, result analysis, and method design.

Class Activation Mapping (CAM) and its gradient-based variants offer an important way to address this issue. These methods project class-related responses back to the image space. They then produce heat maps that match the prediction. In this way, internal discriminative information becomes spatial evidence. Compared with a single score output, a heat map shows the attended

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU34306

region more clearly. It can also reveal background reliance, response shift, and excessive focus on local parts. CAM-based methods are no longer limited to visual display. They are also used in error diagnosis, model comparison, and follow-up pipeline design.

Different tasks need different forms of spatial evidence. Image classification focuses on whether the discriminative region matches the target semantics. Object detection requires high-response regions that can become stable location cues. Semantic segmentation pays more attention to region completeness and boundary consistency. As the task moves from classification to detection and segmentation, the research focus shifts from local discrimination to spatial structure. CAM offers a feasible path for mitigating the black-box problem in visual models. It also exposes shared limits, including weak coverage, broken responses, and blurred boundaries. These issues form the main thread of the following discussion.

2. Image classification

Image classification requires the model to assign a category label to an input image. CAM-based methods usually generate a class-related heat map after forward inference. The heat map shows the spatial region used by the current prediction [1-3]. The basic pipeline includes category score computation, extraction of target-class weights, response map generation, and upsampling to the image space [1, 2, 4]. Research in this task mainly focuses on three points. The first is whether the discriminative region lies on the object. The second is whether the heat map stays consistent with the classification score [4-7]. The third is whether similar categories can be separated effectively [8-11].

Recent studies first examine the heat-map generation mechanism itself. Relevance-CAM extends explanation to layers at arbitrary network depth. It shows that intermediate layers can also provide stable class-related responses [1]. LayerCAM keeps shallow details through position-sensitive gradients and multi-layer fusion. In image occlusion experiments, masking its high-confidence regions reduced Top-1 accuracy from 68.74% to 48.26%. This result was lower than 50.36% for Grad-CAM and 50.07% for Grad-CAM++. Average confidence also dropped to 48.12%. This value was lower than 50.24% for Grad-CAM and 49.99 for Grad-CAM++ [2]. LIFT-CAM proposed by Jung and Oh further introduces a Shapley approximation into CAM weight learning. On ImageNet, PASCAL VOC (VOC), and MS COCO (COCO), SHAP-CAM100 reports Average Drop values of 27.48, 16.71, and 18.27. These results are better than 28.16, 22.93, and 23.98 from SHAP-CAM1. The cosine similarity between LIFT-CAM and SHAP-CAM10k weights reaches 0.980, 0.918, and 0.924 on the three datasets [4]. These methods mainly improve the faithfulness and readability of classification heat maps.

A second line of work targets limited object coverage. TS-CAM combines token semantics and attention maps in the Vision Transformer (ViT). On CUB-200-2011, it achieves 71.3% Top-1 Localization Accuracy (Top-1 Loc), 83.8% Top-5 Localization Accuracy (Top-5 Loc), and 87.7% Ground-Truth-Known Localization Accuracy (GT-Known Loc). In the same table, VGG16-CAM reaches only 44.2% Top-1 Loc. In an ILSVRC ablation study, TS-CAM also reaches 53.4% Top-1 Loc, 64.3% Top-5 Loc, and 67.6% GT-Known Loc [3]. CREAM strengthens full-object responses through foreground-background context embedding and soft clustering. In cross-dataset averages, CREAMBase improves CAM by 5.8% and CAM-IVR by 4.0% [12]. Chen and Sun proposed LPCAM. It replaces the single dependence on classifier weights in traditional CAM with local prototypes. In VOC ablation experiments, the mean Intersection over Union (mIoU) rises from 48.8 for CAM to 54.9. Recall rises from 65.2 to 77.2. Seed mask results on MS COCO also improve by

DOI: 10.54254/2755-2721/2026.GU34306

2.7% and 2.2% [9]. These results show that a heat map focused only on the strongest local response cannot support stable analysis. Wider object coverage is still needed.

A third group of studies focuses on fine-grained classification and explanation quality. CAPE reformulates CAM as a probabilistic ensemble problem. On CUB, mu-CAPE(TS) reports Average Drop in Confidence (AD), Average Increase in Confidence (IC), AD in Deletion (ADD), and AD, Coherency, and Complexity (ADCC) scores of 10.3, 48.5, 74.2, and 84.4. All of these values are better than 21.2, 27.9, 67.4, and 78.8 from the original CAM. CAPE(PF) reduces the overlap mIoU between the first two heat-map classes to 13.4, while the original CAM records 75.9 [5]. ID-ExpO directly includes insertion-deletion metrics in training. The authors report that 57.8% of samples improve both insertion and one-minus-deletion. This proportion is higher than 41.9% for ExpO-S, 38.7% for ExpO-F, and 34.6% for cross-entropy training alone [6]. Finer-CAM targets regions that separate similar categories. On CUB, it raises Localization Accuracy (Loc.) for Grad-CAM from 0.582 to 0.629. It also raises Loc. for Layer-CAM from 0.625 to 0.682. On Birds525, it improves Grad-CAM RD@0.05 from 0.174 to 0.192 and RD@0.1 from 0.245 to 0.260 [7]. Prompt-CAM continues this fine-grained direction. On the CUB validation set, it reaches 0.61 for Insertion and 0.09 for Deletion. These results are better than 0.52 and 0.17 for Grad-CAM. They are also better than 0.54 and 0.13 for Layer-CAM. In manual recognition experiments, the proportion of valid category features identified by PROMPT-CAM is 60.49%. This value is higher than 39.14% for TesNet and 30.39% for ProtoConcepts [8]. CAM research in classification has therefore moved from asking where the model looks to asking how much it covers, how finely it separates, and whether the response agrees with the prediction.

CAM has clear strengths in image classification. It is easy to integrate. Most methods can be used without rebuilding the classifier [1, 2, 8]. Evaluation metrics are also relatively mature. Average Drop in Confidence (AD), Average Increase in Confidence (IC), Insertion, and Deletion can directly measure heat-map quality [2, 4, 5]. AD in Deletion (ADD), AD, Coherency, and Complexity (ADCC), and Localization Accuracy (Loc.) are also used to assess explanation quality [7, 8]. The limits are also clear. Classification supervision constrains only category prediction. It does not directly enforce full spatial coverage. For this reason, heat maps often focus on a few highly discriminative regions [3, 12, 9]. In fine-grained scenes, shared regions and truly distinctive local features are often intertwined. The resolution of explanation maps is still limited [5, 7]. Some methods may also reduce classification accuracy when local interpretability is strengthened. For example, Prompt-CAM reports 73.2% CUB accuracy with a DINO backbone. This value is lower than 78.6% from linear probing [8]. These results show that CAM can already support black-box diagnosis and result checking in classification. In later tasks, the main issue shifts toward location cues and structural expression.

3. Object detection

Object detection needs both category and location outputs. In this task, CAM first provides a class-related heat map. High-response regions are then converted into proposal boxes or proposal regions. These outputs are further refined by multiple instance learning, proposal screening, box regression, retraining, or semi-supervised detection modules. Final evaluation usually uses mean Average Precision (mAP), Correct Localization (CorLoc), Top-1 Loc, Top-5 Loc, and GT-Known Loc. Detection pays more attention to object coverage, response continuity, background suppression, and stable box conversion. When a heat map focuses too much on local parts, IoU and box-level localization accuracy usually drop directly.

DOI: 10.54254/2755-2721/2026.GU34306

In weakly supervised detection, CAM is often used as an upstream signal for proposal generation and proposal screening. Dong et al. proposed LBBA-boosted WSOD. It introduces a class-agnostic box adjuster into the weakly supervised detection pipeline. On PASCAL VOC 2007, the method achieves 56.5 mAP and 72.3 CorLoc. Under multi-scale testing, the results are 56.6 mAP and 72.5 CorLoc. In the OICR-based setting, mAP rises from 51.4 to 55.8. CorLoc rises from 64.9 to 71.6 [10]. Sui et al. proposed SoS-WSOD. It links weak labels, pseudo boxes, and semi-supervised detection into a three-stage pipeline. With a VGG16 backbone, mAP50 rises to 60.3 on VOC 2007 and 57.7 on VOC 2012. With ResNet50, the results further increase to 64.4 and 61.9. On MS COCO, the method reaches 16.6 mAP50:95, 32.8 mAP50, and 15.2 mAP75 [11]. These results show a direct link between detection performance, proposal quality, pseudo-box quality, and later box refinement.

Studies that use heat maps directly for box-level localisation also provide important evidence. TS-CAM reaches 71.3 Top-1 Loc, 83.8 Top-5 Loc, and 87.8 GT-Known Loc on CUB-200-2011. These values are much higher than 44.2, 52.2, and 58.0 from CAM. On ILSVRC, the corresponding results are 53.4, 64.3, and 67.6 [3]. CREAM introduces foreground and background context embedding into the reactivation process. In threshold-free evaluation, its average gains over CAM are 8.6 on CUB, 5.1 on ILSVRC, and 3.8 on OpenImages. The overall average gain is 5.8. CREAMReg reaches 70.44 Top-1 Loc, 85.67 Top-5 Loc, and 90.98 GT-Known Loc on CUB [12]. SPOL uses multiplicative fusion of shallow and deep features together with pseudo-supervised segmentation. It achieves 80.12 Top-1 Loc, 93.44 Top-5 Loc, and 96.46 GT-Known Loc on CUB. On ImageNet-1K, it reaches 59.14, 67.15, and 69.02 [12]. SLT-Net adopts a separate localisation-classification structure. Under the VGG16 setting, it reports 51.2 Top-1 Loc, 62.4 Top-5 Loc, and 67.2 GT-Known Loc. Under ResNet50, the corresponding Top-1 Loc is 56.2 and GT-Known Loc is 68.5 [12]. GenPromp introduces generative prompts into localisation. On CUB, it reaches 87.0 Top-1 Loc, 96.1 Top-5 Loc, and 98.0 GT-Known Loc. On ImageNet-1K, it reports 65.2, 73.4, and 75.0 [12]. These results show that more complete heat-map structure usually leads to better box-level localisation.

CAM has clear value in object detection. Image-level labels can already provide initial location priors. Heat maps can enter proposal generation, proposal screening, and box refinement with relatively low integration cost. The limits are also concentrated. When high-response regions lean toward local discriminative parts, broken responses, background leakage, and threshold sensitivity often cause box drift, undersized boxes, and missed objects. Detection pipelines also depend on proposal quality, regression modules, and retraining strategies. Once the heat-map quality is poor, the room for later improvement also becomes smaller.

4. Image segmentation

Image segmentation outputs pixel-level semantic labels. In this task, CAM usually starts with an initial response map. A pseudo mask is then constructed. A segmentation network is trained afterward. Final evaluation is usually based on mIoU. This pipeline requires stronger foreground coverage, lower cross-class leakage, better boundary closure, and better background suppression. A small initial response weakens the object region. Cross-class activation contaminates pseudo labels. Boundary diffusion then lowers the final segmentation accuracy.

The quality of the initial CAM directly determines the upper bound of pseudo-label quality. On PASCAL VOC 2012, AdvCAM increases seed mIoU from 48.8 to 55.6 and pseudo-mask mIoU from 66.3 to 69.9 on the Inter-pixel Relation Network (IRN) route. Final val/test mIoU reaches 68.1/68.0 [13]. MCTformer reaches 71.9 val mIoU and 71.6 test mIoU on VOC. It also achieves

DOI: 10.54254/2755-2721/2026.GU34306

42.0 mIoU on MS COCO [14]. On VOC, ReCAM raises CAM with ResNet-50 from 48.8 to 54.8. It also raises the IRN route from 66.3 to 70.9. Under the DeepLabV2 setting, ReCAM reaches 68.5 val mIoU and 68.4 test mIoU. ReCAM-M further reaches 71.8 and 72.2 [14]. These results show that better initial response coverage usually leads to a higher upper bound for later pseudo labels.

Category confusion and missing context are also central problems. CTI achieves 74.1 val mIoU and 73.2 test mIoU on VOC. On COCO, it reaches 45.4 val mIoU [15]. CPAL raises AMN to 72.5 val mIoU and 72.9 test mIoU on VOC. It also raises CLIP-ES to 74.5 and 74.7. On COCO, it brings a further gain of 1.4 points [15]. From SAM to CAMs introduces Segment Anything Model (SAM) priors into pseudo-label construction. The full setting reaches 76.1 mIoU. The mask-level result reaches 81.7 [15]. CAM research in segmentation has therefore moved from local salient regions to pseudo-label correction, region completion, and boundary alignment.

CAM in image segmentation has a low annotation cost. Image-level labels can already generate initial seed regions. Its main limits are local focus, threshold sensitivity, cross-class interference, and blurred boundaries. Compared with classification and detection, segmentation places higher demands on CAM. The heat map must express complete regions. Local discriminative areas alone cannot support stable pixel-level prediction.

5. Existing limitations and future prospects

CAM has formed basic application paths in image classification, object detection, and image segmentation. Yet current results still show several shared limits. First, response regions often lean toward highly discriminative local parts. Object coverage is therefore insufficient. In classification, this leads to incomplete evidence. In detection, it causes small proposal boxes or box shift. In segmentation, it appears as missing foreground and broken boundaries. Second, heat-map quality is sensitive to thresholds, feature-layer choice, and post-processing strategy. Results can vary noticeably under different settings. Third, stability is still limited in multi-object scenes, occluded scenes, and small-object scenes. Cross-class interference and background leakage have not been fully resolved. Fourth, some methods rely on extra modules for region completion, pseudo-label correction, or boundary refinement. The overall pipeline is therefore long. Reproduction cost is also high. Fifth, the current evaluation system is still not unified. The classification part often uses Average Drop, Insertion, and Deletion. The detection part pays more attention to localisation accuracy and mAP. The segmentation part mainly uses mIoU. A common benchmark across tasks is still lacking.

Future research can continue in four directions. The first is to improve response completeness and reduce area loss caused by local focus. Possible approaches include cross-layer feature fusion, region consistency constraints, and inter-class repulsion constraints. Their effects can be observed directly through Top-1 Loc, GT-Known Loc, seed mIoU, and mask mIoU. The second is to shorten the pipeline and reduce reliance on extra post-processing modules. Heat-map generation, region expansion, and pseudo-label correction should be integrated into a more unified framework when possible. After the pipeline is shortened, training epochs, inference time, and memory use can all serve as direct indicators. The third is to improve stability in complex scenes. Occlusion, low contrast, small objects, and dense objects should be reported separately. Relevant indicators can include response drift rate, box shift rate, boundary error, or small-object recall. The fourth is to build a cross-task evaluation framework. Under the same backbone and the same training budget, studies should report classification faithfulness, detection localisation accuracy, and segmentation pseudo-label quality together. This design can provide a more stable comparison baseline. It is also closer to real application settings.

DOI: 10.54254/2755-2721/2026.GU34306

From the application perspective, the next direction of CAM is already clear. Classification needs stronger category discrimination. Detection needs more stable location cues. Segmentation needs more complete region expression. If later work can improve the stability of all three aspects at the same time, reduce module stacking, and unify evaluation, the role of CAM in mitigating the black-box problem of visual models will become clearer. Its engineering value will also continue to grow.

6. Conclusion

CAM provides observable spatial evidence for the black-box problem in visual models. It has formed different application paths in image classification, object detection, and image segmentation. In classification, it is mainly used to verify discriminative evidence. In detection, it mainly generates location cues. In segmentation, it supports pseudo-label expansion and region completion. Existing studies show that CAM can improve the readability of model analysis. It can also support localisation and segmentation. Its main limits still lie in local focus, limited object coverage, incomplete boundaries, and unstable results. Future work still needs to improve region completeness, robustness in complex scenes, pipeline simplicity, and evaluation consistency.

References

[1] J. R. Lee, S. Kim, I. Park, T. Eo, D. Hwang, Relevance-CAM: Your model already knows where to look, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2021)
[2] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, LayerCAM: Exploring hierarchical class activation maps for localization. IEEE Trans. Image Process. 30, 5875–5888 (2021)
[3] 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. CoRR abs/2103.14862 (2021)
[4] H. Jung, Y. Oh, Towards better explanations of class activation mapping. CoRR abs/2102.05228 (2021)
[5] T. F. Chowdhury, K. Liao, V. M. H. Phan, M.-S. To, Y. Xie, K. Hung, D. Ross, A. van den Hengel, J. W. Verjans, Z. Liao, CAPE: CAM as a probabilistic ensemble for enhanced DNN interpretation. CoRR abs/2404.02388 (2024)
[6] Y. Yoshikawa, T. Iwata, Explanation-based training with differentiable insertion/deletion metric-aware regularizers. CoRR abs/2310.12553 (2023)
[7] Z. Zhang, J. Gu, Z. Zhang, K. S. Mehrab, C. V. Stewart, T. Berger-Wolf, Y. Su, W.-L. Chao, Finer-CAM: Spotting the difference reveals finer details for visual explanation. CoRR abs/2501.11309 (2025)
[8] A. Chowdhury, D. Paul, Z. Mai, J. Gu, Z. Zhang, K. S. Mehrab, E. Campolongo, D. Rubenstein, C. Stewart, A. Karpatne, T. Berger-Wolf, Y. Su, W.-L. Chao, Prompt-CAM: Making vision transformers interpretable for fine-grained analysis. CoRR abs/2501.09333 (2025)
[9] Z. Chen, Q. Sun, Extracting class activation maps from non-discriminative features as well. CoRR abs/2303.10334 (2023)
[10] B. Dong, Z. Huang, Y. Guo, Q. Wang, Z. Niu, W. Zuo, Boosting weakly supervised object detection via learning bounding box adjusters. CoRR abs/2108.01499 (2021)
[11] L. Sui, C.-L. Zhang, J. Wu, Salvage of supervision in weakly supervised object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 45, 10394–10408 (2023)
[12] J. Xu, J. Hou, Y. Zhang, R. Feng, R.-W. Zhao, T. Zhang, X. Lu, S. Gao, CREAM: Weakly supervised object localization via class re-activation mapping. CoRR abs/2205.13922 (2022)
[13] J. Lee, E. Kim, S. Yoon, Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. CoRR abs/2103.08896 (2021)
[14] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, D. Xu, Multi-class token transformer for weakly supervised semantic segmentation. CoRR abs/2203.02891 (2022)
[15] F. Tang, Z. Xu, Z. Qu, W. Feng, X. Jiang, Z. Ge, Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. CoRR abs/2403.07630 (2024)