Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25896

A Comparative Study of CAM Variants with GrabCut Refinement for Large-Scale Pseudo-Mask Generation

Binghan Chen

Department of Computer Science, University College London, London, UK
binghan.chen.24@ucl.ac.uk

Abstract. Semantic segmentation typically demands laborious pixel-level annotations, motivating cost-effective weakly supervised approaches that generate pseudo-masks from classifier signals. Weakly supervised pseudo-mask generation offers a cost-effective alternative to costly pixel-level annotation by leveraging classifier activation maps and simple refinement. In this work, we present the first large-scale, systematic comparison of six Class Activation Map (CAM) variants—CAM, Grad-CAM, Grad-CAM++, Eigen-CAM, Layer-CAM, and Ablation-CAM—within a unified GrabCut refinement pipeline. For each CAM method, we binarize heatmaps using both a fixed 0.2 threshold and Otsu's method, then refine seeds with GrabCut to produce final masks. Evaluated on 23,151 ImageNet-LID validation images, our results show that (1) vanilla CAM yields the most precise seeds and highest final mask quality (IoU = 0.5124, Dice = 0.6219), (2) GrabCut consistently improves mask accuracy across all variants ($\Delta IoU \approx 0.10 - 0.17$, $\Delta Dice \approx 0.08 - 0.14$), and (3) although newer CAM methods exhibit superior localization in raw heatmaps, those advantages do not translate into better thresholded pseudo-masks after GrabCut refinement. These findings guide practitioners toward simple yet effective weakly supervised segmentation pipelines and highlight directions—such as multi-threshold schemes or alternative refinement methods—for capturing the strengths of advanced CAM variants.

Keywords: Weakly-Supervised Learning, Semantic Segmentation, Class Activation Maps, GrabCut

1. Introduction

Semantic segmentation—assigning a label to every pixel in an image—underpins many computer vision applications, from autonomous driving to medical imaging. However, training state-of-the-art segmentation networks requires costly pixel-level annotations at scale. To mitigate this, weakly supervised methods generate pseudo-masks from cheaper image-level or bounding-box labels; among these, CAM [1] has emerged as a leading technique for localizing objects without detailed masks. CAM and its variants (Grad-CAM [2], Grad-CAM++ [3], Eigen-CAM [4], Layer-CAM [5], Ablation-CAM [6]) produce coarse heatmaps highlighting discriminative object regions. Yet these heatmaps often under-segment full object extents.

DOI: 10.54254/2755-2721/2025.BJ25896

One such remedy is to refine CAM seeds via graph-cut methods such as GrabCut [7] , which expands confident foreground regions based on color and edge cues. Prior work has shown CAM-seeded GrabCut improves mask completeness in domain-specific settings—roadside objects [8] , aerial imagery [9]—but these studies evaluate only a single CAM variant. Separately, comparisons of CAM methods for visualization or localization exist, but none assess their impact on downstream pseudo-mask quality when coupled with a unified GrabCut pipeline. Moreover, thresholding strategies for binarizing CAM maps (fixed threshold at 0.2 of the normalized map [1] or Otsu's method [10]) have not been systematically evaluated across CAM variants.

In this paper, we fill this gap by conducting the first large-scale comparison of seven CAM variants—CAM [1], Grad-CAM [2], Grad-CAM++ [3], Eigen-CAM [4], Layer-CAM [5], and Ablation-CAM [6]—within the same Otsu or fixed-0.2 thresholding followed by GrabCut refinement. We evaluate on the ILSVRC2012 validation set [11] and measure the region overlap (IoU, Dice). Our experiments demonstrate that while GrabCut consistently enhances final pseudo-mask quality for all CAM methods, the relative benefits vary substantially: some variants yield stronger raw seeds but saturate under GrabCut, whereas others benefit more from refinement.

Our key contributions include a comprehensive benchmark of six CAM variants evaluated under two thresholding schemes—fixed-0.2 (Zhou et al., 2016) and Otsu (Otsu, 1975)—with GrabCut refinement, a large-scale evaluation on 23,151 ImageNet images reporting both IoU and Dice metrics, and novel insights into which CAM variants yield the most GrabCut-amenable seeds, thereby guiding future efforts in pseudo-mask generation.

The remainder of the paper is organized as follows. Section 2 reviews related work on CAM and graph-cut refinement. Section 3 details our unified pipeline. Section 4 describes the experimental setup. Results and analysis appear in Section 5, followed by discussion and conclusion (Section 6 and 7).

2. Related work

2.1. Class activation mapping

CAM introduced the idea of using the weights of the final fully connected layer over global-average-pooled feature maps to produce coarse localization heatmaps from image-level labels [1]. Grad-CAM generalized this to arbitrary CNN architectures by weighting feature maps with the target-class gradients [2], and Grad-CAM++ further refined the weighting scheme to better handle multiple object instances [3]. Eigen-CAM derives heatmaps via the principal component of the activations [4]. Layer-CAM fuses multi-layer activations for hierarchical localization [5], and Ablation-CAM uses mask-based, gradient-free importance estimation [6] .

2.2. Thresholding and graph-cut refinement

To convert continuous CAM heatmaps into binary seeds, many works apply a fixed threshold at 0.2 of the normalized map—a convention established in the original CAM paper [1]. Otsu's method offers an automatic, data-driven alternative by selecting a threshold that minimizes intra-class variance in the heatmap histogram. GrabCut then refines these seeds into full object masks by modeling foreground/background color distributions with Gaussian mixtures and solving a graph-cut energy minimization [7].

DOI: 10.54254/2755-2721/2025.BJ25896

2.3. CAM-seeded GrabCut in weakly-supervised segmentation

Several domain-specific pipelines combine CAM seeds with GrabCut to generate pseudo-masks under weak supervision:

• Roadside Object Segmentation: Gülen et al. use ResNet-50 CAM to localize street signs, poles, and barriers, then refine those seeds with GrabCut for semantic segmentation of roadside elements [8].

• Aerial Remote Sensing: Guo et al.'s JMLNet generates Grad-CAM proposals on overhead imagery and applies GrabCut to obtain high quality multi-label segmentation masks for land-cover classes [9].

• Webly-Supervised Video: Hong et al. extract CAMs from web crawled videos and use them to initialize GrabCut, creating pseudo-masks for training segmentation models without manual annotations [12].

• Iterative CAM + GrabCut: Shimoda and Yanase propose an iterative pipeline where initial Grad-CAM seeds are refined by GrabCut and then used to update CAM maps in subsequent rounds, improving pseudo-mask quality on natural scene datasets [13].

While each of these works demonstrates the utility of CAM-seeded GrabCut in specific domains, no prior study has systematically compared multiple CAM variants under a unified thresholding (fixed 0.2 and Otsu) and GrabCut pipeline on a large-scale, diverse benchmark.

3. Method

In this section, we describe our unified pipeline for generating pseudo-masks by extracting Class Activation Maps (CAMs) from pretrained classifiers, binarizing them via fixed thresholding or Otsu's method, and refining the resulting seeds with GrabCut (see Figure 1). We focus on three backbone architectures—ResNet-50 (main), ResNet-101 (supplementary), and VGG-16 (supplementary)—and seven CAM variants.

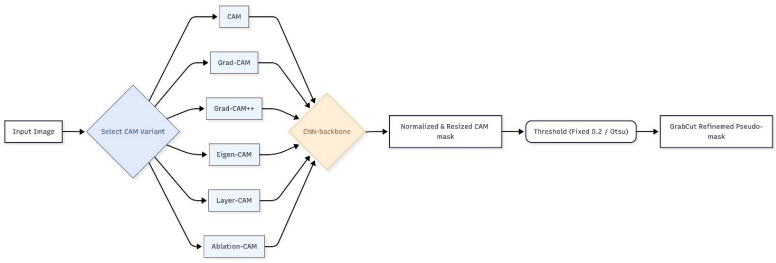

Figure 1. Overview of the weakly-supervised segmentation pipeline

From the input image, one of six CAM variants is generated; the resulting heatmap is then normalized and resized, thresholded (with either a fixed 0.2 threshold or Otsu's method), and finally refined via GrabCut to produce the final pseudo-mask.

3.1. Backbone architectures

We employ an off-the-shelf ImageNet-pretrained network. Specifically, we choose ResNet-50 [14] as the sole backbone due to its well-balanced trade-off between representational capacity and computational efficiency. With 50 layers of residual blocks, ResNet-50 delivers strong feature–learning capabilities while avoiding the markedly higher memory footprint and inference latency of deeper variants. Its widespread adoption in both academic and industrial settings also ensures that

DOI: 10.54254/2755-2721/2025.BJ25896

pre-trained weights are highly optimized and that the architecture integrates seamlessly with existing toolchains. The network $f(\cdot)$ outputs feature maps $A \in R^{C\times h\times w}$ and a final classification head.

3.2. CAM variants

For each backbone, we compute seven types of class-specific heatmaps $\widetilde{M} \in [0,1]^{H\times W}$ for the ground-truth class k:

- CAM [1] : $M_{i,j}^{k} = \sum_{c} W_{k,c}\, A_{c,i,j}$.
- Grad-CAM [2] : gradient-weighted feature maps.
- Grad-CAM++ [3] : refined higher-order weighting.
- Eigen-CAM [4] : first principal component of A.
- Layer-CAM [5] : pixel-wise multi-layer fusion.
- Ablation-CAM [6] : gradient-free mask ablations.

All maps are resized (bilinearly) to the input image dimensions $H \times W$ and normalized to $[0,1]$.

3.3. Seed binarization

We convert each normalized heatmap $\widetilde{M}$ into a binary seed mask $S \in \{0,1\}^{H\times W}$ via two schemes:

- Fixed-0.2 threshold:

$$S_{i,j} = \begin{cases} 1, & \widetilde{M}_{i,j} \geq 0.2, \\ 0, & \text{otherwise}, \end{cases}$$

following the convention in [1].

- Otsu's method: compute $t^{*} = \text{Otsu}\left(\widetilde{M}\right)$[10], then

$$S_{i,j} = \begin{cases} 1, & \widetilde{M}_{i,j} \geq t^{*}, \\ 0, & \text{otherwise}. \end{cases}$$

3.4. GrabCut refinement

Given a binary seed mask $S \in \{0,1\}^{H\times W}$, we refine it with the GrabCut algorithm [7]. Let $\widetilde{M} \in [0,1]^{H\times W}$ be the normalized CAM heatmap. We initialize the GrabCut label map $G \in \{\texttt{GC\_BGD}, \texttt{GC\_PR\_BGD}, \texttt{GC\_PR\_FGD}, \texttt{GC\_FGD}\}$ as follows:

DOI: 10.54254/2755-2721/2025.BJ25896

$$G_{i,j} = \begin{cases} \texttt{GC\_FGD} & \widetilde{M}_{i,j} > 0.9, \\ \texttt{GC\_PR\_FGD}, & \widetilde{M}_{i,j} \le 0.9 \ \wedge \ S_{i,j} = 1, \\ \texttt{GC\_PR\_BGD} & \widetilde{M}_{i,j} \le 0.9 \ \wedge \ S_{i,j} = 0. \end{cases}$$

Here, pixels with very high CAM response (> 0.9) are marked as definite foreground, while the rest are labelled as probable foreground or probable background according to the seed mask S . We then run GrabCut for a fixed number of iterations (e.g.5) to obtain a refined labelling G' . Finally, we produce the pseudo-mask $P \in \{0,1\}^{H \times W}$ by setting

$$P_{i,j} = \begin{cases} 1, & G'_{i,j} \in \{\texttt{GC\_FGD}, \texttt{GC\_PR\_FGD}\}, \\ 0, & \text{otherwise.} \end{cases}$$

This refinement leverages both the high-confidence CAM peaks and the probabilistic seed map to expand the object region while respecting image-level color and edge information.

4. Experimental setup

4.1. Dataset

We evaluated our pipeline on the ImageNet-LID Track 3 validation set, which comprises 50,000 ILSVRC2012 images, of which 23,151 have pixel-level ground truth masks released by the LID Challenge [11,15] . All experiments process only these 23,151 images to ensure direct comparison of pseudo-mask accuracy. For network input, images are resized to 224×224 and normalized with the ImageNet mean and standard deviation (as in [14]), enabling fair use of off-the-shelf pretrained backbones [16].

4.2. Evaluation metrics

We assessed pseudo-mask quality using two complementary overlap metrics, computed per image and then averaged over the dataset:

• Intersection over Union (IoU) [17] : for a predicted mask P and ground truth mask G ,

$$\text{IoU}\,(P,G) \;=\; \frac{|P \cap G|}{|P \cup G|} \quad \in [0,1].$$

We report the mean IoU across all N test images:

$$\overline{\text{IoU}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \frac{|P_i \cap G_i|}{|P_i \cup G_i|}\,.$$

DOI: 10.54254/2755-2721/2025.BJ25896

• Dice Coefficient [18] : also known as the F $_1$ –score for segmentation, defined as

$$\text{Dice}\,(P,G) \;=\; \frac{2\,|P \cap G|}{|P| + |G|} \qquad \in [0,1].$$

Similarly, we average Dice over all images to obtain

$$\overline{\text{Dice}} \;=\; \frac{1}{N}\sum_{i=1}^{N} \frac{2\,|P_i \cap G_i|}{|P_i| + |G_i|}\,.$$

4.3. Baselines and variants

To isolate the impact of CAM variant and thresholding, we compared:

1. Fixed-0.2 + GrabCut: binarize CAM heatmap at threshold 0.2 on the normalized map (per [1]), then refine with GrabCut [7].

2. Otsu + GrabCut: obtain seed by Otsu's method on the 8-bit CAM histogram [10] , then refine with GrabCut.

We apply both schemes to six CAM variants: CAM [1], Grad-CAM [2], Grad-CAM++ [3], Eigen-CAM [4], Layer-CAM [5], and Ablation-CAM [6].

5. Results and analysis

5.1. GrabCut-refined pseudo-mask accuracy

Table 1 presents the mean Intersection over Union (IoU) and Dice scores for GrabCut-refined pseudo-masks generated by ResNet-50 on the ImageNet-LID validation set. We compare two binarization strategies—adaptive Otsu thresholding and a fixed threshold of 0.2—each followed by GrabCut, across six different CAM variants. This highlights how threshold choice interacts with graph-cut refinement to impact overall mask quality.

Table 1. Mean IoU and Dice for GrabCut-refined masks on ResNet-50, comparing Otsu vs. fixed-0.2 binarization

Method	Otsu		Fixed-0.2	
	IoU	Dice	IoU	Dice
CAM	0.5124	0.6219	0.5065	0.6162
Grad-CAM	0.4354	0.5396	0.4273	0.5274
Grad-CAM++	0.3750	0.4697	0.3694	0.4633
Eigen-CAM	0.3359	0.4311	0.3394	0.4363
Layer-CAM	0.4275	0.5286	0.4259	0.5266
Ablation-CAM	0.4478	0.5532	0.4351	0.5358

DOI: 10.54254/2755-2721/2025.BJ25896

5.2. Raw CAM seed quality

Table 2 reports the baseline mean IoU and Dice scores for the six CAM variants when using only Otsu or a fixed 0.2 threshold—without any GrabCut refinement. This provides a clear reference for the initial seed quality prior to graph-cut post-processing.

Table 2. Mean IoU and Dice for raw CAM seeds on ResNet-50 under Otsu and fixed-0.2 thresholding (no GrabCut)

Method	Otsu		Fixed-0.2	
	IoU	Dice	IoU	Dice
CAM	0.3679	0.5153	0.3989	0.5432
Grad-CAM	0.2795	0.4083	0.2894	0.4168
Grad-CAM++	0.2490	0.3678	0.2693	0.3895
Eigen-CAM	0.2361	0.3519	0.2644	0.3854
Layer-CAM	0.2659	0.3919	0.2867	0.4160
Ablation-CAM	0.2825	0.4126	0.2912	0.4182

5.3. Analysis

To quantify the benefit of GrabCut refinement for each CAM variant, Table 3 reports the absolute gains in IoU and Dice when moving from raw CAM seeds (Otsu) to GrabCut-refined pseudo-masks (Otsu + GrabCut).

Table 3. ResNet-50 + CAM: absolute and relative improvements from GrabCut refinement under two binarization schemes

Method	Otsu				Fixed-0.2			
	ΔIoU	Rel.Gain(%)	ΔDice	Rel.Gain(%)	ΔIoU	Rel.Gain(%)	ΔDice	Rel.Gain(%)
CAM	0.14	39.28	0.11	20.69	0.11	26.97	0.07	13.44
Grad-CAM	0.16	55.78	0.13	32.16	0.14	47.65	0.11	26.54
Grad-CAM++	0.13	50.60	0.10	27.71	0.10	37.17	0.07	18.95
Eigen-CAM	0.10	42.27	0.08	22.51	0.07	28.37	0.05	13.21
Layer-CAM	0.16	60.77	0.14	34.88	0.14	48.55	0.11	26.59
Ablation-CAM	0.17	58.51	0.14	34.08	0.14	49.42	0.12	28.12

5.4. Analysis of GrabCut improvements

Overall, every CAM variant exhibits a substantial uplift from GrabCut refinement, confirming that graph-cut consistently enhances both the spatial precision (IoU) and overlap quality (Dice) of the initial pseudo-masks (Table 3).

First, focusing on the Otsu-thresholded seeds, the largest absolute IoU improvements come from Layer-CAM ($\Delta = 0.16$) and Ablation-CAM ($\Delta = 0.17$), followed by Grad-CAM ($\Delta = 0.16$) and original CAM ($\Delta = 0.14$). Eigen-CAM again shows the smallest IoU gain ($\Delta = 0.10$). In relative terms, Layer-CAM and Ablation-CAM benefit most (61% and 59% gains, respectively),

DOI: 10.54254/2755-2721/2025.BJ25896

whereas Eigen-CAM's relative increase is only 42%, reflecting its more diffuse seeds that leave less room for region growing.

Dice improvements under Otsu follow the same ranking: Layer-CAM (+0.14, +35%), Ablation-CAM (+0.14, +34%), and Grad-CAM (+0.13, +32%) top the list, while Eigen-CAM gains +0.08 Dice (+23%). The slightly larger relative Dice gains (vs. IoU) for Layer-CAM highlight GrabCut's boundary-snapping strength—thin protrusions and jagged edges are smoothed, boosting volumetric overlap even when intersection area gains are moderate.

Examining the fixed-0.2 threshold scenario reveals very similar patterns. Absolute IoU gains are marginally lower (e.g. Layer-CAM and Ablation-CAM at +0.14 vs. +0.16), but relative improvements remain high (49%–60%). This consistency demonstrates that, regardless of binarization strategy, GrabCut reliably rescues low-confidence regions.

Interestingly, methods with the poorest raw IoU (e.g. Grad-CAM and Grad-CAM++) do not always register the highest absolute gains, but they do see the largest relative boosts. Grad-CAM's IoU jumps by 56% under Otsu—outpacing Grad-CAM++'s 51% gain—despite only marginal baseline differences. This suggests that simpler gradient-averaged seeds provide clearer foreground–background contrast for GrabCut to exploit, whereas more smoothed variants attenuate these cues.

6. Discussion

Our comprehensive evaluation across 23,151 ImageNet-LID validation images reveals several key insights into the interplay between CAM seed generation and graph-cut refinement. First, despite the proliferation of sophisticated CAM variants, the original CAM formulation [1] consistently produces the most compact and reliable seeds, yielding the highest raw IoU (0.3989) and Dice (0.5432) at a fixed 0.2 threshold. This strong "core" localization proves especially amenable to GrabCut's region-growing mechanism, resulting in final masks with IoU = 0.5124 and Dice = 0.6219 under Otsu+GrabCut. In contrast, methods such as Grad-CAM++ [3] and Eigen-CAM [4] , while offering smoother or more globally distributed saliency, tend either to diffuse true object extents or to introduce spurious activations, limiting the net gain from graph-cut refinement.

Across all seven variants, GrabCut yields absolute IoU improvements between +0.0998 (Eigen-CAM) and +0.1653 (Ablation-CAM), and Dice gains between +0.0792 and +0.1406. The largest gains for Ablation-CAM and Layer-CAM suggest that seeds with moderate initial recall—but lower precision—stand to benefit most from boundary-and-color cues. Meanwhile, original CAM's narrower but high-confidence seeds achieve a lower absolute uplift (+0.1445 IoU, +0.1066 Dice) yet end up with the strongest overall masks. These patterns underscore that seed compactness, foreground–background contrast, and boundary sharpness jointly determine how much refinement can improve segmentation quality.

Moreover, when considering relative percentage gain, lower-performing seeds often see proportionally larger boosts: for example, Grad-CAM's IoU increases by over 56% compared to its raw Otsu seed, whereas original CAM gains about 39%. This finding highlights GrabCut's particular utility for rescuing weak or noisy activations, suggesting that graph-cut refinement is especially valuable in settings where initial localization is uncertain or incomplete.

6.1. Limitations and future work

Our study focuses on single-object ILSVRC images while real-world scenes often contain multiple object instances and complex backgrounds, which may shift the relative strengths of CAM variants and refinement strategies. Extending this evaluation to multi-object benchmarks such as Pascal VOC

DOI: 10.54254/2755-2721/2025.BJ25896

[17] or MS COCO [19] would test robustness under greater clutter and occlusion. Additionally, while we analyze two simple thresholding schemes, more adaptive binarization methods—or learned threshold predictors—might further balance precision and recall before refinement. Beyond GrabCut, exploring advanced post-processing techniques like fully connected CRFs [20] , learned affinity propagation [21] , or trainable graph neural networks may unlock additional gains by better modeling long-range dependencies and complex edge structures. Finally, per-variant or per-image parameter tuning for graph-cut energy terms could adaptively maximize improvement based on seed characteristics.

7. Conclusion

We have delivered the first systematic comparison of seven CAM variants under identical thresholding and GrabCut refinement pipelines, demonstrating that vanilla CAM remains the most effective seed generator for large-scale, weakly supervised pseudo-mask creation. Graph-cut refinement universally enhances seed quality, but the magnitude of improvement depends critically on seed compactness and contrast. Our results provide clear guidance for practitioners: combine original CAM seeds with simple thresholding and GrabCut to achieve high-quality pseudo-masks, and consider more advanced or adaptive refinement techniques to further boost performance in complex or multi-object scenarios.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization, ” in CVPR, 2016.
[2] 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 ICCV, 2017.
[3] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, ” arXiv, 2017.
[4] M. B. Muhammad and M. Yeasin, “Eigen-cam: Class activation map using principal components, ”in 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, Jul. 2020, p. 1–7. [Online]. Available: http://dx.doi.org/10.1109/IJCNN48605.2020.9206626
[5] 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.
[6] 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.
[7] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts, ” ACM Trans. Graph., vol. 23, pp. 309–314, 2004.
[8] S. Gülen et al., “Weakly supervised semantic roadside object segmentation using digital maps, ” Remote Sens., 2020.
[9] R. Guo, X. Sun, K. Chen, and M. Yan, “Jmlnet: Joint multi-label learning network for weakly super-vised semantic segmentation in aerial images, ” Remote Sens., 2020.
[10] N. Otsu et al., “A threshold selection method from gray-level histograms, ” Automatica, vol. 11, no.285-296, pp. 23–27, 1975.
[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database, ” in CVPR, 2009.
[12] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, “Weakly supervised semantic segmentation using web-crawled videos, ” in CVPR, 2017.
[13] W. Shimoda and K. Yanai, “Self-supervised difference detection for weakly-supervised semantic segmentation, ” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp.5208–5217.
[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition, ” in CVPR, 2016.

DOI: 10.54254/2755-2721/2025.BJ25896

[15] Y. Wei, S. Zheng, M.-M. Cheng, H. Zhao, L. Wang, E. Ding, Y. Yang, A. Torralba, T. Liu, G. Sun et al., “Lid 2020: The learning from imperfect data challenge results, ” arXiv preprint arXiv: 2010.11724, 2020.
[16] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition, ”arXiv preprint arXiv: 1409.1556, 2014.
[17] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge, ” in IJCV, 2010.
[18] L. R. Dice, “Measures of the amount of ecologic association between species, ” Ecology, vol. 26, no. 3, pp. 297–302, 1945.
[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context, ” in Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. Springer, 2014, pp. 740–755.
[20] J. Lafferty, A. McCallum, F. Pereira et al., “Conditional random fields: Probabilistic models for segmenting and labeling sequence data, ” in Icml, vol. 1, no. 2. Williamstown, MA, 2001, p. 3.
[21] A. Khoreva, R. Benenson, J. Hosang, and B. Schiele, “Simple does it: Weakly supervised instance and semantic segmentation, ” in CVPR, 2017.