Beyond Softmax: Dual-Branch Sigmoid Architecture for Accurate Class Activation Maps

Yoojin Oh
mydianaoh@ewha.ac.kr
Junhyug Noh
junhyug@ewha.ac.kr

Department of Artificial Intelligence
Ewha Womans University
Seoul, Republic of Korea

Abstract

Class Activation Mapping (CAM) and its extensions have become indispensable tools for visualizing the evidence behind deep network predictions. However, by relying on a final softmax classifier, these methods suffer from two fundamental distortions: *additive logit shifts* that arbitrarily bias importance scores, and *sign collapse* that conflates excitatory and inhibitory features. We propose a simple, architecture-agnostic *dual-branch sigmoid* head that decouples localization from classification. Given any pretrained model, we clone its classification head into a parallel branch ending in per-class sigmoid outputs, freeze the original softmax head, and fine-tune only the sigmoid branch with class-balanced binary supervision. At inference, softmax retains recognition accuracy, while class evidence maps are generated from the sigmoid branch – preserving both magnitude and sign of feature contributions. Our method integrates seamlessly with most CAM variants and incurs negligible overhead. Extensive evaluations on fine-grained tasks (CUB-200-2011, Stanford Cars) and WSOL benchmarks (ImageNet-1K, OpenImages-30K) show improved explanation fidelity and consistent Top-1 Localization gains – without any drop in classification accuracy. Code is available at https://github.com/finallyupper/beyond-softmax.

1 Introduction

Deep neural networks have achieved remarkable success across a wide range of visual recognition tasks, from fine-grained classification [27] to large-scale object detection [18]. However, their decision processes remain largely opaque, limiting trust in safety-critical domains such as medical imaging and autonomous driving. Class Activation Mapping (CAM) [32] was proposed to bridge this gap by projecting the weighted sum of high-level feature maps back onto the input image, yielding a heatmap that highlights discriminative regions for a given class. While effective, CAM requires architectural constraints (Global Average Pooling & fully-connected layers) and exposes only linear combinations of feature maps.

Grad-CAM [19] and Grad-CAM++ [5] generalize CAM to arbitrary network architectures – be they convolutional backbones with deep MLP heads or transformer encoders with `[CLS]` tokens – by using gradients of the target logit to compute per-channel importance

scores. A host of further variants have since been developed: Score-CAM [28] and Ablation-CAM [17] eschew gradients entirely, measuring score perturbations under masking or ablation; Layer-CAM [13] and Eigen-CAM [15] fuse multi-layer activations or principal components; and attention-based methods directly visualize ViT attention maps [1, 10].

Despite these advances, most CAM-style explanations inherit two fundamental distortions from the ubiquitous softmax classifier: (i) *additive logit shift*, whereby adding a constant to all class logits leaves softmax probabilities unchanged but arbitrarily biases importance scores; and (ii) *sign collapse*, whereby softmax depends only on relative ordering of logits and discards the absolute sign of feature contributions. We show in Section 3.2 that these artifacts can dramatically mislead localization maps.

To overcome these issues, we decouple localization from classification via a lightweight, architecture-agnostic *dual-branch sigmoid* head. Starting from any pretrained model, we clone its classification head into a parallel sigmoid branch – freeze the original softmax branch and fine-tune only the sigmoid branch with class-balanced binary supervision. At inference, we retain softmax for recognition and use the sigmoid branch to generate class evidence maps with absolute, signed feature importance. This plug-in approach works across any CAM variants and delivers consistent gains in two scenarios: (1) *explanation fidelity* on fine-grained tasks (CUB-200-2011 [27] and Stanford Cars [14]), reducing Average Drop and boosting % Increase in Confidence and (2) *weakly supervised object localization* on ImageNet-1K [18] and OpenImages-30K [4] – improving Top-1 Loc, MaxBoxAccV2, and PxAP without harming classification – across multiple CAM methods.

In summary, our contributions are:

- A formal analysis of two softmax-induced distortions – *additive logit shift* and *sign collapse* – that undermine CAM-style explanations.
- A dual-branch sigmoid head that restores absolute, signed feature importance for arbitrary classifier architectures.
- Extensive experiments showing enhanced explanation fidelity on CUB-200-2011 and Stanford Cars, and WSOL improvements on ImageNet-1K and OpenImages-30K, all with negligible overhead.

2 Related Work

Gradient-Based Methods. Early pixel-level saliency maps compute the gradient of a class score with respect to each input pixel [22]. SmoothGrad [24] reduces noise by averaging over noisy inputs, and Integrated Gradients [25] accumulates gradients along a path from a baseline. CAM [32] first extended saliency to feature maps via Global Average Pooling and a linear head. Grad-CAM [19] and Grad-CAM++ [5] generalized this to arbitrary architectures by averaging channel gradients, with Grad-CAM++ using higher-order derivatives for multiple-instance localization. Layer-CAM [13] fuses activation-gradient products across layers, and XGrad-CAM [9] derives weights from axiomatic properties.

Gradient-Free and Perturbation-Based Methods. To avoid gradient saturation, Score-CAM [28] uses the forward pass with feature-map masks to measure class scores, and Ablation-CAM [17] quantifies score drops when ablating channels. RISE [16] averages random input masks, while Extremal Perturbations [8] optimize sparse masks to highlight important regions. These perturbation methods often yield sharper maps but require multiple forward passes.

Statistical and Structural Methods. Eigen-CAM [15] leverages the first principal component of the feature tensor as a heatmap. Transformer-specific approaches – Attention Rollout [1] and TS-CAM [10] – visualize aggregated self-attention. Relevance propagation techniques such as LRP [2] and DeepLIFT [20] distribute the prediction score back through network layers.

Weakly-Supervised Object Localization. WSOL methods seek to predict object bounding boxes or masks using only image-level labels. Early works [6, 23, 29, 30] train models by adapting CAM heatmaps to cover entire objects. More recent work, such as Rethinking CAM [3], improves the CAM pipeline itself through refined thresholds and evaluation protocols. Unlike post-hoc explanations, WSOL integrates localization objectives or auxiliary branches during training to directly encourage complete object coverage.

Most CAM-style methods derive channel or pixel weights from softmax logits (or their perturbations) and thus inherit *additive logit shifts* and *sign collapse* distortions. In contrast, our dual-branch sigmoid head trains per-class binary classifiers to recover absolute magnitude and polarity of feature contributions, and plugs seamlessly into any CAM or WSOL pipeline to produce more faithful localization maps.

3 Our Approach

We propose a dual-branch sigmoid head that decouples localization from classification by restoring absolute, signed feature importance in CAM-style heatmaps. Our approach unfolds in three parts: first, Sec. 3.1 reviews CAM fundamentals; next, Sec. 3.2 analyzes softmax-induced distortions; and finally, Sec. 3.3 details our dual-branch sigmoid architecture along with its training and inference procedures.

3.1 Preliminaries: CAM Definitions

Given an input image $\mathbf{x}$, let $F_i \in \mathbb{R}^{P\times Q}$ be the i-th channel feature map of the final convolutional layer, with $i = 1,\dots,N$. A classification head h maps the feature tensor $\mathbf{F} \in \mathbb{R}^{N\times P\times Q}$ to logits

$$\ell_k = h_k(\mathbf{F}), \quad k = 1,\dots,C, \tag{1}$$

with softmax probabilities:

$$y_k = \frac{\exp(\ell_k)}{\sum_{j=1}^{C}\exp(\ell_j)}. \tag{2}$$

Vanilla CAM. In the original CAM [32], h is a Global Average Pooling (GAP) layer followed by a fully-connected (FC) layer. Specifically,

$$\ell_k = \sum_{i=1}^{N} w_{i,k}\bar{F}_i + b_k, \qquad \bar{F}_i \triangleq \frac{1}{PQ}\sum_{p=1}^{P}\sum_{q=1}^{Q} F_i^{(p,q)}. \tag{3}$$

where $w_{i,k}$ and b_k are the FC weights and bias. The (linear) class activation map for class k is

$$M_k(p,q) \;=\; \sum_{i=1}^{N} w_{i,k}\,F_i^{(p,q)}. \tag{4}$$

Gradient-Based CAM. To remove architectural constraints, gradient-based CAM variants [5, 19] first compute intermediate importance scores:

$$\alpha_{i,k} = \frac{1}{PQ} \sum_{p=1}^{P} \sum_{q=1}^{Q} \frac{\partial \ell_k}{\partial F_i^{(p,q)}}, \tag{5}$$

which are then post-processed – *e.g.*, directly as $w_{i,k} = \alpha_{i,k}$ in Grad-CAM, refined by higher-order derivatives in Grad-CAM++ [5], or fused across layers in Layer-CAM [13] – to obtain the final weights for Eq. (4). Many variants additionally apply an elementwise ReLU to obtain a non-negative heatmap:

$$M_k(p,q) \leftarrow \mathrm{ReLU}\big(M_k(p,q)\big). \tag{6}$$

We adopt this ReLU-applied map as the default for visualization and evaluation; however, for analytical clarity (e.g., in Sec. 3.2) we refer to the *linear* score M_k in Eq. (4) without the ReLU.

Other Variants. Beyond vanilla CAM and gradient-based extensions, many CAM methods [9, 15, 17, 28, 31] adhere to the same two-step recipe. First, they derive per-channel importance scores $w_{i,k}$ from the model's softmax logits or scores – using gradients, higher-order derivatives, or perturbations of inputs or feature maps. Second, they linearly combine these weights with the feature maps to yield the final heatmap as Eq. (4). For example, Score-CAM [28] masks the input with each feature map in turn and measures the resulting change in logit

$$\alpha_{i,k} = \ell_k(\mathbf{x} \odot \mathrm{mask}_i) \; - \; \ell_k(\mathbf{x}), \tag{7}$$

then normalizes $\{\alpha_{i,k}\}$ to obtain $w_{i,k}$.

3.2 Softmax-Induced Distortions

As reviewed above, all CAM variants ultimately form a heatmap by linearly combining feature maps with per-channel weights (Eq. (4)). For this linear combination to be semantically valid, the weights should satisfy two tacit requirements: (i) *relative magnitude semantics* – $w_{i,k}$ should reflect how important channel i is for class k so that weights are comparable across channels; and (ii) *sign semantics* – the sign of $w_{i,k}$ should indicate whether the corresponding activation contributes evidence (positive) or inhibition (negative) to class k.

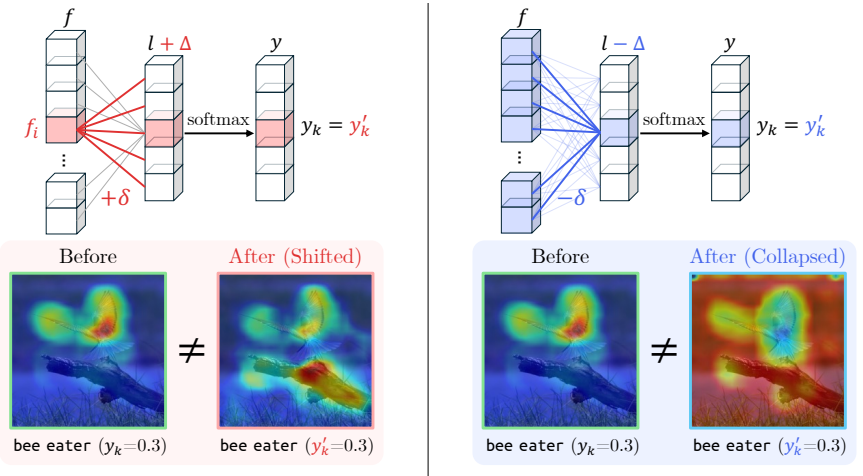
However, when weights are derived from *softmax*-based scores, softmax's invariances break these assumptions: *additive logit shifts* arbitrarily bias magnitudes, and *sign collapse* conflates evidence and inhibition. Figure 1 illustrates these effects with concrete examples.

1) Additive Logit Shift. Softmax is invariant to adding the same constant to all class logits. Let $\ell \in \mathbb{R}^C$ and $\mathbf{1} \in \mathbb{R}^C$ be the all-ones vector. For any $\Delta \in \mathbb{R}$, define $\ell' = \ell + \Delta \mathbf{1}$. Then

$$y'_k = \frac{e^{\ell_k + \Delta}}{\sum_{c=1}^{C} e^{\ell_c + \Delta}} = \frac{e^{\ell_k} e^{\Delta}}{e^{\Delta} \sum_{c=1}^{C} e^{\ell_c}} = y_k. \tag{8}$$

Now apply a *uniform* shift to all FC weights connected to the i-th feature map across classes:

$$w'_{i,k} = w_{i,k} + \delta \quad (\forall k), \qquad w'_{j,k} = w_{j,k} \;\; (j \neq i).$$



(a) Additive Logit Shift (b) Sign Collapse

Figure 1: Softmax-induced distortions in CAM-based localization. (a) *Additive Logit Shift*: adding a constant δ to all feature weights leaves the softmax probability y_k unchanged but disproportionately amplifies feature f_i in the heatmap. (b) *Sign Collapse*: subtracting δ flips formerly positive feature weights to negative without affecting y_k, causing previously highlighted regions to vanish. In both cases, identical classification outputs produce drastically different localization maps.

Under the GAP+FC head in Eq. (3), each logit becomes

$$\ell'_k = \sum_{j=1}^{N} w'_{j,k} \bar{F}_j + b_k = \ell_k + \delta \bar{F}_i,$$

so softmax probabilities are unchanged by (8). However, the linear CAM combination in Eq. (4) changes *spatially*:

$$M'_k(p,q) = \sum_{j=1}^{N} w'_{j,k} F_j^{(p,q)} = \sum_{j=1}^{N} w_{j,k} F_j^{(p,q)} + \delta F_i^{(p,q)} = M_k(p,q) + \delta F_i^{(p,q)}. \quad (9)$$

Thus – even with identical class probabilities – the heatmap can be arbitrarily brightened or dimmed where F_i is large (and likewise after applying ReLU when used).

2) Sign Collapse. Softmax depends only on the *relative differences* among logits and ignores their absolute sign. For example, subtracting a large constant from every FC weight in Eq. (3),

$$w'_{i,k} = w_{i,k} - \delta, \quad \delta \gg 0, \quad (10)$$

shifts all logits uniformly (so y_k remains unchanged by the shift invariance in Eq. (8)), yet it inverts CAM contributions: channels that were positive become inhibitory, thereby misleading the localization.

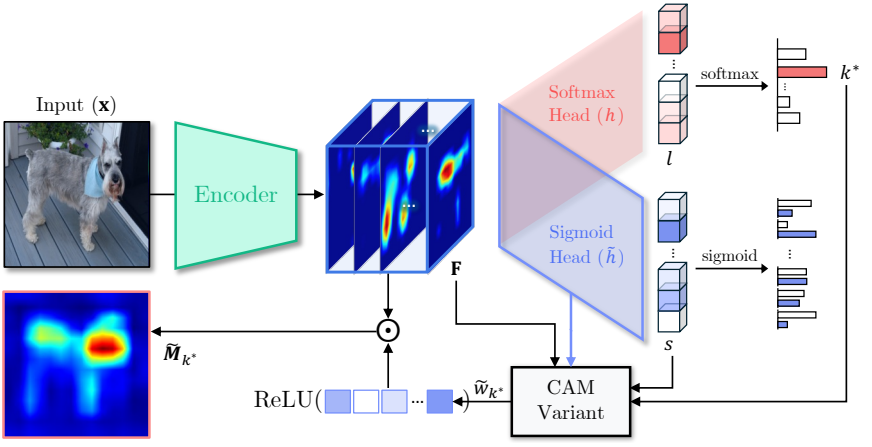


Figure 2: Inference pipeline of the dual-branch sigmoid CAM. After feature extraction, the frozen softmax head predicts the class label k^*. In parallel, any CAM variant computes per-channel importance scores $\tilde{w}_{k^*}$ (via weights or gradients) for s_{k^*}, which are rectified by clamping to positive values. These positive-only scores are then linearly combined with the feature maps to produce the final class evidence heatmap $\tilde{M}_{k^*}$.

3.3 Dual-Branch Sigmoid Head

To disentangle these distortions, we introduce a *dual-branch sigmoid* head that decouples localization from classification.

Head Replication. Starting from a pretrained classifier, we copy its head h into a new branch $\tilde{h}$ with identical architecture (GAP+FC or MLPs), but fresh parameters. The sigmoid branch outputs class-wise scores:

$$s_k = \sigma\big(\tilde{h}_k(\mathbf{F})\big), \quad \sigma(z) = 1/(1+e^{-z}), \quad k = 1,\ldots,C. \tag{11}$$

The original softmax head and backbone remain frozen.

Binary Supervision. Each image is a positive sample for exactly one class and a negative sample for the other $C-1$ classes, inducing a strong imbalance (*e.g.*, 1:999 on ImageNet). We therefore train the sigmoid head $\tilde{h}$ with class-balanced binary cross-entropy:

$$\mathcal{L} = -\frac{1}{B}\sum_{n=1}^{B}\Big[\underbrace{\big(1-\tfrac{1}{C}\big)\log s_{y^{(n)}}}_{\text{positive}} + \underbrace{\tfrac{1}{C}\sum_{k\neq y^{(n)}}\log(1-s_k)}_{\text{negatives}}\Big]. \tag{12}$$

Sigmoid Branch Semantics. Each sigmoid head $s_k = \sigma(\tilde{\ell}_k)$ is trained as an *independent* binary classifier for class k with BCE (Eq. (12)). Because classes are optimized independently, $\tilde{w}_{i,k}$ is determined solely by its effect on s_k rather than by cross-class normalization or logit ranks. This breaks softmax invariances (additive logit shifts and sign collapse) and

restores both the *magnitude* and *polarity* needed for CAM-style maps. For example, under a GAP+FC head (as in vanilla CAM), the sigmoid logit can be written as $\tilde{\ell}_k = \tilde{b}_k + \sum_{i=1}^{N} \tilde{w}_{i,k} \bar{F}_i$ (*c.f.*, Eq. (3)), and it satisfies:

- **Relative importance.** The BCE objective increases $\tilde{w}_{i,k}$ for channels whose activations raise s_k on positives and decreases it otherwise, yielding within-class comparability: larger $|\tilde{w}_{i,k}|$ denotes stronger influence on s_k for class k.
- **Sign semantics (evidence-only mapping).** Since $\sigma(\cdot)$ is strictly increasing, increasing $\bar{F}_i$ raises s_k iff $\tilde{w}_{i,k} > 0$ (evidence) and lowers it iff $\tilde{w}_{i,k} < 0$ (inhibition), thereby restoring the polarity lost under softmax. Leveraging this at inference, we retain only positive contributions – using $\max(0, \tilde{w}_{i,k})$ (or $\max(0, \tilde{\alpha}_{i,k})$) – and combine them with feature maps as in Eq. (4) to produce class-k evidence maps without negative leakage.

Inference. During testing, we first predict the class via the softmax branch and obtain $k^* = \arg\max_k y_k$. For localization, we apply the chosen CAM variant to the *sigmoid* branch – *i.e.*, compute per-channel scores $\tilde{w}_{i,k^*}$ from the sigmoid logit s_{k^*} rather than the softmax logit ℓ_{k^*} (*c.f.*, Eqs. (5), (7)). We then retain only positive evidence by clamping negatives to zero and combine the resulting weights with feature maps as in Eq. (6) to obtain $\tilde{M}_{k^*}$. This drop-in substitution (softmax $\rightarrow$ sigmoid) makes the procedure applicable to any CAM variant and yields signed, distortion-free maps. Figure 2 illustrates the complete inference pipeline.

4 Experiments

We evaluate our dual-branch sigmoid head on two complementary tasks: (1) explanation fidelity in fine-grained classification, which measures how faithfully heatmaps highlight the subtle, class-specific cues that distinguish similar categories; and (2) weakly-supervised object localization (WSOL) on large-scale datasets, which tests whether our method can recover full object extents from image-level supervision. For vanilla CAM both softmax and sigmoid heads employ a GAP+FC design; for all other variants we preserve the original classifier head architectures. Full implementation details and additional results are provided in the Appendix.

4.1 Fine-Grained Explanation Fidelity

Fine-grained classification requires distinguishing very similar categories based on subtle visual cues. Explanation fidelity metrics measure whether heatmaps correctly highlight these small, discriminative regions, and thus reflect the true model reasoning.

Datasets. We evaluate two fine-grained benchmarks:

- **CUB-200-2011** [27]: 6,033 train, 5,755 test images of 200 bird species.
- **Stanford Cars** [14]: 8,144 train, 8,041 test images of 196 car models.

Metrics. We use masking-based *Average Drop* (%) and *% Increase in Confidence* [5]. Following [5], the explanation map is constructed by element-wise multiplying the class-conditional saliency map (upsampled to the input size) with the original image as $E_{k^*} = \tilde{M}_{k^*} \circ \mathbf{x}$, where $\circ$ denotes the Hadamard product, $\tilde{M}_{k^*}$ is the saliency map for predicted class

Backbone	Method	CUB-200-2011		Stanford Cars	
		% Avg Drop (↓)	% Inc Conf (↑)	% Avg Drop (↓)	% Inc Conf (↑)
VGG-16	CAM	45.36	7.08	31.45	9.68
	+ Ours	37.20 (-8.16)	18.86 (+11.78)	43.58 (+12.13)	10.74 (+1.06)
	Grad-CAM	38.88	11.36	56.10	8.11
	+ Ours	35.66 (-3.22)	22.06 (+10.70)	9.96 (-46.14)	19.46 (+11.35)
	Grad-CAM++	31.11	14.46	39.63	11.71
	+ Ours	33.98 (+2.87)	21.75 (+7.29)	9.46 (-30.17)	18.46 (+6.75)
	XGrad-CAM	38.56	11.10	50.59	9.14
	+ Ours	38.04 (-0.52)	20.99 (+9.89)	9.69 (-40.90)	20.53 (+11.39)
	Layer-CAM	37.19	11.27	47.51	9.70
	+ Ours	41.58 (+4.39)	19.50 (+8.23)	10.51 (-37.00)	19.39 (+9.69)
ResNet-50	CAM	44.17	14.76	48.86	4.32
	+ Ours	13.86 (-30.31)	18.42 (+3.66)	3.77 (-45.09)	40.62 (+36.30)
	Grad-CAM	46.38	13.07	29.53	14.56
	+ Ours	19.07 (-27.31)	13.22 (+0.15)	3.66 (-25.87)	41.41 (+26.85)
	Grad-CAM++	43.15	14.50	24.68	17.31
	+ Ours	18.88 (-24.27)	13.72 (-0.78)	3.65 (-21.03)	41.62 (+24.31)
	XGrad-CAM	46.38	13.07	29.53	14.56
	+ Ours	19.07 (-27.31)	13.22 (+0.15)	3.66 (-25.87)	41.41 (+26.85)
	Layer-CAM	43.69	13.91	22.64	18.90
	+ Ours	19.07 (-24.62)	13.22 (-0.69)	3.66 (-18.98)	41.39 (+22.49)

Table 1: Fine-grained explanation fidelity on CUB-200-2011 and Stanford Cars. For % Average Drop (lower is better) and % Increase in Confidence (higher is better), improved values are shown in blue and worsened values in red; parentheses indicate the change relative to the baseline.

k^*, and $\mathbf{x}$ is the input image. These quantify whether the regions highlighted by a heatmap preserve or boost the target class score, serving as proxies for explanation faithfulness.

Backbones & Methods. We use VGG-16 [21] and ResNet-50 [11, 12] backbones. To demonstrate generality, we evaluate five CAM variants – CAM [32], Grad-CAM [19], Grad-CAM++ [5], XGrad-CAM [9], and Layer-CAM [13] – each with and without our sigmoid branch. This comprehensive setup shows how the sigmoid add-on improves diverse explanation techniques under fine-grained conditions, particularly for gradient-based methods.

Results. Table 1 demonstrates that our sigmoid branch markedly improves explanation fidelity across both backbones. On CUB-200-2011 with VGG-16, average drop is reduced by 0.52-8.16 percentage points (*e.g.*, CAM: 45.36 → 37.20, −8.16) while confidence increase rises by 7.29-11.78 points. Stanford Cars shows similar trends: most methods cut average drop by 30-46 points (*e.g.*, Grad-CAM: 56.10 → 9.96, −46.14), and boost confidence by 1-11 points. ResNet-50 yields even larger gains – CAM's drop falls from 44.17 to 13.86 (−30.31) and its confidence jump from 14.76 to 18.42 (+3.66), with analogous improvements for other variants. Only Grad-CAM++ on CUB and one Layer-CAM case exhibit minor reversals; overall, our approach sharply reduces drop and elevates confidence without exception or trade-offs.

Backbone	Method	ImageNet-1K				OpenImages-30K	
		Top1 Cls	Top1 Loc	GT Loc	MBAv2	Top1 Cls	PxAP
VGG-16	CAM	66.56	43.13	59.54	60.00	70.00	58.17
	+ Ours	–	43.14	59.57	60.25	–	60.11
	Δ	–	(+0.01)	(+0.03)	(+0.25)	–	(+1.94)
	Grad-CAM	69.61	37.04	49.02	52.49	70.12	55.25
	+ Ours	–	44.68	59.43	60.17	–	56.53
	Δ	–	(+7.64)	(+10.41)	(+7.68)	–	(+1.28)
ResNet-50	CAM	73.89	47.69	60.80	63.53	74.02	59.36
	+ Ours	–	49.44	62.68	64.73	–	60.49
	Δ	–	(+1.75)	(+1.88)	(+1.20)	–	(+1.13)
	Grad-CAM	69.01	41.14	55.28	57.04	74.02	60.56
	+ Ours	–	40.34	53.61	57.98	–	60.24
	Δ	–	(-0.80)	(-1.67)	(+0.94)	–	(-0.32)
InceptionV3	CAM	69.72	42.16	57.69	63.51	56.22	62.25
	+ Ours	–	44.11	60.12	64.98	–	63.65
	Δ	–	(+1.95)	(+2.43)	(+1.47)	–	(+1.40)
	Grad-CAM	67.79	34.57	46.35	52.30	68.56	49.64
	+ Ours	–	33.72	45.18	53.06	–	48.16
	Δ	–	(-0.85)	(-1.17)	(+0.76)	–	(-1.48)

Table 2: WSOL results on ImageNet-1K and OpenImages-30K. For each base method we shade the baseline row in gray; "+ Ours" rows report updated scores with their Δ shown in parentheses (blue for gains, red for drops).

4.2 Weakly-Supervised Object Localization

WSOL benchmarks assess a model's ability to localize entire objects using only image-level labels. Because traditional CAM-based explanations often focus on the most discriminative parts, WSOL on large-scale datasets reveals whether our sigmoid branch restores complete object coverage.

Datasets. We use two standard WSOL benchmarks:

- **ImageNet-1K** [18]: 1.28*M* train, 50*K* val images over 1,000 classes, with bounding boxes used only for evaluation.
- **OpenImages-30K** [4]: 29,819 train, 2,500 val, and 5,000 test images with pixel-wise masks.

Metrics. On ImageNet-1K, we report Top-1 classification accuracy, Top-1 localization, GT-known localization (IoU$\geq$ 0.5), and MaxBoxAccV2 (MBAv2) [7]. On OpenImages-30K, we report Top-1 classification and pixel-level average precision (PxAP) over the masks [7]. These metrics together ensure that our sigmoid branch maintains recognition performance while improving localization.

Backbones & Methods. We test three widely-used architectures: VGG-16 [21], ResNet-50 [11, 12], and InceptionV3 [26]. As base localization methods, we choose CAM [32] and Grad-CAM [19], then integrate our sigmoid branch into each. This setup evaluates how well our add-on improves diverse architectures and CAM variants.

Results. Table 2 summarizes our WSOL performance. With VGG-16, CAM + Ours yields slight gains in Top-1 Loc (+0.01), GT Loc (+0.03), MBAv2 (+0.25) and a larger boost in PxAP (+1.94), while Grad-CAM + Ours delivers substantial improvements (+7.64 Top-1 Loc, +10.41 GT Loc, +7.68 MBAv2, +1.28 PxAP). For ResNet-50, CAM + Ours improves all metrics, whereas Grad-CAM + Ours trades minor drops in Top-1 Loc (−0.80), GT Loc (−1.67) and PxAP (−0.32) for a MBAv2 gain (+0.94). InceptionV3 + CAM + Ours shows consistent gains, and InceptionV3 + Grad-CAM + Ours experiences small declines in Top-1 Loc (−0.85), GT Loc (−1.17) and PxAP (−1.48) alongside a +0.76 MBAv2 increase. Overall, our sigmoid branch consistently enhances CAM-based localization, with only minor trade-offs in select Grad-CAM cases.

4.3 Qualitative Results

Figure 3 shows two ImageNet WSOL examples – one with VGG-16 and one with ResNet-50. Our sigmoid branch consistently extends heatmaps to cover the full object rather than just the most discriminative part. Additional qualitative results are provided in the Appendix.

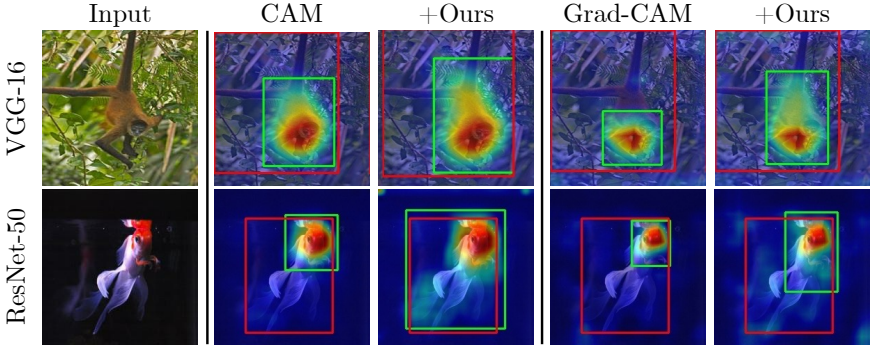


Figure 3: Qualitative WSOL on ImageNet-1K for VGG-16 (top) and ResNet-50 (bottom). From left to right: input, CAM, CAM+Ours, Grad-CAM, Grad-CAM+Ours. The boxes in green and red represent the predictions and ground truths of localization.

5 Conclusion

We have identified two inherent flaws – *additive logit shift* and *sign collapse* – in all softmax-based CAM explanations, and shown how they can lead to misleading localization. To remedy this, we introduced a dual-branch sigmoid head that restores absolute and signed feature importance without touching the backbone or softmax branch. Our approach is architecture-agnostic and compatible with most CAM variants. Extensive experiments on fine-grained explanation and WSOL benchmarks demonstrate that our method consistently improves explanation fidelity and localization robustness, all without degrading classification accuracy.

Acknowledgements

We thank Wonho Bae and Jinhwan Seo for insightful early discussions that helped us refine and articulate the initial problem statement of this work. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)).

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[2] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[3] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 618–634. Springer, 2020.

[4] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale Interactive Object Segmentation with Human Annotators. In *CVPR*, 2019.

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[6] Junsuk Choe and Hyunjung Shim. Attention-Based Dropout Layer for Weakly Supervised Object Localization. In *CVPR*, 2019.

[7] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating Weakly Supervised Object Localization Methods Right. In *CVPR*, 2020.

[8] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[9] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

[10] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2886–2895, 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *CVPR*, 2018.

[13] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[14] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[15] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

[16] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[17] Harish Guruprasad 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*, pages 983–991, 2020.

[18] Olga Russakovsky, Jia Deng, Hao SU, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and LI Fei-Fei. Imagenet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[20] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[21] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Arxiv:1409.1556*, 2014.

[22] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[23] Krishna Kumar Singh and Yong Jae Lee. Hide-And-Seek: Forcing a Network to Be Meticulous for Weakly-Supervised Object and Action Localization. In *ICCV*, 2017.

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[26] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *CVPR*, 2015.

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[28] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia 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*, pages 24–25, 2020.

[29] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object Region Mining With Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In *CVPR*, 2017.

[30] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial Complementary Learning for Weakly Supervised Object Localization. In *CVPR*, 2018.

[31] Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation. *arXiv preprint arXiv:2501.11309*, 2025.

[32] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.