CLAP: CLASSIFICATION AND LOCALIZATION WITH OBJECT-ALIGNED PROTOTYPE EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Part-prototype networks make image classification transparent by expressing predictions through spatial matches to learned prototypes. Yet classification supervision rewards any discriminative cue, so prototypes may respond to a small object part or correlated background rather than faithfully represent the target object. This limits the use of their spatial responses for object localization. We introduce CLAP, a training framework that makes prototype evidence object-aligned using only image-level labels. A lightweight class-agnostic foreground–background head produces pseudo-masks, and two alignment losses encourage collective coverage of the predicted foreground while suppressing responses on predicted background. Classification and localization are then computed from the same class-associated spatial prototype responses, allowing the prototype pathway to retain its discriminative role while gaining localization capability. We instantiate CLAP with ProtoPNet and ProtoPool. On part-prototype benchmarks, CLAP substantially improves GT-known localization while preserving classification accuracy with no additional cost at inference. On standard WSOL benchmarks, it achieves 97.5% GT-known localization on CUB-200-2011 and 74.9% on ILSVRC, remaining competitive with specialized WSOL methods. These results show that part prototypes can provide a unified, intrinsically traceable basis for recognition and localization, establishing prototype-based WSOL as a promising direction for future part-prototype architectures.

1 INTRODUCTION

Deep neural networks have achieved strong performance in large-scale visual recognition (LeCun et al., 2015; Russakovsky et al., 2015) and have been applied to high-stakes domains such as medical imaging and autonomous driving (Esteva et al., 2017; Bojarski et al., 2016). As these models are increasingly deployed in consequential settings, regulatory frameworks such as the EU's GDPR mandate that systems making consequential automated decisions provide meaningful information about their underlying logic (European Parliament and Council, 2016; Goodman & Flaxman, 2017). These concerns have motivated two complementary lines of research: post-hoc explanation methods and evaluations of their reliability (Selvaraju et al., 2017; Adebayo et al., 2018; Hooker et al., 2019), and inherently interpretable models whose reasoning is transparent by construction (Rudin, 2019). Part-prototype networks are a prominent example of the latter, in which classification is formed by comparing image regions against learned prototypical parts (Chen et al., 2019). Part-prototype classifiers expose *which* visual patterns support a prediction and *where* they match the input. Representative part-prototype networks learn prototypes from image-level class supervision and classification-oriented prototype objectives (Chen et al., 2019; Rymarczyk et al., 2022; Ma et al., 2024). Because these objectives do not explicitly distinguish target-object evidence from correlated background, they do not preclude prototypes from exploiting background cues. Such background reliance may not impair classification, but it limits the extent to which prototype responses can serve as reliable localization cues.

Spatial annotations could provide such object–background supervision, but are expensive to collect. Under an ILSVRC annotation protocol, drawing and verifying a bounding box took a median of 34.5 seconds (Papadopoulos et al., 2017); similarly, Open Images V4 provides image-level labels for 19.8K concepts but bounding boxes for only 600 object classes (Kuznetsova et al., 2020). Weakly supervised object localization (WSOL) instead seeks to recover object location using only image-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

level class labels (Choe et al., 2020; Zhang et al., 2022). This gain in scalability, however, comes with a fundamentally weaker training signal: without spatial annotations, the model is optimized primarily for image classification rather than localization. These objectives are not fully aligned, since correct classification can be achieved from any sufficiently predictive cue, including a small object part or correlated background, whereas localization requires recovering the complete spatial extent of the target object while excluding the background. Consequently, classification-based localization maps often concentrate on the most discriminative object regions and overlook less discriminative but still relevant parts (Zhang et al., 2018; Kim et al., 2022). Existing WSOL methods mitigate this issue to some extent through erasure-based strategies that encourage the discovery of complementary object regions (Singh & Lee, 2017; Zhang et al., 2018), background suppression (Wu et al., 2022), separating classification from localization (Zhang et al., 2020; Sabaghi & Oramas, 2026), or leveraging vision-language models Zhao et al. (2023). However, in top-performing methods like TriLite Sabaghi & Oramas (2026) or GenPromp Zhao et al. (2023), classification and localization use dedicated networks or sub-networks, meaning that their localization maps do not reflect classification reasoning.

We address this gap with CLAP, which, to our knowledge, is the first part-prototype WSOL framework in which classification and localization are derived from the same intrinsic prototype responses. To make this relationship explicit, we remove the standard fully connected classifier, whose learned weights can introduce unequal and off-class prototype contributions. Each class logit is instead defined as a unit-weight sum of the peak responses of its class-assigned prototypes, while localization aggregates the corresponding spatial response maps. A lightweight class-agnostic FG/BG head, trained with a C2AM-style contrastive objective (Xie et al., 2022), provides foreground–background pseudo-masks during training. These masks supervise two alignment losses that suppress prototype responses on predicted background and encourage their collective coverage of the foreground. The mask head is discarded at inference, leaving a single prototype pathway for both classification and localization, with fewer parameters and lower latency than the corresponding prototype baselines.

We instantiate CLAP with ProtoPNet and ProtoPool and evaluate it in two complementary settings. On standard part-prototype benchmarks, we show that localization can be added while preserving classification accuracy, and quantify the resulting changes in parameter count and inference latency. On standard WSOL benchmarks, we compare directly with dedicated WSOL methods and demonstrate competitive localization performance.

Contributions.

- We introduce CLAP, a prototype-based WSOL framework in which a class score and its localization map are read from the same class-assigned prototype responses. Thus, both the prediction and its spatial evidence are intrinsically traceable to the prototypes that support them (Section 3.1).
- We introduce an annotation-free alignment objective that uses a lightweight FG/BG head during training to align prototype evidence with object extent. By removing fully-connected layers (FC-free), our design keeps the total parameter count below that of the corresponding prototype baselines even during training; discarding the branch at inference yields a single prototype pathway with reduced parameters and latency (Section 3.2).
- Across part-prototype and WSOL benchmarks, CLAP improves the localization of ProtoPNet and ProtoPool by up to 65.8 pp while preserving classification accuracy and remains competitive with dedicated WSOL methods on CUB-200-2011 and ILSVRC (Sections 5.1 and 5.2).

2 Related Work

Weakly supervised object localization (WSOL). This approach aims to localize objects using only image-level supervision. CAM (Zhou et al., 2016) derives localization maps directly from activations, but these maps often concentrate on the most discriminative object regions. Subsequent methods improve object coverage through spatial masking and erasure (Singh & Lee, 2017; Zhang et al., 2018; Choe & Shim, 2019), background suppression (Wu et al., 2022), class re-activation (Xu et al., 2022), feature alignment (Kim et al., 2022), and transformer-based spatial modeling (Gao

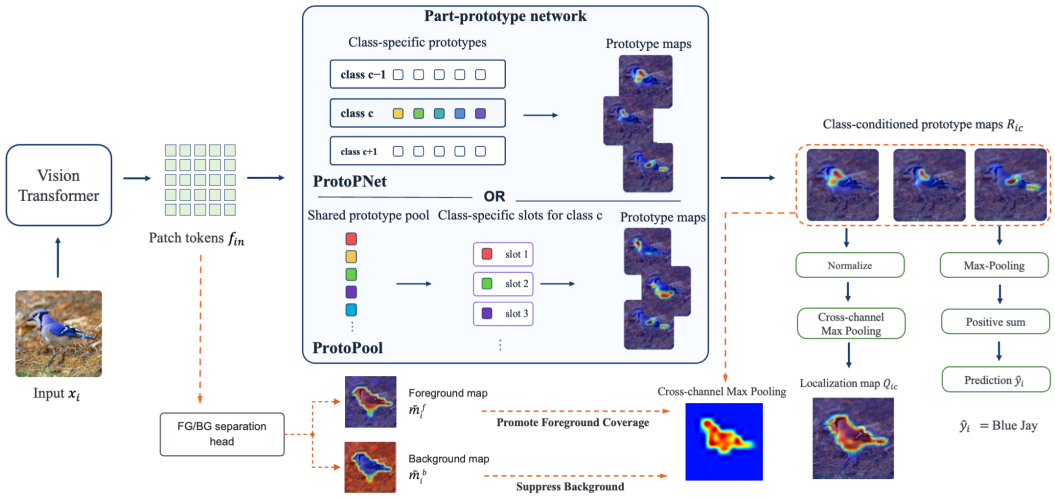


Figure 1: **Overview of CLAP.** The prototype pathway produces class logits and localization maps from the same class-associated response maps. During training, the FG/BG head supplies detached pseudo-masks for background exclusion and foreground coverage; the head is removed at inference.

et al., 2021; Chen et al., 2022; Bai et al., 2022). Another line of work learns class-agnostic objectness or decouples localization from recognition (PSOL (Zhang et al., 2020), C2AM (Xie et al., 2022), and TriLite (Sabaghi & Oramas, 2026)), while GenPromp (Zhao et al., 2023) incorporates a large generative prior. Many strong WSOL methods therefore derive classification and localization from separate networks or specialized subnetworks. Although this separation can substantially improve localization, the resulting bounding box is not directly tied to the image regions used for classification, and the classification evidence does not necessarily determine the predicted localization.

Part-prototype networks. These networks provide spatially grounded classification evidence by comparing image regions with learned prototypical parts. ProtoPNet (Chen et al., 2019) introduced this paradigm by composing class predictions from prototype similarities. Later work organized prototypes into decision trees (Nauta et al., 2021), enabled prototype sharing (Rymarczyk et al., 2021), introduced deformable matching (Donnelly et al., 2022), learned sparse positive decision rules (Nauta et al., 2023), or incorporated prototypes into transformer architectures (Xue et al., 2024; Ma et al., 2024). XProtoNet additionally uses spatial selection maps to refine prototype evidence (Kim et al., 2021). These methods improve the interpretability or selectivity of classification evidence, but do not explicitly encourage prototype responses to cover the object extent required for localization.

CLAP connects these two lines of research by extending part-prototype classification to WSOL while grounding localization in the same prototype evidence that supports the class prediction.

3 Method

3.1 Problem Formulation

Let $\mathcal{D}_{\text{tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{\text{tr}}}$ be the training set with image-level labels $y_i \in \{1, \ldots, C\}$. Given an image, the objective is to predict both its class and a bounding-box covering the corresponding object. Figure 1 shows CLAP, which consists of a prototype pathway shared by classification and localization and a training-only FG/BG head. Given a backbone (DeiT-S or DINOv2-S in our experiments), Phase I trains the lightweight, class-agnostic FG/BG head on frozen features. In Phase II, its pseudo-masks align class-associated prototype responses with the estimated object extent while the prototype pathway learns to classify. The FG/BG head is removed at inference, leaving one prototype pathway for both outputs.

3.2 FG/BG HEAD TRAINING

Image-level labels do not identify whether a spatial response belongs to the object or its context. In Phase I, we freeze the backbone and train a lightweight class-agnostic FG/BG head r_ψ that predicts foreground and background maps $m_i^{\mathrm{f}}(n) \in [0,1]$ and $m_i^{\mathrm{b}}(n) = 1 - m_i^{\mathrm{f}}(n)$. Following C2AM (Xie et al., 2022), we pool backbone features $\mathbf{f}_{in}$ with each predicted map and optimize

$$\begin{aligned} \mathbf{v}_i^r &= \frac{\sum_n m_i^r(n)\mathbf{f}_{in}}{\max\{\epsilon, \sum_n m_i^r(n)\}}, \qquad r \in \{\mathrm{f}, \mathrm{b}\}, \\ \mathcal{L}_{\mathrm{ctr}} &= \lambda_{\mathrm{sep}}\mathcal{L}_{\mathrm{sep}}(\{\mathbf{v}_i^{\mathrm{f}}\}, \{\mathbf{v}_i^{\mathrm{b}}\}) + \lambda_{\mathrm{fg}}\mathcal{L}_{\mathrm{pull}}(\{\mathbf{v}_i^{\mathrm{f}}\}) \\ &\quad + \lambda_{\mathrm{bg}}\mathcal{L}_{\mathrm{pull}}(\{\mathbf{v}_i^{\mathrm{b}}\}). \end{aligned} \tag{1}$$

$\mathcal{L}_{\mathrm{sep}}$ separates foreground and background descriptors, whereas $\mathcal{L}_{\mathrm{pull}}$ promotes within-region consistency. The resulting foreground and background maps serve as pseudo-masks for the Phase II alignment loss.

3.3 PROTOTYPE EVIDENCE AND OBJECT ALIGNMENT

CLAP is an add-on training framework for any part-prototype classifier that exposes spatial prototype response maps and their class associations. For image $\mathbf{x}_i$ and class c, let $R_{ick} \in \mathbb{R}^{H\times W}$ denote the two-dimensional response map of the kth of K class-associated prototypes. These maps correspond directly to class-specific prototypes in ProtoPNet and to class-assigned prototypes selected from ProtoPool's shared pool. CLAP uses the same maps for classification, localization, and object alignment without changing how the underlying architecture computes them. For classification, CLAP applies global max pooling to each response map and rectifies the result, $a_{ick} = [\mathrm{GMP}(R_{ick})]_+$, then forms the class logit as the sum of the corresponding prototype activations, $\ell_{ic} = \sum_{k=1}^{K} a_{ick}$.

Here, GMP takes the maximum over both spatial dimensions. This fixed, bias-free readout replaces the conventional fully connected classifier, removing its parameters and preventing off-class or unequally weighted prototype contributions. For localization, we retain the spatial structure of the same response maps. Let $\overline{R}_{ick}$ denote the per-map min–max normalization of the positive responses in R_{ick}. The class-specific localization map is the elementwise maximum $Q_{ic} = \max_{k=1,\dots,K} \overline{R}_{ick}$, which preserves complementary prototype responses and makes each localized region traceable to a prototype contributing to the class score. For Phase II alignment, we aggregate the unnormalized ground-truth-class responses as $q_i = \max_{k=1,\dots,K} R_{i,y_i,k}$ and retain its positive evidence $q_i^+ = [q_i]_+$. Both maxima are evaluated independently at every spatial position. We align q_i^+ with the pseudo-masks m_i^{b} for background exclusion and m_i^{f} for foreground coverage. Let n index positions on this spatial grid. The alignment loss is as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{align}} &= \lambda_{\mathrm{BG}}\mathcal{L}_{\mathrm{BG}} + \lambda_{\mathrm{FG}}\mathcal{L}_{\mathrm{FG}}, \\ \mathcal{L}_{\mathrm{BG}} &= \frac{1}{B}\sum_{i=1}^{B} \frac{\sum_n m_i^{\mathrm{b}}(n) q_i^+(n)}{\max\{\epsilon, \sum_n m_i^{\mathrm{b}}(n)\}}, \\ \mathcal{L}_{\mathrm{FG}} &= \frac{1}{B}\sum_{i=1}^{B} \frac{\sum_n m_i^{\mathrm{f}}(n)\left[m_i^{\mathrm{f}}(n) - q_i^+(n)\right]_+^2}{\max\{\epsilon, \sum_n m_i^{\mathrm{f}}(n)\}}. \end{aligned} \tag{2}$$

$\mathcal{L}_{\mathrm{BG}}$ suppresses positive class evidence on predicted background. In contrast, $\mathcal{L}_{\mathrm{FG}}$ penalizes only a *shortfall*: it encourages evidence to reach the soft foreground confidence $m_i^{\mathrm{f}}(n)$ but does not suppress stronger evidence. This asymmetry reflects the differing roles of the pseudo-masks: background evidence is always undesirable, whereas low-confidence foreground regions provide only a weak coverage target. Thus, $\mathcal{L}_{\mathrm{FG}}$ never caps a response above $m_i^{\mathrm{f}}(n)$; any suppression there arises from the complementary background term. Because both terms act on the pre-normalization evidence map, they directly align the prototype responses used for classification and localization.

3.4 TRAINING AND INFERENCE

Warm-up. Phase II begins with the backbone and pretrained FG/BG head frozen. We optimize the non-backbone prototype components, including learned class–prototype assignments when present,

using the cross-entropy classification loss $\mathcal{L}_{\mathrm{CE}}$, the underlying model's native prototype-learning objective $\mathcal{L}_{\mathrm{proto}}$ (e.g., clustering and separation), and the alignment loss $\mathcal{L}_{\mathrm{align}}$ (Eq. 2). This establishes discriminative prototype responses under a fixed pseudo-mask signal.

Joint training. We then unfreeze the FG/BG head and selected backbone layers for joint optimization:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{proto}} + \mathcal{L}_{\mathrm{ctr}} + \mathcal{L}_{\mathrm{align}}, \tag{3}$$

where $\mathcal{L}_{\mathrm{align}}$ (Eq. 2) uses the FG/BG pseudo-masks to align prototype evidence. During joint optimization, $\mathcal{L}_{\mathrm{ctr}}$ (Eq. 1) continues to train the FG/BG head as backbone features adapt.

Inference. At inference, the auxiliary FG/BG head is discarded. The model predicts $\hat{y}_i = \arg\max_c \ell_{ic}$ and uses the corresponding prototype-derived heatmap Q_i for localization. Following standard WSOL practice, Q_i is thresholded and the tightest bounding box enclosing the resulting foreground regions is returned as the predicted object location.

4 Experimental setup

4.1 Datasets and Evaluation Protocol

We evaluate CLAP in both the conventional part-prototype classification setting and the weakly supervised object localization setting.

Part-prototype classification. We use CUB-200-2011 (Wah et al., 2011), containing 5,994 training and 5,794 test images from 200 bird species, and Stanford Cars (Krause et al., 2013), containing 8,144 training and 8,041 test images from 196 car categories. To ensure direct comparison with prior part-prototype methods (Chen et al., 2019; Rymarczyk et al., 2022), we follow their object-cropped protocol: training images are cropped using the annotated bounding boxes and augmented offline by approximately $30\times$, while test images are cropped and resized to 224×224.

Weakly supervised object localization. For WSOL, we use the original uncropped images from CUB-200-2011 and ILSVRC (ImageNet-1K). Following the evaluation protocol of Choe et al. (2020), the CUB-200-2011 split comprises 5,994 weakly supervised training images, an additional validation set of 1,000 images with bounding-box annotations for model selection and localization-threshold calibration, and the 5,794 official test images for final evaluation. ILSVRC contains approximately 1.28M training images from 1,000 classes, with its 50K-image validation set used for evaluation. In both datasets, model training uses only image-level class labels.

Evaluation metrics. We report GT-known localization accuracy (*GT-k*) and Top-k localization accuracy for $k \in \{1, 5\}$, following the standard WSOL evaluation convention Choe et al. (2020). A localization is considered correct when the predicted bounding box attains an intersection over union (IoU) of at least 0.5 with a ground-truth box. Top-k Loc additionally requires the ground-truth class to appear among the model's top-k predictions and its corresponding bounding box to satisfy the same IoU criterion.

4.2 Implementation Details

All images are resized to 224×224. We instantiate CLAP with pretrained DeiT-Small (Touvron et al., 2021) or DINOv2-S (Oquab et al., 2024) backbones. Training uses Adam with batch size 128 and learning rate 10^{-3}, except on ILSVRC where we use AdamW with batch size 256 and learning rate 3×10^{-5}. We set $\lambda_{\mathrm{BG}} = 1$ and $\lambda_{\mathrm{FG}} = 1$ for DeiT-S or 0.2 for DINOv2-S. Checkpoints are selected by validation classification accuracy only. Full dataset- and architecture-specific settings are provided in Appendix A.

Table 1: Classification accuracy (Cls) on cropped test images and GT-known localization (GT-k) at IoU 0.5 on uncropped test images from CUB-200-2011 and Stanford Cars. All methods use a DeiT-Small backbone. Inference time is measured on an NVIDIA RTX 4080 Super (batch = 32; 200 runs after 50 warm-up runs).

Method	Params	Inf.	CUB-200-2011		Stanford Cars	
	(M)	(ms/img)	Cls	GT-k	Cls	GT-k
ProtoPool Rymarczyk et al. (2022)	22.6	0.57	83.0	34.2	86.9	62.5
ProtoPNet (Chen et al., 2019)	23.0	0.58	81.0	27.2	89.8	61.9
ProtoPFormer (Xue et al., 2024)	23.7	0.89	83.4	44.0	89.9	61.9
ProtoViT (Ma et al., 2024)	28.3	3.89	**85.3**	34.9	**91.8**	62.7
CLAP–ProtoPool (pool=200, K=10)	22.2	0.53	83.6	**94.0**	86.7	98.4
CLAP–ProtoPNet (K=10)	22.3	0.57	82.9	93.0	87.7	**99.4**

Table 2: Comparison with state-of-the-art WSOL methods on CUB-200-2011 and ILSVRC. For CLAP rows, parameters are listed as CUB-200-2011/ILSVRC.

Method	Backbone	Params (M)	CUB-200-2011			ILSVRC		
			Top-1	Top-5	GT-k	Top-1	Top-5	GT-k
CAM (Zhou et al., 2016)	VGG-16	19.6	41.1	50.7	55.1	42.8	54.9	59.0
TS-CAM (Gao et al., 2021)	DeiT-S	22.1	71.3	83.8	87.7	53.4	64.3	67.6
SCM (Bai et al., 2022)	DeiT-S	22.1	76.4	91.6	96.6	56.1	66.4	68.8
BAS (Zhai et al., 2024)	ResNet-50	25.6	76.8	90.0	95.4	57.5	68.6	72.0
C2AM (Xie et al., 2022)	DN161+EB7	95	81.8	91.1	92.9	59.6	67.1	68.5
GenPromp (Zhao et al., 2023)	SD+EB7	1017	87.0	96.1	98.0	65.2	73.4	75.0
TriLite (Sabaghi & Oramas, 2026)	DeiT-S	22.9	69.2	88.2	95.1	59.2	69.1	71.7
TriLite (Sabaghi & Oramas, 2026)	DINOv2-S	22.9	**87.3**	**96.7**	**98.5**	**65.5**	**75.6**	**77.9**
CLAP–ProtoPool	DeiT-S	22.6/42.4	71.3	86.3	94.2	54.8	65.8	70.4
CLAP–ProtoPNet	DeiT-S	22.7/27.3	74.9	90.4	96.2	59.2	69.6	72.8
CLAP–ProtoPool	DINOv2-S	22.6/42.4	76.5	93.2	96.8	53.7	67.2	71.7
CLAP–ProtoPNet	DINOv2-S	22.7/27.3	82.9	95.0	97.5	59.9	70.5	74.9

5 RESULTS

We validate CLAP along two complementary axes: (i) comparison with existing part-prototype networks, jointly assessing classification and localization (Section 5.1), and (ii) comparison with dedicated WSOL methods (Section 5.2). The first axis examines the localization capacity of existing part-prototype networks and quantifies how CLAP affects their classification accuracy, parameter count, and inference latency. The second axis evaluates whether CLAP is competitive with WSOL-specific methods. In each case we follow the evaluation protocols established by each axis.

5.1 PART-PROTOTYPE NETWORKS

We first check how existing part-prototype networks localize objects without any modification, and then measure how CLAP changes both classification accuracy and localization quality relative to these unmodified baselines. We consider two widely used part-prototype networks, ProtoPNet and ProtoPool. Both were originally proposed with CNN backbones; we re-implement them with a Vision Transformer backbone instead, since the coarser spatial resolution of CNN feature maps constrains achievable localization accuracy, whereas ViT patch tokens preserve finer spatial granularity better suited to part-level localization. Table 1 compares classification accuracy with prototype-based methods on cropped CUB-200-2011 and Stanford Cars test images. *ProtoPool* is our re-run of Rymarczyk et al. (2022) (pool = 200, K=10) under identical conditions, and *ProtoPNet* is our re-implementation of Chen et al. (2019) (K=10); both provide direct classification-only references under the same cropped protocol.

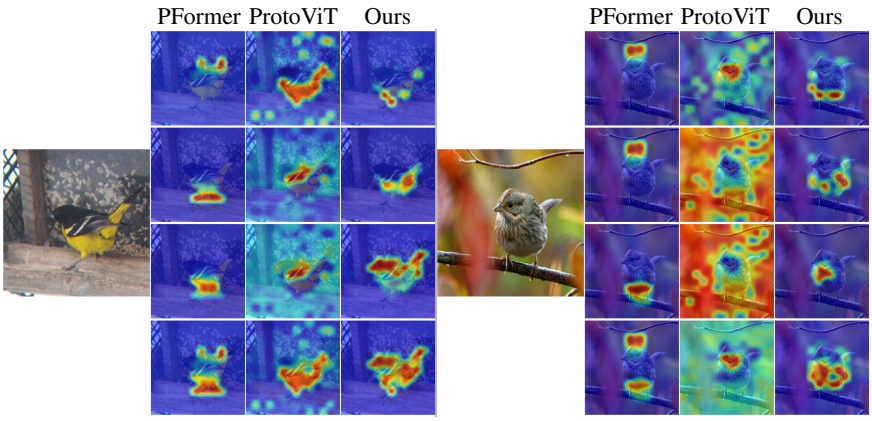
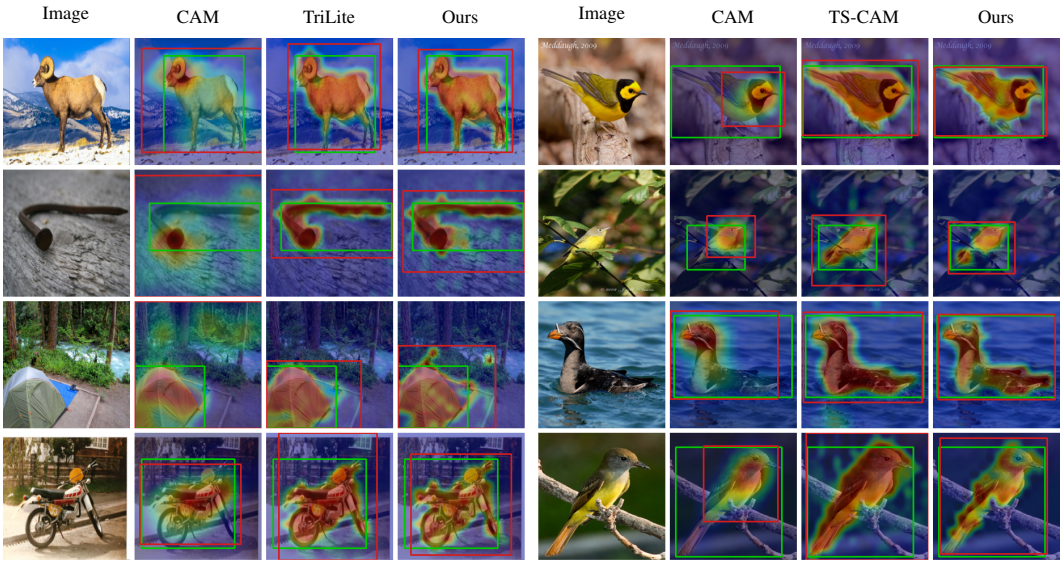


Figure 2: Prototype similarity maps for two CUB-200-2011 test images. Rows 1–3 show the three highest-scoring class-assigned prototype activations; row 4 shows their pixel-wise maximum.



(a) ILSVRC (b) CUB-200-2011

Figure 3: Qualitative WSOL comparison on ILSVRC and CUB-200-2011. Each group shows the input followed by the normalized localization maps from CAM, TriLite (DINOv2-S), and CLAP–ProtoPNet (DINOv2-S).

Table 1 shows that, across the four model–dataset pairs, CLAP improves GT-known localization by 35.9–65.8 percentage points, with classification changes ranging from -2.1 to $+1.9$ points, while maintaining baseline-level latency. We further compare against the more recent part-prototype networks ProtoViT and ProtoPFormer, whose localization performance remains comparable to that of the baseline methods. Figure 2 complements these results by visualizing the response maps of the three most activated prototypes, showing that their evidence often remains concentrated on limited object regions. Together, these results suggest that limited object coverage persists across recent part-prototype approaches.

For ProtoViT, we use the official pretrained weights; for ProtoPFormer, which has no public checkpoint, we train and evaluate using the authors' released code. Since part-prototype models are typically trained on object-cropped images, we evaluate localization on uncropped test images, introducing the same train/test domain shift for all methods. ProtoPFormer is the exception, as it is trained and evaluated on uncropped images and is therefore not directly comparable in this respect.

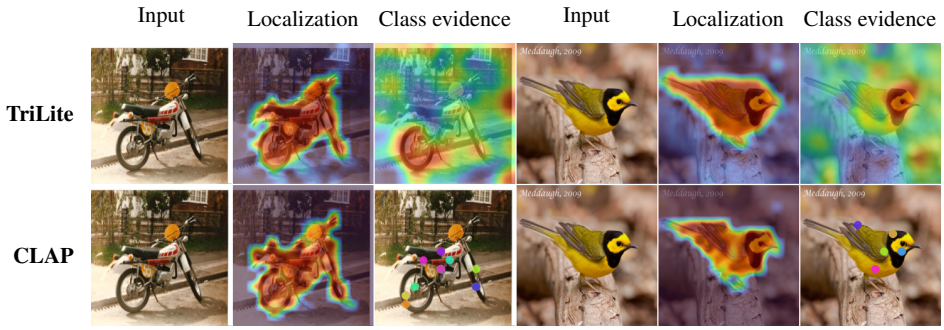


Figure 4: Qualitative comparison of localization and class evidence. Top: TriLite localization and LeGrad evidence for its CLS-token classifier. Bottom: CLAP prototype localization and exact max-pooled prototype sites. Each dot marks the spatial maximum of a target-class prototype.

5.2 Weakly Supervised Object Localization

Table 2 and Figure 3 show that CLAP is competitive with specialized WSOL methods both quantitatively and qualitatively, while deriving classification and localization from the same prototype evidence. Under the matched DeiT-S backbone, CLAP–ProtoPNet improves over the strong TriLite baseline in GT-known localization on both CUB-200-2011 and ILSVRC. Using DINOv2-S further raises GT-known to 97.5% and 74.9%, respectively, bringing CLAP within 0.5 and 0.1 percentage points of GenPromp despite the latter's 1,017M-parameter generative model.

Relative to GenPromp and DINOv2-S TriLite, the shortfall is larger for Top-1 and Top-5 localization than for GT-known. Top-k localization requires both a correct class prediction and successful localization, whereas GT-known evaluates only the localization map for the ground-truth class. The discrepancy therefore suggests that classification errors account for a substantial part of the remaining gap. In CLAP, each class score is computed directly from class-associated prototype responses. This constrained scoring mechanism may sacrifice some recognition capacity relative to a dedicated discriminative classifier, which may partly explain the larger gap in Top-k localization. Results with ProtoPNet and ProtoPool further indicate that the alignment mechanism is not tied to a particular prototype parameterization. Because it requires only spatial prototype response maps and their class associations, it can in principle be applied to newer part-prototype architectures that expose these quantities. The presented results therefore provide empirical evidence that prototype-aligned WSOL is a viable direction, rather than defining its performance ceiling; stronger backbones and more advanced part-prototype architectures offer clear avenues for further gains.

Although TriLite and CLAP yield similar localization maps, Figure 4 shows a key difference in classification evidence. TriLite uses separate classification and localization branches, so spatial evidence for its CLS-token decision requires a post-hoc method such as LeGrad (Bousselham et al., 2025), which can highlight regions different from its localization map. In contrast, CLAP's prototype peaks are the exact patchwise maxima forming the target-class logit, directly coupling classification and localization through the same pathway.

5.3 Ablation Study

We isolate how the foreground–background signal is transferred to the prototype pathway by independently enabling the background-exclusion and foreground-coverage terms in Equation 2. As shown in Table 3, on both datasets removing the alignment loss entirely yields the lowest localization performance, while enabling both terms jointly yields the highest. The relative contribution of each term, however, differs between CUB and ILSVRC. On CUB, which consists predominantly of close-up bird images, background exclusion alone already improves localization by encouraging prototypes to concentrate on the foreground region; since this term does not guarantee full foreground coverage, adding the foreground-coverage term yields a further substantial gain. On ILSVRC, whose images exhibit greater diversity in object scale, background exclusion alone drives most of the improvement, with foreground coverage providing an additional but comparatively

Table 3: Controlled loss ablation for CLAP–ProtoPNet with DeiT-S. The background-exclusion loss $\mathcal{L}_{\mathrm{BG}}$ and foreground-coverage loss $\mathcal{L}_{\mathrm{FG}}$ are enabled independently. ProtoPNet uses $K{=}10$ on CUB-200-2011 and $K{=}20$ on ILSVRC.

		CUB-200-2011			ILSVRC		
$\mathcal{L}_{\mathrm{BG}}$	$\mathcal{L}_{\mathrm{FG}}$	Top-1	Top-5	GT-k	Top-1	Top-5	GT-k
		25.2	29.8	30.5	41.1	47.6	49.5
✓		52.4	62.7	64.9	54.5	64.3	67.2
	✓	52.1	63.1	67.5	44.2	51.3	53.4
✓	✓	74.9	90.4	96.2	59.2	69.6	72.8

smaller gain. These results indicate that although both datasets benefit from the joint alignment loss, the relative reliance on background- versus foreground-driven signals is dataset dependent.

6 Closing Remarks

Summary. We introduced CLAP to show that part-prototype classifiers can be guided toward the object of interest while retaining the discriminative role of their prototypes. By aligning class-associated prototype responses with foreground–background cues, CLAP encourages the learned evidence to cover the target object and avoid contextual background. The same prototype pathway can therefore support both image classification and object localization, rather than treating localization as a separate prediction or a post-hoc explanation. Experiments with ProtoPNet and ProtoPool show that CLAP substantially improves localization while preserving classification performance and remains competitive with specialized WSOL methods. These findings broaden the role of part-prototype networks from classification evidence to a unified basis for recognition and localization. More broadly, CLAP establishes prototype-based WSOL as a promising direction, with newer and more expressive part-prototype architectures offering a clear path toward further advances.

Limitations. Because CLAP builds on an existing part-prototype network, it inherits limitations of the underlying model that are not directly addressed by the proposed alignment objective. In particular, although CLAP employs prototype orthogonality regularization to encourage diversity in the latent space, such regularization does not guarantee that different prototypes capture semantically distinct object parts, and redundant prototypes attending to the same concepts may still persist (Zhu et al., 2025). Consequently, limited semantic diversity or coverage in the base prototype set may still remain in CLAP. Furthermore, the alignment signal relies on class-agnostic pseudo-masks produced by the FG/BG head. Incomplete foreground coverage, background leakage, or activation on salient non-target objects can propagate to the learned prototype responses. This sensitivity may be especially pronounced in scenes containing multiple foreground objects.

Future work. A promising direction is to extend part-prototype-based WSOL to multi-instance localization (Xu et al., 2023). Because prototypes act as localized semantic cue detectors, grouping their spatial responses may help distinguish multiple instances of the same class. Moreover, feature-derived class prototypes have been used for weakly supervised object localization or segmentation and could provide complementary semantic supervision if integrated into part-prototype networks alongside the proposed FG/BG head (Tang et al., 2024). Finally, integrating the framework with newer part-prototype architectures that expose spatial prototype responses and explicit class associations could provide greater discriminative capacity (Xue et al., 2024; Yu & Yang, 2026).

AI use statement

In this work, we used generative AI tools solely to polish author-written text, including grammatical corrections and improvements to clarity. We did not use generative AI tools to develop the method, design or conduct experiments, generate or analyze data, or produce the reported results. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

Reproducibility statement

The principal information needed to reproduce our experiments is summarized in Section 4.2. Appendix A provides the fine-grained dataset- and architecture-specific configurations, including optimization settings, loss weights, prototype capacities, and training schedules. To further facilitate reproducibility, the supplementary material includes the source code and configuration files together with clear instructions for training and evaluating the proposed method.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. Weakly supervised object localization via transformer with implicit spatial calibration. In *Computer Vision – ECCV 2022*, volume 13669 of *Lecture Notes in Computer Science*, pp. 612–628. Springer, 2022. doi: 10.1007/978-3-031-20077-9_36. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136690603.pdf.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. URL https://arxiv.org/abs/1604.07316.

Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. Legrad: An explainability method for vision transformers via feature formation sensitivity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20336–20345, 2025. doi: 10.1109/ICCV51701.2025.01891. URL https://openaccess.thecvf.com/content/ICCV2025/html/Bousselham_LeGrad_An_Explainability_Method_for_Vision_Transformers_via_Feature_Formation_ICCV_2025_paper.html.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html.

Zhiwei Chen, Changan Wang, Yabiao Wang, Guannan Jiang, Yunhang Shen, Ying Tai, Chengjie Wang, Wei Zhang, and Liujuan Cao. LCTR: On awakening the local continuity of transformer for weakly supervised object localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):410–418, 2022. doi: 10.1609/aaai.v36i1.19918. URL https://ojs.aaai.org/index.php/AAAI/article/view/19918.

Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2219–2228, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Choe_Attention-Based_Dropout_Layer_for_Weakly_Supervised_Object_Localization_CVPR_2019_paper.html.

Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3133–3142, 2020. doi: 10.1109/CVPR42600.2020.00320. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Choe_Evaluating_Weakly_Supervised_Object_Localization_Methods_Right_CVPR_2020_paper.html.

Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable ProtoPNet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pp. 10265–10275, 2022. doi: 10.1109/CVPR52688.2022.01002. URL https://openaccess.thecvf.com/content/CVPR2022/html/Donnelly_Deformable_ProtoPNet_An_Interpretable_Image_Classifier_Using_Deformable_Prototypes_CVPR_2022_paper.html.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. doi: 10.1038/nature21056.

European Parliament and Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

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*, pp. 2886–2895, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Gao_TS-CAM_Token_Semantic_Coupled_Attention_Map_for_Weakly_Supervised_Object_ICCV_2021_paper.html.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017. doi: 10.1609/aimag.v38i3.2741.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Eunji Kim, Siwon Kim, Minkyu Seo, and Sungroh Yoon. XProtoNet: Diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15719–15728, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Kim_XProtoNet_Diagnosis_in_Chest_Radiography_With_Global_and_Local_Explanations_CVPR_2021_paper.html.

Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14258–14267, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Kim_Bridging_the_Gap_Between_Classification_and_Localization_for_Weakly_Supervised_CVPR_2022_paper.html.

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*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77. URL https://openaccess.thecvf.com/content_iccv_workshops_2013/W19/html/Krause_3D_Object_Representations_2013_ICCV_paper.html.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. doi: 10.1007/s11263-020-01316-z.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

Chiyu Ma, Jon Donnelly, Wenjun Liu, Soroush Vosoughi, Cynthia Rudin, and Chaofan Chen. Interpretable image classification with adaptive prototype-based vision transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 41447–41493, 2024. doi: 10.52202/

079017-1312. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/48dfc849640344e2d58df0b5bb78c33b-Abstract-Conference.html.

Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14933–14943, 2021. doi: 10.1109/CVPR46437.2021.01469. URL https://openaccess.thecvf.com/content/CVPR2021/html/Nauta_Neural_Prototype_Trees_for_Interpretable_Fine-Grained_Image_Recognition_CVPR_2021_paper.html.

Meike Nauta, Rickard Schouten, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2744–2753, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Nauta_PIP-Net_Patch-Based_Intuitive_Prototypes_for_Interpretable_Image_Classification_CVPR_2023_paper.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4930–4939, 2017. doi: 10.1109/ICCV.2017.528. URL https://openaccess.thecvf.com/content_iccv_2017/html/Papadopoulos_Extreme_Clicking_for_ICCV_2017_paper.html.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

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. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://www.image-net.org/challenges/LSVRC/.

Dawid Rymarczyk, Anke van der Linden, Jacek Tabor, and Bartosz Zieliński. ProtoPShare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1420–1430, 2021. doi: 10.1145/3447548.3467245. URL https://dl.acm.org/doi/10.1145/3447548.3467245.

Dawid Rymarczyk, Łukasz Struski, Michał Górszczak, Koryna Lewandowska, Jacek Tabor, and Bartosz Zieliński. Interpretable image classification with differentiable prototypes assignment. In *Computer Vision – ECCV 2022*, volume 13672 of *Lecture Notes in Computer Science*, pp. 351–368. Springer, 2022. doi: 10.1007/978-3-031-19775-8_21. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/3240_ECCV_2022_paper.php.

Arian Sabaghi and José Oramas. TriLite: Efficient weakly supervised object localization with universal visual features and tri-region disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2602.23120.

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*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Krishna Kumar Singh and Yong 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*, pp. 3524–3533, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Singh_Hide-And-Seek_Forcing_a_ICCV_2017_paper.html.

Feilong Tang, Zhongxing Xu, Zhaojun Qu, Wei Feng, Xingjian Jiang, and Zongyuan Ge. Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3324–3334, 2024. doi: 10.1109/CVPR52733.2024.00320. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tang_Hunting_Attributes_Context_Prototype-Aware_Learning_for_Weakly_Supervised_Semantic_Segmentation_CVPR_2024_paper.html.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Pingyu Wu, Wei Zhai, and Yang Cao. Background activation suppression for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14248–14257, 2022. doi: 10.1109/CVPR52688.2022.01385. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wu_Background_Activation_Suppression_for_Weakly_Supervised_Object_Localization_CVPR_2022_paper.html.

Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2AM: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 989–998, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xie_C2AM_Contrastive_Learning_of_Class-Agnostic_Activation_Map_for_Weakly_Supervised_CVPR_2022_paper.html.

Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. CREAM: Weakly supervised object localization via class RE-activation mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9437–9446, 2022. doi: 10.1109/CVPR52688.2022.00922. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xu_CREAM_Weakly_Supervised_Object_Localization_via_Class_RE-Activation_Mapping_CVPR_2022_paper.html.

Rui Xu, Yong Luo, Han Hu, Bo Du, Jialie Shen, and Yonggang Wen. Rethinking the localization in weakly supervised object localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5484–5494, 2023. doi: 10.1145/3581783.3611959.

Mengqi Xue, Qihan Huang, Haofei Zhang, Jingwen Hu, Jie Song, Mingli Song, and Canghong Jin. ProtoPFormer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1516–1524, 2024. URL https://www.ijcai.org/proceedings/2024/168.

Anni Yu and Yu-Bin Yang. Prism: Prototype-based reasoning with inter-modal semantic mining for interpretable image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2853–2863, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_PRISM_Prototype-based_Reasoning_with_Inter-modal_Semantic_Mining_for_Interpretable_Image_CVPR_2026_paper.html.

Wei Zhai, Pingyu Wu, Kai Zhu, Yang Cao, Feng Wu, and Zheng-Jun Zha. Background activation suppression for weakly supervised object localization and semantic segmentation. *International Journal of Computer Vision*, 132(3):750–775, 2024. doi: 10.1007/s11263-023-01919-2. URL https://link.springer.com/article/10.1007/s11263-023-01919-2.

Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13460–13469, 2020. doi: 10.1109/CVPR42600.2020.01347. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Zhang_Rethinking_the_Route_Towards_Weakly_Supervised_Object_Localization_CVPR_2020_paper.html.

Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5866–5885, 2022. doi: 10.1109/TPAMI.2021.3074313. URL https://arxiv.org/abs/2104.07918.

Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1325–1334, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Adversarial_Complementary_Learning_CVPR_2018_paper.html.

Tao Zhao, Gang Yang, Xin Su, Guannan Li, Hui Ma, Jin Zhang, Haihong Shi, Jieping Ye, and Zheng Ma. Generative prompt model for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6351–6361, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhao_Generative_Prompt_Model_for_Weakly_Supervised_Object_Localization_ICCV_2023_paper.html.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016. doi: 10.1109/CVPR.2016.319. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html.

Zhijie Zhu, Lei Fan, Maurice Pagnucco, and Yang Song. Interpretable image classification via non-parametric part prototype learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9762–9771, 2025. doi: 10.1109/CVPR52734.2025.00912. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhu_Interpretable_Image_Classification_via_Non-parametric_Part_Prototype_Learning_CVPR_2025_paper.html.

A Implementation Details

Shared setup. All images are resized to 224×224. We first initialize the FG/BG head and then train the full model; the Phase II checkpoint with the best validation classification accuracy is used for evaluation.

Optimization. Table 4 gives the configuration-specific Phase II base learning rate, batch size, and epoch schedule. Prototype vectors and add-on layers use three times the base learning rate. During full-backbone fine-tuning, the backbone multiplier is 0.01, except for DINOv2-S on ILSVRC, where it is 0.001 for ProtoPool and 0.003 for ProtoPNet. The FG/BG head uses the base rate, except for DeiT-S on ILSVRC, where it uses 0.1 times the base rate. We use Adam with weight decay 10^{-3}; on ILSVRC, we instead use AdamW with weight decay 10^{-2}. The joint-stage learning rate is decayed by 0.1 every five epochs for object-cropped experiments, by 0.5 every 40 epochs for CUB WSOL, and by 0.5 every 10 epochs for ILSVRC WSOL.

Loss weights and prototype capacity. We use $\lambda_{\mathrm{BG}} = 1$ throughout. For object-cropped experiments, λ_{FG} is 1; for WSOL, it is 1 with DeiT-S and 0.2 with DINOv2-S. On CUB WSOL, ProtoPool uses a pool of 200 prototypes and $K = 10$ class slots, while ProtoPNet uses $K = 10$. On ILSVRC WSOL, ProtoPool uses a pool of 1,000 prototypes and both models use $K = 20$.

Table 4: **Configuration-specific Phase II training hyperparameters.** LR is the base learning rate; BB mult. is the backbone learning-rate multiplier; Epochs reports warm-up/joint epochs.

Setting	Model	Backbone	BB mult.	LR	Batch	Epochs
CUB (cropped)	ProtoPool	DeiT-S	0.01	10^{-3}	128	10/100
CUB (cropped)	ProtoPNet	DeiT-S	0.01	10^{-3}	128	20/60
Cars (cropped)	ProtoPool	DeiT-S	0.01	10^{-3}	128	10/100
Cars (cropped)	ProtoPNet	DeiT-S	0.01	10^{-3}	128	10/100
CUB (WSOL)	ProtoPool/ProtoPNet	DeiT-S	0.01	10^{-3}	128	20/100
CUB (WSOL)	ProtoPool	DINOv2-S	0.01	10^{-4}	128	20/100
CUB (WSOL)	ProtoPNet	DINOv2-S	0.01	3×10^{-4}	128	20/100
ILSVRC (WSOL)	ProtoPool/ProtoPNet	DeiT-S	0.01	3×10^{-5}	256	5/30
ILSVRC (WSOL)	ProtoPool	DINOv2-S	0.001	3×10^{-5}	256	5/30
ILSVRC (WSOL)	ProtoPNet	DINOv2-S	0.003	10^{-5}	256	5/30

B Additional Qualitative Results

Localization maps. Figures 5 and 6 provide additional examples under the same visualization protocol as the main qualitative comparison. Each four-column group contains the input image and the localization maps from CAM, TriLite, and CLAP; green and red boxes denote ground-truth and predicted boxes.

Prototype-level evidence. Figure 7 visualizes the individual target-class prototype responses underlying CLAP's score. For each selected prototype, the dot marks its spatial maximum, i.e., the location that enters the max-pooled prototype activation.

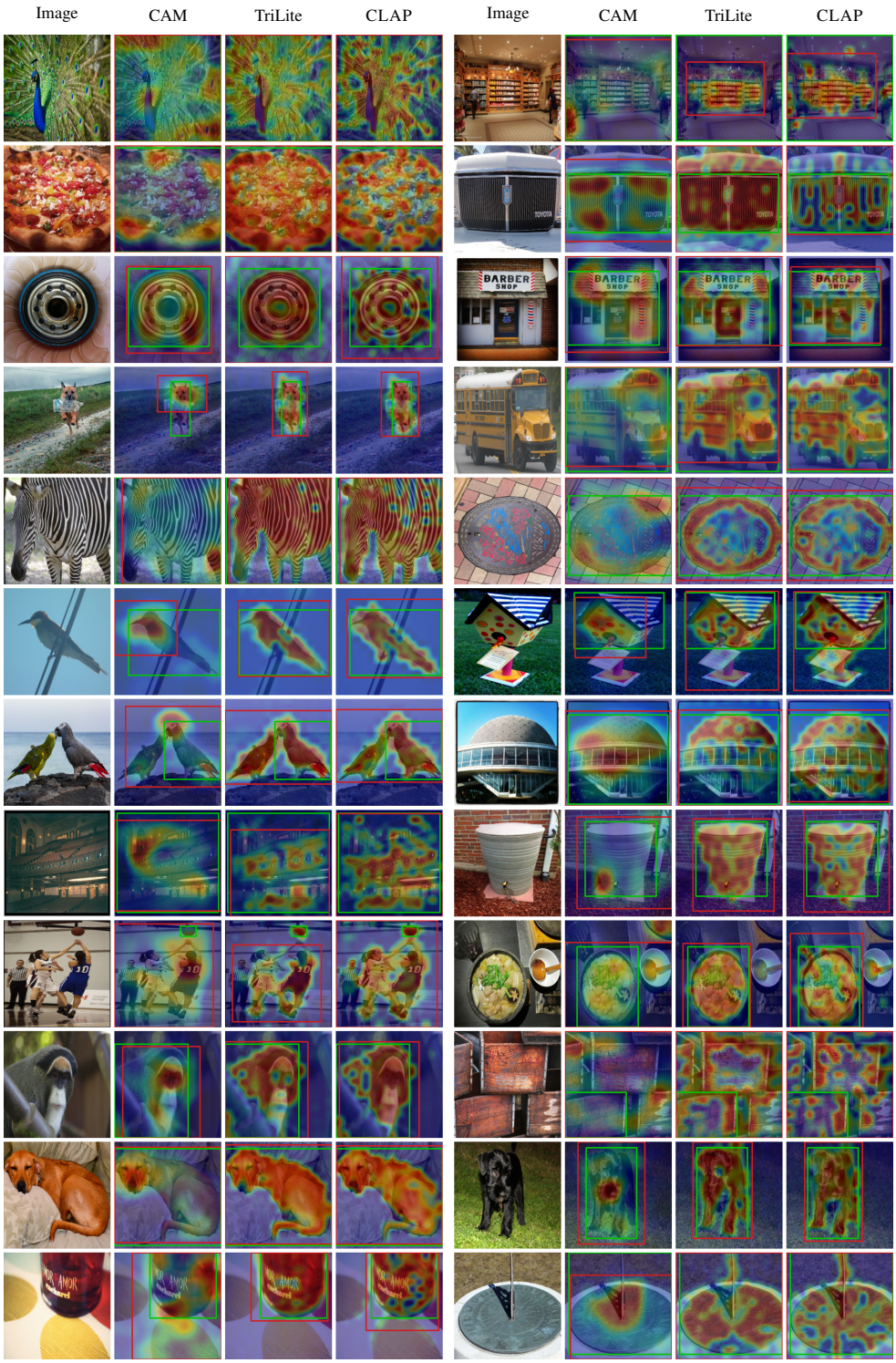


Figure 5: Additional qualitative WSOL results on ILSVRC-1K. Each four-column group shows an input image followed by localization maps from CAM, TriLite, and CLAP. Green and red boxes indicate ground-truth and predicted boxes, respectively.



Figure 6: Additional qualitative WSOL results on CUB-200-2011. Each four-column group shows an input image followed by localization maps from CAM, TriLite, and CLAP. Green and red boxes indicate ground-truth and predicted boxes, respectively.

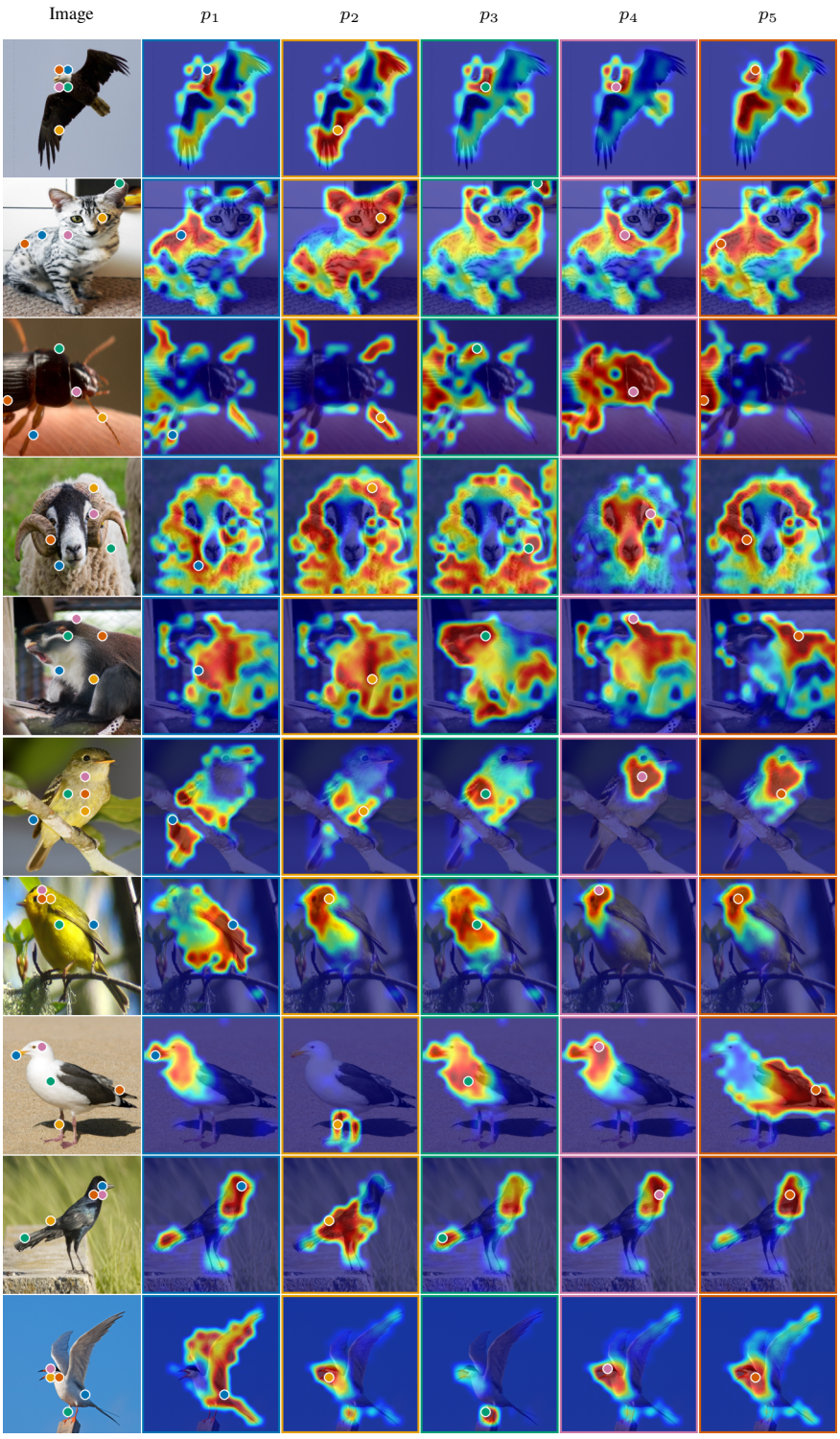


Figure 7: Selected target-class prototype response maps on ILSVRC (top) and CUB-200-2011 (bottom). Dots identify the max-pooled sites.