Label-Agnostic Category Discovery

Supplementary Material

6. Additional Details

6.1. Details of Benchmark Datasets

In our setting, we retrieve semantically aligned labeled subsets $\mathcal{X}_S$ from a large-scale visual lexicon to guide $\ell^\star$-consistent representation learning on unlabeled task data $\mathcal{X}_U$. To comprehensively evaluate the generality and robustness of the proposed approach, we adopt six widely used benchmarks (CUB-200, Oxford Flowers-102, UEC-Food100, Oxford-IIIT Pet, MIT Indoor67, and Caltech-256) as unlabeled domains. These datasets span diverse semantic granularities, domains, and appearance variations, providing a broad testbed for $\ell^\star$-consistent learning. The rationale is as follows:

- CUB-200 (Birds): A canonical fine-grained visual categorization dataset with subtle inter-class differences; ideal for assessing fine-grained semantic alignment and lexicon retrieval.
- Oxford Flowers-102: A plant dataset with substantial intra-class variation in color, shape, and texture; suitable for testing appearance consistency on fine-grained natural categories.
- UEC-Food100: Contains diverse food categories with strong contextual/domain shifts (e.g., utensils, plating, ingredients), challenging domain robustness and semantic retrieval in unstructured scenes.
- Oxford-IIIT Pet: Covers both cats and dogs with hierarchical semantics from coarse (species) to fine (breed), enabling assessment of coarse-to-fine semantic consistency.
- MIT Indoor67: Scene-centric with complex spatial layouts and frequent object co-occurrences, useful for evaluating scene-level semantic alignment and structural understanding.
- Caltech-256: General-purpose object recognition across a broad semantic range (natural and man-made objects), providing a wide semantic space for open-world category discovery.

ImageNet offers a well-curated ontology derived from WordNet, covering over 1,000 object categories with balanced semantics and clear hierarchical relations. Its label taxonomy aligns well with human concept organization, making it an ideal foundation for defining lexicon levels $\ell^\star$ and measuring semantic distances between categories. Moreover, ImageNet images typically depict centered objects with minimal background clutter, facilitating clean and representative class embeddings that serve as reliable anchors for semantic retrieval.

6.2. Details of Evaluation Protocol

To quantitatively assess the performance of our method under the proposed setting, we employ three commonly used clustering evaluation metrics: Clustering Accuracy (Acc), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). All metrics are computed based on the predicted cluster assignments of unlabeled samples and their corresponding ground-truth categories. Since the predicted cluster indices are permutation-invariant, we first determine the optimal one-to-one correspondence between predicted clusters and ground-truth classes using the Hungarian algorithm. Below we provide the definitions of each metric.

Clustering Accuracy (Acc). Accuracy measures the proportion of correctly matched samples between predicted clusters and ground-truth categories after optimal permutation. Formally,

$$\mathrm{Acc} = \frac{1}{n} \sum_{x_i \in \mathcal{X}_S} \mathbb{1}\big\{ y_i = \pi_x\big(\hat{y}_i^K\big) \big\}, \tag{23}$$

where π_x denotes the optimal assignment mapping obtained by the Hungarian algorithm, and K is the ground-truth (or estimated) number of categories. This metric directly reflects the alignment consistency between predicted partitions and true labels.

Normalized Mutual Information (NMI). NMI evaluates the amount of shared information between the predicted clusters C and the ground-truth labels Y. It is defined as:

$$\mathrm{NMI}(Y, C) = \frac{2 \times I(Y; C)}{H(Y) + H(C)}, \tag{24}$$

where $I(Y; C)$ is the mutual information between Y and C, and $H(\cdot)$ denotes the entropy. NMI ranges from 0 to 1, where higher values indicate stronger mutual dependence and better clustering quality. It is insensitive to label permutation and robust to unbalanced category distributions.

Adjusted Rand Index (ARI). ARI measures the similarity between two partitions by considering all pairs of samples and counting how consistently they are assigned to the same or different clusters in both partitions. It adjusts the Rand Index (RI) by accounting for chance grouping:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}, \tag{25}$$

where RI is the Rand Index and $\mathbb{E}[\mathrm{RI}]$ denotes its expected value under random labeling. ARI takes values in $[-1, 1]$, where higher values indicate better agreement between predicted clusters and ground-truth classes after chance correction.

Table 6. Summary of unlabeled task datasets. The mix spans fine-grained objects (birds, flowers, pets), context-rich categories (food), scenes (Indoor67), and generic objects (Caltech-256), enabling a comprehensive evaluation of visual-lexicon retrieval and $\ell^\star$-consistent representation learning across semantic levels and domains.

Dataset	**Domain**	**#Classes**	**#Images**
CUB-200-2011	Birds (FGVC)	200	5,794
Oxford Flowers-102	Flowers (FGVC)	102	1,020
UEC-Food100	Food (objects in context)	100	14,361
Oxford-IIIT Pet	Animals (hierarchical)	37	3,680
MIT Indoor67	Scenes	67	15,620
Caltech-256	Generic objects	257	30,607

Table 7. Clustering performance under different number of labeled categories. The first row reports results using the ground-truth number of categories. From the second row onward, $\mathcal{C}_S$ denotes the labeled category space size, and "Est." indicates the estimated number of clusters.

$\mathcal{C}_S$	CUB-200				Pet				Flower			
	Est.	ACC	NMI	ARI	Est.	ACC	NMI	ARI	Est.	ACC	NMI	ARI
(ground truth)		48.0	0.75	0.37		76.3	0.87	0.72		53.4	0.89	0.54
10	13	6.6	0.49	0.06	16	42.7	0.80	0.47	12	11.8	0.55	0.10
20	21	9.9	0.55	0.10	38	75.0	0.87	0.72	46	42.8	0.79	0.35
30	30	13.8	0.59	0.13	38	75.0	0.87	0.72	47	43.8	0.80	0.36
50	57	23.5	0.66	0.19	53	73.2	0.86	0.71	60	54.9	0.84	0.44
100	106	36.9	0.71	0.29	125	47.7	0.81	0.49	100	78.1	0.90	0.68
150	197	48.2	0.75	0.37	155	42.3	0.79	0.44	170	71.0	0.91	0.69
200	203	48.2	0.75	0.38	178	39.2	0.79	0.41	187	69.0	0.91	0.68
300	335	45.8	0.76	0.36	273	29.5	0.76	0.30	334	53.4	0.89	0.54

Together, these three metrics comprehensively evaluate clustering quality from complementary perspectives and providing a reliable assessment of the proposed method's performance.

7. Additional Experiments

7.1. Empirical Observation for Estimating the Number of Categories

When examining the relationship between the number of the labeled categories $|\mathcal{C}_S|$ and the estimated number of clusters $\tilde{K}$, an interesting and consistent phenomenon emerges across all datasets. As $|\mathcal{C}_S|$ gradually increases from an insufficient to an appropriate scale, the gap $|\tilde{K} - |\mathcal{C}_S||$ typically exhibits a sudden drop. This behavior indicates that the model's internal estimation of category granularity becomes aligned with the external labeled category space, suggesting that $\mathcal{C}_S$ has reached a scale that matches the intrinsic category structure of the data.

In our experiments, this sharp decline in the curve coincides with a noticeable stabilization of clustering performance metrics such as ACC, NMI, and ARI. Before the drop, $\tilde{K}$ tends to lag behind $|\mathcal{C}_S|$, leading to a large inconsistency; the clusters are under-partitioned and coarse, which results in relatively low accuracy and normalized mutual information. After the drop, $\tilde{K}$ quickly approaches $|\mathcal{C}_S|$, and the performance metrics rise rapidly, forming a plateau phase that reflects a stable clustering state. Beyond this point, further enlarging $|\mathcal{C}_S|$ often leads to over-clustering, where the model begins to split coherent categories, causing both the gap and performance measures to fluctuate or decline.

The six datasets studied in this work all exhibit this clear turning pattern. For CUB-200, the gap shrinks abruptly around $|\mathcal{C}_S| = 150$–200, aligning with the point where ACC and NMI are stable. For Pet, the gap minimizes near $|\mathcal{C}_S| = 30$, after which performance slightly drops due to over-clustering. For Flower, the gap reaches zero at $|\mathcal{C}_S| \approx 100$, corresponding to the ground-truth category number. Food shows a smaller but visible decline around $|\mathcal{C}_S| = 50$–100, where NMI stabilizes near 0.6. Caltech256 displays a wide plateau between $|\mathcal{C}_S| = 276$ and 305, suggesting that the estimated and labeled category spaces become consistent in this range. For Indoor, a similar align-

Table 8. Clustering performance under different labeled category space sizes $\mathcal{C}_S$ on the Food, Caltech256, and Indoor datasets. The first row reports results using the ground-truth number of categories; from the second row, "Est." is the estimated number of clusters for each dataset.

$\mathcal{C}_S$	CUB-200				Pet				Flower			
	Est.	ACC	NMI	ARI	Est.	ACC	NMI	ARI	Est.	ACC	NMI	ARI
(ground truth)		44.9	0.60	0.27		70.5	0.84	0.59		52.1	0.66	0.40
10	11	13.0	0.35	0.08	15	15.0	0.48	0.07	16	34.2	0.52	0.24
20	15	16.5	0.40	0.11	44	26.6	0.63	0.21	29	46.7	0.60	0.35
30	47	29.9	0.52	0.21	75	36.3	0.70	0.32	38	51.4	0.63	0.38
50	81	41.7	0.58	0.26	107	44.6	0.75	0.39	71	51.6	0.66	0.39
100	165	43.5	0.62	0.27	209	66.5	0.82	0.57	97	48.7	0.67	0.37
150	183	42.4	0.62	0.26	208	66.5	0.82	0.57	156	41.0	0.67	0.30
200	198	41.3	0.62	0.25	276	70.7	0.84	0.58	160	40.6	0.67	0.30
300	349	32.8	0.63	0.20	305	70.3	0.84	0.58	402	22.3	0.64	0.16

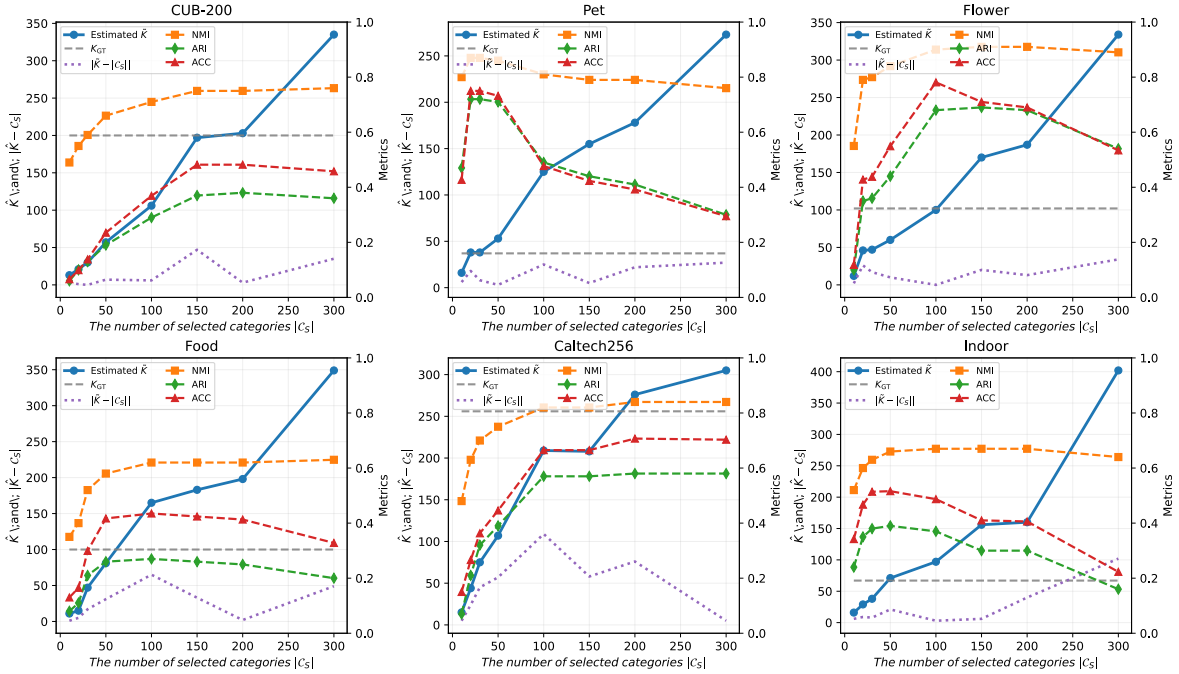

Figure 5. The relationship between the number of selected known classes and the estimated number of categories. The estimated number of categories generally shows a positive correlation with the number of selected known categories. However, as the estimated number of categories approaches the ground truth, it becomes largely unaffected by the number of selected known categories.

ment occurs near $|\mathcal{C}_S| = 100$, where both $|\hat{K} - |\mathcal{C}_S||$ and performance metrics stabilize.

These observations demonstrate that the sudden reduction in the gap provides an intuitive and reliable signal for identifying the appropriate labeled category space size. Visually, this point corresponds to the moment when the blue curve ($\hat{K}$) converges toward the gray reference line ($|\mathcal{C}_S|$ or K_{GT}) in the plots, while the colored performance curves (ACC, NMI, ARI) begin to flatten. Therefore, the turning point of $|\hat{K} - |\mathcal{C}_S||$ serves as a direct and data-driven indicator of a suitable $\mathcal{C}$ without requiring any prior knowledge of the true category number.

7.2. Fine-grained Analysis of Novel-class Estimation

In the main experiments, we already observed a characteristic trend: as the number of labeled classes increases, our estimate of the number of novel classes first decreases rapidly

Table 9. Effect of warm-up epochs on performance.

epochs	CUB-200			Pet		
	ACC	NMI	ARI	ACC	NMI	ARI
0	49.8	0.776	0.40	86.5	0.90	0.80
1	51.3	0.760	0.40	87.0	0.90	0.81
5	51.2	0.760	0.40	87.3	0.90	0.80
10	52.0	0.760	0.40	86.8	0.90	0.81

Table 10. Effect of labeled-batch size on performance. Unlabeled batch size is fixed at 128.

L_{batch}	CUB-200			Pet		
	ACC	NMI	ARI	ACC	NMI	ARI
32	47.5	0.74	0.37	82.6	0.89	0.78
64	51.1	0.76	0.40	84.6	0.90	0.79
128	52.0	0.76	0.40	86.8	0.90	0.81
256	52.0	0.76	0.41	86.3	0.90	0.80

Table 11. Effect of temperature τ on performance.

τ	CUB-200			Pet		
	ACC	NMI	ARI	ACC	NMI	ARI
0.7	42.4	0.71	0.31	83.5	0.87	0.75
0.4	52.2	0.76	0.41	86.1	0.90	0.80
Linear	52.0	0.76	0.41	86.3	0.90	0.80

Table 12. Fine-grained study of the relationship between the number of labeled classes and the estimated number of novel classes on the *Flower* dataset. The ground-truth number of novel classes is 102.

# Labeled classes	Estimated novel classes	ACC	NMI	ARI
80	111	78.0	0.91	0.70
90	123	76.9	0.91	0.70
100	100	78.1	0.90	0.68
110	129	76.3	0.91	0.70
120	137	75.9	0.91	0.71

and then stabilizes around the ground-truth value. However, in those settings the step size of the labeled-class count was relatively coarse, which makes it difficult to see how the estimate behaves in the neighborhood of a "good" labeled-class configuration. To obtain a more fine-grained view, we further conduct a dedicated study on the *Flower* dataset, where we densely vary the number of labeled classes around the promising region identified before.

Table 12 reports the results. The first column lists the number of labeled classes used as anchors, and the second column shows the corresponding estimate of the number of novel classes. The last three columns give the clustering ACC, NMI, and ARI. We observe that when moving from 80 to 90 labeled classes, the estimated number of novel classes is still clearly over-counted (111 and 123 versus the ground-truth 102). Once the number of labeled classes reaches 100, the estimate drops sharply and becomes very close to the true value: our method predicts 100 novel classes, which is within about 2% of the ground-truth 102. Further increasing the number of labeled classes to 110 and 120 again leads to over-estimation (129 and 137 classes), while the clustering metrics change only mildly. This fine-grained experiment confirms that there exists a suitable regime of labeled-class coverage in which the semantic structure provided by labeled anchors is just rich enough to support accurate discovery and counting of novel classes, and that our estimator is particularly stable and precise in this regime.

8. Hyperparameter Sensitivity Analysis

To evaluate the robustness of our method with respect to key training hyperparameters, we conducted a series of ablation studies on CUB-200 and Oxford-Pet datasets.

Warm-up Epochs. Tab. 9 shows the effect of varying the warm-up epochs from 0 to 10. A moderate warm-up (1–5 epochs) leads to stable improvements on both datasets, suggesting that a short warm-up phase helps the model stabilize contrastive learning and avoid noisy gradients at early stages. Further increasing the warm-up duration yields only marginal gains, indicating convergence stability.

Batch Size of Retrieved Labeled Data. In Tab. 10, we fix the unlabeled batch size at 128 and vary retrieved batch size from 32 to 256. Performance improves notably from 32 to 128, demonstrating that using a sufficiently large labeled batch enhances the diversity of retrieved supervision signals. Beyond 128, the results saturate, implying that our method is not overly sensitive to this parameter.

Temperature Coefficient. Tab. 11 presents results with fixed $\tau = 0.7$ and 0.4, as well as a linear-decay schedule. A lower temperature ($\tau = 0.4$) consistently yields better clustering accuracy and alignment, highlighting that stronger sharpening of similarities improves representation discrimination. The linear schedule performs comparably, showing that gradual annealing achieves a good balance between stability and sharpness.

Overall, these experiments confirm that our approach is robust to hyperparameter variations within reasonable ranges and does not rely on fine-tuned settings for consistent performance.

9. Effect of Visual Lexicon Granularity

To analyze the sensitivity of our framework to the granularity of the ImageNet/WordNet label space used as a visual lexicon, we compare two ways of constructing category prototypes. In the first setting (denoted as $\mathrm{g}^{\star}-1$), we map each of the 1,000 ImageNet leaf categories to its

Table 13. Impact of ImageNet label granularity on the visual lexicon (I). $g^\star-1$ uses the WordNet parent categories of ImageNet-1K classes (530 prototypes), while $g^\star$ uses the original 1K leaf categories as visual anchors.

	CUB-200			Pet			Flower		
granularity	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
$g^\star-1$	47.6	0.74	0.37	76.0	0.85	0.69	80.6	0.91	0.73
$g^\star$	52.0	0.76	0.40	86.8	0.90	0.81	83.6	0.92	0.74

Table 14. Impact of ImageNet/WordNet label granularity on the visual lexicon (II). Notation is the same as in Tab. 13.

	Food			Caltech256			Indoor		
granularity	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
$g^\star-1$	39.1	0.57	0.24	67.2	0.82	0.53	50.7	0.66	0.40
$g^\star$	40.0	0.58	0.25	70.5	0.84	0.58	52.4	0.68	0.41

parent node in WordNet, which results in a coarser lexicon of about 530 parent categories that are used as visual anchors for matching unlabeled samples. In the second setting (denoted as $g^\star$), we directly use the original 1,000 ImageNet leaf categories as visual anchors, providing a finer-grained set of prototypes. The quantitative results on six benchmarks are summarized in Tables 13 and 14. Across all datasets, using the full 1K leaf-category lexicon leads to consistently better performance than using the merged parent-category lexicon, with particularly notable gains on fine-grained datasets such as Pet and Caltech256 in terms of both ACC and ARI. These observations indicate that our framework can effectively exploit fine-grained semantic structure in the visual dictionary, and that collapsing ImageNet categories to higher-level WordNet parents tends to discard discriminative information and slightly degrades the quality of the discovered clusters.

10. Qualitative Results

In our visualization analysis, we first retrieve a set of semantically similar images from ImageNet according to the given unlabeled samples. By performing supervised learning on these retrieved images, the model learns to correctly recognize the underlying semantic categories related to the unlabeled data. Subsequently, even without ground-truth labels, our method enables the model to accurately localize the relevant objects in the unlabeled domain. Notably, we observe that the pretrained model tends to represent the entire image, often being distracted by complex backgrounds, shadows, or irrelevant textures. After introducing the supervised learning on the retrieved similar categories, the model not only learns to focus on the primary objects in labeled images but also transfers this attention to the unlabeled samples, highlighting the target objects more precisely, as shown in Fig. 6.

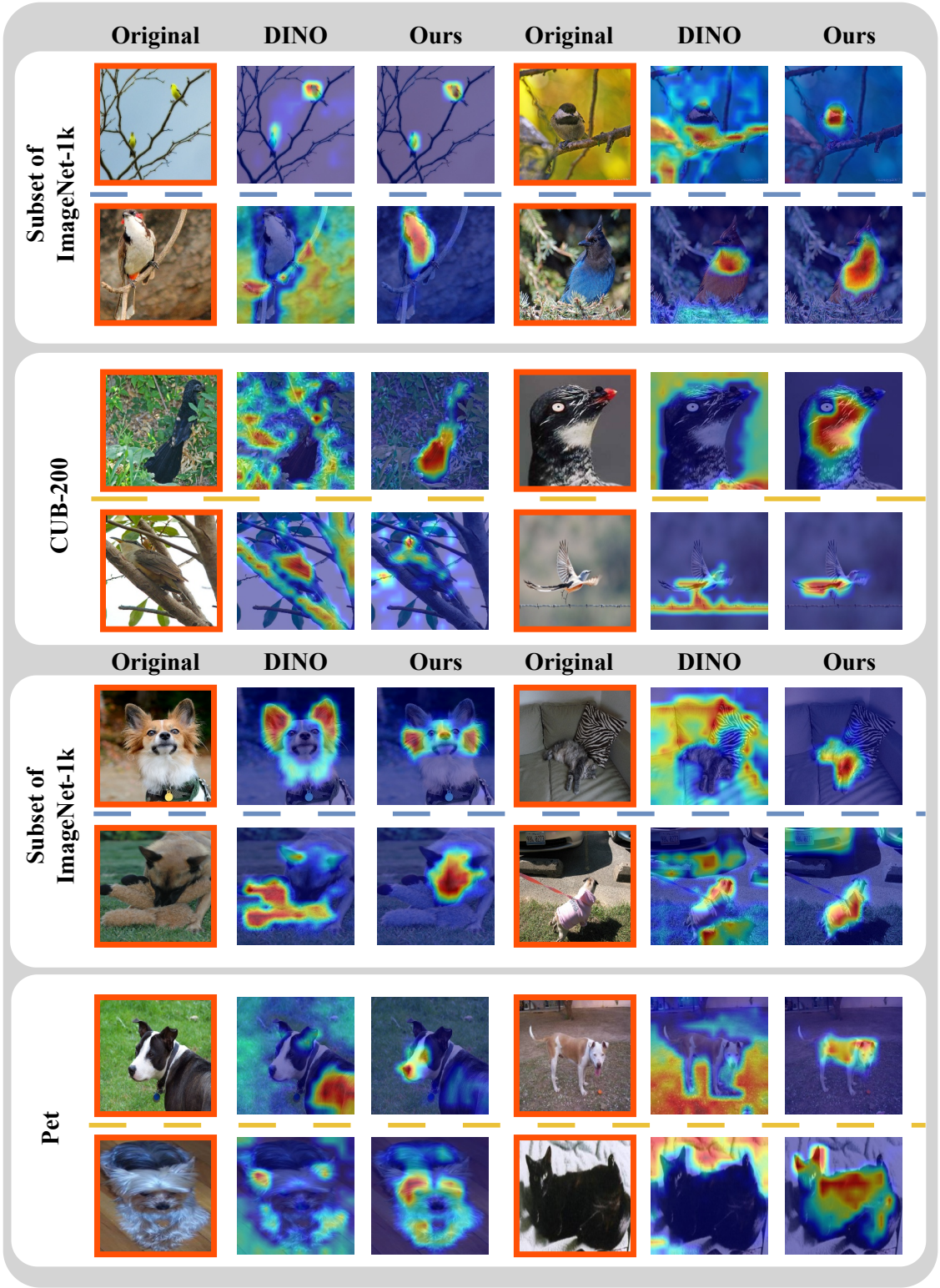

Figure 6. Visualization results using ScoreCAM to highlight the most discriminative regions that contributed to the model's decision. Warmer colors indicate areas with stronger activation.