Subtoken Image Transformer (SiT) for Generalized Category Discovery

Anonymous ICCV submission

Paper ID

Abstract

Generalized Category Discovery (GCD) aims to classify known and novel categories by learning discriminative features that capture fine-grained inter-class distinctions while remaining invariant to irrelevant variations. Existing methods propose loss functions or use pseudo-labels for better clustering and alignment but overlook the impact of image tokenization in Vision Transformer (ViT). We hypothesize that a network should prioritize computational resources for discriminative regions to enhance feature representation. We propose Subtoken Image Transformer (SiT), which enhances ViT tokenization by dividing tokens into subtokens in discriminative regions, enabling finer granularity and improved feature representation. SiT is fine-tuned from a pretrained ViT, leveraging attention-based random region proposal during training, while a separate selection network identifies key regions during inference. Experiments on fine-grained GCD benchmarks (CUB, FGVC-Aircraft, and Stanford Cars) and coarse-grained GCD benchmarks (CIFAR-10 and ImageNet-100) demonstrate SiT's superiority over state-of-the-art methods, revealing semantically critical patches essential for fine-grained discrimination. The code will be publicly released upon publication.

1. Introduction

Fine-grained classification addresses the critical challenge of distinguishing highly similar subcategories within a broader class, such as identifying bird species [49], vehicle models [25], or plant species [45]. However, traditional fine-grained classification heavily relies on labeled datasets, which is labor-intensive and time-consuming. Hence, Vaze *et al*. [46] propose the Generalize Category Discovery (GCD) task that learns from labeled fine-grained data and discovers novel categories in unlabeled data without prior knowledge of the number of subcategories of new groups. GCD requires learning discriminative representations that can effectively capture subtle inter-class differences in labeled data while also generalizing to unlabeled data. Existing methods try to address the task either by optimizing learning objective function [10, 33, 38, 47, 48, 52, 55] or generating weak pseudo-labels through sample relationships [14, 18, 35, 37, 38, 50, 54, 65, 65, 67].

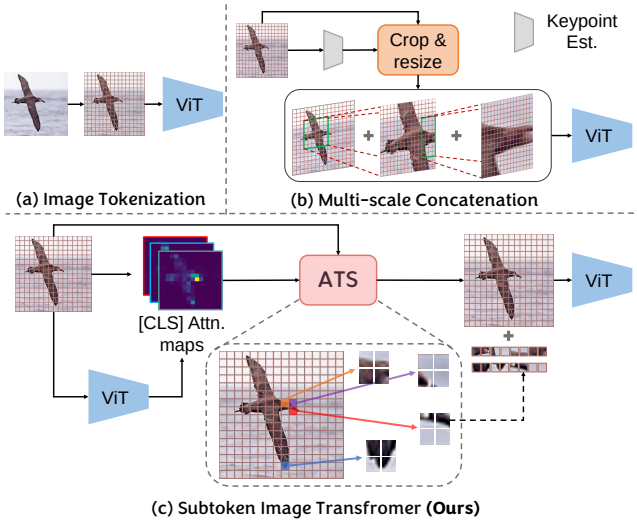

Figure 1. Motivation of SiT. Compared with multi-scale input that crops and resizes the predicted bounding box regions, SiT relies on the attention map for localizing fine-grained details on generic objects. Instead of region-level resizing, SiT performs an Attention-based Token-level Subdivision (ATS) that reduces ineffective tokens, focusing more on fine-grained discriminative regions.

However, previous methods overlook the impact of image tokenization in Vision Transformer (ViT), which plays a crucial role in determining how visual information is processed and represented. ViT's default square patch tokenization allocates equal computational resources to all image regions, including those irrelevant to distinguishing fine-grained categories. Also, small regions lead to small number of tokens while fine-grained classification often requires focusing on minute details (such as subtle shape differences in a bird's leg), which may not be adequately captured by a uniform patch-based tokenization. This limitation prevents models from effectively attending to and modeling small yet critical regions that are essential for distinguishing visually similar subcategories.

Motivated by this limitation, we propose a novel tokenization strategy that dynamically allocates more tokens to the most important regions, ensuring finer granularity in areas crucial for fine-grained discrimination. Instead of treating all image regions equally, our approach adaptively subdivides tokens in discriminative regions, allowing the model to capture subtle yet critical details while reducing redundancy in less informative areas. However, a key challenge lies in determining which regions are important for the task without explicit location supervision.

To tackle this challenge, we propose Subtoken Image Transformer (SiT), a novel tokenization method that dynamically refines token representation by allocating more tokens to discriminative regions, ensuring finer granularity where needed. An alternative approach would involve using auxiliary bounding boxes, *e.g.*, detecting the head of a bird, to crop and create multi-scale inputs, which is restricted to human knowledge. In contrast, SiT selectively increases resolution only in semantically important tokens, enhancing feature representation while maintaining flexibility in capturing complex/irregular regions. A comparison with existing methods is shown in Fig. 1.

Specifically, SiT first fine-tunes a pretrained ViT, and then leverages Attention-based Token Subdivision (ATS) to refine tokenization dynamically. During training, SiT utilizes attention maps to identify and randomly propose highly probable discriminative regions for token subdivision, allowing the model to learn fine-grained representations while maintaining generalization across varying subdivisions. This process enables ViT to adapt to subdivided tokens at inference time. During inference, instead of random subdivision, SiT employs an auxiliary selection network that deterministically identifies key regions for token division.

In summary, we make four contributions:

- We introduce **SiT (Subtoken Image Transformer)**, a novel tokenization strategy that dynamically refines token representation by allocating more tokens to discriminative regions.
- We propose **Attention-based Token Subdivision**, which uses attention maps from ViT to probabilistically sample token division locations during training. We also introduce **Selection Network** during testing that deterministically identifies key regions for token division.
- Extensive experiments on multiple fine-grained datasets demonstrate that **SiT significantly enhances generalized category discovery**, outperforming existing methods in distinguishing visually similar subcategories.

2. Related Works

Dynamic Token Scaling. Vision Transformers (ViTs), exemplified by CLIP [36], SigLIP [62], and the DINO family [6, 32], partition images into non-overlapping tokens and leverage self-attention for feature extraction, contrasting with CNN-based convolutional feature hierarchies. Recent approaches address visual feature extraction through distinct architectural strategies. LLaVA-NeXT [28] proposes grid-based image decomposition for visual-language tasks, processing subregions independently to enhance local feature extraction. Swin Transformer [29] and Pyramid vision transformer [51] employ multi-scale feature aggregation through shifted windowing, though they introduce computational redundancy while overlooking intra-image heterogeneity. These methods demonstrate the importance of spatial adaptation but lack dynamic resource allocation mechanisms for region-specific processing demands. In this work, we propose an Attention-based Token Subdivision (ATS), dynamically subdividing tokens to amplify discriminative features while maintaining pretrained ViT efficiency.

Fine-grained Localization. Fine-grained localization pinpoints discriminative features critical for inter-class differentiation, primarily via keypoint detection and class activation maps (CAMs). Keypoint detection identifies semantic landmarks to encode structural attributes, widely applied in facial alignment [3, 27, 42, 63], human pose estimation [5, 15, 58], and object analysis (e.g., vehicles [40, 60, 61]). These keypoints enhance tasks like face and gait recognition [24, 44] by providing complementary geometric priors. However, defining consistent keypoints for generic objects (e.g., industrial produ) remains challenging due to the lack of anatomical priors, risking overemphasis on unimportant features. CAM-based methods [11, 22, 70] generate class-specific saliency maps to identify crucial regions, but are limited in propagating them into feature refinement. To bridge this gap, we propose an attention-driven token selection method that identifies and amplifies discriminative tokens, optimizing fine-grained classification without relying on predefined keypoints.

Generalized Category Discovery (GCD). GCD tackles the task of jointly recognizing known classes and discovering novel categories in unlabeled data, as formalized by Vaze *et al*. [46] and Cao *et al*. [4]. Unlike semi-supervised learning (SSL), which assumes unlabeled data belong to labeled classes [7, 31, 34, 39, 59], GCD extends novel category discovery (NCD) [17, 19, 20, 66, 68, 69] by requiring models to simultaneously leverage labeled data and partition unlabeled novel subcategories. This creates a fine-grained benchmark demanding precise separation of subtle inter-class distinctions. Current approaches follow two paradigms: (1) Prototype-based methods, which align features using class anchors for both known and novel categories [1, 9, 10, 21, 23, 43, 53, 54, 56, 57, 64]; and (2) Local similarity-based methods, which cluster samples or generate pseudo-labels via pairwise relations [14, 18, 35, 37, 38, 47, 50, 54, 65, 65, 67]. However, these works priori-

tize optimizing clustering metrics or loss functions, neglecting to analyze how intrinsic visual saliency (*e.g.*, discriminative regions) impacts GCD performance. We revisit this problem by investigating: (1) Which object regions/tokens are pivotal for fine-grained differentiation? and (2) Can targeted token subdivision enhance feature discernibility?

3. Methods

3.1. Preliminary

Image Tokenization. Given an image tensor $x \in \mathbb{R}^{C\times H\times W}$, ViTs partition x into $N = \frac{H}{P} \times \frac{W}{P}$ non-overlapping patches, where P is the patch size. All patches are stacked into an image patch sequence $I \in \mathbb{R}^{N\times C\times P\times P}$. Then I is flattened in its last three dimensions, and a learnable projection matrix $\mathbf{W_p} \in \mathbb{R}^{CP^2\times D}$ (*i.e.*, patch embedding layer) transforms I into patch embeddings, combined with positional embeddings $E_{\text{pos}} \in \mathbb{R}^{N\times D}$ to generate the input token sequence $z_0 \in \mathbb{R}^{N\times D}$ for the transformer:

$$z_0 = \mathbf{W_p} I + E_{\text{pos}}. \tag{1}$$

Attention Map. For input tokens $\mathbf{z}_l \in \mathbb{R}^{N\times D}$ at transformer block l, the self-attention mechanism computes queries $\mathbf{Q} = \mathbf{z}_l\mathbf{W}_Q$, keys $\mathbf{K} = \mathbf{z}_l\mathbf{W}_K$, and values $\mathbf{V} = \mathbf{z}_l\mathbf{W}_V$, where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D\times d_k}$ are learnable weights. The attention map $\mathbf{A}_l \in \mathbb{R}^{N\times N}$ derived as:

$$\mathbf{A}_l = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right). \tag{2}$$

Conventionally, the attention map from the CLS token $\mathbf{A}_{l,\text{CLS}} \in \mathbb{R}^N$ represents the global attention for the model. Each element $\mathbf{A}_{l,\text{CLS}}(i,j)$ quantifies how many tokens i attend to token j, reflecting the model's focus on discriminative regions (*e.g.*, object parts). Higher values indicate stronger semantic relevance between tokens. ViTs exhibit hierarchical attention patterns across layers, with deeper blocks increasingly focusing on semantically fine-grained details, while shallow layers prioritize on low-level textures or global context [8, 13].

For Multi-Head Self-Attention (MHSA), attention maps are gathered from all heads $\hat{\mathbf{A}}_{l,\text{CLS}} \in \mathbb{R}^{N_{\text{head}}\times N}$ where N_{head} is the number of head. We illustrate how we select the most important attention map in Sec. 3.3.

3.2. Attention-based Token Subdivision (ATS)

Defining consistent keypoints for generic objects remains challenging due to the lack of anatomical priors, risking overemphasis on non-discriminative features. We introduce ATS to leverage knowledge from attention maps.

Image Subdivision. As shown in Fig. 2, given an image patch sequence I and its attention map $\mathbf{A}^h_{l,\text{CLS}}$ from head h,

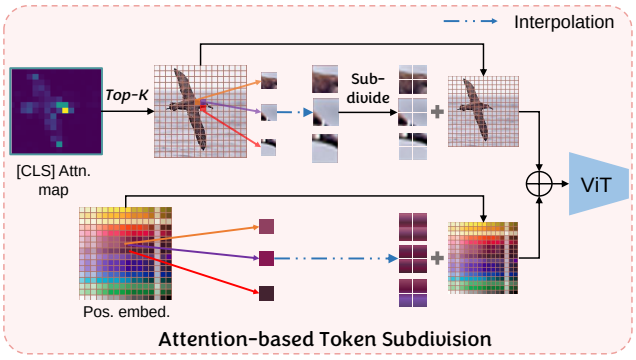

Figure 2. Overview of ATS. We select the top-K patches based on the attention map for subdivision. We select the neighboring positional embedding for interpolation purpose.

we locate the top K patch set $T \in \mathbb{R}^{K\times(C\times P\times P)}$ sorted by the attention scores. For each patch, we generate $f \times f$ number of subtokens. We first resize each patch by

$$\hat{T}_i = \text{Upscale}(T_i, \text{factor} = f), \quad \hat{T}_i \in \mathbb{R}^{C\times(fP)\times(fP)}. \tag{3}$$

Then we divide the interpolated patch into smaller patches for all $m, n \in \{0, \ldots, f-1\}$ in an $f \times f$ grid. Specifically, one subtoken is defined as:

$$\hat{T}_i^{(m,n)} = \hat{T}_i[:, mP:(m+1)P, nP:(n+1)P], \tag{4}$$

where all subtokens are denoted as:

$$T' = \bigcup_{i=1}^{K}\bigcup_{m=0}^{f-1}\bigcup_{n=0}^{f-1} \hat{T}_i^{(m,n)} \in \mathbb{R}^{Kf^2\times(C\times P\times P)}. \tag{5}$$

Put it simply, subtokens are created by upscaling and dividing each patch. T' concatenates with I to perform input patch sequence to patch embedding layer:

$$I^{in} = \text{Concat}(I, T'). \tag{6}$$

Note that we do not drop T after obtaining T'. We believe that T and T' can provide multi-scale information for fine-grained classification. The effect of dropping T can be found in Sec. 4.3.

Positional Embedding Interpolation. Positional embeddings (PE) are also subdivided in a similar fashion to image patches. However, while image patches can be upscaled due to their spatial resolution, a positional embedding at a particular location is a vector and cannot be directly upscaled without remaining unchanged across the newly introduced positions.

To address this, we expand each positional embedding into a spatial structure by using its neighboring positional embeddings. Given an original positional embedding $E_{\text{pos},i}$,

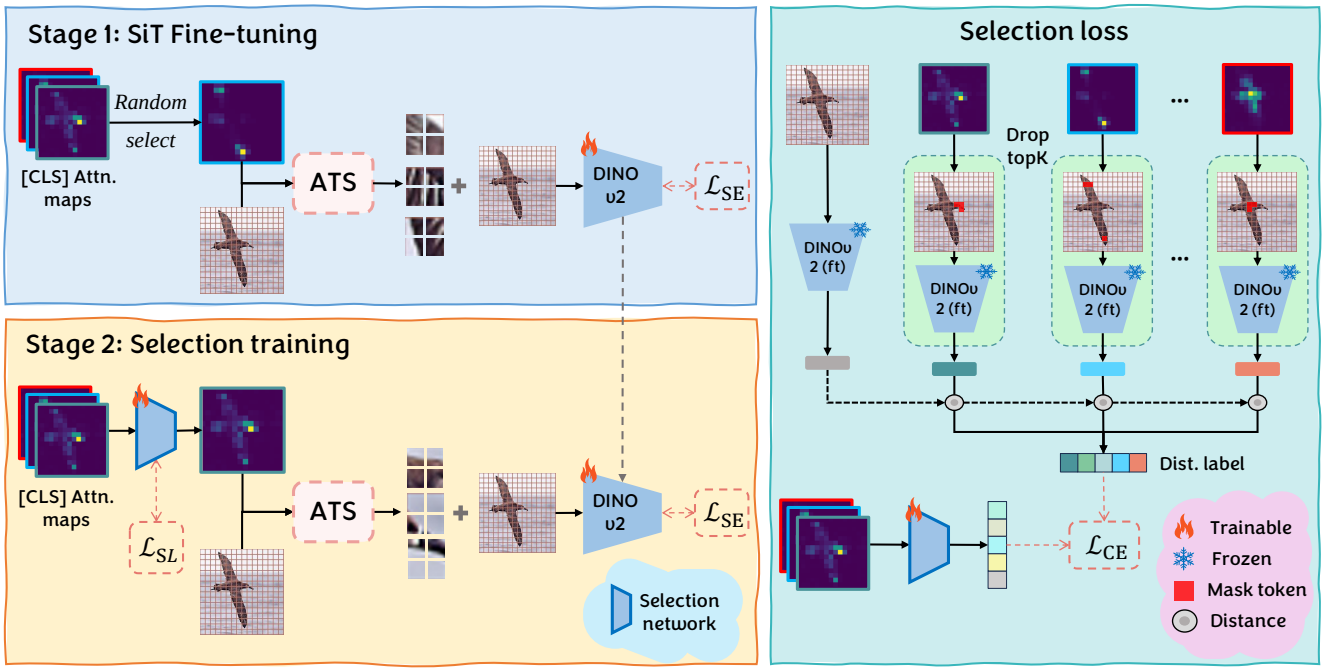

Figure 3. Two-Stage Fine-tuning Framework. Stage 1 performs attention shift adaptation through the randomized selection of attention maps for SiT fine-tuning. Stage 2 introduces a selection network that predicts attention map importance probabilities using selection loss (right), computed as feature degradation distance with maximum-distance head indices as pseudo-labels for selection probability prediction.

we generate an upscaled embedding patch using spatial interpolation:

$$\hat{E}_{\text{pos},i} = \text{Upscale}(\{E_{\text{pos},j} \mid j \in \mathcal{N}(i)\}, f), \tag{7}$$

where $\mathcal{N}(i)$ represents the set of eight neighboring adjacent positional embeddings. After interpolation, we divide $\hat{E}_{\text{pos},i}$ into smaller positional embeddings in an $f \times f$ grid, similar to image patch subdivision:

$$\hat{E}^{(m,n)}_{\text{pos},i} = \hat{E}_{\text{pos},i}[:, mP:(m+1)P, nP:(n+1)P], \tag{8}$$

for all $m, n \in \{0, \ldots, f-1\}$. We then gather all subdivided positional embeddings into the final expanded set:

$$E^{T'}_{\text{pos}} = \bigcup_{i=1}^{K} \bigcup_{m=0}^{f-1} \bigcup_{n=0}^{f-1} \hat{E}^{(m,n)}_{\text{pos},i} \in \mathbb{R}^{Kf^2 \times D}. \tag{9}$$

Finally, the expanded positional embeddings are concatenated with the original positional embeddings to maintain spatial consistency in the input sequence:

$$z_0^{\text{new}} = \mathbf{W}_{\mathbf{p}} I^{\text{in}} + \text{Concat}(E_{\text{pos}}, E^{T'}_{\text{pos}}). \tag{10}$$

This method ensures that newly introduced PE retains meaningful spatial structure.

3.3. Two-stage Fine-Tuning

A key challenge in subtoken division is to select the most informative regions during training. Random selection risks overlooking discriminative features, while a natural alternative—using attention maps—can be too restrictive. Selecting the top-K patches based on the average attention map across all heads results in low diversity, as it repeatedly emphasizes a narrow subset of regions. To address this, we propose a two-stage fine-tuning strategy that introduces randomness in region selection during training and refines the selection process to ensure consistency during inference.

- **Stage 1: Randomized Top-K Selection.** Instead of averaging attention maps across all heads, we randomly sample from individual attention heads during training. This encourages diverse region selection, exposing the model to different semantic patterns and enhancing its ability to learn robust discriminative features.
- **Stage 2: Head Selection for Inference.** A mismatch arises because inference relies on the averaged attention map, which may not align with the head-specific selections used during training. To resolve this, we introduce a selection network that learns to identify the most informative attention head, ensuring consistency between training and inference while preserving focus on critical regions.

To address the above challenges, we propose a two-stage fine-tuning framwork: 1) SiT fine-tuning and 2) Selection training, as shown in Fig. 3 (left).

SiT Fine-Tuning (Stage 1). The first-stage fine-tuning of the SiT aligns the model's attention mechanisms with the target dataset distribution by simulating diverse attention

behaviors. We stochastically sample attention maps $A^h_{l,\text{CLS}}$ from layer l and head h to construct dynamic token sequences, addressing three key challenges: (1) interpolated positional information for subtokens, preserving spatial relationships across scales; (2) integrated subtokens for attention focus (*i.e.*, $I + T'$) to balance global and local attention; and (3) enriching token diversity through head-wise stochastic sampling. We follow Rastegar *et al.* [38] to fine-tune the backbone using SelEx loss. This stage ensures robust subtokens feature extraction while preserving pre-trained knowledge for stage 2 refinement.

Selection Training (stage2). Building upon Stage 1's fine-tuned model, whose attention maps are better aligned with the target distribution, Stage 2 identifies semantically critical regions for more precise token subdivision for further fine-tuning. We introduce a lightweight selection network, prioritizing attention maps that focus on discriminative regions. The selection network consists of an attention map branch and an image feature branch that takes $\hat{\mathbf{A}}_{l,\text{CLS}}$ and image feature from the ViT as input and predicts the selection probability $Y_{pred} \in \mathbb{R}^{N_\text{head}}$. Since ground-truth labels for optimal region selection are unavailable, we propose a self-supervised selection loss as shown in the right side of Fig. 3. The idea of selection loss is that the proximity between original features (*i.e.*, I only) and degraded features (obtained by discarding tokens T and T') indicates the discriminative power of the discarded tokens. High proximity suggests T and T' contribute minimally to fine-grained distinctions, while low proximity implies T and T' encode critical fine-grained cues, necessitating their retention.

We take the fine-tuned DINOv2 from stage 1 and proceed original features x_{ori} and degraded feature $x^h_{drop} \in \mathbb{R}^{N_\text{head} \times D}$ for attention map $\hat{\mathbf{A}}^h_{l,\text{CLS}}$ of head h. Let $Y_\text{true} \in \mathbb{R}^{N_\text{head}}$ represents the distance between x_{ori} and x_{drop}, the selection loss is defined as:

$$\mathcal{L}_\text{SL} = -\frac{1}{N_\text{head}} \sum_{h=1}^{N_\text{head}} \text{argmax}(Y^{(h)}_\text{true}) \log \left(Y^{(h)}_\text{pred} \right), \quad (11)$$

where $Y^{(h)}_\text{true}$ is the ground-truth one-hot encoded label. $\text{argmax}(Y^{(h)}_\text{true})$ represents the index of x_{drop} with maximum distance towards x_{ori}. We continue to fine-tune DINOv2 and train the selection layer. The total loss for stage 2 is:

$$\mathcal{L}_\text{all} = \mathcal{L}_\text{SE} + \beta \mathcal{L}_\text{SL}, \quad (12)$$

where β is a hyperparameter to control the weight of $\mathcal{L}_\text{SL}$.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our approach on fine-grained datasets: CUB-200 [49], FGVC-Aircraft [30], and Stanford-Cars [25]. In addition, we demonstrate the versatility of our method on coarse-grained datasets: CIFAR10 [26], and ImageNet-100 [12]. This comprehensive evaluation underscores the broader applicability of our approach beyond fine-grained classification tasks. Detailed statistics of the datasets are provided in the Supplementary.

Implementation Details. We follow SelEx [38] to set up known, novel categories for all datasets and use DINOv2 [32] pretrained on ImageNet 22K. We use the batch size of 128 for training and set the same loss hyperparameters as SelEx. We fine-tune the last two blocks of DINOv2. We use the bilinear function as the interpolation function. We set $K = 10\%$ (*i.e.*, top 10% of image tokens) for CUB, and $K = 1\%$ for other datasets. $f = 2$ for scale factor and learning rate $lr = 0.1$ for fine-tuning DINOv2 in stage 1 and stage 2, $lr = 1e^{-4}$ for selection network. β is set to 1.

4.2. Comparison with State-of-the-Art

Fine-grained Image Classification. Our method is evaluated against baseline approaches on three fine-grained datasets, as summarized in Tab. 1. The results demonstrate the superior capability of our method in both all and novel category classification, highlighting its effectiveness for fine-grained recognition. The performance improvements can be attributed to SiT's ability to provide crucial fine-grained semantic tokens across multiple scales, enabling the model to prioritize discriminative details without the need to modify the loss function, architecture, or fine-tuning strategy. Notably, the larger performance gain on novel categories (3.4% versus 0.3% on known classes) underscores our method's reduced susceptibility to overfitting and enhanced generalization to unseen objects.

Generic Image Classification. We also validate our method on generic image classification tasks that focus on coarse-grained objects in Tab. 2 using the same backbone DINOv1 [6] as the baseline for a fair comparison. Our method demonstrates the superior performance of CIFAR10 and ImageNet-100 compared with SelEx. Note that SiT is designed for fine-grained classification, our method still has performance gains in generic objects, which highlights the robustness of SiT. The result also demonstrates that detail regions are also important for generic image classification even though the coarse-grained objects have a larger diversity than fine-grained objects.

4.3. Ablation Studies

We comprehensively evaluate the effects of method components and hyperparameters in this section. We present additional ablations in the Supplementary.

Effects of Token Selection Methods. We benchmark our token selection strategy against three baselines: (1) random sampling K tokens, (2) averaged attention maps with top-K

Table 1. **Comparison with state-of-the-art for fine-grained image classification.** Our method outperforms all other baseline methods in all three categories (*All*, *Known*, *Novel*), especially significant improvement in the *Novel*, indicating the effectiveness of SiT. Bold and underlined numbers indicate the best and second-best accuracies, respectively. [Keys: *: reported from [47]. †: reported from [65]].

		CUB-200			FGVC-Aircraft			Stanford-Cars			Average		
	Method	All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	ORCA† [4]	36.3	43.8	32.6	31.6	32.0	31.4	31.9	42.2	26.9	33.3	39.3	30.3
	GCD [46]	51.3	56.6	48.7	45.0	41.1	46.9	39.0	57.6	29.9	45.1	51.8	41.8
	GPC [67]	52.0	55.5	47.5	43.3	40.7	44.8	38.2	58.9	27.4	44.5	51.7	39.9
	XCon [16]	52.1	54.3	51.0	47.7	44.4	49.4	40.5	58.8	31.7	46.8	52.5	44.0
	SimGCD [54]	60.3	65.6	57.7	54.2	59.1	51.8	53.8	71.9	45.0	56.1	65.5	51.5
	PIM [9]	62.7	75.7	56.2	-	-	-	43.1	66.9	31.6	-	-	-
	PromptCAL [65]	62.9	64.4	62.1	52.2	52.2	52.3	50.2	70.1	40.6	55.1	62.2	51.7
	DCCL [35]	63.5	60.8	64.9	-	-	-	43.1	55.7	36.2	-	-	-
	AMEND [2]	64.9	75.6	59.6	52.8	61.8	48.3	56.4	73.3	48.2	58.0	70.2	52.0
	μGCD [47]	65.7	68.0	64.6	53.8	55.4	53.0	56.5	68.1	50.9	58.7	63.8	56.2
	SPTNet [50]	65.8	68.8	65.1	59.3	61.8	58.1	59.0	79.2	49.3	61.4	69.9	57.5
	CMS [10]	68.2	76.5	64.0	56.0	63.4	52.3	56.9	76.1	47.6	60.4	72.0	54.6
	GCA [33]	68.8	73.4	66.6	52.0	57.1	49.5	54.4	72.1	45.8	58.4	67.5	54.0
	InfoSieve [37]	69.4	77.9	65.2	56.3	63.7	52.5	55.7	74.8	46.4	60.5	72.1	54.7
	TIDA [52]	-	-	-	54.6	61.3	52.1	54.7	72.3	46.2	-	-	-
	SelEx [38]	73.6	75.3	72.8	57.1	64.7	53.3	58.5	75.6	50.3	63.0	71.9	58.8
DINOv2	GCD* [46]	71.9	71.2	72.3	55.4	47.9	59.2	65.7	67.8	64.7	64.3	62.3	65.4
	SimGCD* [54]	71.5	78.1	68.3	63.9	69.9	60.9	71.5	81.9	66.6	69.0	76.6	65.3
	μGCD* [47]	74.0	75.9	73.1	66.3	68.7	65.1	76.1	91.0	68.9	72.1	78.5	69.0
	SelEx [38]	<u>87.4</u>	<u>85.1</u>	<u>88.5</u>	<u>79.8</u>	**82.3**	<u>78.6</u>	<u>82.2</u>	<u>93.7</u>	<u>76.7</u>	<u>83.1</u>	<u>87.0</u>	<u>81.3</u>
	SiT (Ours)	**91.8**	**86.3**	**94.6**	**80.8**	<u>81.3</u>	**80.5**	**83.8**	**94.9**	**78.5**	**85.5**	**87.3**	**84.7**

selection, and (3) multi-scale concatenation via CapeX [41]. The multi-scale baseline crops body and facial regions using keypoint-derived bounding boxes resizes subregions to input resolution $H \times W$, and concatenates all tokens with interpolated positional embeddings. Details of multi-scale implementation are provided in Supplementary. The performance comparison is shown in Tab. 3a.

Compared with the multi-scale concatenation, our method achieves superior performance with much fewer tokens during evaluation, underscoring the importance of focusing on crucial regions—excessive tokens containing non-essential information may introduce noise and degrade model performance. Our method outperforms random sampling and averaged attention map with the same number of tokens, demonstrating that selecting the crucial region for subdivision is important. Fig. 4 reveals a consistent focus on discriminative regions (e.g., avian wingtips), providing interpretable insights into subcategory recognition.

Effects of Hyperparameters. Tab. 3b reveals the relationship between scale factor f and top-K selection. We conduct the ablation using the stage 1 fine-tuned model for comparison. The optimal performance occurs at $f = 2$ with $K = 10\%$, suggesting moderate scaling enables effective utilization of more patches. Smaller f values better accommodate larger K by maintaining patch diversity, while larger f requires a more conservative K selection to avoid redundant overlapping patches.

Table 2. **Comparison with baseline method for coarse-grained image classification.** Bold numbers show the best accuracies. Our method has a consistent performance for the three experimental settings (*All*, *Known*, *Novel*), demonstrating that our method is also suitable for coarse-grained classification.

	CIFAR-10			ImageNet-100			Average		
Method	All	Known	Novel	All	Known	Novel	All	Known	Novel
SelEx [38]	95.9	**98.1**	94.8	83.1	93.6	77.8	89.5	95.6	86.3
SiT (Ours)	**96.7**	97.5	**96.3**	**83.9**	**94.0**	**78.9**	**90.8**	**95.8**	**87.6**

Effects of Token Masking. We evaluate the effects of token masking strategies to understand the contribution of different types of tokens. Tab. 4 demonstrates the complementary roles of selected tokens T and subtokens T'. Our method achieves peak performance with all tokens while masking T' alone degrades performance by 2.1% in *Novel* category. Compared with masking T, the performance drastically drops when masking T', demonstrating the contribution of T'. Crucially, simultaneous masking of both T and T' causes catastrophic performance collapse, revealing their synergistic relationship: T provides stable base pat-

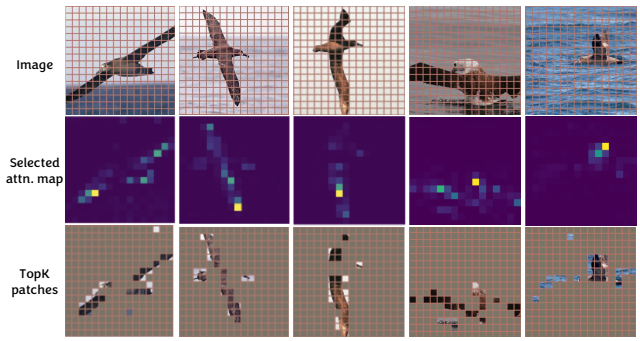

Figure 4. Visualization of attention map selection ($K = 10\%$). SiT demonstrates consistent semantic region selection across instances through the token subdivision. Zoom in for a better effect.

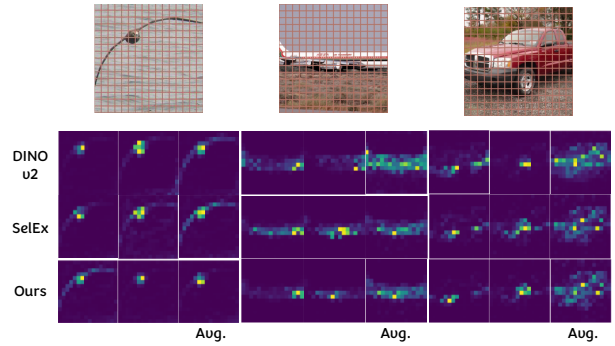

Figure 5. Attention map analysis of DINOv2, SelEx, and our SiT. The rightmost column displays cross-head averaged attention map. Our method exhibits concentrated activation patterns in compact regions compared to baseline approaches, suggesting enhanced localization precision for discriminative features.

Method	# Tokens	All	Known	Novel
SelEx [38]	256	87.4	85.1	88.5
Random sampling	356	82.4	85.2	81.0
Averaged attn. map	356	89.0	84.6	91.2
Multi-scale	768	87.8	86.1	88.8
SiT (Ours)	356	**91.8**	**86.3**	**94.6**

(a) **Effects of token selection methods.**

f	$K(\%)$	# Tokens	All	Known	Novel
2	1	264	90.3	84.0	93.5
2	5	304	90.5	85.6	93.1
2	10	356	91.8	86.3	94.6
3	1	274	90.4	84.2	93.6
3	5	364	90.3	84.0	93.5
3	10	481	90.7	85.9	93.3

(b) **Effects of scale factor f and top K.**

Table 3. Ablation studies on effects of the patch selection method, hyperparameters on CUB. (a) Our selection network performs better than other methods. (b) Higher f with larger K values increase token redundancy and computational cost.

Method	# Tokens	All	Known	Novel
SiT (Ours)	356	91.8	86.3	94.6
Mask T	331	90.9	84.5	94.1
Mask T'	256	89.7	84.6	92.4
Mask $T + T'$	231	68.6	65.1	70.3

Table 4. Effects of masking different tokens on CUB. The huge difference in performance indicates the importance of T' and T.

terns while T' captures fine-grained details. Masking T results in a 0.5% drop in Novel accuracy, indicating that subtokens significantly boost novel class recognition.

Performance Comparison of Stage 1 and 2. Tab. 5a presents a comparison of the evaluation metrics between Stage 1 and Stage 2. We use the averaged attention map for Stage 1 evaluation. Notably, our Stage 1 results already outperform the baseline. After incorporating the selection training in Stage 2, the performance further improved. These results indicate that the selection layer effectively prioritizes critical attention maps during Stage 2, leading to consistent performance gains among all datasets. Our attention map selection method plays a key role in enhancing the model's ability to capture discriminative semantic features across varying instances and subcategories.

Fine-grained Details Analysis. Our quantitative analysis reveals consistent selection patterns for fine-grained regions. As shown in Fig. 4, the method consistently identifies discriminative semantic regions across pose-varying instances (*e.g.*, wing trailing edges in avian species), demonstrating effective localization of class-critical visual patterns. The subdivision process preserves these key regions while maintaining natural context transitions. Such behavior is crucial for tasks that rely on precise recognition of subtle visual cues, thereby enhancing overall fine-grained classification performance.

Fig. 5 compares attention maps of SelEx [38], pretrained DINOv2 [32], and our method. While both SelEx and our approach focus on fine-grained regions compared to DINOv2, our method localizes smaller, more precise areas, highlighting its superior ability to isolate discriminative details. This refined localization directly correlates with performance improvements, as concentrating on critical regions reduces irrelevant signal interference. Furthermore, the averaged attention maps of baseline methods exhibit diluted focus on key regions due to the aggregation of multi-head attention scores, as the differences in averaged attention maps are nuanced. This observation elucidates why attention map averaging underperforms relative to our method in Tab. 3a: our method preserves fine-grained focus through

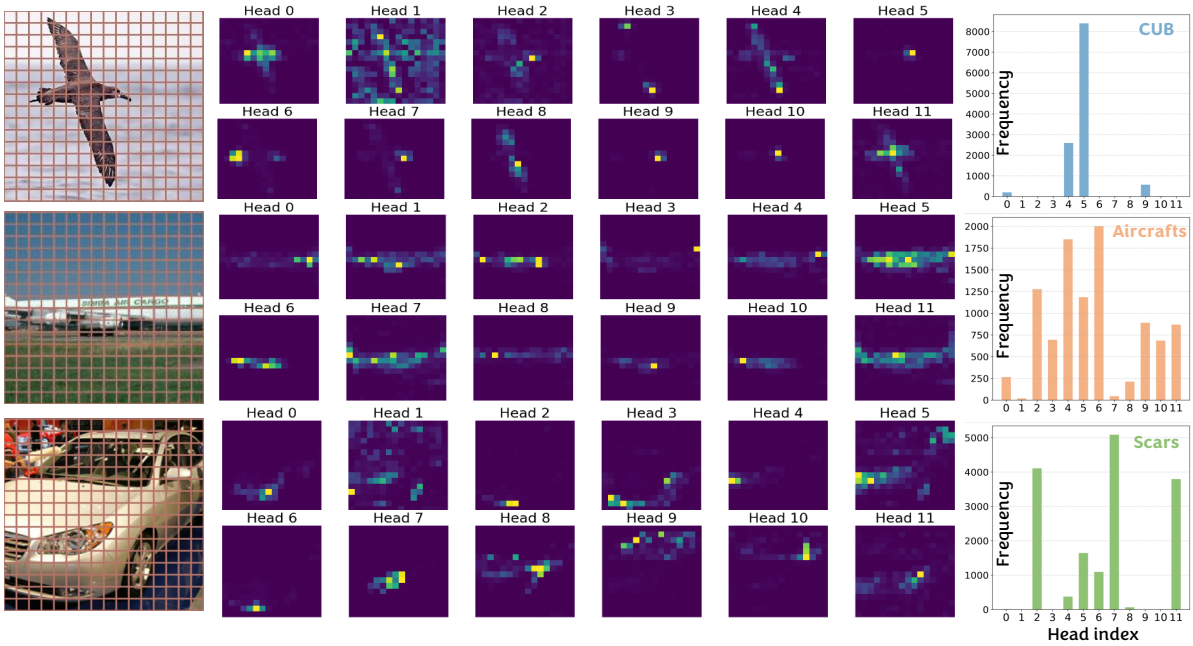

Figure 6. Distribution of head selection on fine-grained evaluation set. The pattern of head selection distribution implies the important factor of different heads for the target datasets. We select one example from each dataset to visualize the attention pattern of each head in the last layer of DINOv2.

Method	All	Known	Novel
SelEx [38]	87.4	85.1	88.5
Stage 1	91.4	86.2	93.9
Stage 2	91.8 (+0.4%)	86.3 (+0.1%)	94.6 (+0.7%)

(a) Stage 1 and 2 Comparison on CUB.

Method	All	Known	Novel
SelEx [38]	79.8	82.3	78.6
Stage 1	80.0	81.0	79.3
Stage 2	80.8 (+0.8%)	81.3 (+0.3%)	80.5 (+0.8%)

(b) Stage 1 and 2 Comparison on Aircraft.

Method	All	Known	Novel
SelEx [38]	82.2	93.7	76.7
Stage 1	83.0	93.6	78.0
Stage 2	83.8 (+0.8%)	94.9 (+1.1%)	78.5 (+0.5%)

(c) Stage 1 and 2 Comparison on Scars.

Table 5. Two-stage fine-tuning performance comparison across fine-grained datasets. The results indicate the effectiveness of the proposed two-stage training.

attention map selection, enhancing feature discriminability. More attention map visualizations and comparisons are provided in the Supplementary.

Fig. 6 analyzes attention head selection across fine-grained datasets, highlighting domain-specific critical regions. For bird species recognition, high-frequency attention heads focus on wingtip areas, whereas aircraft prioritize turbine components, and vehicles emphasize headlight structures. We notice that a subset of attention heads remains entirely unselected across all categories, indicating potential redundancy or providing ineffective information. This underscores the potential feasibility of optimizing ViTs through head pruning for future works.

5. Conclusion

This work presents SiT, a novel Subtoken Vision Transformer adaptation that enhances fine-grained category discovery through dynamic computational allocation. By developing attention-based token subdivision and selection mechanisms, our method enables localized resolution enhancement in discriminative regions while maintaining global contextual understanding. Extensive validation on fine-grained and coarse-grain benchmarks demonstrates SiT's superior performance over existing approaches. The learned attention patterns reveal semantically meaningful regions aligned with domain expertise, providing interpretable evidence for model decisions. The proposed two-stage fine-tuning strategy effectively enhances the model's capability of fine-grained representation and focuses on crucial region, bridges pretrained representations and downstream tasks without architectural modifications. Future extensions could explore SiT's applicability in multimodal learning scenarios and its integration with emerging efficient transformer variants.

References

[1] Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, QianYing Wang, and Ping Chen. Generalized category discovery with decoupled prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 2

[2] Anwesha Banerjee, Liyana Sahir Kallooriyakath, and Soma Biswas. Amend: Adaptive margin and expanded neighborhood for efficient generalized category discovery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 6

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, 2017. 2

[4] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2022. 2, 6

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 5

[7] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 2009. 2

[8] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 3

[9] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 6

[10] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 6

[11] Arpita Chowdhury, Dipanjyoti Paul, Zheda Mai, Jianyang Gu, Ziheng Zhang, Kazi Sajeed Mehrab, Elizabeth G Campolongo, Daniel Rubenstein, Charles V Stewart, Anuj Karpatne, et al. Prompt-cam: A simpler interpretable transformer for fine-grained analysis. *arXiv preprint arXiv:2501.09333*, 2025. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. On-the-fly category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[15] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

[16] Yixin Fei, Zhongkai Zhao, Siwei Yang, and Bingchen Zhao. Xcon: Learning with experts for fine-grained category discovery. In *British Machine Vision Conference*, 2022. 6

[17] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[18] Fulin Gao, Weimin Zhong, Zhixing Cao, Xin Peng, and Zhi Li. Opengcd: Assisting open world recognition with generalized category discovery. *arXiv preprint arXiv:2308.06926*, 2023. 1, 2

[19] Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[20] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[21] Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *arXiv preprint arXiv:2304.06928*, 2023. 2

[22] 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*, 2021. 2

[23] Hyungmin Kim, Sungho Suh, Daehwan Kim, Daun Jeong, Hansang Cho, and Junmo Kim. Proxy anchor-based unsupervised learning for continuous generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[24] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. Keypoint relative position encoding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[25] 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*, 2013. 1, 5

[26] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009. 5

[27] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 2

[30] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[31] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018. 2

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 5, 7

[33] Jona Otholt, Christoph Meinel, and Haojin Yang. Guided cluster aggregation: A hierarchical approach to generalized category discovery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 1, 6

[34] Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020. 2

[35] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 6

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 2021. 2

[37] Sarah Rastegar, Hazel Doughty, and Cees G. M. Snoek. Learn to categorize or categorize to learn? self-coding for generalized category discovery. In *Advances in Neural Information Processing Systems*, 2023. 1, 2, 6

[38] Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees GM Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *European Conference on Computer Vision*. Springer, 2024. 1, 2, 5, 6, 7, 8

[39] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Semi-supervised learning with scarce annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 2020. 2

[40] N Dinesh Reddy, Minh Vo, and Srinivasa G Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 2

[41] Matan Rusanovsky, Or Hirschorn, and Shai Avidan. Capex: Category-agnostic pose estimation from textual point explanation. *arXiv preprint arXiv:2406.00384*, 2024. 6

[42] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2019. 2

[43] Zhaorui Tan, Chengrui Zhang, Xi Yang, Jie Sun, and Kaizhu Huang. Revisiting mutual information maximization for generalized category discovery. *arXiv preprint arXiv:2405.20711*, 2024. 2

[44] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 2

[45] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 1

[46] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 6

[47] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems 37*, 2023. 1, 2, 6

[48] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems*, 2023. 1

[49] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. 2011. 1, 5

[50] Hongjun Wang, Sagar Vaze, and Kai Han. SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *Proceedings of the International Conference on Learning Representations*, 2024. 1, 2, 6

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 2

[52] Yu Wang, Zhun Zhong, Pengchong Qiao, Xuxin Cheng, Xiawu Zheng, Chang Liu, Nicu Sebe, Rongrong Ji, and Jie Chen. Discover and align taxonomic context priors for open-world semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2023. 1, 6

[53] Ye Wang, Yaxiong Wang, Yujiao Wu, Bingchen Zhao, and Xueming Qian. Beyond known clusters: Probe new prototypes for efficient generalized class discovery. *arXiv preprint arXiv:2404.08995*, 2024. 2

[54] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16590–16600, 2023. 1, 2, 6

[55] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 1

[56] Ruixuan Xiao, Lei Feng, Kai Tang, Junbo Zhao, Yixuan Li, Gang Chen, and Haobo Wang. Targeted representation alignment for open-world semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[57] Muli Yang, Liancheng Wang, Cheng Deng, and Hanwang Zhang. Bootstrap your own prior: Towards distribution-agnostic novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[58] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 2

[59] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

[60] Yuxiang Yang, Junjie Yang, Yufei Xu, Jing Zhang, Long Lan, and Dacheng Tao. Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 2022. 2

[61] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*, 2021. 2

[62] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2

[63] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 2016. 2

[64] Lu Zhang, Lu Qi, Xu Yang, Hong Qiao, Ming-Hsuan Yang, and Zhiyong Liu. Automatically discovering novel visual categories with self-supervised prototype learning. *arXiv preprint arXiv:2208.00979*, 2022. 2

[65] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 6

[66] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 2021. 2

[67] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 6

[68] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[69] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[70] 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*, 2016. 2