Subtoken Image Transformer (SiT) for Generalizable Fine-grained Recognition

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present Subtoken Image Transformer (SiT), a novel image tokenization method designed to enhance fine-grained visual recognition in Vision Transformers (ViTs). We hypothesize that allocating more representational capacity to semantically informative regions improves the network's ability to capture subtle inter-class differences. SiT achieves this by dynamically subdividing image tokens into subtokens in discriminative regions, enabling finer feature representations without increasing global computational costs. SiT builds upon a pretrained ViT backbone and employs attention-guided region proposals during training. During inference, a lightweight selection network identifies key regions for token subdivision. To assess the effectiveness of fine-grained capturing and generalization of SiT, we adopt Generalized Category Discovery (GCD) as a challenging evaluation protocol due to its requirement to classify known and novel categories by learning discriminative features that capture fine-grained inter-class distinctions while remaining invariant to irrelevant variations. 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

Vision Transformers (ViTs) [8] have emerged as a powerful architecture for visual recognition, treating an image as a sequence of patch tokens and applying self-attention to model global context. Typically, ViTs tokenize images by uniformly dividing them into fixed-size square patches, assigning each patch an equal share of computational resources and representational capacity. While this uniform tokenization works well for coarse-grained recognition tasks, it falls short in fine-grained scenarios where subtle visual cues such as texture differences or small object parts are critical. ViT's default square patch tokenization allocates equal computational resources to all image regions, including those irrelevant to distinguishing fine-grained categories. Moreover, small or subtle regions may correspond to too few tokens, limiting the model's ability to capture minute but semantically important 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.

Recent efforts have explored adaptive tokenization. For instance, SapiensID [15] introduces Retina Patch, which concatenates multi-scale image tokens to enhance important regions via a keypoint predictor. However, it lacks mechanisms for dynamically reallocating computational resources based on regional importance and may introduce unnecessary tokens. MSViT [12] improves efficiency

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

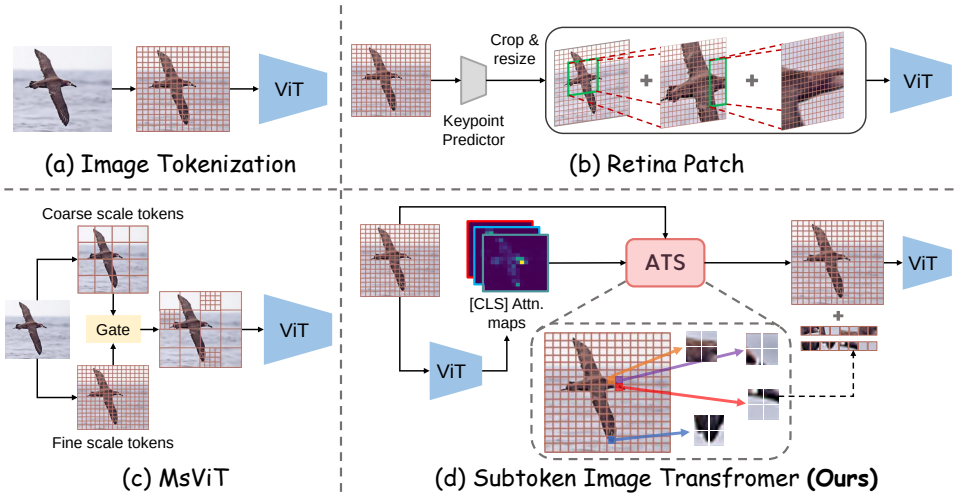

Figure 1: Motivation of SiT. SiT uses attention maps to localize fine-grained details instead of relying on multi-scale input and coarse region-level resizing. It performs Attention-based Token-level Subdivision (ATS) to reduce ineffective tokens and focus on discriminative regions.

by selecting between coarse and fine token resolutions via a gating mechanism, yet it remains constrained to binary scale choices and cannot flexibly adjust granularity within a region. Motivated by these limitations, 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 called Retina Patch [15] uses 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 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 three 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 introduce a **Selection Network** during testing that deterministically identifies key regions for token division.
- Extensive experiments on multiple fine-grained datasets and challenging Generalized Category Discovery (GCD) task demonstrate that **SiT significantly enhances fine-grained recognition on unseen categories**, outperforming existing methods in distinguishing visually similar subcategories.

2 Related Works

Dynamic Token Scaling. Vision Transformers (ViTs), exemplified by CLIP [22], SigLIP [32], and the DINO family [3, 21], 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. Swin Transformer [19] and Pyramid vision transformer [28] employ multi-scale feature aggregation through shifted windowing, though they introduce computational redundancy while overlooking intra-image heterogeneity. Kim et al. [15] introduces Retina Patch with multi-scale tokens concatenation for ViT. These methods demonstrate the importance of spatial adaptation but lack dynamic resource allocation mechanisms for region-specific processing demands. MSViT [12] proposes a mix-scale tokens using a learnable gating to choose between coarse or fine tokens in every region, but fails to provide flexible choices in subtokenization. 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 [2, 9, 29, 14, 25] and class activation maps (CAMs). However, defining consistent keypoints for generic objects (e.g., industrial products) [30, 31] remains challenging due to the lack of anatomical priors, risking overemphasis on unimportant features. CAM-based methods [6, 13, 35] 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. [26] and Cao et al. [1]. GCD extends novel category discovery (NCD) [11, 33, 34, 10] by requiring models to simultaneously leverage labeled data and partition unlabeled novel subcategories. We adapt the GCD task to evaluate the fine-grained capturing capability of SiT and investigate two aspects: (1) Which object 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 $z_0 \in \mathbb{R}^{N\times D}$.

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 to generate attention map $\mathbf{A}_l \in \mathbb{R}^{N\times N}$. 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, 4].

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.

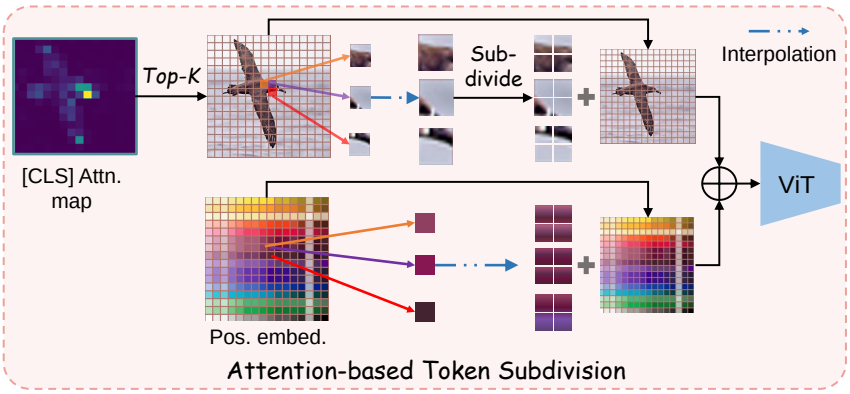

Figure 2: Overview of ATS. We select the top-K patches based on the attention map for subdivision. We select the neighboring positional embedding for interpolation purpose.

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, 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

$$\hat{T}_i = \text{Upscale}(T_i, \text{factor}=f), \quad \hat{T}_i \in \mathbb{R}^{C\times(fP)\times(fP)}. \tag{1}$$

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:

$$\hat{T}_i^{(m,n)} = \hat{T}_i[:, mP:(m+1)P, nP:(n+1)P], \tag{2}$$

where all subtokens are denoted as:

$$T' = \bigcup_{i=1}^{K}\bigcup_{m=0}^{f-1}\bigcup_{n=0}^{f-1} \hat{T}_i^{(m,n)} \in \mathbb{R}^{Kf^2\times(C\times P\times P)}. \tag{3}$$

Put it simply, subtokens are created by upscaling and dividing each patch. T' concatenates with I to perform the input patch sequence I^{in} to the patch embedding layer. 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 $E_{\text{pos},i}$, we generate an upscaled embedding patch using spatial interpolation:

$$\hat{E}_{\text{pos},i} = \text{Upscale}(\{E_{\text{pos},j} \mid j \in \mathcal{N}(i)\}, f), \tag{4}$$

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:

$$\hat{E}^{(m,n)}_{\text{pos},i} = \hat{E}_{\text{pos},i}[:, mP:(m+1)P, nP:(n+1)P], \tag{5}$$

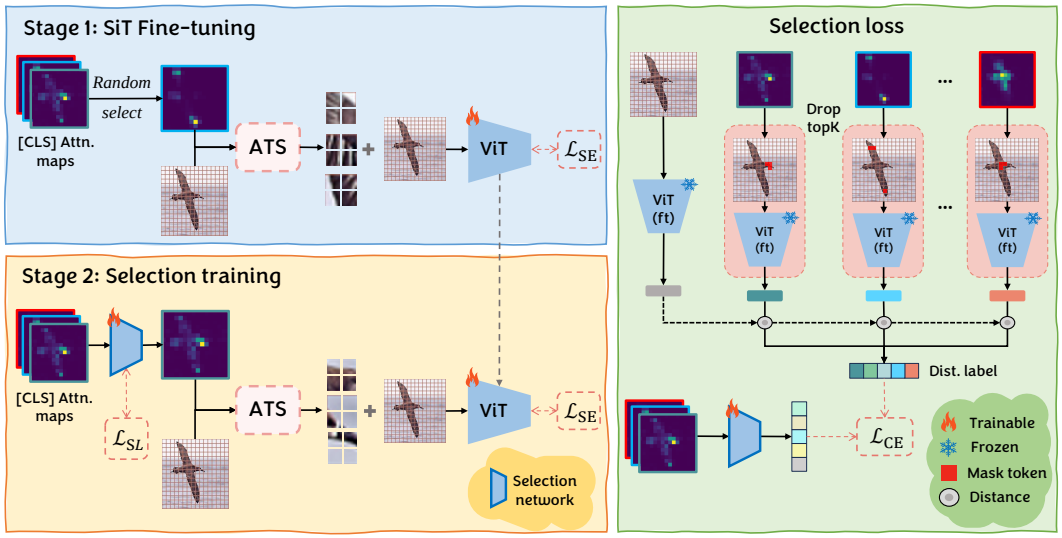

Figure 3: Two-Stage Fine-tuning Framework. Stage 1 performs attention shift adaptation through the randomized selection of attention maps for SiT fine-tuning. Stage 2 introduces a selection network that predicts attention map importance probabilities using selection loss (right), computed as feature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

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

$$E_{\text{pos}}^{T'} = \bigcup_{i=1}^{K} \bigcup_{m=0}^{f-1} \bigcup_{n=0}^{f-1} \hat{E}_{\text{pos},i}^{(m,n)} \in \mathbb{R}^{Kf^2 \times D}. \tag{6}$$

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

$$z_0^{\text{new}} = \mathbf{W}_{\mathbf{p}} I^{\text{in}} + \text{Concat}(E_{\text{pos}}, E_{\text{pos}}^{T'}). \tag{7}$$

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 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).

- **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 during training. To address this, we introduce a selection network that identifies the most informative attention head, ensuring consistency between training and inference while preserving focus on critical regions.

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 $A_{l,\text{CLS}}^{h}$ 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. This stage ensures robust subtokens feature extraction while preserving pretrained knowledge for stage 2 refinement.

Selection Training (Stage 2). 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 ViT 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:

$$\mathcal{L}_{\text{SL}} = -\frac{1}{N_{\text{head}}} \sum_{h=1}^{N_{\text{head}}} \text{argmax}(Y^{(h)}_{\text{true}}) \log \left(Y^{(h)}_{\text{pred}} \right), \tag{8}$$

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 further fine-tune ViT and train the selection layer. The total loss for stage 2 is shown as follows, where β is a hyperparameter to control the weight of $\mathcal{L}_{\text{SL}}$:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{SE}} + \beta \mathcal{L}_{\text{SL}}. \tag{9}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our approach on fine-grained datasets: CUB-200 [27], FGVC-Aircraft [20], and Stanford-Cars [16]. In addition, we demonstrate the versatility of our method on coarse-grained datasets: CIFAR10 [17], and ImageNet-100 [7]. This comprehensive evaluation underscores the broader applicability of our approach beyond fine-grained classification tasks. Detailed statistics of the datasets are provided in the Appendix.

Baseline Setup. We compare SiT with classic ViT [5], Retina Patch [15] and MsViT [12] which apply different strategies towards image tokenization. For a fair comparison, we use the same loss function and hyperparameters proposed by SelEx [23]. We use CapeX [24] as the keypoint predictor for Retina Patch. Details of the Retina Patch implementation are provided in the Appendix.

Implementation Details. We follow SelEx [23] to set up known, novel categories for all datasets and use DINOv2 [21] pretrained on ImageNet 22K [18] and DINOv1 [3] pretrained on ImageNet 1K. We use the batch size of 128 for training and set the same loss hyperparameters as SelEx. We use DINOv2 and DINOv1 as pretrained ViT and fine-tune the last two blocks. We use the bilinear function as the interpolation function. We set $K = 10\%$ (*i.e.*, top 10% of image tokens) for CUB, $K = 2\%$ for Aircraft, and $K = 1\%$ for SCars. $f = 2$ for scale factor and learning rate $lr = 0.1$ for fine-tuning ViT 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 the all and novel category classifications, highlighting its effectiveness for fine-grained recognition. Compared to Retina Patch, our method achieves better performance with

Table 1: **Comparison with baseline methods for fine-grained image classification.** Our method outperforms baseline methods in most settings (*All*, *Known*, *Novel*), with 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 [23].]

	Method	CUB-200			FGVC-Aircraft			Stanford-Cars			Average		
		All	Known	Novel	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	ViT*	73.6	75.3	72.8	57.1	**64.7**	53.3	58.5	75.6	50.3	63.0	71.9	58.8
	MsViT	73.5	74.9	72.8	55.6	64.5	51.2	52.6	74.0	42.3	60.6	71.1	55.4
	Retina Patch	71.6	73.5	70.7	52.9	57.7	50.5	52.0	72.9	41.9	58.8	68.0	54.4
	SiT (Ours)	**75.7**	**76.4**	**75.4**	**57.5**	64.1	**54.2**	**59.3**	**76.0**	**52.1**	**64.2**	**72.2**	**60.6**
DINOv2	ViT*	87.4	85.1	88.5	79.8	**82.3**	78.6	82.2	93.7	76.7	83.1	87.0	81.3
	MsViT	88.3	85.7	90.0	79.9	79.6	80.0	81.9	93.1	76.5	83.4	86.1	82.2
	Retina Patch	87.8	86.1	88.8	73.2	74.0	72.8	80.4	93.8	73.9	80.5	84.6	78.5
	SiT (Ours)	**91.8**	**86.3**	**94.6**	**80.8**	81.3	**80.5**	**83.8**	**94.9**	**78.5**	**85.5**	**87.3**	**84.7**

Table 2: **Comparison with baseline methods 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 its applicability to coarse-grained classification.

Method	CIFAR-10			ImageNet-100			Average		
	All	Known	Novel	All	Known	Novel	All	Known	Novel
DINOv1	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**

fewer tokens by focusing on crucial regions, highlighting that excessive, non-essential tokens may introduce noise and are less compatible with pretrained ViTs. While MSViT improves computational efficiency, its fine-scale tokens offer limited benefits for fine-grained recognition tasks. Both MsViT and Retina Patch exhibit degraded performance on GCD, revealing that they either fail to generalize to pretrained ViT architectures or offer limited additional fine-grained information. 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.

Coarse-grained Image Classification. We also validate our method on generic image classification tasks that focus on coarse-grained objects in Tab. 2 using the backbone DINOv1 [3] as the baseline to show the effectiveness of our methods with different backbones. Our method demonstrates the superior performance of CIFAR10 and ImageNet-100 compared with DINOv1 fine-tuned with the same loss function. 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 detailed 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

Effects of Token Selection Methods. We compare our token selection strategy against two sampling strategies: (1) random sampling of K tokens, (2) averaged attention maps with top-K selection. The performance comparison is shown in Tab. 3a. 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 intra-class recognition.

Effects of Hyperparameters. Tab. 3b presents the effects of scale factor f and top-K selection. We conduct the ablation using the stage 1 fine-tuned SiT for comparison. The optimal performance occurs at $f = 2$ with $K = 10\%$, suggesting moderate scaling enables effective utilization of more

Table 3: **Ablation studies of SiT on CUB.** (a) Effects of token selection methods; (b) Effects of f and K; (c) Effects of Masking tokens.

(a) **Token selection methods.**

Method	# Tokens	All	Known	Novel
DINOv2	256	87.4	85.1	88.5
Random	356	82.4	85.2	81.0
Average	356	89.0	84.6	91.2
Ours	356	91.8	86.3	94.6

(b) **Scale factor f and top K.**

f	$K(\%)$	All	Known	Novel
2	1	90.3	84.0	93.5
2	10	91.8	86.3	94.6
3	1	90.4	84.2	93.6
3	10	90.7	85.9	93.3

(c) **Masking tokens.**

Method	All	Known	Novel
Ours	91.8	86.3	94.6
w/o T	90.9	84.5	94.1
w/o T'	89.7	84.6	92.4
w/o $T + T'$	68.6	65.1	70.3

Table 4: **Two-stage fine-tuning performance comparison On Aircraft and Scars.** The results indicate the effectiveness of the proposed two-stage training.

(a) Stage 1 and 2 Comparison on Aircraft.

Method	All	Known	Novel
DINOv2	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 Scars.

Method	All	Known	Novel
DINOv2	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%)

Table 5: **Comparison of inference efficiency and resource usage.** Token count is averaged per image. All measurements are taken with batch size 128 on a NVIDIA A6000 GPU.

Model	**# Tokens**	**Runtime (ms)**	**Memory (MB)**	**FLOPs (G)**
DINOv2	256	393	1078	2854.9
MSViT	256	422	1175	2859.6
Retina Patch	711	1744	4525	10583.2
SiT ($K = 10\%$)	356	1064	2144	6885.9
SiT ($K = 2\%$)	276	877	2144	6010.8
SiT ($K = 1\%$)	264	855	2145	5879.5

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.

Effects of Token Masking. We analyze token masking strategies through Tab. 3c, revealing complementary roles of tokens T and subtokens T'. Our method achieves optimal performance using all tokens. Masking T' alone causes 2.1% *Novel* accuracy drop versus 0.5% when masking T, confirming T''s greater contribution. Crucially, masking both triggers catastrophic collapse, demonstrating their synergy: T establishes base patterns while T' encodes fine details. The significant 2.1% novel class recognition gap highlights T''s critical role in handling unseen categories.

Performance Comparison of Stage 1 and 2. Tab. 4 compares both stages, where Stage 1 with averaged attention maps already surpasses the baseline. Stage 2's selection training brings further gains, demonstrating its effectiveness in prioritizing critical patterns. While Stage 2's absolute improvement is smaller, it refines Stage 1's diverse attention by selecting informative tokens, enhancing both discriminative power (through subcategory-specific semantics) and interpretability (via salient pattern highlighting). This synergy shows that Stage 1 establishes attention diversity while Stage 2 optimizes semantic focus for fine-grained recognition.

Computation Cost. Despite the increased computational cost compared to traditional ViT, our SiT models maintain accessible runtime and memory usage. The higher FLOPs mainly result from additional fine-grained token refinement that is essential for achieving strong performance on fine-grained recognition tasks. Importantly, the resource demand remains significantly lower than Retina Patch, making SiT a practical and scalable solution for real-world deployment.

Fine-grained Details Analysis. Quantitative results reveal consistent selection of discriminative regions (Fig. 4), precisely localizing class-critical patterns (*e.g.*, avian wing trailing edges) while preserving contextual continuity. The method adapts to domain-specific features: attention heads

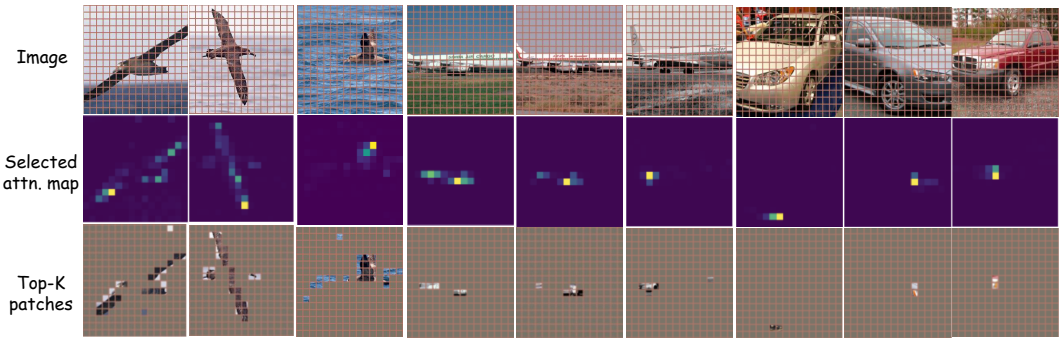

Figure 4: **Visualization of attention map selection.** $K = 10\%$ for CUB, 2% for Aircraft, and 1% for SCars. SiT demonstrates consistent semantic region selection across instances through the token subdivision. Zoom in for a better effect.

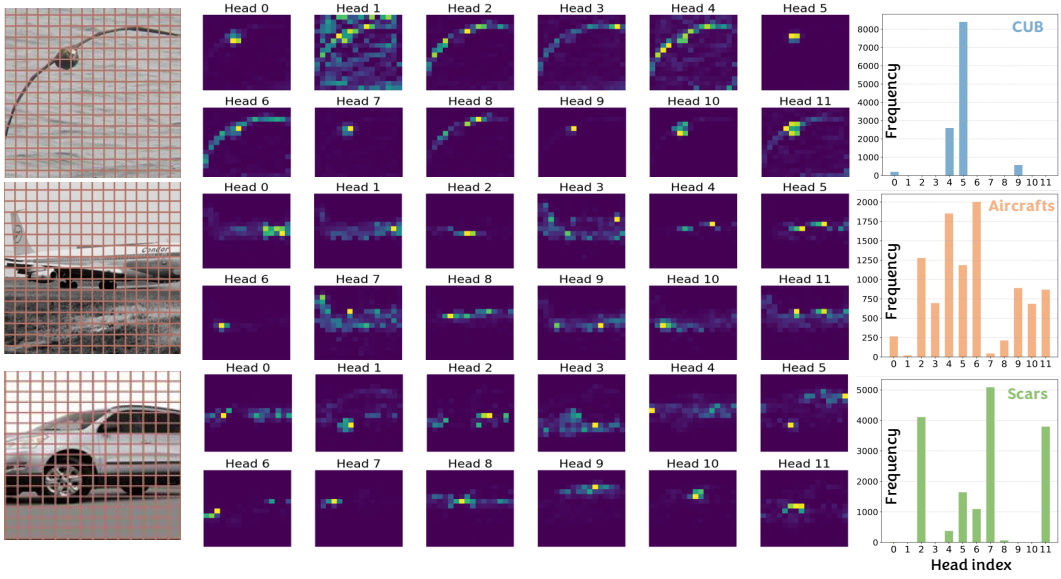

Figure 5: **Distributions of head selection frequency.** The head selection distribution pattern implies the varying importance (or roles) of different heads for the target datasets. We select one example from each dataset to visualize the attention pattern of each head in the last layer of DINOv2.

focus on wingtips for birds, turbines for aircraft, and headlights for vehicles. Fig. 5 analyzes attention head selection across fine-grained datasets, highlighting domain-specific critical regions. Notably, permanently unselected heads across categories suggest architectural redundancy, indicating that potential head pruning could optimize ViTs. This targeted selection enhances fine-grained recognition by emphasizing subtle but decisive visual cues. Additional visualizations in the Appendix.

5 Conclusion

This work presents SiT, a novel Subtoken Vision Transformer that enhances fine-grained recognition through dynamic image tokenization. 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-grained 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 regions, bridges pretrained representations and downstream tasks without architectural modifications.

References

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

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

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

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

[5] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*. Springer.

[6] 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. 2025. Prompt-CAM: A Simpler Interpretable Transformer for Fine-Grained Analysis. *arXiv preprint arXiv:2501.09333* (2025).

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

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

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

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

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

[12] Jakob Drachmann Havtorn, Amélie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. 2023. Msvit: Dynamic mixed-scale tokenization for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 838–848.

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

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

[15] Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. 2025. SapiensID: Foundation for Human Recognition. *arXiv preprint arXiv:2504.04708* (2025).

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

[17] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images*. Technical Report. University of Toronto, Toronto, Ontario.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*.

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

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

[21] 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. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024).

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

[23] Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees GM Snoek. 2024. SelEx: Self-expertise in Fine-Grained Generalized Category Discovery. In *European Conference on Computer Vision*. Springer.

[24] Matan Rusanovsky, Or Hirschorn, and Shai Avidan. 2024. CapeX: Category-Agnostic Pose Estimation from Textual Point Explanation. *arXiv preprint arXiv:2406.00384* (2024).

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

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

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

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

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

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

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

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

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

[34] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. 2021. 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*.

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