Training-free Discriminative Patch Mining for Robust Few-Shot Recognition with CLIP

Zhenzhang Ye[1], Duolikun Danier[1], Bo Zhao[2], and Hakan Bilen[1]

[1] The University of Edinburgh
{zye2, duolinkun.danier, hbilen}@ed.ac.uk
[2] Shanghai Jiao Tong University
bozhao@sjtu.edu.cn

Abstract. Few-shot classification performance of Contrastive Language-Image Pre-training (CLIP) varies widely across datasets, especially when class names provide weak or ambiguous semantic priors. This issue is alleviated in vision-only fine-grained methods, as they identify discriminative local features but require training on the full dataset. We thus introduce a training-free approach that integrates discriminative patches into CLIP to reduce reliance on textual prompts. Our method identifies patches with high intra-class consistency and low inter-class ambiguity, forming a Class-Discriminative Patch Set (CDPS) for each category. Using CDPS, we enhance the recognition ability of CLIP through a hybrid classifier combining global image-text alignment with local patch-based similarity. Extensive experiments on diverse benchmarks show that CDPS injects fine-grained discriminatory power into CLIP and yields more robust few-shot recognition performance against ambiguous class names.

Keywords: Vision-Language Model · Few-Shot Recognition · Training-Free Adaptation

1 Introduction

Learning from limited labeled data is a fundamental requirement for scalable and label-efficient visual intelligence. Few-shot learning seeks to enable models to generalize to novel and rare concepts with minimal supervision, thereby reducing reliance on costly manual annotations. Recent advances leverage large-scale vision-language models (VLMs), such as CLIP [40], ALIGN [17], and SigLIP [66], which are trained on web-scale image-text pairs to align visual and textual representations in a shared embedding space. Despite strong generalization capability, VLMs are particularly sensitive to the quality and semantics of textual prompts.

This sensitivity becomes obvious when class names are obscure or lack human-interpretable meaning. In such cases, textual prompts may misalign with the pre-trained text encoder, leading to degraded recognition performance. As shown in Fig. 1, REAL-Prompt [33] demonstrates that replacing Latin bird names in the Semi-Aves dataset [51] with more common English names exceptionally improves accuracy. Conversely, datasets such as FGVCAircraft [28], whose class

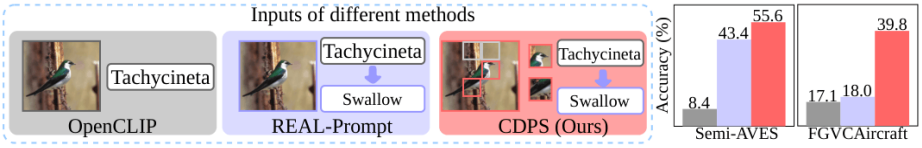

Fig. 1: Prompt engineering vs. CDPS. **Left:** Conceptual comparison: standard VLM (OpenCLIP [40]) uses the class name directly, prompt engineering (REAL-Prompt [33]) uses the synonym, and our CDPS method uses discriminative patches to compensate for ambiguous prompts. **Right:** Empirical results show prompt engineering's effectiveness is inconsistent (significant gain on common synonyms like Semi-Aves, but limited gain on technical names like FGVCAircraft). In contrast, our CDPS consistently provides improvements across both scenarios via patch feature integration.

names consist primarily of technical identifiers (e.g., "*727-200*"), exhibit limited gains due to the absence of semantically meaningful synonyms.

A natural remedy is to enrich the descriptive quality of textual prompts. Prompt Engineering methods [29–31, 33, 37, 39] construct manual templates or leverage Large Language Models (LLM) to generate detailed class descriptions. However, these approaches often lack sufficient discriminative power for fine-grained recognition [39] and typically require extensive manual trial-and-error [26]. Prompt Learning [4, 72, 73] automates this process by introducing learnable tokens optimized jointly with the model. While circumventing the need of manual prompting, such methods remain constrained by the representational capacity of the text embedding space, due to its lack of discriminative granularity that is critical for subtle, fine-grained features.

These limitations motivate revisiting visual discriminative modeling. Prior to the emergence of VLMs, extensive research in fine-grained visual classification (FGVC) [59] focused on discovering discriminative object parts and forming mid-level representations from localized regions. Such methods employ object detection [12, 44], segmentation [24, 27, 68], or attention mechanisms [11, 52] to highlight key semantic regions. Although these vision-only representations avoid ambiguities introduced by textual prompts, they typically require full-dataset supervision and intensive dataset-specific training. Consequently, their reliance on abundant annotations limits their generalization ability in few-shot settings and complicates integration with multimodal representations [3, 60].

To address these challenges, we aim to integrate the capability to identify discriminative object parts into VLMs to reduce reliance on textual prompts, while preserving the few-shot and multimodal strengths. Specifically, we extract patch embeddings from the VLMs image encoder (e.g., the ViT backbone in CLIP [40]) and select discriminative patches according to two criteria: (1) *high intra-class consistency* — patches that frequently occur within the same class; and (2) *low inter-class ambiguity* — patches that rarely appear in other classes. The selected patches constitute the **Class-Discriminative Patch Set (CDPS)** for each class. Leveraging CDPS, we construct a patch-based classifier that measures similarity between the patches of a query image and the

class-specific CDPS. We further adaptively fuse this patch-level similarity into text-to-image similarity to form a **Hybrid Classifier** that remains robust regardless of diverse levels of semantic density in class names. We conduct extensive experiments on fine-grained benchmarks with varying degrees of class-name ambiguity and demonstrate that our approach consistently outperforms state-of-the-art training-free baselines and remains competitive to training-based ones. Note that we use training-free in the sense that no backbone, or adapter parameters are optimized.

In summary, our contributions are threefold:

- We introduce a training-free method to adapt VLMs for fine-grained few-shot recognition by mining discriminative local patches.
- We propose the **Class-Discriminative Patch Set (CDPS)**, which captures intra-class consistency and inter-class distinctiveness for robust recognition under ambiguous class names.
- We design a hybrid classifier, which integrates CDPS into text similarity, achieving strong and stable performance across diverse datasets.

2 Related Work

2.1 Vision-language Pre-training and Patch Selection

The study on alignment between pictures and natural languages can be dated back to early work [48]. The subsequent works on multi-modal alignment focus on integrating natural language into visual features [10,50] or learning from interactive natural language feedback [63]. A recent trend on aligning visual and textual features aims to jointly learn cross-modality encoders from large-scale image-text pairs, such as CLIP [40], ALIGN [17], and SigLIP [66], showing impressive perfomance across various tasks including zero-shot image recognition [38,40], semantic segmentation [25,67,74], text-to-image generation [41,58], shape generation [46] and video understanding [42,54]. In this work, we focus on CLIP due to its widespread adoption as a backbone for vision-language tasks. Mining local patches in CLIP has also been explored in prior studies [35, 62], primarily to improve pre-training efficiency. In contrast, our method identifies patches that are discriminative for downstream recognition tasks.

2.2 Mid-level Representations

Before the emergence of VLMs, research in fine-grained classification primarily focused on capturing discriminative semantic parts of objects to construct mid-level representations. Early methods achieved this part localization through the integration of external modules. These included using Regional Proposal Networks [44] to predict object bounds, Fully Convolutional Networks [27] to transfer learned representations to segmentation masks, and dedicated navigator networks [65] for locating discriminative regions. To incorporate the correlation between discriminative parts, RACN [11] generates region attention maps

via a recurrent visual attention model. Subsequent research aimed to further refine the region attention maps [14, 36, 71] or improve efficiency [70] through multi-level approaches. Despite their notable performance improvement, these methods require extensive supervision and a computationally intensive training for each dataset, limiting their generalizability. While our work shares a similar intuition, focusing on discriminative local information, we address considerably more challenging constraints: training-free and few-shot scenarios.

2.3 Few-shot Fine-grained Recognition

The task of fine-grained recognition was introduced decades ago [18]. Various learning-based methods can be referred to the surveys [59, 61]. Several methods [3, 53, 55, 60] achieve remarkable results in few-shot scenarios but lack the cross-dataset generalizability.

Recently, the focus has shifted toward adapting VLMs to few-shot recognition due to its promising generalizability. **Prompt Learning** has emerged as a dominant paradigm among training-based approaches. CoOP [73] is the first work to incorporate the prompt learning into the adaptation of CLIP to downstream vision tasks, and its successor, CoCoOP [72], extends this capability to unseen classes. Furthermore, PLOT [4] applies optimal transport to formulate the features as a probability distribution, yielding a more robust alignment. MaPLe [19] further explores the multi-level prompt coupling across both vision and language branches. **Adapter-based methods** [13, 21, 47, 69] is another branch of training-based methods parallel to prompt learning. It introduces an external, small set of parameters appended to VLMs and adjusts these small weights to align image and text embeddings on the distribution of downstream dataset. For example, CLAP [47] optimizes an adaptation of the general augmented Lagrangian method. LDC [21] learns to eliminate inter-class confusion in logits with designed modules. Although their success in some datasets is remarkable, these methods remain fundamentally anchored to the original textual prompts. Consequently, their effectiveness is bottlenecked when class names in the downstream dataset are semantically ambiguous, limiting further performance gains regardless of how well the prompts or adapters are trained.

Training-free methods offer an efficient alternative to downstream tasks. **LLM-based training-free methods** focus on improving textual quality without parameter updates. Instead, they usually leverage an additional backbone to improve the textual prompts. For example, CuPL [37] guides LLMs to generate category descriptions with designed prompts, while MPVR [31] comprises a system prompt, in-context example demonstration and metadata to instruct a LLM for category-specific descriptions. Similarly, recent work [45] leverages GPT [2] to construct hierarchical prompts for classification. However, the required LLM in these methods may suffer from hallucinations [16] or generate descriptions that lack the discriminative power for fine-grained tasks [39].

Most closely related to our work are the **training-free methods without any additional backbones**. Several approaches have explored this paradigm: TIMO [22] builds a mutual guidance mechanism to align visual and textual

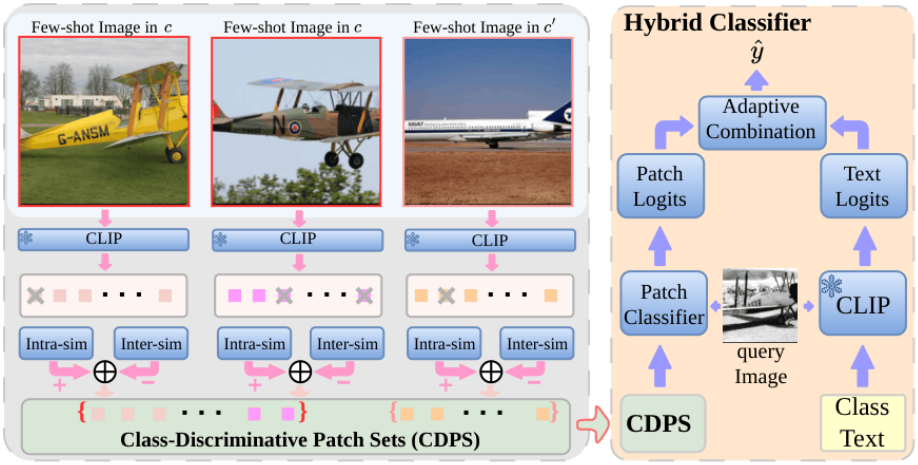

Fig. 2: Overview of construction and application of CDPS. **Left**: we utilize the intermediate patch embeddings from a pre-trained CLIP image encoder to construct CDPS by filtering patches based on the attention map, intra-class consistency and inter-class ambiguity. **Right**: CDPS can enhance few-shot classification via hybrid classifier which combines textual and patch-level logits.

modalities at the same time. GDA [57] applies Gaussian Discriminant Analysis to refine downstream classification of CLIP. ECALP [23] proposes a graph-based approach for label-efficient adaptation and inference. However, these methods remain vulnerable to the semantic ambiguity of class names. When the initial textual prompts from the downstream dataset lack semantic meanings, the resulting alignment remains suboptimal. Unlike prior training-free CLIP adaptation methods that operate mainly on global image-text embeddings or label propagation, CDPS explicitly mines local visual evidence that is class-consistent and inter-class discriminative.

3 Method

The critical concept underpinning our approach is the extraction of CDPS from the few-shot images, which contains the semantical patch embeddings for each class. This process is executed without any fine-tuning or parameter updates. The constructed CDPS then serves as the basis for the hybrid classifier to enhance the performance of image recognition. An overview of our approach is illustrated in Fig. 2.

In the following sections, we focus on CLIP [40] due to its widespread use. We start with an overview of CLIP in Sec. 3.1, followed by the construction of CDPS in Sec. 3.2. We detail the hybrid classifier in Sec. 3.3.

3.1 CLIP Preliminary

We hereby introduce CLIP and define the notations used throughout the subsequent discussion. In the standard zero-shot classification involving C distinct classes, a text prompt T_c is constructed for each class $c \in \{1, \ldots, C\}$. The model employs an image encoder $\mathcal{E}$ and a text encoder $\mathcal{T}$ to map their respective inputs into a shared embedding space. The predicted label $\hat{y}$ for an image x is determined by selecting the maximum cosine similarity between the class token embedding and the class text embeddings:

$$\hat{y} = \arg\max_c \mathbf{z}_{\mathrm{T}}(x, T_c), \text{where } \mathbf{z}_{\mathrm{T}}(x, T_c) := \mathrm{Softmax}(\cos(\mathcal{E}(x), \mathcal{T}(T_c))). \tag{1}$$

In addition to the class token embedding, the Vision Transformer (ViT) [8] based image encoder provides patch feature representations and attention maps that are essential for our method. Considering the final transformer block in ViT, it outputs a sequence $\mathcal{V}(x)$ of feature vectors corresponding to the patch tokens: $\mathcal{V}(x) := \{v_p\}_{p=1}^{P}$. Here, each $v_p \in \mathbb{R}^F$ is the patch embedding of the p-th patch token. P is the number of patches, and F is the patch feature dimension (e.g. $P = 196, F = 768$ in ViT-B-16). The multi-head self-attention mechanism yields an attention map $A \in \mathbb{R}^{(P+1)\times(P+1)}$. Each entry $A_{i,j}$ represents the attention score from token i to token j. In the following discussion, we focus on the attention map relative to the class token in the final layer. Since the class token in the final layer aggregates high-level semantic information [40], the corresponding attention map captures the semantic relevance of each local patch feature to the global visual concept.

3.2 Construction of CDPS

The central idea of our method is that classification performance should be improved by focusing on discriminative local features. Therefore, instead of relying on the complete set of local patches, we believe that only a subset of class-discriminative ones is truly informative to recognize the image x.

The identification of a patch embedding v_p as class-discriminative is contingent on two concurrent conditions:

- *High Intra-Class Consistency*: the patch embedding should appear frequently within the same class.
- *Low Inter-Class Ambiguity*: the patch embedding should rarely appear in images from other classes.

Taking the FGVCAircraft dataset [28] as a case study: a patch capturing the specific shape of an airplane's wing should appear frequently in the same category. Conversely, a patch representing a generic blue sky could also show high intra-class consistency, but it must be rejected due to its lack of discriminative power between different aircraft categories. We now demonstrate the process of constructing CDPS, which contains patches satisfy these dual criteria.

(a) Attention Weight (b) Connected Regions (c) CDPS

Fig. 3: Visualization of selected patches in each step. We illustrate the selection process using two examples: (a) initial patches selected by attention map (blue); (b) extraction of the connected regions (green) to isolated patches; (c) the final discriminative patches (red) to construct CDPS.

Preliminary Filtering. As many patches usually are not class-discriminative or essential for CLIP, we introduce a preliminary filtering step by exploiting the attention map from the final layer associated with the class token to mitigate the influence of unnecessary patches. Specifically, the top ρ-percentile patches with the largest value in the attention map are selected, as they contribute most significantly to the class token. However, several isolated noisy patches are retained as shown in Fig. 3 (a). To filter isolated noisy regions, we keep only the N_{cp} largest connected regions among the selected patches as shown in Fig. 3 (b). This step removes fragmented areas and preserves spatially coherent regions that are most relevant to the prediction.

The resulting region comprises patches that are both spatially connected and highly necessary to the class token, which is denoted as a filter mask $\mathcal{M}(x)$. It is then utilized to prune the complete patch set $\mathcal{V}(x)$: $\mathcal{V}_{\mathcal{M}}(x) = \mathcal{V}(x) \cap \mathcal{M}(x)$. It is necessary to note that the filter mask $\mathcal{M}(x)$ is used only for patch embedding selection, not direct manipulation of the input image. The complete unmasked image still passes through the encoder, based on the assumption that the contextual information provided by the unnecessary patches is valuable for ViT to compute context-aware patch embeddings.

Patch Score. After preliminary filtering, we rank the remaining patch embeddings based on the two selection criteria (Intra-Class Consistency and Inter-Class Ambiguity). Considering a specific patch embedding v, its similarity to an image can be defined as its maximum similarity among all (filtered) patch embeddings within that image. Intuitively, the value measures whether the image contains a patch that closely matches v. Based on this, we extend the formulation to a collection of images $X_c = \{x_i\}$ from the same class c, namely the *patch-to-class* (P2C) similarity, which is the average of the patch-to-image similarities across all images in the class:

$$\mathrm{sim}_{\mathrm{P2C}}(v, X_c) := \frac{1}{|X_c|} \sum_{x \in X_c} \max_{v_p \in \mathcal{V}_{\mathcal{M}}(x)} \cos(v, v_p). \tag{2}$$

A high patch-to-class similarity value indicates that patch embedding appears frequently within the images of class c. This interpretation naturally leads to the definition of discriminative scores. Given a patch v extracted from an image

x of class c, its discriminative scores is defined as:

$$\text{score}_{\text{P}}(v) := \text{sim}_{\text{P2C}}(v, X_c) - \frac{\omega}{|C|} \sum_{c' \neq c}^{C} \text{sim}_{\text{P2C}}(v, X_{c'}), \tag{3}$$

where ω is a weighting parameter used to balance the importance of the two criteria. The first term promotes features that satisfy the high intra-class consistency criterion, while the second one enacts the penalty if the feature appears frequently in other classes as well. When the dataset has many categories, evaluating the penalty term across all labels becomes computationally expensive. To mitigate this, we propose an efficient approximation that focuses on the hard negative classes. Specifically, we identify a subset of N_c categories that exhibit high similarity to a target class c, which is determined by the zero-shot classification in Eq. (1). For each few-shot image, we compute the textual logits $\mathbf{z}_{\text{T}}$. The N_c categories associated with the largest mean logits (excluding the ground truth c) are selected. This data-driven selection ensures that the discriminative score penalizes patch features that contribute to high-probability misclassifications, thereby refining the decision boundaries where the model is most prone to inter-class interference.

Class-Discriminative Patch Set. Using the derived score, the final Class-Discriminative Patches for an entire class X_c is the union of the top-K patches selected from every image in it: $\mathcal{V}_{\mathcal{D}}^{c} := \bigcup_{x \in X_c} \{v_p\}_{p=1}^{K}$.

3.3 Hybrid Classifier

Once CDPS $\mathcal{V}_{\mathcal{D}}^{c}$ has been constructed for each class c from few-shot images, we can leverage these sets to classify a query image $\tilde{x}$. Given any two patch sets $\mathcal{V}$ and $\tilde{\mathcal{V}}$, analogous to Eq. (2), we can define the *set-to-set* (S2S) similarity :

$$\text{sim}_{\text{S2S}}(\tilde{\mathcal{V}}, \mathcal{V}) := \frac{1}{|\tilde{\mathcal{V}}|} \sum_{\tilde{v} \in \tilde{\mathcal{V}}} \max_{v_p \in \mathcal{V}} \cos(\tilde{v}, v_p). \tag{4}$$

The predicted label $\hat{y}$ for a query image $\tilde{x}$ is then determined by maximizing the logits computed from the similarity between its (filtered) patch set and CDPS:

$$\hat{y} = \arg\max_{c} \mathbf{z}_{\text{P}}(\tilde{x}, \mathcal{V}_{\mathcal{D}}^{c}), \text{where } \mathbf{z}_{\text{P}}(\tilde{x}, \mathcal{V}_{\mathcal{D}}^{c}) := \text{Softmax}\big(\text{sim}_{\text{S2S}}(\mathcal{V}_{\mathcal{M}}(\tilde{x}), \mathcal{V}_{\mathcal{D}}^{c})\big), \tag{5}$$

i.e., $\mathbf{z}_{\text{P}}$ is the logits from patches, analogous to $\mathbf{z}_{\text{T}}$ in Eq. (1).

In our preliminary experiments, we observed a dependency of this method's efficacy on the characteristics of the downstream dataset. When the class names are ambiguous or when classification relies heavily on subtle local features, the patch-based classification Eq. (5) outperforms the standard zero-shot approach from Eq. (1). Conversely, when the class names are easily distinguished by high-level semantics, the standard zero-shot approach tends to be the better choice. This observation aligns with the empirical evidence on the Semi-Aves and FGVCAircraft datasets in Fig. 1.

However, quantifying the semantic meaning of class names in a dataset is an intractable problem, as it is intrinsically tied to the pre-training distribution of the CLIP backbone. Besides, as the downstream few-shot set scales, the relative reliability of fixed textual prompt may diminish compared to direct visual evidence. To this end, we propose a hybrid classifier that adaptively combines the textual prompt and the local patch features:

$$\hat{y} = \arg\max_{c} \lambda \mathbf{z}_{\mathrm{T}}(\tilde{x}, T_c) + (1-\lambda)\mathbf{z}_{\mathrm{P}}(\tilde{x}, \mathcal{V}_{\mathcal{D}}(X_c)). \tag{6}$$

Here, $\lambda \in [0, 1]$ balances the contributions of textual semantics and discriminative patch features. Since the ambiguity of class names varies across downstream datasets, we determine λ automatically using the available few-shot training set. Specifically, we adopt a leave-one-out strategy: each few-shot sample is treated as validation data in turn, and we perform a grid search over $\lambda \in [0, 1]$ with a step size of 0.1. The value of λ that yields the best average validation accuracy is selected. This procedure enables the classifier to adapt to different scenarios. A smaller λ means the prediction relies more on the discriminative patches, indicating that either class names are ambiguous or the discriminative patches are more reliable due to more available image shots.

4 Experiments

This section introduces the experimental settings, followed by a comprehensive performance evaluation and ablation analysis of the proposed method. The code can be found in `https://github.com/VICO-UoE/CDPS`.

4.1 Experimental Setting

Datasets. As our method is designed to mine discriminative patches to enhance the classification, we focus on the challenging real-world tasks requiring fine-grained visual discrimination. Following prior research [21–23], we conduct extensive experiments on 12 datasets known for their subtle inter-class differences: UCF101 [49], Caltech101 [9], SUN397 [64], Semi-Aves [51], Flowers102 [32], FGVCAircraft [28], EuroSAT [15], DTD [6], OxfordPets [34], Food101 [1], StanfordCars [20], and ImageNet [7]. Additionally, we evaluate the distribution generalization of the hybrid classifier on ImageNet-V2 [43] and ImageNet-Sketch [56] as out-of-domain distributions (OOD) benchmarks.

Implementation Details. We utilize the OpenCLIP ViT-B-16 [5] as the frozen backbone, as its transformer architecture provides the high-resolution patch embeddings essential for constructing CDPS. The results from OpenCLIP ViT-L-14 and ViT-B-32 can be found in the appendix. All the hyperparameters are consistent across all the experiments. For CDPS construction, we set $\rho = 40\%$ and $N_{cp} = 2$ for preliminary filtering, $N_c = 5$ and $\omega = 0.1$ for the patch score computation, and select top $K = 20$ patches from every few-shot image. We present the performance across $\{1, 2, 4, 8, 16\}$ shots. It is worth noting that the 1-shot

Table 1: The average accuracy over 12 datasets. Bold and underlined values indicate the best and second-best results respectively.

Shots	Training-based		Training-free				
	CLAP	LDC	TipA	GDA	ECALP	TIMO	Ours
1	69.69	**70.64**	61.23	60.91	61.99	65.29	<u>69.77</u>
2	72.61	<u>72.56</u>	63.46	64.90	62.94	67.20	**73.90**
4	74.73	<u>75.34</u>	65.26	68.49	64.94	70.33	**76.21**
8	77.29	<u>77.97</u>	67.34	71.83	67.07	72.84	**78.30**
16	79.34	**80.86**	69.11	74.59	69.23	75.52	<u>80.28</u>

setting is challenging for CDPS strategy, as it relies on intra-class consistency, which requires at least two samples per class. However, we report the results of this setting for completeness.

Compared Methods. We compare against the state-of-the-art training-free methods that rely solely on the few-shot downstream images without any additional backbones, including: TipA [69], GDA [57], TIMO [22] and ECALP [23]. For a comprehensive evaluation, we also include two training-based approaches: CLAP [33] and LDC [21].

4.2 Validation of CDPS

In this section, we demonstrate that the patches identified by CDPS contain non-trivial and semantic features that are effective for the downstream recognition task. We validate their utility in the hybrid classifier through extensive quantitative and qualitative analyses on few-shot recognition tasks, complemented by a stability test on OOD datasets.

Quantitative Validation. The hybrid classifier provides a rigorous approach to verify if CDPS are sufficiently discriminative to compensate for the ambiguous textual prompt. We summarize the average results over 12 datasets in Tab. 1. Overall, our approach consistently outperforms the training-free baselines across all shot settings. When compared to training-based approaches, our method achieves superior results in the low-shot settings while remaining highly competitive as the number of available samples increases. As illustrated by the per-dataset accuracy in Fig. 4, CDPS successfully captures fine-grained visual cues, bolstering the overall robustness of the hybrid classifier.

CDPS compensates for the ambiguous textual prompt. Our method achieves notable accuracy improvement on FGVCAircraft, StanfordCars, DTD and Caltech101, superior to both training-free and training-based counterparts as shown in Fig. 4. These gains are due to the datasets whose class names are highly technical or semantically sparse, such as "*Audi V8*" versus "*Audi 100*" in StanfordCars, where textual prompts provide insufficient discriminative information. On the other hand, CDPS addresses the issue of polysemous class names, i.e., one name might encompass different concepts. For instance, "*banded*" in DTD or "*tick*" in Caltech101 can refer to multiple visual interpretations. By integrating patch-

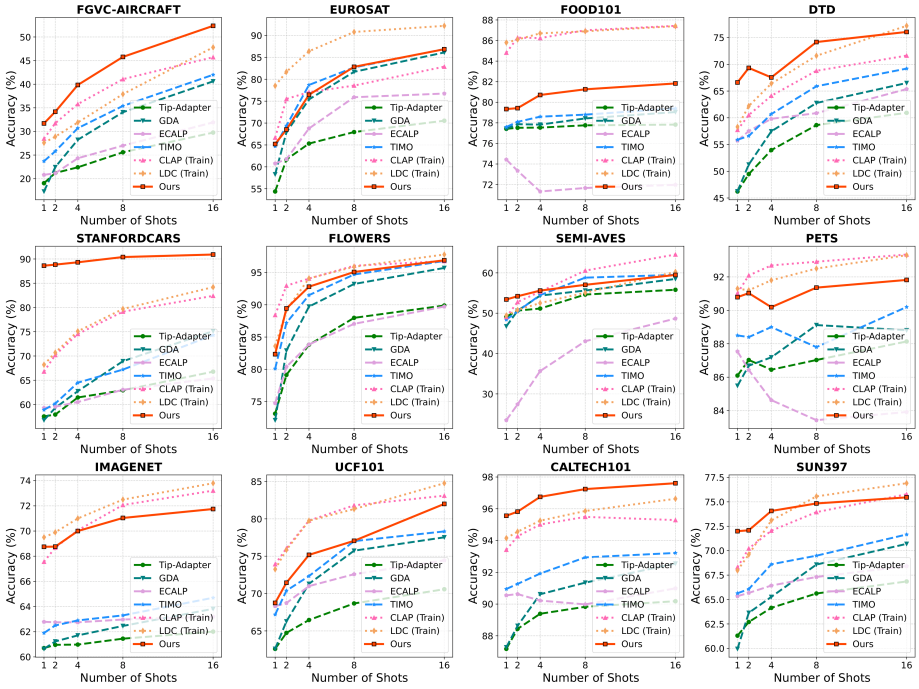

Fig. 4: Few-shot recognition results across 12 datasets. CLAP and LDC are training-based, while the rest are training-free.

level visual features, the hybrid classifier can easily figure out the specific visual concept. These results substantiate that the patch features from CDPS serve as a vital compensation for textual limitations.

CDPS is consistent across informative semantics. Even in datasets where class names are highly descriptive, our method still outperforms other training-free methods. Comparing to training-based approaches, our hybrid classifier maintains a highly competitive performance with the singular exception of the Food101 dataset. We hypothesize that the class names in Food101 are sufficiently descriptive and semantically distinct, providing a strong textual prior. Consequently, the marginal utility of CDPS is reduced in such semantics-meaningful scenarios. While training-based methods naturally exhibit better performance as the shot count increases, our method remains a highly efficient and robust alternative, circumventing the high computational cost associated with fine-tuning.

Qualitative Demonstration. We present visual evidence of CDPS in Fig. 5 where we select a few-shot image from each dataset and present the identified patches in CDPS under 4-shot setting.

CDPS successfully identifies discriminative patches satisfying the dual criteria. For instance, our method consistently locates features critical for fine-grained distinction, such as identifying an aircraft by the specific shape of its engine or a

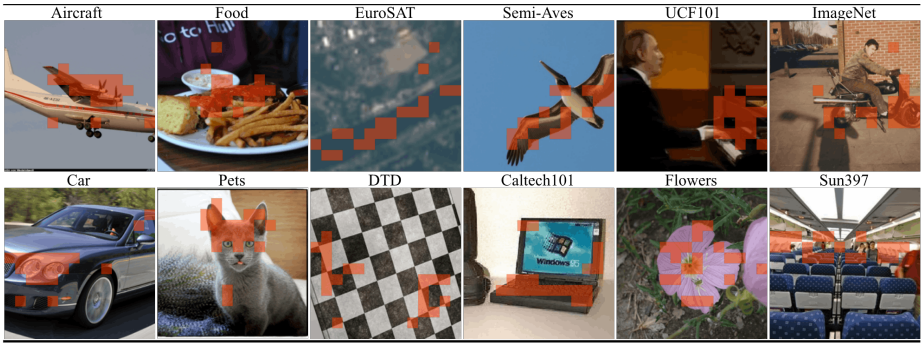

Fig. 5: Visualization of CDPS across different Datasets. The top-K discriminative patches are highlighted by red boxes.

car by its unique front grille design. It also effectively filters the patches violating our dual criterion. Patches corresponding to pommes frites within the "*pork rib*" category from Food101 are rejected, as other categories also contain fries. Similarly, human-related patches in UCF101, ImageNet and Sun397 are ignored due to low intra-class consistency. While they are visually prominent, they lack the intra-class consistency required to define the corresponding categories like "*playing piano*", "*vespa*" or "*cabin*". Instead, CDPS correctly prioritizes the semantic patches for these classes. Our method also demonstrates versatility on different levels of abstraction. In EuroSAT, our method manages to locate patches related to the river, which are the essential markers as the primary discriminative features. It also identifies the corner- and edge-like structure in the texture dataset DTD, a common discriminative feature consistent with conventional computer vision. These qualitative results validate that CDPS successfully leverages patch embeddings to locate the discriminative features.

OOD task. To evaluate distributional robustness, we construct CDPS using 16-shot samples from ImageNet and evaluate performance of the hybrid classifier on ImageNet-V2 and ImageNet-Sketch with results in Tab. 2. Our method generalizes well to ImageNet-Sketch but shows modest comparison on ImageNet-V2. We hypothesize that this behavior stems from the shape bias inherent in high-level embeddings extracted from the final ViT layers. While such structural features are well suited to stylized images in Sketch, they may insufficiently capture the textural cues needed to handle the distribution shift in ImageNet-V2. To verify this assumption, we extract and fuse features from intermediate layers. Specifically, we obtain attention maps and features from layers 2, 6, and 12, and combine them using a simple linear combination. This straightforward strategy yields a 1.32% improvement. A more systematic investigation of feature-fusion strategies may further improve performance.

Table 2: Comparison of generalization ability. All methods use the 16 shots from ImageNet and test on ImageNet-V2 and ImageNet-Sketch.

Tar.	Train.-based		Training-free			
	CLAP	LDC	TipA	GDA	ECALP	Ours
V2	64.06	65.52	63.31	65.04	**65.72**	63.66
S	47.66	49.11	48.69	48.96	54.66	**56.03**

Table 3: Ablation study on preliminary filter (Fil.) and adaptive combination (Adp.) in the hybrid classifier.

Fil.	Adp.	Air	Car	UCF	SAT
-	-	31.41	86.85	69.23	76.80
✓	-	32.52	86.87	75.95	80.67
-	✓	38.47	88.08	71.05	77.80
✓	✓	**39.84**	**88.46**	**77.24**	**80.89**

4.3 Ablation Studies

Contribution of Key Components. We investigate the individual contributions from two core components: preliminary filtering and adaptive combination. To evaluate preliminary filtering, we bypass it and perform a direct top-K selection based on the score over the entire patch set. To assess the adaptive combination, we replace it with a static $\lambda = 0.5$, assigning equal importance to textual and patch-based logits. Experiments in Tab. 3 are conducted across FGVCAircraft, StanfordCars, UCF101, and EuroSAT in a 4-shot setting.

Our analysis reveals that the impact of these components is dataset-dependent, implying the semantic ambiguity from their class names. When the datasets have semantically informative class names (e.g., UCF101 and EuroSAT), the preliminary filtering is more impactful. In these scenarios, classification often centers on localized entities instead of complex backgrounds. Filtering the spatial noise can help CDPS to focus on the primary semantic regions. Conversely, if the datasets contain technical class names with limited semantic textual prompts (e.g., FGVCAircraft and StanfordCars), the adaptive combination contributes more performance gains. This empirical result suggests that our adaptive λ enables the classifier to rely more on visual features when textual prompts are less discriminative. By integrating these complementary strategies, the hybrid classifier achieves an adaptive balance between textual prompts and visual discrimination, ensuring robust performance across diverse downstream datasets.

The selection of λ. The λ is determined through a cross-validation grid search on available few-shot samples. We further validate that this simple strategy provides an interpretable and near-optimal selection of λ.

The selected λ successfully balances the relative reliability of textual prompts and visual discrimination. We report the λ values across four representative datasets and different shot settings in Tab. 4. For datasets involving technical names (e.g., FGVCAircraft) or those exhibiting highly localized visual patterns (e.g., Flowers102), the selected λ is lower, emphasizing on the discriminative patches from CDPS. In contrast, textual prompts become more reliable for datasets with high semantic class names (e.g., Food101 and UCF101), consistent with the observed higher λ value. Furthermore, as the number of shots increases, the value of λ tends to decrease. This trend suggests that with a larger few-shot

Table 4: The λ selected by the hybrid classifier.

Shots	Air	Flower	Food	UCF
2	0.5	0.5	0.9	0.7
4	0.6	0.5	0.8	0.7
8	0.4	0.4	0.7	0.5
16	0.1	0.3	0.6	0.4

Table 5: Comparison between emperically optimal λ and the selected λ. We present the absolute difference between them and the accuracy improvement (%) by using emperically optimal λ.

Shots	2	4	8	16	Avg.
λ Difference	0.09	0.07	0.13	0.11	0.10
Acc. Improv.	0.81	0.47	0.28	0.32	0.47

set, the patch-based logits become increasingly reliable, allowing the classifier to shift its confidence toward localized visual features over textual prompts.

The selected λ is close to the empirically optimal. We compare the cross-validated λ against an empirically optimal one, obtained through an exhaustive grid search (step size 0.01) directly on the test set. This configuration serves as the empirical performance upper bound for our method. As reported in Tab. 5, the gap between our cross-validated λ and the optimal is marginal across all shots. This proximity suggests that our selection strategy consistently identifies a near-optimal weighting parameter. While a slight performance degradation is observed, it remains within a negligible value, only 0.47% difference on average.

These two findings validate our intuition regarding the role of λ and verify the efficacy of the leave-one-out cross-validation strategy in adapting the hybrid classifier to diverse settings and downstream datasets.

Sensitivity Analysis of Hyperparameters. The construction of CDPS requires 5 hyperparameters: the attention percentile ρ, the connected region numbers N_{cp}, the potential misclassification subset size N_c, the weighting parameter ω for inter-class penalty, and the final discriminative patch number K. We show the ablation on ρ, N_{cp} and K for FGVCAircraft in Fig. 6, with remaining ablation results provided in Appendix.

As illustrated in Fig. 6, performance fluctuations in the 2-shot setting are more pronounced due to the limited number of few-shot images. When more images are available, we observe that both low and high attention percentile (ρ) or number of connected regions (N_{cp}) cause a performance degradation. While considering the number of discriminative patches K, increasing the number generally yields higher accuracy as CDPS captures more comprehensive visual patches. However, this results in increased computational overhead and the improvement becomes limited with $K = 30$ due to the inclusion of non-discriminative patches. We identify $K = 20$ as the optimal point, providing a robust balance between classification performance and inference efficiency. These results demonstrate that while CDPS is robust across a range of values, moderate constraints on parameter selection are essential to maintain high-fidelity semantic focus.

Ablation on Intra-/Inter-class Similarity. We demonstrate the importance of both terms in Eq. (3) through the results in Tab. 6. As shown, using either term alone leads to degraded performance, whereas combining them achieves the best results. This confirms the necessity of jointly considering intra-class consistency and inter-class ambiguity when computing the patch score.

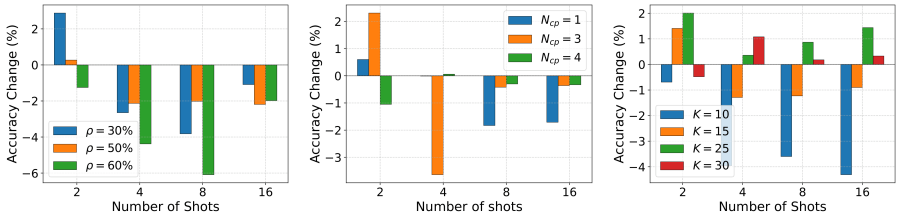

Fig. 6: Sensitivity analysis of key hyperparameters under different shots settings. We report the accuracy change relative to our default configuration ($\rho = 40\%$, $N_{cp} = 2$, $K = 20$). Left: Impact of the attention percentile ρ. Middle: Impact of the connected region number N_{cp}. Right: Impact of the discriminative patch number K.

Table 6: Ablation of intra-/inter-class similarity; values show accuracy drop.

Shots	w/o	DTD	Air	Pet	Bird	Avg. Drop
4	inter	2.06	5.15	2.27	2.78	3.07
	intra	1.40	6.98	3.73	6.57	4.67
8	inter	2.06	3.22	1.62	3.62	2.63
	intra	1.36	6.05	3.65	5.81	4.22

Table 7: Time and memory comparison on ImageNet. Time includes the preparation and inference.

Shots	Metric	TipA	GDA	TIMO	Ours
8	Time(s)	799.7	317.2	305.7	**80.6**
	Mem.(G)	**16.0**	16.2	18.2	29.0
16	Time(s)	876.6	512.6	522.5	**120.8**
	Mem.(G)	**17.7**	18.4	18.5	31.4

Time and Memory Overhead. Tab. 7 compares the preparation time, including CDPS construction for our method and augmentation for GDA and TIMO, as well as the inference time on ImageNet. For GDA and TIMO, inference accounts for only about 1% of the total runtime. In contrast, our set-to-set similarity computation compromises approximately 70% of the total runtime. Nevertheless, this computation is highly parallelizable. By leveraging additional memory, our approach achieves faster overall runtime.

5 Conclusion

We addressed the performance degradation of CLIP when the class names do not provide sufficient semantic information by identifying discriminative patches from few-shot samples. CDPS enables a training-free hybrid classifier combining textual prompts and local patch features. We demonstrated that CDPS improves few-shot performance and complements textual prompts, especially in scenarios where class names are ambiguous. Our classifier is limited to employing individual patches, does not model spatial layouts of patches, though some contextual information is encoded in individual patches due to the global attention mechanisms in ViT. Finally, our method could fail when the attention map is unable to locate the correct region, leading to non-discriminative patch selections.

6 Acknowledgement

This project was supported by the EPSRC Visual AI grant EP/T028572/1 and NSFC-62306046.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: ECCV (2014)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS **33**, 1877–1901 (2020)
3. Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.Z., Guo, J.: Your" flamingo" is my" bird": Fine-grained, or not. In: CVPR. pp. 11476–11485 (2021)
4. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: Plot: Prompt learning with optimal transport for vision-language models. arXiv preprint arXiv:2210.01253 (2022)
5. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR. pp. 2818–2829 (2023)
6. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPRW. pp. 178–178. IEEE (2004)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. NeurIPS **26** (2013)
11. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: CVPR. pp. 4438–4446 (2017)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. pp. 580–587 (2014)
13. Gondal, M.W., Gast, J., Ruiz, I.A., Droste, R., Macri, T., Kumar, S., Staudigl, L.: Domain aligned clip for few-shot classification. In: WACV. pp. 5721–5730 (2024)
14. He, X., Peng, Y., Zhao, J.: Fast fine-grained image classification via weakly supervised discriminative localization. IEEE Transactions on Circuits and Systems for Video Technology **29**(5), 1394–1407 (2018)
15. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
16. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
17. Jia, C., Yang, Y., Xia, R., Wu, Y., Girshick, R., Kulkarni, V., He, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)

18. Johnson, K.E., Eilers, A.T.: Effects of knowledge and development on subordinate level categorization. Cognitive Development **13**(4), 515–545 (1998)
19. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR. pp. 19113–19122 (2023)
20. Krause, J., Deng, J., Stark, M., Fei-Fei, L.: Collecting a large-scale dataset of fine-grained cars (2013)
21. Li, S., Liu, F., Hao, Z., Wang, X., Li, L., Liu, X., Chen, P., Ma, W.: Logits deconfusion with clip for few-shot learning. In: CVPR. pp. 25411–25421 (2025)
22. Li, Y., Guo, J., Qi, L., Li, W., Shi, Y.: Text and image are mutually beneficial: Enhancing training-free few-shot classification with clip. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5039–5047 (2025)
23. Li, Y., Su, Y., Goodge, A., Jia, K., Xu, X.: Efficient and context-aware label propagation for zero-/few-shot training-free adaptation of vision-language model. arXiv preprint arXiv:2412.18303 (2024)
24. Lin, D., Shen, X., Lu, C., Jia, J.: Deep lac: Deep localization, alignment and classification for fine-grained recognition. In: CVPR. pp. 1666–1674 (2015)
25. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: CVPR. pp. 15305–15314 (2023)
26. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys **55**(9), 1–35 (2023)
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
28. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
29. Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K., O'Connor, N.E.: Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In: ICCV. pp. 262–271 (2023)
30. Menon, S., Vondrick, C.: Visual classification via description from large language models. arXiv preprint arXiv:2210.07183 (2022)
31. Mirza, M.J., Karlinsky, L., Lin, W., Doveh, S., Micorek, J., Kozinski, M., Kuehne, H., Possegger, H.: Meta-prompting for automating zero-shot visual recognition with llms. In: ECCV. pp. 370–387. Springer (2024)
32. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (2008)
33. Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., Kong, S.: The neglected tails in vision-language models. In: CVPR. pp. 12988–12997 (2024)
34. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: CVPR (2012)
35. Pei, G., Chen, T., Wang, Y., Cai, X., Shu, X., Zhou, T., Yao, Y.: Seeing what matters: Empowering clip with patch generation-to-selection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24862–24872 (2025)
36. Peng, Y., He, X., Zhao, J.: Object-part attention model for fine-grained image classification. IEEE TIP **27**(3), 1487–1500 (2017)
37. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: ICCV. pp. 15691–15701 (2023)
38. Qian, Q., Hu, J.: Online zero-shot classification with clip. In: ECCV. pp. 462–477. Springer (2024)

39. Qu, X., Gou, G., Zhuang, J., Yu, J., Song, K., Wang, Q., Li, Y., Xiong, G.: Proapo: Progressively automatic prompt optimization for visual classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25145–25155 (2025)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
41. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022)
42. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: CVPR. pp. 6545–6554 (2023)
43. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: ICML. pp. 5389–5400. PMLR (2019)
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS. vol. 28 (2015)
45. Ren, Z., Su, Y., Liu, X.: Chatgpt-powered hierarchical comparisons for image classification. NeurIPS **36**, 69706–69718 (2023)
46. Sanghi, A., Chu, H., Lambourne, J.G., Wang, Y., Cheng, C.Y., Fumero, M., Malekshan, K.R.: Clip-forge: Towards zero-shot text-to-shape generation. In: CVPR. pp. 18603–18613 (2022)
47. Silva-Rodriguez, J., Hajimiri, S., Ben Ayed, I., Dolz, J.: A closer look at the few-shot adaptation of large vision-language models. In: CVPR. pp. 23681–23690 (2024)
48. Smith, M.: Text, speech, and vision for video segmentation: the informedia project. In: AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision (1995)
49. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
50. Srivastava, S., Labutov, I., Mitchell, T.: Joint concept learning and semantic parsing from natural language explanations. In: Proceedings of the 2017 conference on empirical methods in natural language processing. pp. 1527–1536 (2017)
51. Su, J.C., Maji, S.: The semi-supervised inaturalist-aves challenge at fgvc7 workshop. arXiv preprint arXiv:2103.06937 (2021)
52. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: ECCV. pp. 805–821 (2018)
53. Tang, L., Wertheimer, D., Hariharan, B.: Revisiting pose-normalization for fine-grained few-shot recognition. In: CVPR. pp. 14352–14361 (2020)
54. Tang, M., Wang, Z., Liu, Z., Rao, F., Li, D., Li, X.: Clip4caption: Clip for video caption. In: Proceedings of the 29th ACM international conference on multimedia. pp. 4858–4862 (2021)
55. Tsutsui, S., Fu, Y., Crandall, D.: Meta-reinforced synthetic data for one-shot fine-grained visual recognition. NeurIPS **32** (2019)
56. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. NeurIPS **32** (2019)
57. Wang, Z., Liang, J., Sheng, L., He, R., Wang, Z., Tan, T.: A hard-to-beat baseline for training-free clip-based adaptation. arXiv preprint arXiv:2402.04087 (2024)
58. Wang, Z., Liu, W., He, Q., Wu, X., Yi, Z.: Clip-gen: Language-free training of a text-to-image generator with clip. arXiv preprint arXiv:2203.00386 (2022)
59. Wei, X.S., Song, Y.Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J., Belongie, S.: Fine-grained image analysis with deep learning: A survey. IEEE TPAMI **44**(12), 8927–8948 (2021)

60. Wei, X.S., Wang, P., Liu, L., Shen, C., Wu, J.: Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. IEEE TIP **28**(12), 6116–6125 (2019)
61. Wei, X.S., Wu, J., Cui, Q.: Deep learning for fine-grained image analysis: A survey. arXiv preprint arXiv:1907.03069 (2019)
62. Wei, Z., Pan, Z., Owens, A.: Efficient vision-language pre-training by cluster masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26815–26825 (2024)
63. Weston, J.E.: Dialog-based language learning. NeurIPS **29** (2016)
64. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. pp. 3485–3492. IEEE (2010)
65. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: ECCV. pp. 420–435 (2018)
66. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Siglip: Multimodal embedding vision and language using sigmoid loss. arXiv preprint arXiv:2303.15343 (2023)
67. Zhang, D., Liu, F., Tang, Q.: Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In: ICCV. pp. 24677–24687 (2025)
68. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: ECCV. pp. 834–849. Springer (2014)
69. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
70. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: ICCV. pp. 5209–5217 (2017)
71. Zheng, H., Fu, J., Zha, Z.J., Luo, J., Mei, T.: Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE TIP **29**, 476–488 (2019)
72. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
73. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. In: IJCV (2022)
74. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: CVPR. pp. 11175–11185 (2023)