CLIP-driven Coarse-to-fine Semantic Guidance for Fine-grained Open-set Semi-supervised Learning

Xiaokun Li, Yaping Huang*, Qingji Guan*
Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University
22110102@bjtu.edu.cn, yphuang@bjtu.edu.cn, qjguan@bjtu.edu.cn

Abstract

Fine-grained open-set semi-supervised learning (OSSL) investigates a practical scenario where unlabeled data may contain fine-grained out-of-distribution (OOD) samples. Due to the subtle visual differences among in-distribution (ID) samples, as well as between ID and OOD samples, it is extremely challenging to separate the ID and OOD samples. Recent Vision-Language Models, such as CLIP, have shown excellent generalization capabilities. However, it tends to focus on general attributes, and thus is insufficient to distinguish the fine-grained details. To tackle the issues, in this paper, we propose a novel CLIP-driven coarse-to-fine semantic-guided framework, named CFSG-CLIP, to progressively focus on the distinctive fine-grained clues. Specifically, CFSG-CLIP comprises a coarse-guidance branch and a fine-guidance branch derived from the pre-trained CLIP model. In the coarse-guidance branch, we design a semantic filtering module to initially filter and highlight local visual features guided by cross-modality features. Then, in the fine-guidance branch, we further design a visual-semantic injection strategy, which embeds category-related visual cues into the visual encoder to further refine the local visual features. By the designed dual-guidance framework, local subtle cues are progressively discovered to distinct the subtle difference between ID and OOD samples. Extensive experiments demonstrate that CFSG-CLIP achieves competitive performance on multiple fine-grained datasets. The source code is available at https://github.com/LxxxxK/CFSG-CLIP.

1. Introduction

Open-set semi-supervised learning (OSSL) extends traditional semi-supervised learning by addressing more realistic scenarios, where the unlabeled data contains both in-distribution (ID) and out-of-distribution (OOD) samples drawn from different distributions [40]. These OOD samples cause the decision boundary of the model to deviate from the distribution of the original labeled data, thereby reducing its robustness in practical applications.

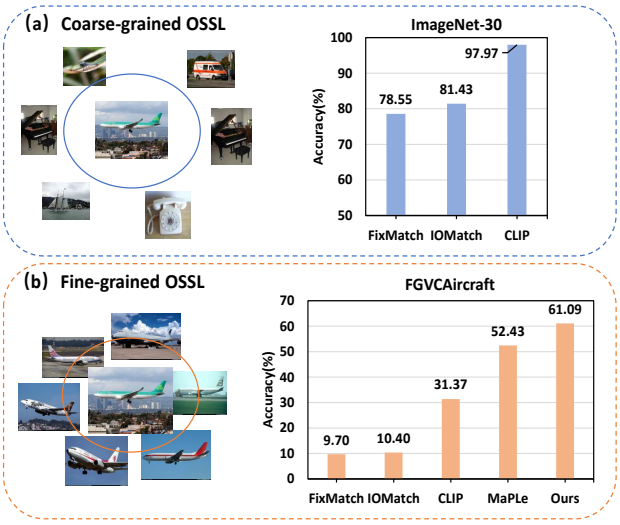

Figure 1. Comparison of open-set semi-supervised learning (OSSL) methods on coarse-grained and fine-grained classification tasks. (a) On the coarse-grained ImageNet-30 dataset, some outstanding methods (*e.g.*, FixMatch [30], IOMatch [22]) achieve excellent performance. CLIP [27] trained on the large-scale image-text pairs dataset achieves high performance as expected. (b) On the fine-grained FGVCAircraft dataset, the generalization capabilities of CLIP remain effective but limited, as it tends to focus on general attributes while failing to capture the fine-grained details.

Most existing OSSL methods [6, 11, 22, 28, 30, 35, 36, 40] verify their effectiveness mainly on coarse-grained datasets such as MNIST [20], CIFAR-10/100 [17] and ImageNet [5]. However, the more challenging and realistic fine-grained OSSL task remains under-explored. Large variance in the same subcategory and small variance between different subcategories lead to a reduction in the effectiveness of semi-supervised (SSL) methods that rely on pseudo-labeling [21, 25] or consistency regularization [19, 29, 33]. Even worse, as shown in Fig. 1, the tight decision bound-

* Corresponding author.

aries between samples prevent OSSL methods based on the maximum softmax probability [4, 9] and ad-hoc thresholds [10, 26, 28] from accurately separating unlabeled ID from OOD samples [35], resulting in low performance.

Recently, Vision-Language Models [12, 27, 38], such as CLIP [27], capture the relationships between images and texts by performing contrastive learning on large-scale image-text pairs, enabling CLIP to perform well on downstream tasks by directly matching class embeddings with image features. Subsequently, a series of variants [13, 44, 45] based on prompt learning effectively improve the adaptability of CLIP in downstream tasks through textual prompts or visual prompts. However, these methods aim to capture the global semantic features within the whole images, thus it is not sufficient to focus on the regions of interest to enable a fine-grained understanding.

For solving the above issues, in this paper, we propose a novel CLIP-driven coarse-to-fine semantic-guided framework, called CFSG-CLIP, by progressively filtering out irrelevant interference and focusing the distinctive fine-grained clues, which is crucial for accurately separating ID and OOD samples for fine-grained OSSL task. CFSG-CLIP comprises a coarse-guidance branch and a fine-guidance branch derived from the pre-trained CLIP model. In the coarse-guidance branch, we design a semantic filtering module to initially discard irrelevant regions and retain local visual features guided by cross-modality textual prompts features and global visual features. Then, in the fine-guidance branch, to further refine the local visual features, we propose a novel visual-semantic injection strategy that embeds category-related visual cues into the visual encoder, thus providing a more concise guidance to locate fine-grained clues among different subcategories. By the designed dual-guidance framework, local subtle cues are progressively discovered to distinct the subtle difference between fine-grained ID and OOD samples. We conduct comprehensive experiments on several fine-grained datasets to verify the effectiveness of our proposed framework.

In summary, the main contributions of this work are summarized as follows:

- We propose a novel CLIP-driven coarse-to-fine semantic-guided framework (CFSG-CLIP) for the fine-grained OSSL task. CFSG-CLIP initially explores category-relevant visual features (coarse-guidance branch) and further refines fine-grained local cues (fine-guidance branch), which can progressively exclude irrelevant parts and highlight distinctive fine-grained clues.
- In the coarse-guidance branch, we design a semantic filtering module to initially capture global and local visual features that are semantically relevant to the categories by cross-modality textual prompts and global visual features.
- In the fine-guidance branch, we propose a visual-semantic injection strategy to embed the category-related visual semantic cues, derived from the coarse-guidance branch, into the visual encoder of the fine-guidance branch, enabling the visual encoder to focus on more fine-grained details.
- Extensive experiments on multiple fine-grained datasets demonstrate the superiority of CFSG-CLIP over the state-of-the-art CLIP driven methods.

2. Related Work

Open-set Semi-supervised Learning (OSSL) is a more realistic and practical setting of SSL where the unlabeled data not only contains ID samples but also includes OOD samples. The key issue in OSSL is to reduce the negative impact of OOD samples in the unlabeled data while fully utilizing the unlabeled ID samples. Existing works handle OOD samples by discarding [4, 11, 40], weighting [8], or treating them as negative samples [28] to enhance model performance and robustness. IoMatch [22] and SCOMatch [36] categorize OOD samples into a new class and jointly optimize a (K+1)-way classification head with ID samples.

The mentioned efforts mainly focus on handling the impact of coarse-grained OOD samples under ideal assumptions, often overlooking the fine-grained attributes between ID and OOD samples. Their specially designed approaches make incorrect judgments when faced with tight feature distribution boundaries, resulting in reduced performance.

Contrastive Language-Image Pre-training (CLIP) [27] utilizes web-scale image-text pairs to align images and texts in the same feature space through contrastive learning, which can be transferred to various downstream tasks with different vocabularies. Based on CLIP, several studies employ visual adapters [7, 43], language prompt learning [3, 44, 45] and vision prompt learning [1, 13] to adapt the pre-trained CLIP model to downstream few-shot classification and OOD detection tasks. In addition, some studies [14, 37, 42] further propose multimodal prompt learning methods to fine-tune CLIP. Despite the significant performance gains achieved by these methods, they typically perform cross-modal interaction by computing the similarity between global features of each modality, thereby neglecting the fact that local features in CLIP contain a lot of information irrelevant to the category semantics. To capture the local features, several approaches, including FILIP [39], LoCoOp [24], GalLoP [18], learn fine-grained representations of image patches and language tokens through token-wise maximum similarity matching.

Unlike the mentioned efforts, we design a coarse-to-fine framework that initially selects local visual features using both textual features and global visual features, and then iteratively aggregates those local features as priors to further guide the visual encoder to focus on fine-grained clues.

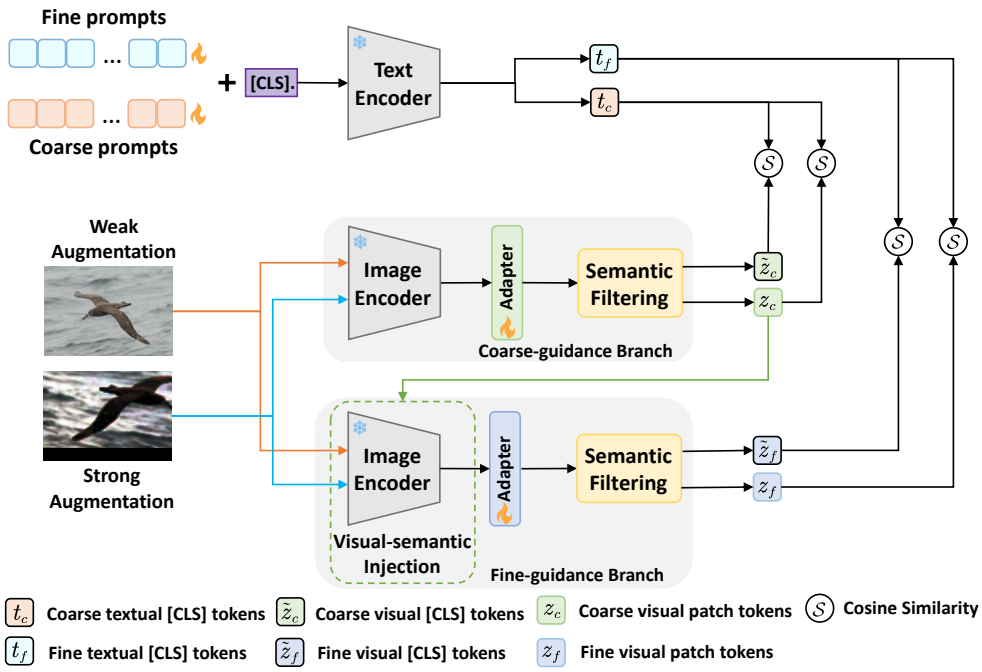

Figure 2. Overview of the proposed CFSG-CLIP framework. CFSG-CLIP is composed of a coarse-guidance branch and a fine-guidance branch based on the pre-trained CLIP model. In the coarse-guidance branch, we design a semantic filtering module to initially capture global and local visual features. In the fine-guidance branch, we design a visual-semantic injection strategy to embed category-related visual cues into the visual encoder for further refining the local visual features. For brevity, we omit the SSL training process.

3. Methodology

In this section, we present the proposed CLIP-driven coarse-to-fine semantic-guided framework (CFSG-CLIP) for fine-grained open-set semi-supervised learning (OSSL) task, where the key issue is to accurately separate ID and OOD samples. By filtering out irrelevant interference and generating fine-grained visual features, OOD samples can be effectively identified and removed.

3.1. Problem Setting

Fine-grained open-set semi-supervised learning (OSSL) focuses on a more complex scenario where fine-grained out-of-distribution (OOD) samples are mixed with in-distribution (ID) samples in the unlabeled data. We divide the training data into two datasets: a small set of labeled samples $\mathcal{D}^l = \{(x_1^l, y_1^l), (x_2^l, y_2^l), ..., (x_N^l, y_N^l)\}$ and a large set of unlabeled samples $\mathcal{D}^u = \{(x_1^u), (x_2^u), ..., (x_U^u)\}$. For an M-way classification problem, $y^l \in \{1, \ldots, M\}$ is a one-hot label over M classes. Unlike SSL tasks, where unlabeled data is assumed to only contain ID samples, the presence of OOD samples in OSSL means that the label of x_j^u is not guaranteed to be one of M known classes. Following [22], we adopt the same SSL strategy as FixMatch [30], where a batch of B labeled samples is used for supervised training, a batch of μB unlabeled samples performs self-training, with μ being a hyperparameter that controls the ratio of unlabeled to labeled samples in a batch.

3.2. Overall Architecture

Fig. 2 illustrates an overview of our proposed CFSG-CLIP framework, which comprises a coarse-guidance branch and a fine-guidance branch based on the pre-trained CLIP model, to progressively filter out irrelevant parts and focus on the distinctive fine-grained clues.

Specifically, given an unlabeled image with corresponding strong and weak augmentations, we first utilize a parameter-fixed visual encoder with a learnable adapter to extract image features. We then employ a semantic filtering module to initially capture image-level global visual features and patch-level local visual features, guided by coarse prompts in the coarse-guidance branch. In the fine-guidance branch, we apply a visual-semantic injection strategy, embedding the category-related local visual features extracted in the coarse-guidance branch into the visual encoder of the fine-guidance branch. This guides the visual encoder to focus more on fine-grained details. Compared to the coarse prompts, which capture the overall semantics of the image, fine prompts specifically align with the detailed features extracted by the fine-guidance branch.

3.3. Semantic Filtering Module

Recent vision-language models, such as CLIP, have shown their effectiveness in image recognition tasks by performing

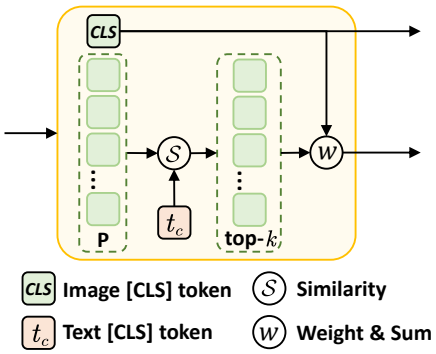

Figure 3. The architecture of Semantic Filtering module. It captures global and local visual features that are semantically relevant to the categories by cross-modality textual prompts and global visual features.

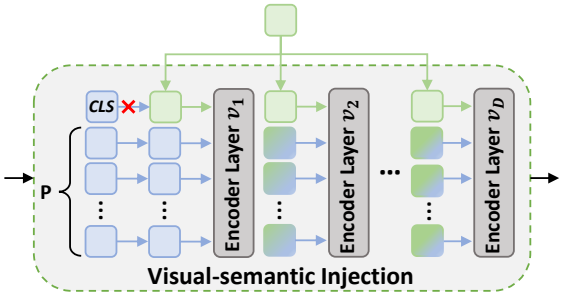

Figure 4. Illustration of visual-semantic injection strategy.

contrastive learning on large-scale image-text pairs. However, they tend to focus on general attributes, making them insufficient for capturing fine-grained details. To address this, we design a semantic filtering module that leverages both textual and visual features to obtain global and local visual features. Unlike previous studies [18, 24, 32] that only employ textual features to guide the visual encoder in screening category-related local visual features, we not only filter local regions in the coarse-guidance branch but also leverage image-level global visual features to further refine the local visual features in the fine-guidance branch. Thanks to the plug-and-play nature of the semantic filtering module, it can be easily utilized in both branches. The only difference lies in the cross-modality textual features used to filter patch-level features, which come from coarse and fine prompts, respectively.

We take the coarse-guidance branch as an example to introduce the module design, which is depicted in Fig. 3. Formally, given an image x, the visual encoder $\mathcal{V}$ with the corresponding adapter $\mathcal{A}_c$ (two linear layers with a ReLU) produces a set of normalized visual features $\mathcal{Z} = \{\tilde{z}_c, z_{c_1}, z_{c_2}, ..., z_{c_P}\} \in \mathbb{R}^{(P+1)\times d}$, where $\tilde{z}_c$ represents the coarse features of the [CLS] token and $z_{c_i} \in \{z_{c_1}, z_{c_2}, ..., z_{c_P}\}$ are the patch-level visual features of the image. P is the number of patches in one image and d is the dimension of the features. Inspired by CoOp [45], we apply a coarse prompt $p_c^m = \{v_1, v_2, ..., v_n, \mathcal{C}^m\}$ to align the global and local visual features, where $v_1, ..., v_n$ are n learnable vectors of the same dimension, and $\mathcal{C}^m$ is the word embedding of the m-th class name. As shown in Fig. 3, the textual features $t_c^m = \mathcal{T}(p_c^m)$ generated by the textual encoder $\mathcal{T}$ are used to calculate the similarity s_{c_i} with the i-th patch-level local visual feature as follows:

$$s_{c_i} = \text{sim}(z_{c_i}, t_c^m), \tag{1}$$

$$\mathcal{K} = \{i \in P : \text{rank}(s_{c_i}) \leq k\}, \tag{2}$$

where the $\mathcal{K}$ is the set of top-k patch-level visual features, and the $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

Although textual features can assist the visual encoder in filtering local visual semantic features, some irrelevant local features may still be incorrectly identified as category-related features in fine-grained scenes. Therefore, we further compute the similarity between the image-level global features and the patch-level local features to ensure that the local visual features are consistent with the overall semantic information of the image. The similarity weights $\mathcal{W} = \{w_{c_i}\}_{i=1}^k$ for the top-k patch tokens are computed as follows:

$$w_{c_i} = \frac{\exp(\text{sim}(z_{c_i}, \tilde{z}_c))}{\sum \exp(\text{sim}(z_{c_i}, \tilde{z}_c))}, i \in \mathcal{K}. \tag{3}$$

Then, the top-k patch-level visual features z_c are aggregated as follows:

$$z_c = \sum_{i \in \mathcal{K}} w_{c_i} z_{c_i}. \tag{4}$$

Finally, the coarse-guidance branch simultaneously outputs the image-level global features $\tilde{z}_c$ and the aggregated coarse patch-level features z_c.

3.4. Visual-semantic Injection

We initially capture image-level global and patch-level local features in the coarse-guidance branch, but the local visual features are insufficiently refined because the selection of local features is based on global textual and visual features. To guide the visual encoder to focus more on the fine-grained details of the image, we further design a visual-semantic injection strategy in the fine-guidance branch. Specifically, the local visual features extracted in the coarse-guidance branch replace the original learnable [CLS] token embedding of image and are injected into the specific layers of the visual encoder as visual-semantic priors. Through the self-attention mechanism, the visual semantic cues are propagated to patch-level visual features. Fig. 4 illustrates the overall flowchart of the proposed visual-semantic injection strategy.

Formally, we discard the original image [CLS] token embedding and inject the local visual features z_c into the first D blocks of the T transformer blocks within the visual encoder $\mathcal{V}$:

$$[\_, E_j] = \mathcal{V}_j([\text{proj}(z_c), E_{j-1}]) \quad j = 1, 2, \ldots, D, \quad (5)$$

$$[\tilde{z}_{f_j}, E_j] = \mathcal{V}_j([\text{proj}(z_c), E_{j-1}]) \quad j = D+1, \ldots, T, \quad (6)$$

where $[.,.]$ refers to the concatenation operation and $\text{proj}(\cdot)$ is a projection layer. T is the number of transformer blocks, and $E_j \in \mathbb{R}^{P \times d}$ is the fixed-size patch embeddings of the transformer blocks. Accordingly, the outputs of the T-th block are the refined [CLS] token features $\tilde{z}_f$ and refined patch-level features z_f.

Since the local visual features z_c from coarse-guidance branch contain visual-semantic priors, they guide the visual encoder in the fine-guidance branch to focus on fine-grained details within the image. Simultaneously, through the interaction of the self-attention mechanism, the patch-level local features further enhance the original features z_c. Afterward, we employ an adapter $\mathcal{A}_f$ to learn the refined visual features with category semantic priors and adopt the semantic filtering module to produce enhanced global features $\tilde{z}_f$ and the refined local visual features z_f. Unlike coarse prompts, we utilize fine prompts $t_f^m = \mathcal{T}(p_f^m)$ to specifically learn and guide more subtle visual features.

3.5. Dual-branch Training

To enhance the reliability of the fine-grained OSSL training process, we adopt the same SSL strategy from FixMatch [30] to optimize the coarse-guidance branch and fine-guidance branch, respectively.

Optimization of Coarse-guidance Branch. For a labeled sample x^l with the corresponding label y^l, we apply a weak transformation $\alpha(\cdot)$. For an unlabeled sample x^u, we apply both weak transformation $\alpha(\cdot)$ and strong transformation $\beta(\cdot)$. These augmentations are fed into the coarse-guidance branch to obtain the corresponding image-level global feature set $\{\tilde{z}_c^l, \tilde{z}_c^{u_w}, \tilde{z}_c^{u_s}\}$ and patch-level local visual feature set $\{z_c^l, z_c^{u_w}, z_c^{u_s}\}$. Then, we apply the features t_c^m of the coarse prompts to separately compute the prediction probabilities using these visual features as:

$$\tilde{p}_c^l = \frac{\exp(\text{sim}(\tilde{z}_c^l, t_c^m)/\tau)}{\sum_{m'} \exp(\text{sim}(\tilde{z}_c^l, t_c^{m'})/\tau)}, \quad (7)$$

$$p_c^l = \frac{\exp(\text{sim}(z_c^l, t_c^m)/\tau)}{\sum_{m'} \exp(\text{sim}(z_c^l, t_c^{m'})/\tau)}, \quad (8)$$

where the $\tilde{p}_c^l$ and p_c^l are the prediction probabilities for the global and local features of x^l, respectively. Similarly, we can obtain the prediction probabilities for global and local features of x^u with different augmentations, denoted as $\tilde{p}_c^{u_w}, \tilde{p}_c^{u_s}, p_c^{u_w}$ and $p_c^{u_s}$. The parameter τ represents the temperature of the softmax function.

In summary, the coarse-guidance branch is optimized using the standard cross-entropy loss $H(\cdot,\cdot)$, defined as:

$$\begin{aligned} L_c = H(y^l, \tilde{p}_c^l) + H(y^l, p_c^l) + \lambda_c (\mathcal{F}(x^u) H(\tilde{p}_c^{u_w}, \tilde{p}_c^{u_s}) \\ + \mathcal{F}(x^u) H(p_c^{u_w}, p_c^{u_s})), \end{aligned} \quad (9)$$

where the $\mathcal{F}(\cdot)$ is the filtering function defined as $\mathcal{F}(x^u) = \mathbb{1}(\max_m(p^u) > \delta)$, which selects the reliable known-class pseudo labels for strong augmentation samples. Here, δ is a predefined threshold and λ_c is the weight of the self-training loss.

Optimization of Fine-guidance Branch. We also feed the augmentation samples into the fine-guidance branch. Unlike aligning coarse local visual features with coarse prompts, we use the features t_f^m of fine-grained prompts to align refined global and local visual features $\{\tilde{z}_f^l, \tilde{z}_f^{u_w}, \tilde{z}_f^{u_s}\}$ and $\{z_f^l, z_f^{u_w}, z_f^{u_s}\}$. The prediction probabilities are calculated as follows:

$$\tilde{p}_f^l = \frac{\exp(\text{sim}(\tilde{z}_f^l, t_f^m)/\tau)}{\sum_{m'} \exp(\text{sim}(\tilde{z}_f^l, t_f^{m'})/\tau)}, \quad (10)$$

$$p_f^l = \frac{\exp(\text{sim}(z_f^l, t_f^m)/\tau)}{\sum_{m'} \exp(\text{sim}(z_f^l, t_f^{m'})/\tau)}, \quad (11)$$

where the $\tilde{p}_f^l$ and p_f^l denote the prediction probabilities for the global and local features of x^l. For brevity, the prediction probabilities for x^u are $\tilde{p}_f^{u_w}, \tilde{p}_f^{u_s}, p_f^{u_w}$ and $p_f^{u_s}$. The optimization objective of the fine-guidance branch is defined as follows:

$$\begin{aligned} L_f = H(y^l, \tilde{p}_f^l) + H(y^l, p_f^l) + \lambda_f (\mathcal{F}(\tilde{u}_f) H(\tilde{p}_f^{u_w}, \tilde{p}_f^{u_s}) \\ + \mathcal{F}(u_f) H(p_f^{u_w}, p_f^{u_s})), \end{aligned} \quad (12)$$

where $\mathcal{F}(\cdot)$ is the same as in Eq. 9, and λ_f is similar to λ_c in controlling the weighting of the self-training loss. We optimize the losses of the two branches separately during model training.

4. Experiments

4.1. Experiments Setting

Datasets. We evaluate the proposed CFSG-CLIP on the following five fine-grained benchmark datasets: CUB-200-2011 [34], Stanford Dogs [15], Stanford Cars [16], FGVCAircraft [23], and Semi-Aves [31].

Dataset Split. For the CUB-200-2011 and Semi-Aves datasets, we follow the data split strategy from [31] for our comparison experiments under the OSSL setting. Since datasets like Stanford Dogs, Stanford Cars, and FGVC Aircraft do not have publicly available OSSL splits, we apply the same split method as [31]. Specifically, we designate the *odd* classes as ID classes and the *even* classes

Table 1. Classification accuracy (%) for CLIP-based methods on four fine-grained benchmark datasets with varying labeled set sizes under the fine-grained OSSL setting. The results are presented as the mean with standard deviation over three runs using different random seeds.

Method	Stanford Dogs		Stanford Cars		CUB-200-2011		FGVCAircraft	
	5	20	5	20	5	20	5	20
CLIP [27]	79.25±0.00	79.25±0.00	75.97±0.00	75.97±0.00	66.00±0.00	66.00±0.00	31.37±0.00	31.37±0.00
CLIP-LORA [41]	83.81±0.37	84.31±0.27	82.71±0.56	82.45±1.30	70.95±0.85	73.40±0.75	40.89±1.73	42.37±0.71
CLIP-Adapter [7]	82.91±0.25	86.02±0.27	84.31±0.02	87.13±0.28	80.03±0.29	84.77±1.02	47.77±0.90	55.79±0.73
CoOp [45]	83.01±0.26	85.68±0.37	85.45±0.31	87.64±0.46	80.10±0.29	85.40±0.37	45.39±0.96	55.43±0.30
LoCoOp [24]	83.08±0.25	86.26±0.11	84.10±0.72	87.83±0.66	79.27±0.45	85.63±0.54	45.53±1.36	54.67±1.59
PLOT [3]	84.46±0.07	87.11±0.09	86.28±0.30	88.59±0.45	81.43±0.66	87.20±0.14	49.59±0.37	58.25±0.93
MaPLe [14]	**85.64±0.15**	87.64±0.20	88.16±0.25	90.34±0.25	83.30±0.33	88.77±0.21	52.43±0.47	64.33±1.21
Ours	85.48±0.21	**89.42±0.16**	**90.38±0.09**	**93.08±0.08**	**84.73±0.17**	**91.75±0.24**	**61.09±0.27**	**73.56±0.58**

Table 2. Open-set classification balanced accuracy (%) for CLIP-based methods on four fine-grained benchmark datasets with varying labeled set sizes under the fine-grained OSSL setting.

Method	Stanford Dogs		Stanford Cars		CUB-200-2011		FGVCAircraft	
	5	20	5	20	5	20	5	20
CLIP [27]	77.17±0.00	77.17±0.00	75.70±0.00	75.70±0.00	64.10±0.00	64.10±0.00	31.08±0.00	31.08±0.00
CLIP-LORA [41]	82.34±0.57	82.67±0.36	82.10±0.56	81.03±1.24	70.52±1.34	72.20±0.45	40.10±1.69	41.57±0.69
CLIP-Adapter [7]	81.63±0.14	84.36±0.25	83.65±0.04	86.50±0.25	81.36±0.60	85.20±0.95	46.87±0.88	53.52±1.35
CoOp [45]	81.65±0.20	84.25±0.41	84.63±0.36	86.84±0.42	81.04±0.06	84.92±0.51	44.55±0.94	54.35±0.30
LoCoOp [24]	81.78±0.23	84.68±0.20	83.25±0.76	87.17±0.64	80.45±0.86	85.46±0.32	44.67±1.35	53.60±1.57
PLOT [3]	82.95±0.07	85.54±0.11	85.58±0.32	87.83±0.34	83.32±0.31	87.16±0.17	48.68±0.37	57.13±0.92
MaPLe [14]	**84.09±0.19**	86.02±0.25	87.43±0.27	89.48±0.25	84.72±0.53	88.66±0.60	51.79±0.43	63.12±1.19
Ours	84.02±0.15	**87.77±0.19**	**89.65±0.05**	**92.34±0.10**	**86.46±0.25**	**90.92±0.24**	**59.92±0.26**	**72.13±0.57**

as OOD classes to minimize domain mismatch between ID and OOD categories. For each ID class, we randomly select 5 or 20 samples from the training set as labeled data, using the remaining training samples as unlabeled data. In the test set, we use the *odd* classes as ID classes, with all samples from each class used to evaluate the model's performance.

4.2. Implementation Details

We adopt a pretrained ViT-B/16 CLIP model as the backbone for both the visual and textual encoders in our model. For the proposed CFSG-CLIP, we set the learnable textual prompt length to 16, following the setting in CoOp [45]. All models are trained for 50 epochs with a batch size of 32 and a learning rate of 0.002 using the SGD optimizer on a single RTX 3090 GPU. The weights λ_c and λ_f of the self-training losses are both set to 1. We use $\delta = 0.95$ to filter out reliable pseudo labels for unlabeled samples and $\mu = 1$ to control the relative sizes of ID and OOD samples in each batch across all experiments. To ensure a fair comparison, we implement CLIP-Adapter [7], CoOp [45], LoCoOp [24], PLOT [3], and MaPLe [14] methods based on the SSL technique from FixMatch, while also employing exponential moving average (EMA) and a distribution alignment strategy [2], following IOMatch [22].

4.3. Comparison with State-of-the-arts

The proposed CFSG-CLIP is compared with state-of-the-art vision-language-model-based feature adaption methods, including CLIP-Adapter [7]), few-shot out-of-distribution detection (LoCoOp [24]), and prompt learning methods (CoOp [45], PLOT [3], MaPLe [14]) for addressing the fine-grained OSSL task. Since these methods are not directly applicable to our task, we re-implement them using the same training strategies of our method.

Table 1 summarizes the comparison results across four datasets with varying numbers of labeled samples (5 and 20). Since the pre-trained CLIP [27] can be used for prediction directly without any fine-tuning, it is considered the baseline method in our experiments. The results of CLIP demonstrate that its limitations in downstream tasks involving fine-grained attributes, as it tends to focus on general features and therefore struggles with fine-grained tasks. We would like to emphasize that CFSG-CLIP outperforms other methods by large margins on three datasets

Table 3. Classification accuracy (%) on the Semi-Aves dataset is reported for two settings: unlabeled data with ID samples U_{in}, and unlabeled data with a mix of ID and OOD samples $U_{in} + U_{out}$.

Method	U_{in}	$U_{in} + U_{out}$
CLIP [27]	10.05±0.00	10.05±0.00
CLIP-Adapter [7]	50.16±1.55	50.10±0.44
CoOp [45]	47.67±0.97	47.93±0.76
LoCoOp [24]	48.04±0.96	47.30±1.30
PLOT [3]	53.68±0.35	54.72±0.96
MaPLe [14]	60.39±0.10	59.68±0.64
Ours	**65.63**±0.27	**63.69**±0.74

Table 4. Ablation studies on the Stanford cars and FGVCAircraft datasets. The '$\mathcal{A}$' stands for adapter, 'SFM' denotes semantic filtering module, 'C' means coarse-guidance branch, 'F' means fine-guidance branch, and 'VSI' is visual-semantic injection strategy. The results are reported based on a single run with seed 1.

Method	Stanford Cars	FGVCAircraft
CoOp+$\mathcal{A}$	86.31	50.42
+SFM (C)	89.15	54.65
+SFM (C)+SFM (F)	90.03	58.01
+SFM (C)+SFM (F)+VSI	90.28	61.07

(*i.e.*, Standford Cars, CUB-200-2011, FGVCAircraft). Especially on the most challenging FGVCAircraft dataset, we achieve a significant performance gain in two cases (8.66% and 9.23% in 5 and 20 labeled samples) compared to the second best method, MaPLe. While we are only slightly behind MaPLe by 0.16% in the setting of 5 labeled samples per class on the Stanford Dogs dataset, we achieve a 1.78% advantage in the setting of 20 labeled samples. Following [22], we also evaluate classification balanced accuracy on open-set test data including both seen and unseen classes as shown in Table 2. Obviously, CFSG-CLIP still has a relatively large performance advantage.

Additionally, we evaluate CFSG-CLIP on the Semi-Aves dataset to assess its effectiveness and generalization capability. We conduct experiments under two settings: one where the unlabeled data contains only ID samples and another where it includes a mix of both ID and OOD samples. Our method achieves significant improvements of 5.24% and 4.01% in these settings, as shown in Table 3. The leading performance highlights the superior capability of CFSG-CLIP in fine-grained OSSL tasks.

4.4. Ablation Study

To verify the effectiveness of CFSG-CLIP, we conduct ablation studies on the Stanford Cars and FGVCAircraft datasets as shown in Table 4 and Table 5. The components of CFSG-CLIP mainly contain the semantic filtering module and the visual-semantic injection strategy. Additionally, we decompose the semantic filtering module to analyze the impact of each component. We use CoOp [45] with an adapter as the baseline for all our modules.

Table 5. Ablation studies for semantic filtering module. The results are reported based on a single run with seed 1.

Method	Stanford Cars	FGVCAircraft
CoOp+$\mathcal{A}$	86.31	50.42
Textual Filtering	88.65	53.57
Visual Weighting	89.15	54.65

Effect of semantic filtering module. We design a semantic filtering module to capture both global and local visual features in the coarse-guidance and fine-guidance branches. In the coarse-guidance branch, semantic filtering module is guided by coarse prompts to initially filter local visual features, resulting in improvements of 2.84% and 4.23% (2nd row). Moreover, applying the semantic filtering module in both branches further enhances performance, achieving a 3.36% improvement on the FGVCAircraft (3rd row).

Effect of visual-semantic injection strategy. We propose a visual-semantic injection strategy to embed the category-related visual-semantic cues, extracted by the coarse-guidance branch, into the visual encoder of the fine-guidance branch. This enhances the model's ability to focus on more fine-grained features. The ablation results show that the visual-semantic injection strategy further improves performance by 0.24% and 3.06% (4th row), respectively.

Effect of global visual feature weighting. In our proposed semantic filtering module, we design a global visual feature weighting scheme to obtain more distinctive features. Since the alignment between text and image is based on global features, using only textual features to filter the local features of the images is insufficient. To verify this, we compare simple textual filtering with our proposed global weighting scheme, and the results are presented in Table 5. We observe that the global visual feature weighting further refines the local visual features filtered by cross-modality textual features, leading to performance gains of 0.5% and 1.08% on two datasets.

4.5. Further Analysis

The depth of visual-semantic injection. Fig. 6 (left) illustrates the effect of injection depth for CFSG-CLIP. Generally, embedding more visual-semantic priors into the visual encoder improves model performance. However, since the visual-semantic priors extracted by the coarse-guidance branch are coarse-grained local visual features, injecting them into deeper layers causes the visual encoder to focus on category-irrelevant details. To balance guidance and refinement of local features, we embed these priors into the middle layers (layers 1 to 7), yielding the best performance.

The number of k. In our method, we apply the seman-

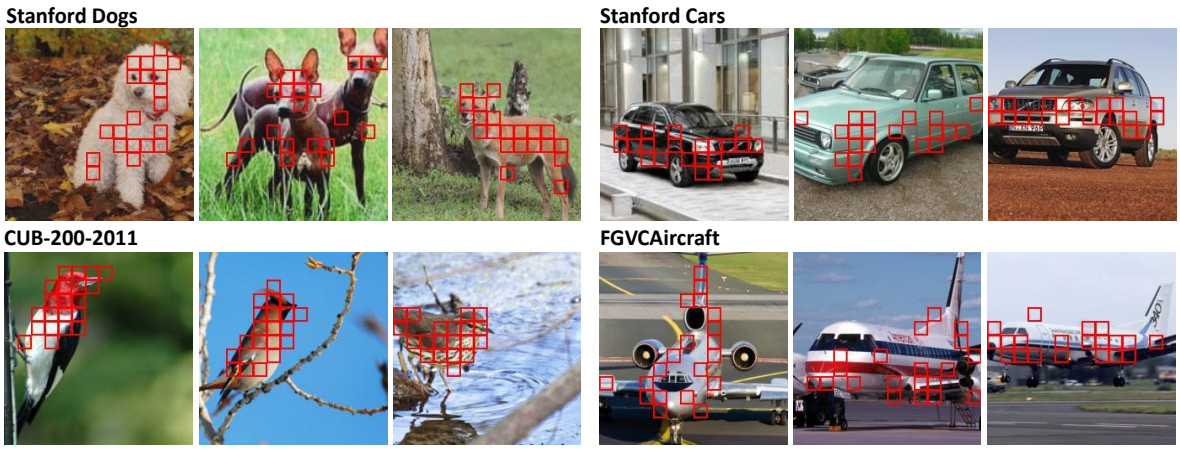

Figure 5. Visualization of patch-tokens extracted by semantic filtering module. We find that the semantic filtering module can correctly extract local visual regions on different fine-grained datasets.

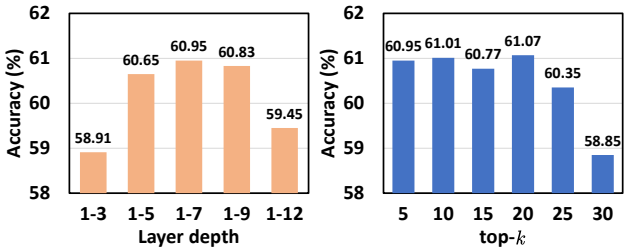

Figure 6. Ablation studies on injection depth (*left*) and top-k (*right*) on the FGVCAircraf datasets. The results are reported based on a single run with seed 1.

tic filtering module in both the coarse-guidance and fine-guidance branches, where the top-k patch tokens are selected to represent category-related local visual features. As shown in Fig. 6 (*right*), our method achieves relatively stable performance within a certain range of k. However, if k is too large, it might introduce unnecessary background noise, leading to a drop in performance. Specifically, CFSG-CLIP achieves the best performance with $k = 20$. Additionally, we visualize the top 20 image patches extracted by the semantic filtering module on four datasets as shown in Fig. 5. It can be clearly observed that even on the more challenging CUB-200-2011 dataset, the selected patches accurately locate the semantic parts of the objects.

The operations of visual-semantic injection. We design a visual-semantic injection strategy that embeds the local visual features extracted by the coarse-guidance branch into the visual encoder of the fine-guidance branch to guide the model's focus on fine-grained clues. To achieve satisfying performance, we choose two different operations, concatenation and replacement, to inject visual-semantic priors into the visual encoder. As shown in Table 6, the performance of replacement is better than that of concatenation. The result demonstrates that, compared to [CLS] tokens that learn image-level global features, our visual-semantic guidance can focus on more local category-relevant features.

Table 6. Evaluation results for different operations of employing local visual features. The results are reported based on a single run with seed 1.

Method	Stanford Cars	FGVCAircraft
Concatenate	89.99	60.47
Replace	90.28	61.07

5. Conclusion

In this paper, we introduce a new framework called CLIP-driven Coarse-to-Fine Semantic Guidance (CFSG-CLIP), designed to progressively filter out category-irrelevant features while focusing on distinguishing fine-grained details for the OSSL task. CFSG-CLIP consists of two branches. In the coarse-guidance branch, we incorporate a semantic filtering module that uses cross-modality textual prompts to capture initial global and local visual features relevant to the category. In the fine-guidance branch, we implement a visual-semantic injection strategy that injects category-relevant visual cues into the visual encoder to further refine local visual features. Thanks to our dual-guidance architecture, subtle local cues are progressively identified, enabling the model to differentiate fine-grained details between ID and OOD samples. Extensive experiments validate the effectiveness of our method.

Limitation. Although our method is specifically designed for fine-grained datasets, we believe it can also be applied to coarse-grained datasets, though this requires further validation. Additionally, its potential for more rare fine-grained categories is still a future direction to explore.

Acknowledgements. This work is supported by National Natural Science Foundation of China (62271042, 62376021, 62302032, 62301220).

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 6

[3] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 2, 6, 7

[4] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3569–3576, 2020. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. Ssb: Simple but strong baseline for boosting performance of open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16068–16078, 2023. 1

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 2, 6, 7

[8] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International conference on machine learning*, pages 3897–3906. PMLR, 2020. 2

[9] Lu Han, Han-Jia Ye, and De-Chuan Zhan. On pseudo-labeling for class-mismatch semi-supervised learning. *arXiv preprint arXiv:2301.06010*, 2023. 2

[10] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14585–14594, 2022. 2

[11] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8310–8319, 2021. 1, 2

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[13] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2

[14] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2, 6, 7

[15] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, 2011. 5

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[18] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. *arXiv preprint arXiv:2407.01400*, 2024. 2, 4

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1

[20] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 1

[21] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 1

[22] Zekun Li, Lei Qi, Yinghuan Shi, and Yang Gao. Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15870–15879, 2023. 1, 2, 3, 6, 7

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

[24] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4, 6, 7

[25] Islam Nassar, Samitha Herath, Ehsan Abbasnejad, Wray Buntine, and Gholamreza Haffari. All labels are not created equal: Enhancing semi-supervision via label grouping and co-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7241–7250, 2021. 1

[26] Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data. In *European Conference on Computer Vision*, pages 134–149. Springer, 2022. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6, 7

[28] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 34:25956–25967, 2021. 1, 2

[29] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 1

[30] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 3, 5

[31] Jong-Chyi Su, Zezhou Cheng, and Subhransu Maji. A realistic evaluation of semi-supervised learning for fine-grained classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12966–12975, 2021. 5

[32] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022. 4

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[35] Erik Wallin, Lennart Svensson, Fredrik Kahl, and Lars Hammarstrand. Prosub: Probabilistic open-set semi-supervised learning with subspace-based out-of-distribution detection. *arXiv preprint arXiv:2407.11735*, 2024. 1, 2

[36] Zerun Wang, Liuyu Xiang, Lang Huang, Jiafeng Mao, Ling Xiao, and Toshihiko Yamasaki. Scomatch: Alleviating overtrusting in open-set semi-supervised learning. *arXiv preprint arXiv:2409.17512*, 2024. 1, 2

[37] Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16076–16084, 2024. 2

[38] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 2

[39] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

[40] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 438–454. Springer, 2020. 1, 2

[41] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1593–1603, 2024. 6

[42] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 2

[43] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 2

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 2

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 4, 6, 7