Holistic Optimal Label Selection for Robust Prompt Learning under Partial Labels

Yaqi Zhao, Haoliang Sun*, Yating Wang, Yongshun Gong, and Yilong Yin

School of Software, Shandong University, Jinan, China

Abstract. Prompt learning has gained significant attention as a parameter-efficient approach for adapting large pre-trained vision-language models to downstream tasks. However, when only partial labels are available, its performance is often limited by label ambiguity and insufficient supervisory information. To address this issue, we propose *Holistic Optimal Label Selection* (**HopS**), leveraging the generalization ability of pre-trained feature encoders through two complementary strategies. First, we design a *local* density-based filter that selects the top frequent labels from the nearest neighbors' candidate sets and uses the softmax scores to identify the most plausible label, capturing structural regularities in the feature space. Second, we introduce a *global* selection objective based on optimal transport that maps the uniform sampling distribution to the candidate label distributions across a batch. By minimizing the expected transport cost, it can determine the most likely label assignments. These two strategies work together to provide robust label selection from both local and global perspectives. Extensive experiments on eight benchmark datasets show that HopS consistently improves performance under partial supervision and outperforms all baselines. Those results highlight the merit of holistic label selection and offer a practical solution for prompt learning in weakly supervised settings. The code is available at https://github.com/Qizhoay/HopS.

Keywords: Partial label learning · Prompt learning · Optimal transport

1 Introduction

Large pre-trained vision-language models (VLMs), such as CLIP [29], have demonstrated remarkable capabilities across a wide range of downstream tasks [9, 21, 30, 32]. Among various fine-tuning paradigms [20], prompt tuning has emerged as an efficient alternative to full model fine-tuning, offering the advantage of adapting large models with minimal additional parameters [54]. By optimizing a small set of learnable prompts while keeping the backbone frozen, prompt learning preserves the generalization strength of the pre-trained model and reduces training costs—making it particularly attractive in resource-constrained or few-shot learning scenarios [51].

* Corresponding author.

The effectiveness of prompt learning can be significantly compromised in weakly supervised settings [52], particularly those involving partial labels [22], where only a subset of candidate labels is provided per instance without explicit ground-truth annotations. This setting is common in real-world applications such as webly supervised learning [28], human-in-the-loop labeling [42], and open-world recognition [41]. The main challenge in such scenarios lies in the label ambiguity, which hampers supervision and often degenerate the generalization performance of learning algorithms, especially for prompt learning with few-shot instances [55].

To address the challenges of partial label learning (PLL), state-of-the-art methods typically select a plausible label from the candidate set using representations learned through contrastive learning (e.g., PICO [36]). However, such approaches are not directly applicable to pre-trained models, as fine-tuning their vision encoders may compromise their zero-shot capabilities [17]. *This presents a fundamental challenge in filling the gap between reliable label selection with frozen vision encoders and effective prompt learning with label disambiguation.*

To overcome this limitation, we propose a holistic label selection strategy that fully leverages the representational generalization of pre-trained encoders for robust prompt learning. Specifically, HopS includes two complementary selection mechanisms. The first is a **Local Density-based Filter** (LDF), which estimates label frequency within a k-nearest neighbor (k-NN) structure in the image feature space and selects the most frequent labels in the neighborhood for a subset, capturing local semantic regularities in a non-parametric manner. LDF effectively exploits the zero-shot capabilities of the encoder and avoids overfitting to noisy labels during the early training stages. Among the subset, the most plausible label is then identified based on the softmax scores. The second is a **Global Optimal Transport Planner** (GOP). It maps a uniform global prior—representing the overall underlying distribution—to the candidate label distributions across a mini-batch. By minimizing the expected transport cost, GOP encourages globally optimal label assignments. The resulting transport plan explicitly characterizes how each instance contributes to the class-wise probability mass in the label distribution, thereby providing a principled and interpretable foundation for identifying and selecting the most relevant label.

By integrating both *local* and *global* perspectives, our framework enables more accurate and stable label selection, mitigating the impact of label ambiguity. We evaluate our method across a variety of vision-language benchmarks under partial supervision and demonstrate consistent improvements over strong baselines. Our key contributions are as follows:

- We provide a comprehensive investigation into leveraging the representational generalization of pre-trained VLMs for prompt learning in the partial supervision setting.
- We propose a novel holistic label selection framework that combines a local density-based filter with a global optimal transport planner.
- We empirically validate the effectiveness of our approach, achieving state-of-the-art performance on multiple benchmarks.

2 Related Work

2.1 Partial Label Learning

PLL [2, 4, 15, 50] assumes that each training instance is associated with a candidate label set, within which only one label is correct but not explicitly specified. To address the ambiguity introduced by these noisy candidates, subsequent research has proposed various strategies, including consistency regularization [8,23,44] and contrastive learning [36]. These methods aim to leverage partial supervision to learn more discriminative representations and improve generalization. Expanding on this direction, [14] introduced a dissimilarity propagation-guided label shrinkage method to refine candidate label sets by eliminating irrelevant labels, thereby enhancing supervision quality. Meanwhile, works such as [48, 49] have begun to explore the more challenging instance-dependent PLL setting, where candidate label sets are generated based on the characteristics of each instance. More recently, [46] and [34] have continued to leverage candidate labels to learn robust representations. Additionally, [22] highlighted the critical role of feature representations and label denoising in effective PLL. A related work, SoLar [37], addresses partial label learning under class imbalance by employing OT to align the estimated long-tailed class prior with a uniform distribution. In contrast, our method leverages OT to align the overall underlying label distribution with the candidate label distributions. Furthermore, unlike conventional PLL approaches, our work explores prompt learning for pre-trained VLMs, which leverage powerful vision encoders with strong zero-shot capabilities, rather than training encoders from scratch.

2.2 Prompt Learning

Prompt learning has emerged as a parameter-efficient paradigm for adapting large-scale pre-trained models to downstream tasks [10]. Rather than fine-tuning the entire model, it introduces auxiliary input prompts—either manually crafted templates or learnable continuous embeddings—to guide model predictions. In the vision-language domain, CoOp [54] first demonstrated that learnable context vectors could serve as effective prompts for adapting CLIP-like models, highlighting their strong transferability. Since then, several studies have enhanced the robustness of prompt learning through strategies such as aligning sample representations in multimodal contexts [1, 35, 45], employing mixture-of-expert prompts [38], and utilizing unsupervised prompt distillation [16, 19]. In addition, prompt learning has achieved significant improvements on various downstream tasks, including text-to-image generation [33] and open-vocabulary semantic segmentation [18]. While these advances have shown promise in fully supervised settings, their effectiveness under partial supervision remains unexplored. To address this gap, we extend robust prompt tuning [11, 26, 43] to the PLL scenario, aiming to bridge weak supervision with prompt-based adaptation in pre-trained VLMs.

3 Preliminary

3.1 Candidate Label Sets in PLL

We define the training dataset as $\mathcal{D} = \{(\mathbf{x}_i, S_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathcal{X}$ is an input instance and $S_i \subseteq \{1, 2, \ldots, C\}$ is the candidate label set for a C-category classification task. Here, all elements $\{s_{i1}, \ldots, s_{il} \ldots, s_{iL}\}$ in S_i contain one ground-truth label and the rest $L-1$ elements are the confused labels, involving false positive labels for the current instance. Different levels of label confusion is corresponding to the number of confused labels in each candidate set.

3.2 Prompt Optimization with The Candidate Set

Prompt optimization (PO) in CLIP replaces manual prompt engineering by learning continuous context vectors in an end-to-end manner, while keeping the pre-trained model parameters frozen to fully leverage the knowledge encoded within them. Let V, T and $\mathbf{t}$ denote the vision encoder, text encoder, and learnable prompt parameters, respectively. $p(s_{ij} \mid \mathbf{x}_i; \mathbf{t})$ denotes the predicted probability of the j-th label within the candidate set S_i, computed via the cosine similarity between the visual and textual embeddings. These embeddings are derived from the visual encoder V, and the text encoder T with the input of the prompt vector $\mathbf{t}$ and the class name embedding $\mathbf{v}_j$. For PO under partial-label supervision, the standard cross-entropy loss is computed over the candidate label set S_i as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{CE}}(\mathbf{x}_i, S_i; \mathbf{t}) &= \sum_{j=1}^{C} -s_{i,j} \log\left(p(s_{ij} \mid \mathbf{x}_i; \mathbf{t})\right) \\ \text{s.t.} \quad \sum_{j \in S_i} s_{i,j} &= 1, \text{ and } s_{i,j} = 0 \text{ for } j \notin S_i. \end{aligned} \tag{1}$$

The prediction probability of the input $\mathbf{x}_i$ is given by:

$$p(s_{ij} \mid \mathbf{x}_i; \mathbf{t}) = \frac{\exp\left(\cos\left(V(\mathbf{x}_i), T(\mathbf{t}, \mathbf{v}_j)\right)\right)}{\sum_{q=1}^{C} \exp\left(\cos\left(V(\mathbf{x}_i), T(\mathbf{t}, \mathbf{v}_q)\right)\right)}. \tag{2}$$

Despite the expressive power of PO, the inherent label ambiguity within each candidate set S_i creates a fundamental gap between partial label learning and fully supervised learning. This introduces two core challenges: (1) effectively disambiguating the candidate labels, and (2) accurately identifying the ground-truth label for each instance.

4 Methodology

To identify the ground-truth label and guide the model in learning effective prompt vectors, we propose a holistic optimal label selection strategy for prompt

learning in VLMs. HopS integrates both local and global perspectives through two elaborated components: a local density-based filter, which selects the "local-consensus" candidate, and a global optimal transport planner, which identifies the "global-harmony" candidate. These components work in concert to robust and complementary label guidance for effective prompt optimization under partial supervision.

4.1 Local Density-Based Filter

The local density of an instance in the image feature space reflects the intrinsic semantic regularities associated with its candidate labels. To leverage this density information for label selection, we identify the k-nearest neighbors of each instance $(\mathbf{x}_i, S_i)$ in the image feature space, denoted as $\{(\mathbf{x}_j, \hat{S}_j)\}_{j=1}^{k}$. Specifically, we build an affinity matrix $\mathcal{A}$ by computing the cosine similarity between image features of all instances, where the features are extracted using the frozen vision encoder of CLIP. Based on $\mathcal{A}$, we retrieve the top-k most similar examples for each instance, thereby capturing local semantic structure. Once the k-nearest neighbors are selected, we then construct a multiset-union candidate set for each instance as:

$$\mathcal{N}_i = \left(\biguplus_{j=1}^{k} \hat{S}_j \right) \biguplus S_i, \tag{3}$$

where $\uplus$ denotes multiset union, preserving label multiplicities. It is worth noting that, since the affinity matrix is pre-computed prior to training, the retrieval step for each instance reduces to a neighbor search with time complexity $\mathcal{O}(n \cdot \log k)$. Given the small number n of training samples in prompt learning, the retrieval cost is negligible.

Next, we compute the frequency of each category c in $\mathcal{N}_i$. Let $f(c)$ denote the relative frequency of element c in the multiset $\mathcal{N}_i$, defined as:

$$f(c) = \frac{|\{l \mid s_{jl} = c\}|}{|\mathcal{N}_i|}, \tag{4}$$

where $|\{l \mid s_{jl} = c\}|$ counts the number of times element c appears in $\mathcal{N}_i$, and $|\mathcal{N}_i|$ is the total number of elements in the multiset, including duplicates.

To enforce label consistency and suppress unreliable candidates, we retain categories whose frequency exceeds a predefined threshold $\tau \in [0, 1]$, thereby forming a consensus-aware mask set for each instance as:

$$\mathcal{M}_i = \{c \in \mathcal{N}_i \mid f(c) \geq \tau\}. \tag{5}$$

The refined candidate set is then obtained by intersecting the mask set $\mathcal{M}_i$ with the original candidate set S_i, i.e., $\mathcal{M}_i \cap S_i$, ensuring that only labels supported by both the instance and its neighbors are retained. To prevent degenerate cases where all candidate labels are eliminated, we revert to the original candidate set S_i if the refined set becomes empty.

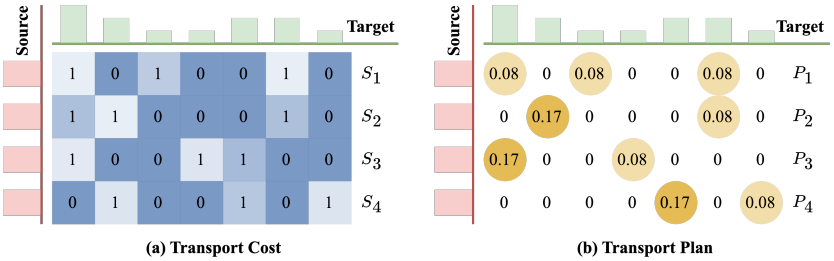

Fig. 1: Illustration of the transport cost matrix and the resulting transport plan between the **uniform source instance distribution** (Source) and the **target candidate label distribution** (Target).

Finally, the most plausible candidate label y^{local} is selected as the one with the highest prediction probability in the refined candidate set, as determined by Eq. (2).

4.2 Global Optimal Transport Planner

To complement the local selection, we introduce a globally consistent label selection mechanism via optimal transport (OT). This formulation matches instances with their candidate labels in a way that aligns with both a uniform prior over instances and the empirical label distribution within the batch, selecting for each instance the label that receives the highest transport mass. As shown in Fig. 1, the matrix on the left illustrates the candidate label sets for four instances in a seven-class classification task, with lighter colors indicating lower transport costs and, consequently, higher label credibility. The right matrix shows the resulting optimal transport plan under the given cost constraints. Here, darker colors indicate greater transported mass, reflecting the planner's preference for allocating probability mass to more credible classes. The transport process strictly adheres to the marginal constraints, ensuring the consistency of source and target distributions before and after transport.

Discrete Transport Formulation. Given a batch of B samples, we define a uniform source distribution over instences, where $\mathbf{r} \in \Delta^B$ is a B-dimensional probability simplex. A candidate-aware marginal distribution is estimated based on candidate frequencies, denoting as $\mathbf{c} \in \Delta^C$. The element in $\mathbf{r}$ and $\mathbf{c}$ are computed as:

$$r_i = \frac{1}{B}, \quad c_l = \frac{1}{B} \sum_{i=1}^{B} \frac{s_{il}}{|S_i|}. \tag{6}$$

The transport cost matrix $\mathbf{M}_{\text{cost}} \in \mathbb{R}^{B \times C}$ in OT is constructed from the similarity between visual and textual representations. This cost formulation encourages semantic alignment between image-label pairs, while discouraging assignments to non-candidate classes by assigning them infinite cost. Specifically,

given an instance $\mathbf{x}_i$, the cost of assigning it to class j is defined as:

$$\mathbf{M}_{\text{cost}}[i,j] = \begin{cases} 1 - \frac{\exp(\cos(V(\mathbf{x}_i), T(\mathbf{t}, \mathbf{v}_j)))}{\sum_{q=1}^{C} \exp(\cos(V(\mathbf{x}_i), T(\mathbf{t}, \mathbf{v}_q)))}, & \text{if } j \in \mathcal{S}_i \\ \infty, & \text{otherwise.} \end{cases} \tag{7}$$

The optimal transport plan $\mathbf{P}$ aims to move probability mass from a uniform distribution over instances to a candidate-aware label distribution with minimal total cost. To ensure differentiability and enable efficient computation, we incorporate an entropy regularization term $H(\mathbf{P}) = -\sum_{i=1}^{B}\sum_{j=1}^{C} P_{ij} \log P_{ij}$ following [5]. The resulting optimal transport objective is formulated as:

$$\begin{aligned} \min_{\mathbf{P} \in \mathbb{R}_+^{B \times C}} \quad & \langle \mathbf{P}, \mathbf{M}_{\text{cost}} \rangle - \varepsilon H(\mathbf{P}) \\ \text{s.t.} \quad & \mathbf{P}\mathbf{1}_C = \mathbf{r},\ \mathbf{P}^\top \mathbf{1}_B = \mathbf{c},\ P_{ij} = 0 \text{ if } s_{ij} = 0. \end{aligned} \tag{8}$$

Here, ε is a hyper-parameter controlling the strength of the entropy regularization, and $\mathbf{1}_C$ and $\mathbf{1}_B$ denote all-one vectors of length C and B, respectively.

Approximation Scheme. To efficiently solve the entropy-regularized optimal transport problem described in Eq. (8) and compute the transport plan, we adopt the well-known Sinkhorn-Knopp algorithm, an efficient iterative method.

$$\mathbf{P}^{\text{ot}} = \text{diag}(\boldsymbol{\alpha}) \cdot \exp\left(-\frac{\mathbf{M}_{\text{cost}}}{\varepsilon}\right) \cdot \text{diag}(\boldsymbol{\beta}), \tag{9}$$

The iterative steps are given as follows:

$$\boldsymbol{\alpha} \leftarrow \mathbf{r} \oslash (\mathbf{M}\boldsymbol{\beta}), \quad \boldsymbol{\beta} \leftarrow \mathbf{c} \oslash (\mathbf{M}^\top \boldsymbol{\alpha}). \tag{10}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are scaling vectors, $\oslash$ denotes element-wise division, $\mathbf{r}$ and $\mathbf{c}$ are the source and target marginal distributions defined above, and $\mathbf{M}$ is the Gibbs kernel [5] derived from the cost matrix.

After computing the optimal transport plan $\mathbf{P}^{\text{ot}} \in \mathbb{R}^{B \times C}$, which encodes the soft assignment between instances and candidate labels, we select y^{global} as the label with the largest transport mass. Unlike y^{local}, this global selection considers not only the candidate constraint but also the semantic alignment captured by the transport plan, thereby improving consistency across training instances.

4.3 Learning Objective and Procedure

The Objective Function. After select the two most plausible labels y^{local} and y^{global}, we conduct prompt optimization and jointly optimize the cross-entropy loss. Given an input image $\mathbf{x}$, the overall loss is defined as:

$$\arg\min_{\mathbf{t}} \ \mathcal{L}_{\text{CE}}(p(y \mid \mathbf{x}; \mathbf{t}), y^{\text{local}}) + \lambda \mathcal{L}_{\text{CE}}(p(y \mid \mathbf{x}; \mathbf{t}), y^{\text{global}}), \tag{11}$$

where λ is the weighting coefficient for the two loss components, the prediction probability $p(y \mid \mathbf{x}; \mathbf{t})$ for each category is computed by Eq. (2).

Training Procedure. As illustrated in Algorithm 1, the HopS performs both LDF and GOP selection to iteratively refine predictions under partial supervision, resulting in a more robust and discriminative prompt. The computational overhead is negligible (see Tab. 4).

Algorithm 1 The Holistic Optimal Selection

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_i, S_i)\}_{i=1}^{n}$, pre-trained encoders V and T
Parameter: $k, \tau, \varepsilon, \lambda$
Output: The prompt vector $\mathbf{t}$

Extract and store all image features $R_{\text{vis}} \leftarrow V(X)$.
Calculate $\mathcal{A}$ between R_{vis} by cosine similarity.
for $batch = 1, 2, \ldots$ **do**
  `// LDF`
  Choose k neighbors to compute $\mathcal{N}_i$ corresponding to each instance by Eq. (3).
  Calculate the frequency $f(\cdot)$ by Eq. (4).
  Construct the consensus-aware mask set and refined candidate set by Eq. (5).
  Identify Y^{local} with the highest prediction probability in the refined candidate set.
  `// GOP`
  Initialize the distribution of $\mathbf{r}$ and $\mathbf{c}$ by Eq. (6).
  Calculate the cost matrix $\mathbf{M}_{\text{cost}}$ with $T(t, v_j)$ and R_{vis}[instance indexes] by Eq. (7).
  for $iter = 1, 2, \ldots$ **do**
    `// Sinkhorn-Knopp Algorithm`
    Approximate transport plan by Eq. (9).
  end for
  Gain the Y^{global} with the largest transport mass.
  Update $\mathbf{t}$ with Eq. (11).
end for

5 Experiments

We conduct experiments on eight datasets, comparing our approach with eight representative loss functions, zero-shot learning, and 16-shot fully supervised prompt tuning using CoOp. All methods are evaluated under varying degrees of label ambiguity. Our method consistently outperforms all baselines, highlighting its effectiveness in prompt learning with partial labels. Extensive experiments and analyses further demonstrate the complementarity of the two label selection strategies.

5.1 Experimental Settings

Basic settings are outlined. Additional details (e.g., hyperparameters) are in the Appendix Sec. A.

Dataset. We adopt eight datasets: Caltech [6], DTD [3], EuroSAT [13], FGVCAircraft [24], Food [25], Flowers [25], OxfordPets [27], and UCF [31]. These datasets encompass a wide spectrum of visual recognition tasks, especially for challenging fine-grained classification.

Confusion Types. We manually corrupt the datasets into partially labeled versions using two strategies: random-uniform (*rand*) and instance-dependent (*insd*). (1) In the *rand* setting, confusing labels are randomly selected into the

candidate set with equal probability for each instance. (2) The *insd* setting depends on instance-level information. Following [48], we adopt a prototype-based label confusion strategy, which uses pre-trained image encoders to identify the top-$(L-1)$ most similar classes—excluding the ground-truth label—as candidate labels. We employ three visual backbones (i.e., ResNet-18/50 [12] trained on ImageNet and CLIP-ResNet50 [29]) to simulate annotators with varying levels of domain expertise.

Confusion Levels. The confusion rate, defined as the ratio of the number of confused labels to L, serves as an indicator of the difficulty in identifying the ground-truth label within the candidate set. The *insd*-confusion and *rand*-confusion levels of the candidate labels are controlled by the size of the candidate label set L, selected from $\{2, 3, 4, 5, 8, 9, 10\}$, with the corresponding confusion rates γ_c being $\{0.50, 0.67, 0.75, 0.80, 0.88, 0.89, 0.90\}$.

Prompt Types. The design of prompts can be categorized into two forms, consistent with CoOp: unified prompt (*uni*) and classified prompts (*cls*). The *uni* prompt shares a common set of context vectors across all classes, making it suitable for general classification tasks. In contrast, *cls* assigns an independent set of context vectors to each class.

Baselines. We compare HopS with five state-of-the-art partial label learning losses: CC [8], RC [8], LWC [40], MSE [7], and EXP [7]. Furthermore, we include three well-known robust learning losses, MAE [39], SCE [39], and GCE [53], as well as three recent partial label learning methods, Papi [46], CroSel [34], and SoLar [37], representing recent advances in the field. All hyperparameters for these methods are carefully tuned following the settings reported in their original publications. The detailed configurations of all compared methods are provided in Appendix Tab. 6.

The Learning Framework. To ensure a fair comparison, we adopt the most influential context optimization method, CoOp, and conduct experiments in both *uni* and *cls* prompts types. All results are reported as the average over three runs using three different random seeds.

5.2 Comparison Results

***Rand*-Confusion Type.** We evaluate all methods under two prompt settings: *uni*-prompt and *cls*-prompts. As shown in Fig. 2, HopS achieves SOTA performance across datasets under varying levels of label confusion. It is worth highlighting that, on Caltech, Flowers, and UCF, HopS achieves performance comparable to fully supervised 16-shot CoOp and significantly outperforms the zero-shot capability of CLIP across all datasets. Moreover, HopS maintains consistently high performance even under a severe label confusion rate of 0.9. This highlights the effectiveness of our holistic label selection strategy in reducing label ambiguity and providing reliable supervision. The detailed numerical results can be found in Appendix Tabs. 7 and 8.

***Insd*-Confusion Type.** Since partial-label losses perform well under the *rand* confusion setting, we further compare HopS against them under instance-dependent partial labels, which more closely resemble real-world scenarios where label noise

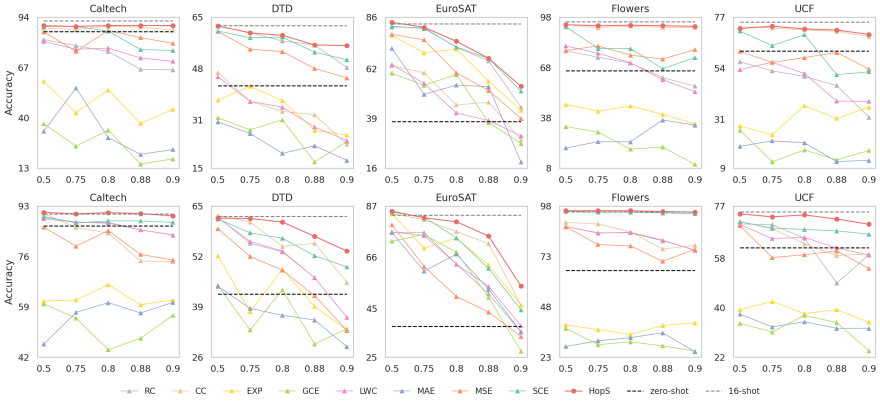

Fig. 2: HopS achieves the best testing accuracy under *uni*-prompt (top) and *cls*-prompts (bottom) across five *rand* confusion rates.

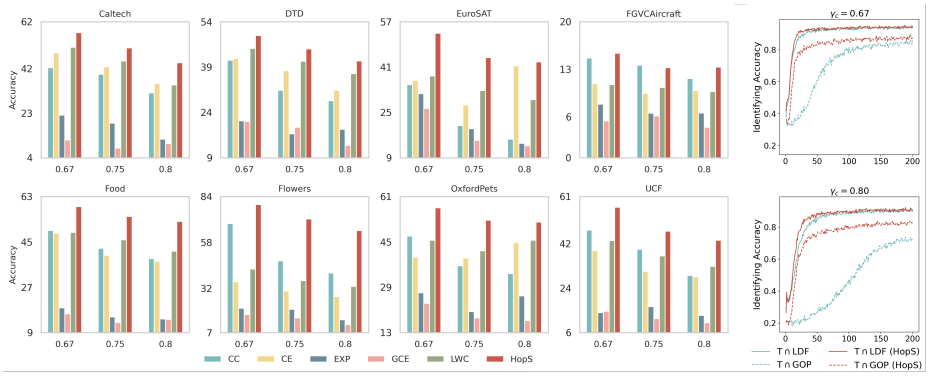

Fig. 3: HopS achieves the best averaged testing accuracy across three confusion rates (left). The guiding effect of LDF on GOP (right).

depends on feature-level similarities. As shown in Fig. 3 (left), HopS consistently outperforms other methods under the *uni*-prompt setting when averaging results over ResNet-18, ResNet-50, and CLIP-ResNet50 simulating backbones. Notably, our method retains a substantial advantage even as the confusion rate rises to 0.8. This demonstrates the robustness of HopS in more challenging and realistic conditions.

In particular, this performance gain holds across settings where candidate labels are derived from annotators with varying levels of domain expertise (i.e., simulated through different backbone encoders that reflect distinct data conditions), indicating that HopS is resilient to the type and quality of label noise. This robustness stems from the complementary nature of LDF and GOP components in HopS. These results indicate that HopS is effective not only in synthetic settings but also in real-world scenarios with instance-dependent ambiguity, such

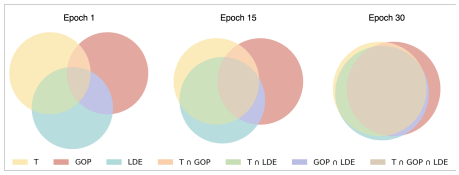

Fig. 4: Overlap of labels identified by LDF and GOP in HopS with the GT **T** on the Food ($\gamma_c = 0.67$).

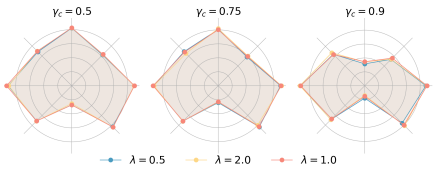

Fig. 5: The performance remains stable across a wide range of λ on **eight** datasets (directions), and λ=1 works well.

as crowd-sourced labels or low-resource domains. *The detailed per-configuration results are reported in Appendix Tabs* 11 *and* 10.

5.3 Further Analysis

Complement Effect of LDF and GOP. To assess the interaction between the LDF and GOP modules, we compare their individual performance and that of the holistic model, HopS, **in identifying the ground-truth (GT) label**. The training accuracy of LDF, GOP, and HopS over 200 epochs, evaluated under the *uni*-prompt on the Food dataset at confusion rates of 0.67 and 0.80, is illustrated in Fig. 3 (right), emphasizing LDF's role in accelerating early optimization by guiding the global module. The four curves compare LDF and GOP used stand-alone versus used as components within HopS. The symbol $\cap$ denotes the proportion of predicted labels from the LDF or GOP that are consistent with the GT label, where T denotes the GT label. Using modules as HopS components yields faster convergence and a higher, more stable plateau, and this advantage becomes increasingly pronounced as the γ_c intensifies. *Additional results are provided in Appendix Figs.* 9-16.

Furthermore, Fig. 4 shows that the proportion of correct labels jointly identified by both modules (i.e., the brown region) steadily increases, indicating a strong complementary effect. We also examine the relative contributions of the two pseudo-label sources by varying the loss weighting coefficient $\lambda \in \{0.5, 1.0, 2.0\}$. As shown in Fig. 5, achieving a balanced combination of local and global signals leads to consistent improvements in performance across different confusions. In contrast, placing excessive weight on either signal—whether local or global—results in performance degradation. This finding underscores the critical importance of maintaining an effective synergy between the two sources of information. *The specific results are provided in Tab.* 12 *in the Appendix.*

Effect of Batch Size. We investigate the effect of varying batch sizes, B, within the set $\{16, 32, 64, 128, 256\}$ for GOP, on the performance of HopS, focusing on the testing accuracy under varying levels of label confusion: 0.50, 0.75, 0.80, 0.88, and 0.90. As illustrated in Fig. 6, the results compare the performance of *uni*-prompt and *cls*-prompts across different confusion rates. For the 47-class small-scale dataset DTD and the 100-class large-scale dataset Caltech, the accuracy trends with respect to B remain consistent across different confusion levels.

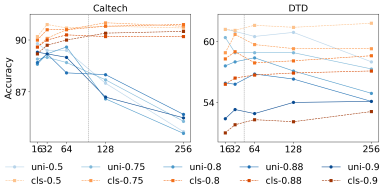

Fig. 6: B close to C is recommended.

Table 1: Testing Accuracy (%) of Various Candidate-Searching Strategies for label refinement in LDF.

γ_c	CoOp	Nei (Hard)	Con	Gra	Soft
0.50	84.70	**86.71**	86.20	83.40	86.70
0.67	80.00	85.60	84.30	83.10	**86.00**
0.80	74.60	84.10	83.80	78.40	**84.80**

Moreover, the most significant fluctuations in accuracy are observed when the batch size is approximately equal to the number of classes. Based on this observation, we recommend choosing a batch size close to the number of classes. *Results for other datasets are provided in the Appendix Fig.* 17 *and Tab.* 13.

Effect of LDF. We compare three candidate-searching strategies for label refinement in partial label learning: (1) Neighbor-based refinement (Nei), which selects pseudo-labels based on feature similarity to nearest neighbors; (2) Confidence-based refinement (Con), which chooses the most confident prediction; and (3) Graph-based propagation (Gra), which propagates labels through a constructed similarity graph, enabling global label consistency by considering relationships between all samples in the dataset and effectively transferring labels across connected nodes. As shown in Tab. 1, the simplest strategy—neighbor-based refinement—yields the best performance, suggesting that in weakly supervised few-shot settings, leveraging local structural similarity can outperform more complex propagation methods. One possible reason for this observation is that neighbor-based methods are inherently more robust to noisy predictions. By grounding pseudo-label selection in local feature consistency, these methods can better capture semantic regularities in the data. In contrast, graph-based methods might cause noise accumulation, especially in low-data regimes, where the constructed graph may not accurately reflect the underlying label manifold.

Additionally, we evaluate two label update strategies: Hard-KNN, which uses majority voting, and Soft-KNN, which uses similarity-weighted voting. Although Soft-KNN slightly outperforms Hard-KNN in some cases, the gains are marginal, indicating that the simple hard voting mechanism is sufficient and more computationally efficient.

Effect of GOP. The GOP module assumes that the candidate label distribution closely resembles the ground-truth label distribution. To assess this consistency, we use cosine similarity, which captures the directional alignment between two distributions and reflects their structural similarity regardless of scale. This metric provides a reliable measure of how well the candidate labels approximate the true semantic distribution. As shown in Fig. 7, the cosine similarity approaches 1 on EuroSAT under four different confusion rates, indicating a strong alignment between the distributions. We opt not to use Kullback–Leibler (KL) divergence due to the prevalence of zero values in the ground-truth distributions, which makes KL computation unstable and ill-defined. In contrast, cosine similarity is robust to sparsity.

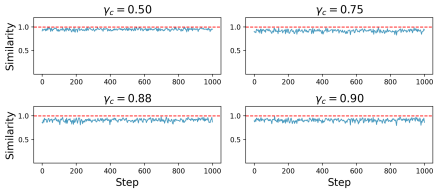

Fig. 7: High cosine similarity between candidate and ground-truth label distributions across batches.

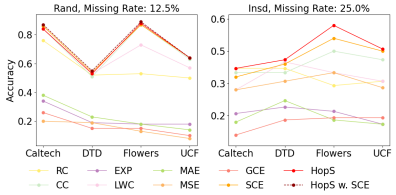

Fig. 8: Testing accuracy under the *uni-prompt* across two noisy conditions with different confusion rates.

Table 2: HopS achieves the best testing accuracy (%) under two confusion types.

γ_c	Caltech						Flowers					
	HopS	CroSel	Papi	HopS	CroSel	Papi	HopS	CroSel	Papi	HopS	CroSel	Papi
0.67	**89.0**	84.0	55.9	**50.3**	39.0	42.0	**92.9**	92.1	78.7	**82.0**	78.0	67.0
0.75	**89.0**	82.7	51.8	**45.3**	25.2	23.5	**92.9**	90.5	70.4	**74.0**	62.0	58.7
0.80	**89.5**	73.9	52.2	**38.9**	11.8	13.4	**93.2**	89.0	57.8	**65.0**	50.2	49.8

Comparison with Full-Data PLL Methods. We conducted a systematic comparison with two recently proposed partial label learning methods, Papi [46] and CroSel [34]. Unlike the 16-shot setting used in previous experiments, Papi and CroSel were trained on the entire datasets to fully exploit their optimal performance, while HopS was still trained under the 16-shot few-shot setting. The experiments were conducted on two challenging datasets, Caltech and Flowers, both characterized by a large number of categories, with Flowers being more fine-grained and exhibiting higher intra-class similarity. As shown in Tab. 2, the left and right sub-columns for each dataset correspond to *rand* and *insd*, respectively. HopS's strong performance across both datasets and three confusion rates, demonstrating its robustness in handling both diverse and fine-grained categories. Due to space limitations, we report only the results with the ResNet-50 backbone; results for other backbones are provided in Appendix Tab. 14.

Settings with Missing Ground-Truth. To gain deeper insights into the robustness of HopS, we evaluate it under more challenging settings where the ground-truth label may be absent from the candidate set, i.e., the noisy partial label learning scenario [47]. We conduct 16-shot experiments on two types of label confusion, *rand* and *insd*, as previously described, with each candidate set S containing three labels. The missing rates of ground-truth labels are 12.5% (2 out of 16 shots) for *rand* and 25% (4 out of 16 shots) for *insd*.

As shown in Fig. 8, under the *insd* setting, HopS consistently uncovers meaningful correlations within the candidate label sets, leading to significantly better performance than other methods. This suggests that HopS is particularly well-suited for real-world scenarios where label noise in SS is instance-dependent. In contrast, under the *rand* setting where such dependencies are absent, HopS per-

Table 3: Testing accuracy (%) under under long-tailed partial-label settings with different confusion types.

Dataset	Method	$\gamma_c = 0.67$				$\gamma_c = 0.75$			
		$insd^e$	$insd^s$	$rand^e$	$rand^s$	$insd^e$	$insd^s$	$rand^e$	$rand^s$
Caltech	SoLar	9.2	7.9	16.1	22.6	6.4	4.4	18.5	19.4
	HopS^u	46.4	50.7	71.7	71.2	35.6	42.0	67.6	67.4
	HopS^c	50.9	52.8	74.6	71.4	42.4	45.9	72.6	70.7
Flowers	SoLar	11.1	14.0	18.1	20.8	9.1	10.6	17.3	20.7
	HopS^u	48.3	51.4	65.4	61.1	39.8	43.8	61.9	53.9
	HopS^c	54.4	53.1	69.9	66.1	46.8	47.5	69.9	59.7

Table 4: Comparison of epoch time.

Method	Time (s)
EXP	0.12
MSE	0.11
GCE	0.11
LWC	0.11
MAE	0.12
SCE	0.11
HopS	0.12

forms slightly worse than the SOTA robust loss. Nevertheless, its performance can be improved by integrating robust loss functions (e.g., the SCE loss). Figure 8 (left) illustrates that HopS, when combined with the SCE loss, achieves the best performance under the *rand* setting, highlighting its flexibility and compatibility with other robust learning techniques. *Detailed values corresponding to Fig.* 8 *can be found in Appendix Tabs.* 15 *and* 16.

Challenging Settings with Long-tail Distribution. To further evaluate robustness under distribution shift, we compare HopS with SoLar [37] on two representative datasets under two long-tailed patterns: exponentially decayed class-frequency distribution (^{e}) and two-level distribution (^{s}). As summarized in Table 3, SoLar suffers a significant performance drop in the data-limited few-shot setting, as it relies on learning sufficiently strong representations. In contrast, HopS consistently outperforms SoLar across all long-tailed configurations, regardless of whether uni-prompts (HopS^u) or cls-prompts (HopS^c) are used.

Running Time of Methods. We measured the training time computational cost of each method, reported as the average time per epoch (in seconds), as shown in Tab. 4. Notably, although HopS introduces both a local module and a global module to enhance label identification, the additional computational overhead remains negligible due to the relatively small dataset size. As a result, HopS achieves superior performance without incurring a significant increase in training cost. *Additional results are provided in Tab.* 17 *of Appendix.*

6 Conclusion

In this work, we propose a holistic label selection framework for prompt learning under partial-label supervision, which improves robustness by combining a local density-based filter with a global optimal transport planner. By leveraging the generalization ability of frozen pre-trained vision-language encoders, HopS enables effective label disambiguation. Extensive experiments on eight benchmark datasets demonstrate consistent and significant improvements over strong baselines, highlighting its practicality and generalizability under weak supervision.

Acknowledgements

The work is supported in part by Shandong Sci-tech SMEs Innovation Project (No. 2024TSGC0740), Natural Science Foundation of China (No. U23A20389), Natural Science Foundation of Shandong Province (No. ZR2024MF101), and Young Expert of Taishan Scholars (No. tsqn202312026).

References

1. Chen, G., Yao, W., Song, X., Li, X., Rao, Y., Zhang, K.: PLOT: prompt learning with optimal transport for vision-language models. In: International Conference on Learning Representations (2023)
2. Chen, Y., Patel, V., Chellappa, R., Phillips, P.: Ambiguously labeled learning using dictionaries. IEEE Transactions on Information Forensics and Security **9**(12), 2076–2088 (2014)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Computer Vision and Pattern Recognition (2014)
4. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. Machine Learning Research **12**, 1501–1536 (2011)
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems **26** (2013)
6. 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: Computer Vision and Pattern Recognition Workshop (2004)
7. Feng, L., Kaneko, T., Han, B., Niu, G., An, B., Sugiyama, M.: Learning with multiple complementary labels. In: International Conference on Machine Learning (2020)
8. Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B., Sugiyama, M.: Provably consistent partial-label learning. Advances in Neural Information Processing Systems **33**, 10948–10960 (2020)
9. Frans, K., Soros, L., Witkowski, O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. Advances in Neural Information Processing Systems **35**, 5207–5218 (2022)
10. Guo, Y., Li, S., Liu, Z., Zhang, T., Chen, C.: A parameter-efficient and fine-grained prompt learning for vision-language models. In: Association for Computational Linguistics. pp. 31346–31359 (2025)
11. Guo, Y., Gu, X.: Joapr: Cleaning the lens of prompt learning for vision-language models. In: Computer Vision and Pattern Recognition. pp. 28695–28705 (2024)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (2016)
13. 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 pp. 2217–2226 (2019)
14. Jia, Y., Yang, F., Dong, Y.: Partial label learning with dissimilarity propagation guided candidate label shrinkage. Advances in Neural Information Processing Systems (2023)
15. Jin, R., Ghahramani, Z.: Learning with multiple labels. Advances in Neural Information Processing Systems **15** (2002)

16. Jin, X., Zhang, H., Wu, Z., et al.: Unsupervised prompt learning for vision-language models. In: Computer Vision and Pattern Recognition (2022)
17. Kumar, A., Raghunathan, A., Jones, R., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (2022)
18. Li, J., Lu, Y., Xie, Y., Qu, Y.: Relationship prompt learning is enough for open-vocabulary semantic segmentation. Advances in Neural Information Processing Systems (2024)
19. Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., Yang, J.: Promptkd: Unsupervised prompt distillation for vision-language models. In: Computer Vision and Pattern Recognition (2024)
20. 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)
21. Luo, Z., Zhao, P., Xu, C., Geng, X., Shen, T., Tao, C., Ma, J., Lin, Q., Jiang, D.: Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In: International Conference on Computer Vision. pp. 11206–11217 (2023)
22. Lv, J., Liu, Y., Xia, S., Xu, N., Xu, M., Niu, G., Zhang, M., Sugiyama, M., Geng, X.: What makes partial-label learning algorithms effective? Advances in Neural Information Processing Systems (2024)
23. Lv, J., Xu, M., Feng, L., Niu, G., Geng, X., Sugiyama, M.: Progressive identification of true labels for partial-label learning. In: International Conference on Machine Learning (2020)
24. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
25. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: Computer Vision, Graphics and Image Processing (2008)
26. Pan, B., Li, Q., Tang, X., Huang, W., Fang, Z., Liu, F., Wang, J., Yu, J., Shi, Y.: Nlprompt: Noise-label prompt learning for vision-language models. In: Computer Vision and Pattern Recognition. pp. 19963–19973 (2025)
27. Parkhi, O., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Computer Vision and Pattern Recognition (2012)
28. Qin, Y., Chen, X., Shen, Y., Fu, C., Gu, Y., Li, K., Sun, X., Ji, R.: Capro: Webly supervised learning with cross-modality aligned prototypes. Advances in Neural Information Processing Systems (2023)
29. Radford, A., Kim, J., 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: International Conference on Machine Learning (2021)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: International Conference on Computer Vision. pp. 10684–10695 (2022)
31. Soomro, K., Zamir, A., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
32. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. In: International Conference on Computer Vision. pp. 13019–13029 (2024)
33. Teo, C., Abdollahzadeh, M., Ma, X., Cheung, N.: Fairqueue: Rethinking prompt learning for fair text-to-image generation. Advances in Neural Information Processing Systems (2024)

34. Tian, S., Wei, H., Wang, Y., Feng, L.: Crosel: Cross selection of confident pseudo labels for partial-label learning. In: Computer Vision and Pattern Recognition (2024)
35. Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
36. Wang, H., Xiao, R., Li, Y., Feng, L., Niu, G., Chen, G., Zhao, J.: Pico: Contrastive label disambiguation for partial label learning. In: International Conference on Learning Representations (2022)
37. Wang, H., Xia, M., Li, Y., Mao, Y., Feng, L., Chen, G., Zhao, J.: Solar: Sinkhorn label refinery for imbalanced partial-label learning. Advances in Neural Information Processing Systems **35**, 8104–8117 (2022)
38. Wang, R., An, S., Cheng, M., Zhou, T., Hwang, S., Hsieh, C.: One prompt is not enough: Automated construction of a mixture-of-expert prompts. In: International Conference on Machine Learning (2024)
39. Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L.: Symmetric cross entropy for robust learning with noisy labels. In: International Conference on Computer Vision (2019)
40. Wen, H., Cui, J., Hang, H., Liu, J., Wang, Y., Lin, Z.: Leveraged weighted loss for partial label learning. In: International Conference on Machine Learning (2021)
41. Wen, S., Brbic, M.: Cross-domain open-world discovery. In: International Conference on Machine Learning (2024)
42. Werner, T., Burchert, J., Stubbemann, M., Schmidt-Thieme, L.: A cross-domain benchmark for active learning. Advances in Neural Information Processing Systems (2024)
43. Wu, C.E., Tian, Y., Yu, H., Wang, H., Morgado, P., Hu, Y.H., Yang, L.: Why is prompt tuning for vision-language models robust to noisy labels? In: International Conference on Computer Vision. pp. 15488–15497 (2023)
44. Wu, D., Wang, D., Zhang, M.: Revisiting consistency regularization for deep partial label learning. In: International Conference on Machine Learning (2022)
45. Wu, M., Cai, X., Ji, J., Li, J., Huang, O., Luo, G., Fei, H., Jiang, G., Sun, X., Ji, R.: Controlmllm: Training-free visual prompt learning for multimodal large language models. Advances in Neural Information Processing Systems (2024)
46. Xia, S., Lv, J., Xu, N., Niu, G., Geng, X.: Towards effective visual representations for partial-label learning. In: Computer Vision and Pattern Recognition (2023)
47. Xu, M., Lian, Z., Feng, L., Liu, B., Tao, J.: Alim: adjusting label importance mechanism for noisy partial label learning. Advances in Neural Information Processing Systems **36**, 38668–38684 (2023)
48. Xu, N., Qiao, C., Geng, X., Zhang, M.: Instance-dependent partial label learning. Advances in Neural Information Processing Systems **34**, 27119–27130 (2021)
49. Yang, F., Cheng, J., Liu, H., Dong, Y., Jia, Y., Hou, J.: Mixed blessing: Class-wise embedding guided instance-dependent partial label learning. arXiv preprint arXiv:2406.10502 (2024)
50. Yu, F., Zhang, M.: Maximum margin partial label learning. In: Asian Conference on Machine Learning (2016)
51. Zeng, F., Cheng, Z., Zhu, F., Wei, H., Zhang, X.: Local-prompt: Extensible local prompts for few-shot out-of-distribution detection. In: International Conference on Learning Representations (2025)
52. Zhang, F., Jiang, W., Shu, J., Zheng, F., Wei, H., et al.: On the noise robustness of in-context learning for text generation. Advances in Neural Information Processing Systems (2024)

53. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in Neural Information Processing Systems (2018)
54. Zhou, K., Yang, J., Loy, C., Liu, Z.: Learning to prompt for vision-language models. In: International Journal of Computer Vision. pp. 2337–2348 (2022)
55. Zhou, Y., Xia, X., Lin, Z., Han, B., Liu, T.: Few-shot adversarial prompt learning on vision-language models. Advances in Neural Information Processing Systems **37**, 3122–3156 (2024)