Navigating Real-World Partial Label Learning: Unveiling Fine-Grained Images with Attributes

Haoran Jiang[1,2,3], **Zhihao Sun**[4], **Yingjie Tian**[2,3,5,6*]

[1] School of Mathematical Sciences, University of Chinese Academy of Sciences
[2] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences
[3] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences
[4] School of Computer Science and Technology, University of Chinese Academy of Sciences
[5] School of Economics and Management, University of Chinese Academy of Sciences
[6] MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation at UCAS
jianghaoran21@mails.ucas.ac.cn, sunzhihao21@mails.ucas.ac.cn, tyj@ucas.ac.cn

Abstract

Partial label learning (PLL), a significant research area, addresses the challenge of annotating each sample with a candidate label set containing the true label when obtaining accurate labels is infeasible. However, existing PLL methods often rely on generic datasets like CIFAR, where annotators can readily differentiate candidate labels and are unlikely to confuse, making it less realistic for real-world partial label applications. In response, our research focuses on a rarely studied problem, PLL on fine-grained images with attributes. And we propose a novel framework called *Shared to Learn, Distinct to Disambiguate* (SoDisam). Within the candidate label set, the categories may exhibit numerous shared attribute features, posing a challenge in accurately distinguishing them. Rather than perceiving it as an impediment, we capitalize on these shared attributes as definitive sources of supervision. This insight guides us to learn attribute space visual representation to focus on the information from these shared attributes. Moreover, we introduce an attribute attention mechanism tailored to harness the remaining distinct attributes. This mechanism directs the originally holistic feature towards specific regions, capturing corresponding discriminative features. In addition, a dynamic disambiguation module is introduced, continuously adjusting the two aforementioned mechanisms and achieve the final disambiguation process. Extensive experiments demonstrate the effectiveness of our approach on fine-grained partial label datasets. The proposed SoDisam framework not only addresses the challenges associated with fine-grained partial label learning but also provides a more realistic representation of real-world partial label scenarios.

Introduction

Partial label learning (Cour, Sapp, and Taskar 2011; Feng et al. 2020; Lv et al. 2020; Wang et al. 2021; Wen et al. 2021; Wu, Wang, and Zhang 2022; Lv et al. 2023) (PLL), as one of the spotlights in the field of weakly-supervised learning (Zhou 2018; Hüllermeier and Beringer 2006; Natarajan et al. 2013; Zhu and Goldberg 2009; Yang et al. 2022; Ishida, Niu, and Sugiyama 2018; Ishida et al. 2017), enables model to learn from each training instance with only a set of ambiguous candidate labels containing the unknown ground truth label. In this way, PLL can reduce the money and time cost and the need of professional expertise when annotating data. Therefore, PLL has various application scenarios, including but not limited to medical disease diagnosis (Song et al. 2016), image annotation (Chen, Patel, and Chellappa 2017) and web mining (Luo and Orabona 2010).

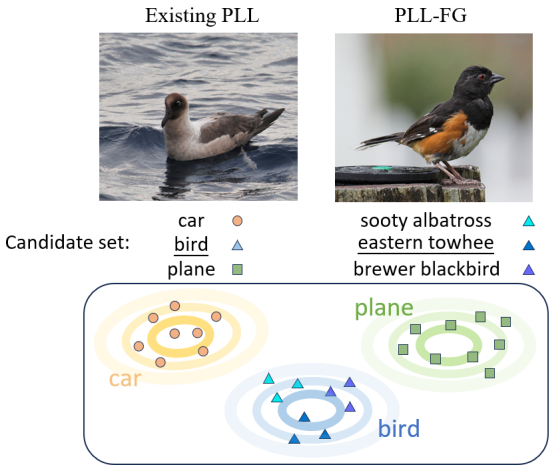

Figure 1: Examples of existing PLL and PLL-FG training samples, with PLL-FG samples closer to real-world scenarios due to more challenging true label annotation.

Previous methods (Wang et al. 2021; Wu, Wang, and Zhang 2022; Lv et al. 2020; Xia et al. 2023; Lyu, Wu, and Feng 2022; Zhang et al. 2021) for deep PLL achieved remarkable performance which even can be comparable to supervised learning under certain experiment settings. However, existing methods based on deep learning mainly focus on synthetic PLL datasets which are generated from generic datasets (e.g., CIFAR-10/100) (Tian, Yu, and Fu 2023). The categories involved in these synthetic PLL datasets include various objects, such as bird, car, plane and so on. One key issue is that there is a significant gap between categories, which does not align with the original intention of PLL, where annotators are unable to distinguish the true label and thus annotate a set of candidate labels.

As an example shown in Fig. 1, consider a sample in the

*Corresponding author.

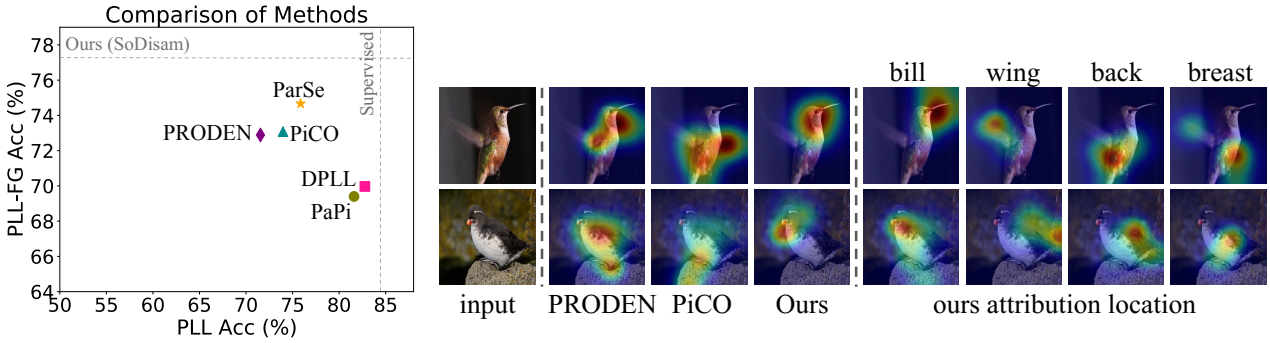

Figure 2: Left: Comparison of existing deep PLL methods on fine-grained CUB dataset with ambiguity level (see description in Experiment) $q = 0.01$ and generic CIFAR-100 dataset with $q = 0.05$. The X-axis represents performance in existing PLL task, while the Y-axis represents that in PLL-FG. Results demonstrate that most methods do not perform equally well in both tasks. Right: Visualization shows existing methods tend to learn irrelevant information for fine-grained datasets. Our approach focuses on attribute features of fine-grained objects, providing effective disambiguation support in PLL.

training set that consists of an image and a candidate label set: {car, bird, plane}. The unknown true label for this sample is 'bird'. However, this sample is not an ideal or representative training example for real-world applications because it is difficult for annotators to be confused between 'car', 'bird' and 'plane'. Therefore, this paper considers a more realistic and challenging problem, *partial label learning on fine-grained images* (PLL-FG). Fine-grained datasets typically comprise images belonging to the same general category, such as the bird dataset CUB (Wah et al. 2011). When annotating such datasets, it becomes challenging for annotators without sufficient domain expertise and extensive experience to accurately label the images. In this scenario, using a candidate label set for data annotation, instead of precise labels, is a beneficial approach to reduce annotation costs while ensuring labeling accuracy.

Fig. 2 shows performances of four powerful deep learning-based approaches and a simple method PRODEN (Lv et al. 2020). We can see the four apporoaches achieve performance close to supervised learning on CIFAR-100. However, their performance on the CUB might be the same or inferior to simpler method PRODEN. Visualizations also indicate that existing methods fail to focus on the discriminative features of fine-grained images, potentially incorporating irrelevant background regions, which makes it challenging to ensure the model's generalization ability and robustness. In PLL-FG, the main challenge lies in the high similarity among candidate label categories, which hinders the model from obtaining precise category supervision due to ambiguity. However, we view this as an opportunity. The high similarity among these ambiguous categories stems from their *shared attributes*, which can *serve as definitive supervision*. The remaining *distinct attributes*, representing distinctions between categories, are *crucial for label disambiguation* and accurate identification of true labels. Based on this, we propose a novel PLL-FG framework, ***Shared to Learn, Distinct to Disambiguate*** (dubbed as SoDisam). Firstly, leveraging shared attributes as supervision inforamtion, our model learns the visual representations of images in the attribute space. These visual representations capture the visual information corresponding to the attributes. Subsequently, the attribute space visual representations of distinct attributes are utilized to guide the original holistic features to focus on these discriminative attribute features, enhancing disambiguation. Lastly, to avoid an excessive number of labels in the original candidate label set, which could result in fewer shared attributes and less discriminative distinct attributes, we introduce a dynamic disambiguation module. This module dynamically adjusts the candidate label set by removing obvious false labels, obtaining more representative shared and distinct attributes. Moreover, this module identifies easy samples that are easily disambiguated, providing accurate category supervision to the model. These three modules organically collaborate, enabling better visual representations guidance for more discriminative holistic features, leading to enhanced disambiguation, finally, again aiding in learning more precise visual representations, establishing a cyclical and mutually reinforcing interaction. Empirically, SoDisam achieves state-of-the-art results on multiple benchmark datasets. The main contributions of our paper are summarized as follows:

- We consider a novel and realistic problem, partial label learning on fine-grained images, and assess its feasibility in real-world scenarios.
- We propose a novel framework named SoDisam, which utilizes shared attributes from ambiguous candidate labels to learn visual representations and employs distinct attributes to guide holistic features towards discriminative attributes for disambiguation.
- We design a dynamic disambiguation module that adjusts the candidate set dynamically and draw easy samples to provide accurate category supervision.
- Empirically, SoDisam effectively concentrates on discriminative attribute features and achieves state-of-the-art performance on three benchmark datasets.

Partial Labeling for Fine-Grained Images

Building upon the motivations presented above, this section delves into the feasibility of implementing partial label

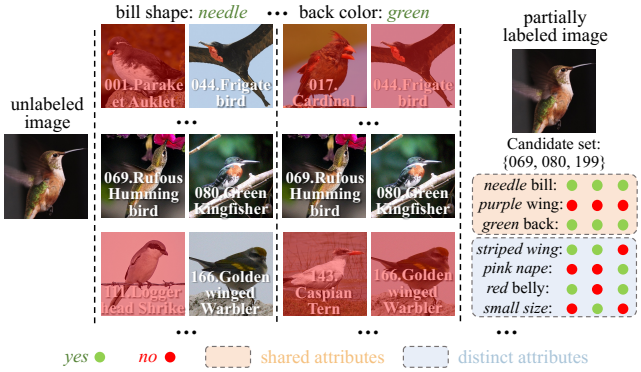

Figure 3: Illustration of the fine-grained image annotation process: utilizing visual appearance in images to generate candidate label sets by attribute-based exclusion of negative labels without requiring extensive domain knowledge.

learning on fine-grained datasets in real-world applications, while also addressing the primary challenges encountered and introducing the key insights of our proposed method.

Feasibility. Auxiliary information (e.g., attributes) has long been commonly used in zero-shot learning (Lampert, Nickisch, and Harmeling 2009; Li et al. 2019, 2022), few-shot learning (Snell, Swersky, and Zemel 2017) and fine-grained image recognition (Chen et al. 2018). Attributes usually are descriptions of different characteristics of a class, such as shape (e.g., circle) and color (e.g., green). With the incorporation of these class attributes, we have discovered that even ordinary and inexperienced annotators can accurately annotate a candidate label set for fine-grained data. This is possible because annotators can observe the images and identify several distinct features, such as 'bill shape: needle' and 'back color: green' as seen in Fig. 3. By utilizing these features, annotators can filter out classes that do not possess these characteristics, while retaining the remaining classes. Consequently, the filtered set of classes represents the partial label candidate set for the image to be annotated.

This annotation process does not require any specialized knowledge, yet it ensures that the true label remains within the final candidate label set, adhering to the requirements of partial label learning. Currently, class attributes are readily available in many datasets, and in the era of large vision-language models, generating and aligning the required attributes automatically is entirely feasible (Xu et al. 2022).

Challenges means opportunities. However, learning from such datasets presents substantial challenges. Since the training set lacks true labels, the model must navigate through ambiguous candidate label sets to identify the true category labels and acquire valuable knowledge for accurate classification. This becomes especially difficult in fine-grained images, where classes often share many common attributes, leading to small inter-class distances and high similarity between samples from different classes. Consequently, the model faces difficulty in learning precise classification boundaries from ambiguous supervision.

Nevertheless, amid these challenges, opportunities for model learning also emerge. The primary challenge of partial label learning lies in learning from uncertain and ambiguous supervisory information. If all classes in candidate sets share several common attributes, these shared attributes must correspond to the specific characteristics possessed by the sample, as illustrated in the Fig. 3. Therefore, these shared attributes can serve as definite supervision information to guide the model to learn the corresponding knowledge. Naturally, focusing on the remaining attributes, namely the distinct attributes which represents the difference between classes, directing attention towards the regions corresponding to these attributes, would significantly aid in label disambiguation. Subsequently, we derive a novel and effective model from this perspective.

The Proposed Framework

Overview. In this section, we propose a novel framework for PLL-FG, named as **SoDisam** from ***Shared to Learn, Distinct to Disambiguate***, comprising three key modules: attribute space visual representation module, attribute attention mechanism, and dynamic disambiguation module. Attribute space visual representation module considers the shared attributes in candidate set as definitive supervisory information and guides the model to focus and learn knowledge from these attributes. And attribute attention module is tailored to harness the remaining distinct attributes, directing the holistic feature attention towards specific regions to learn discriminative representations. Furthermore, we introduce a dynamic disambiguation module, which incessantly adjusts the two aforementioned modules and achieve the final disambiguation process. Eventually, we present the training objective. Before further discussion, we first define the problem and give a brief introduction of the baseline of PLL.

Preliminaries

Definition. Partial label learning is defined similar to ordinary multi-class classification as follows. Given training set $\{(\boldsymbol{x}_i, Y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, where $\mathcal{Y} = \{1, 2 \cdots, K\}$ represents K classes in the label space and $\boldsymbol{x}_i \in \mathcal{X}$ is input labeled with a candidate label set $Y_i \subseteq \mathcal{Y}$ that includes the ground-truth label $y_i \in Y$. The task is to learn a vector function $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), \cdots, g_K(\boldsymbol{x}))$ to obtain class prediction. Each class is annotated with A attributes as the class attribute vector $\{\boldsymbol{z}^k\}_{k \in K}$, where $\boldsymbol{z}^k = [z_1^k, z_2^k, \cdots, z_A^k]$, z_a^k denotes the score of a-th attribute of class k.

Baseline. The basic framework for PLL consists of a backbone network to extract features and a classification head to make the prediction. Given input sample $(\boldsymbol{x}, Y)$, we can acquire a feature map $\boldsymbol{F} \in \mathbb{R}^{\boldsymbol{C} \times \boldsymbol{H} \times \boldsymbol{W}}$ and a holistic feature $\boldsymbol{h} \in \mathbb{R}^C$ by feeding $\boldsymbol{x}_i$ to a backbone encoder and a Global Average Pooling GAP. Subsequently, the feature $\boldsymbol{h}$ are passed through fully connected layers to obtain the classification result, and then train with a PLL loss:

$$\mathcal{L}_{cls}(\boldsymbol{x}, Y) = L(\boldsymbol{g}^{cls}(\boldsymbol{x}), Y), \tag{1}$$

where $\boldsymbol{g}_{cls}(\boldsymbol{x})$ denotes the class probability output of input $\boldsymbol{x}$, $L(\boldsymbol{g}_{cls}(\boldsymbol{x}), Y)$ represents an arbitrary PLL loss, such as exponential loss (EXP) (Ishida et al. 2017), PRODEN loss (Lv et al. 2020).

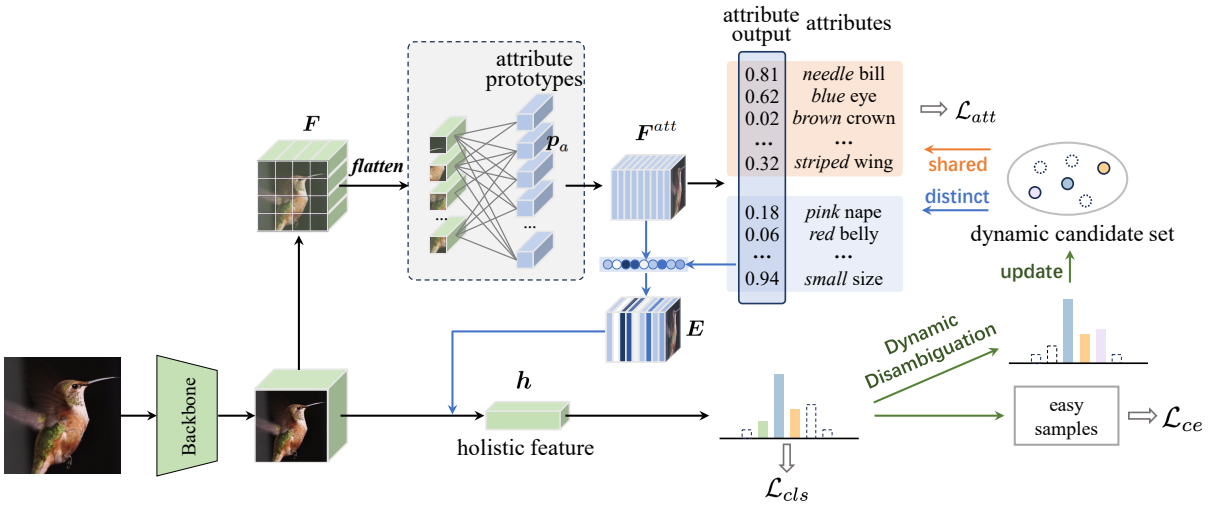

Figure 4: Overview of SoDisam. The *shared attributes* of candidate labels serves as definitive supervision to learn an attribute space visual representation $\boldsymbol{F^{att}}$. Then Attribute Attention Mechanism (blue arrows) tailored to the remaining *distinct attributes*, directs the holistic feature towards specific regions to capture corresponding discriminative features via $\boldsymbol{F^{att}}$. Finally, Dynamic Disambiguation Module (green arrows) updates the candidate set and thoroughly disambiguates easy samples.

However, due to the inherently similar appearances of fine-grained objects, even in supervised learning settings, backbone networks struggle to capture subtle discriminative features in the holistic feature $\boldsymbol{h}$. This can result in the guidance of classification outcomes by irrelevant information like backgrounds (Wei et al. 2021). In the problem of PLL, this issue can be exacerbated. Hence, leveraging the attributes mentioned earlier, we aim to assist the model in learning more discriminative features.

Attribute Space Visual Representation Module: Based on Shared Attributes

In order to glean more discriminative features, it is imperative to guide the network's attention towards a plethora of visual attributes of the target objects, such as bill shape, back color, and more. Leveraging the availability of class attributes annotations $\boldsymbol{z}^k$, we can adopt attribute prediction as a task objective, facilitating the model's transformation of the image's visual features into attribute space visual representations. Despite the absence of precise class information in PLL, we have observed that the more similar categories within the candidate label set, the more they tend to offer a multitude of shared attributes. These *shared attributes* can serve as definite supervision, signifying that an image's category must indeed possess or not possess these attributes.

To extract the attribute space representation of images, we initially flatten the features $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$, where C, H and W denote the channel, height and width respectively, extracted by the backbone, representing the local features of the image as $\boldsymbol{F}_{ij} \in \mathbb{R}^C, i = 1, 2, \cdots, H, j = 1, 2, \cdots, W$. To imbue the feature space with attribute semantics and align it with the visual feature space, we encode the words corresponding to the attributes by language models (e.g., Glove or BERT). Subsequently, through a fully connected layer, we derive attribute prototype matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_A]$, where $\boldsymbol{p}_a \in \mathbb{R}^C$ denote the prototype of a-th attribute in the attribute space. Following that, we can derive attribute-space visual feature at spatial location (i, j) as $\boldsymbol{F}^{att}_{i,j} \in \mathbb{R}^A$ via $\boldsymbol{F}^{att}_{i,j} = \boldsymbol{P}^\top \boldsymbol{F}_{ij}$, where $i = 1, 2, \cdots, H, j = 1, 2, \cdots, W$. By recovering the spatial positions based on (i, j), the complete attribute-space visual representation can be obtained $\boldsymbol{F}^{att} = \{\boldsymbol{F}^{att}_{i,j}\}$. After applying GAP to $\boldsymbol{F}^{att}$ over H and W, the resulting attribute-based vector undergoes further processing through a fully connected layer with parameter $V \in \mathbb{R}^{C\times A}$. This process yields attribute output $\boldsymbol{g}^{att}(\boldsymbol{x}) \in \mathbb{R}^A$ to represent the compatibility score between the input $\boldsymbol{x}$ and each attribute by

$$\boldsymbol{g}^{att}(\boldsymbol{x}) = (\mathrm{GAP}(\boldsymbol{F}^{att}))^\top V. \tag{2}$$

As only the shared attributes are certain, we solely utilize the shared attributes to train this module. Given the candidate label set $Y = \{y_1, y_2, \cdots, y_l\}$, we define two index sets, denoted as $s(Y), d(Y)$, which represent the sets of indexes of shared attributes and distinct attributes, respectively

$$\begin{gathered} s(Y) = \left\{i | \{z^y_i\}_{y\in Y} > 0, i \in [A]\right\} \cup \left\{i | \{z^y_i\}_{y\in Y} = 0, i \in [A]\right\}, \\ d(Y) = [A]/s(Y), \end{gathered} \tag{3}$$

where $[A] = \{1, 2, \cdots A\}$. As indicated, shared attributes are composed of attributes that are either possessed or not possessed by all labels in the candidate set. Based on $s(Y)$, given attribute output $\boldsymbol{g}^{att}$ and class attribute $\boldsymbol{z}^k$, we calculate a class score v as $v^y = \sum_{i\in s(Y)} g^{att}_i(\boldsymbol{x}) \times z^y_i$, and design an attribute loss $\mathcal{L}_{att}$:

$$\mathcal{L}_{att} = -\log \frac{\sum_{y\in Y} \exp(v^y)}{\sum_{y'\in \mathcal{Y}} \exp(v^{y'})}. \tag{4}$$

By optimizing the attribute loss, we encourage the module to learn more accurate attribute space visual representations, improving the ability to capture discriminative features according to attributes.

Attribute Attention Mechanism: Based on Distinct Attributes

Upon achieving attribute space visual representations, we can employ these features to guide the holistic features that were previously unable to capture subtle differences. Therefore, those *distinct attributes* should become a focal point of attention. Since these attributes among the candidate labels are inconsistent, they play a crucial role in distinguishing ambiguous classes within the candidate set.

Practically, we have introduced an attribute attention mechanism (AAM) to steer the holistic feature $\boldsymbol{h}$ towards these attributes, which comprises a conventional channel attention mechanism coupled with guidance derived from the attribute space visual representation $\boldsymbol{F}^{att} \in \mathbb{R}^{A\times H\times W}$ and attribute output $\boldsymbol{g}^{att}(\boldsymbol{x}) \in \mathbb{R}^{A}$ of distinct attributes. First we reshape $\boldsymbol{F}^{att}$ to $\mathbb{R}^{A\times HW}$ and define a guidance vector $\boldsymbol{u} \in \mathbb{R}^{A}$, where $\boldsymbol{g}^{att}$ retains only the components corresponding to distinct attributes, with the remaining components set to 0, to guide the model's attention towards distinct attributes. The vector $\boldsymbol{u}$ is then expanded to $\boldsymbol{U} \in \mathbb{R}^{A\times HW}$ to apply weighting to $\boldsymbol{F}^{att}$ and yield weighted visual representation

$$\boldsymbol{F}^{w} = \boldsymbol{U} \odot \boldsymbol{F}^{att}, \tag{5}$$

where $\odot$ indicates Hadamard product.

Next, we calculate a channel attention map $\boldsymbol{M} \in \mathbb{R}^{A\times A}$ from weighted attribute space visual representation $\boldsymbol{F}_{w}^{att}$ which is obtained by applying matrix multiplication between $\boldsymbol{F}_{w}^{att}$ and $\boldsymbol{F}_{w}^{att\top}$ and a softmax layer, as

$$m_{ji} = \frac{\exp\left(\boldsymbol{F}_{i}^{w} \cdot \boldsymbol{F}_{j}^{w}\right)}{\sum_{i=1}^{A} \exp\left(\boldsymbol{F}_{i}^{w} \cdot \boldsymbol{F}_{j}^{w}\right)}, \tag{6}$$

where m_{ji} measures the i-th channel's impact on the j-th channel. We conduct a matrix multiplication between the $\boldsymbol{M}^{\top}$ and $\boldsymbol{F}^{w}$, followed by reshaping their resultant matrix into $\mathbb{R}^{A\times H\times W}$. Subsequently, the resulting matrix is scaled and subjected to an element-wise summation with $\boldsymbol{F}^{w}$, with a channel mapping and yield the output $\boldsymbol{E} \in \mathbb{R}^{C\times H\times W}$. $\boldsymbol{E}$ encompasses discriminative properties, and then is employed to guide the original feature map as:

$$\boldsymbol{F}' = \lambda_1 \boldsymbol{E} + \boldsymbol{F}, \tag{7}$$

where λ_1 is a trainable scalar starting from 0. This process results in a new feature map $\boldsymbol{F}'$, which further yields refined holistic feature contributing to the ultimate class output with attention to distinct attributes-associated regions.

Dynamic Disambiguation Module

Once the image and candidate label sets are given, the shared attributes and distinct attributes are fixed accordingly. However, when there is substantial ambiguity within the candidate label set, an excessive number of label categories can lead to a reduction in the number of shared attributes, resulting in decreased discriminative power of distinct attributes. These labels often include some obvious mislabeled examples, such as in the label set {lion, cheetah, Samoyed} of a cheetah, where it's evident that 'Samoyed' is an easily recognizable wrong label. By excluding such labels, we can identify more valuable shared attributes and distinct attributes from the remaining labels {lion, cheetah}. Consequently, we maintain a dynamic candidate set $D \subset Y$ for each sample, facilitating the dynamic adjustment of the candidate set in a dynamic disambiguation module (DDM). This allows for the removal of incorrect labels and reconsideration of previously excluded candidate labels. Such an approach ensures that the model benefits from increased supervision by shared attributes and focuses on the more discriminative distinct attributes. Since both removed and reintroduced labels originate from the original candidate label set, i.e., $1 \leq |D| \leq |Y|$, for convenience, assume that at epoch t, $1 < |D| < |Y|$, enabling the removal and reintroduction of labels within this epoch with the constraint that at most one label can be removed or reintroduced at a time.

Update D. For the label we aim to remove, given the constraint of removing only one label at a time, its corresponding output probability within the D are bounded to be the lowest. Moreover, the probability need to fall below a certain threshold to prevent erroneous eliminations. Thus, we establish the following condition,

$$Con_{out}(i) = \begin{cases} i = \arg\min_{i\in D} g_i^{cls}(\boldsymbol{x}), \\ g_i^{cls}(\boldsymbol{x}) < \tau, \end{cases} \tag{8}$$

where τ is a fixed threshold. If $Con_{out}(i)$ is met, i will be removed from D. Similarly, if label j satisfies $Con_{in}(i)$

$$Con_{in}(j) = \begin{cases} i = \arg\max_{i\notin D} g_i^{cls}(\boldsymbol{x}), \\ i \in Y, \\ g_i^{cls}(\boldsymbol{x}) > \tau, \end{cases} \tag{9}$$

then label j will be reintroduced into D.

Draw easy samples. In PLL, the category classifier can solely learn from ambiguous label sets, leading to the loss of precise category supervision. Concurrently, it's evident that the dataset may encompass samples that are easily disambiguated, such as cases with a minimal count of candidate labels or substantial dissimilarity among label categories. Identifying and utilizing such straightforward samples to provide the model with accurate category supervision is highly beneficial. These simple samples often exhibit high-confidence predictions with consistent outputs during training. Inspired by (Li et al. 2023), we introduce a dynamic threshold $\epsilon_t(\boldsymbol{x})$, in which t is the current epoch, for each sample $\boldsymbol{x}$ to draw easy samples, as

$$\epsilon_t(\boldsymbol{x}) = l\epsilon_{t-1}(\boldsymbol{x}) + (1-l)b(\boldsymbol{x}), \epsilon_0 = 0, \tag{10}$$

where l is the factor to control the stability of ϵ, $b(\boldsymbol{x}) = max_{i\in Y} g_i^{cls}(\boldsymbol{x})$ indicates the confidence of prediction is greater than the threshold, it is selected as an easy sample and apply cross-entropy (CE) loss, as

$$\mathcal{L}_{ce} = \begin{cases} -\log b(\boldsymbol{x}), & \text{if } b(\boldsymbol{x}) > \epsilon_t(\boldsymbol{x}) \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Total training objective. Unify the three losses mentioned above together as the final training objective:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{att} + \beta\mathcal{L}_{ce}, \tag{12}$$

where α, β are weighting factors.

Dataset	Method	$q = 0.01$	$q = 0.02$	$q = 0.03$	$q = 0.05$
CUB	EXP	$64.01 \pm 0.64\%$	$62.39 \pm 0.99\%$	$58.39 \pm 1.04\%$	$52.53 \pm 0.87\%$
	PRODEN	$74.39 \pm 0.04\%$	$72.89 \pm 0.15\%$	$70.02 \pm 0.12\%$	$67.21 \pm 0.03\%$
	PiCO	$74.34 \pm 0.84\%$	$73.08 \pm 1.16\%$	$72.44 \pm 0.79\%$	$71.17 \pm 0.72\%$
	DPLL	$69.98 \pm 0.15\%$	$69.80 \pm 0.04\%$	$69.52 \pm 0.07\%$	$68.78 \pm 0.12\%$
	PaPi	$72.10 \pm 0.66\%$	$69.66 \pm 1.04\%$	$66.98 \pm 0.97\%$	$63.54 \pm 1.20\%$
	ParSE	$75.01 \pm 0.09\%$	$74.67 \pm 0.10\%$	$74.35 \pm 0.12\%$	$68.11 \pm 0.12\%$
	SoDisam (ours)	$\mathbf{77.47 \pm 0.19\%}$	$\mathbf{75.67 \pm 0.18\%}$	$\mathbf{74.80 \pm 0.07\%}$	$\mathbf{71.45 \pm 0.22\%}$
Dataset	Method	$q = 0.05$	$q = 0.1$	$q = 0.15$	$q = 0.2$
AWA	EXP	$85.62 \pm 0.97\%$	$83.01 \pm 0.77\%$	$79.40 \pm 1.32\%$	$76.89 \pm 1.67\%$
	PRODEN	$89.64 \pm 0.13\%$	$88.01 \pm 0.17\%$	$87.56 \pm 0.04\%$	$87.43 \pm 0.09\%$
	PiCO	$89.78 \pm 0.46\%$	$86.75 \pm 0.36\%$	$82.31 \pm 0.62\%$	$78.62 \pm 0.76\%$
	DPLL	$89.67 \pm 0.05\%$	$88.77 \pm 0.15\%$	$88.63 \pm 0.09\%$	$87.28 \pm 0.14\%$
	PaPi	$88.25 \pm 0.44\%$	$87.66 \pm 0.83\%$	$85.54 \pm 0.85\%$	$80.64 \pm 0.53\%$
	ParSE	$84.78 \pm 0.76\%$	$84.02 \pm 0.56\%$	$82.25 \pm 0.57\%$	$81.24 \pm 0.57\%$
	SoDisam (ours)	$\mathbf{92.66 \pm 0.08\%}$	$\mathbf{92.39 \pm 0.09\%}$	$\mathbf{91.93 \pm 0.07\%}$	$\mathbf{91.20 \pm 0.04\%}$
Dataset	Method	$q = 0.005$	$q = 0.01$	$q = 0.02$	$q = 0.03$
SUN	EXP	$50.62 \pm 2.25\%$	$47.70 \pm 1.95\%$	$45.42 \pm 1.62\%$	$42.30 \pm 0.99\%$
	PRODEN	$54.84 \pm 0.75\%$	$53.00 \pm 0.42\%$	$52.56 \pm 0.60\%$	$50.89 \pm 0.78\%$
	PiCO	$55.29 \pm 0.64\%$	$52.09 \pm 0.68\%$	$47.67 \pm 1.10\%$	$43.02 \pm 1.32\%$
	DPLL	$56.11 \pm 0.68\%$	$55.87 \pm 0.37\%$	$54.28 \pm 0.49\%$	$53.45 \pm 0.72\%$
	PaPi	$58.32 \pm 0.92\%$	$56.75 \pm 0.63\%$	$53.21 \pm 0.50\%$	$51.62 \pm 0.81\%$
	ParSE	$60.90 \pm 0.82\%$	$59.22 \pm 0.58\%$	$57.75 \pm 0.69\%$	$54.98 \pm 0.47\%$
	SoDisam (ours)	$\mathbf{63.46 \pm 0.19\%}$	$\mathbf{61.69 \pm 0.18\%}$	$\mathbf{60.20 \pm 0.27\%}$	$\mathbf{58.33 \pm 0.16\%}$

Table 1: Comparison with sota methods on CUB, AWA, SUN *w/* different ambiguity q for five trails. Bold indicates the superior.

Experiment

Set up

Datasets and Partial Label Generation. We choose three popular fine-grained datasets with attributes. CUB (Wah et al. 2011) consists of 11,788 images of 200 bird classes with 312 attributes, AWA2 (Lampert, Nickisch, and Harmeling 2009) contains 37,322 images of 50 animal classes with 85 attributes and SUN (Xiao et al. 2010; Patterson and Hays 2012) consists of 108,754 images of 395 scene classes with 102 attributes. We manually corrupt these datasets into partially labeled versions following (Lv et al. 2020).

Baselines. We compare SoDisam with six sota PLL methods: 1) EXP (Ishida et al. 2017), 2) PRODEN (Lv et al. 2020), 3) PiCO (Wang et al. 2021), 4) DPLL (Wu, Wang, and Zhang 2022), 5) ParSE (He et al. 2022), 6) PaPi (Xia et al. 2023). We adopt the same backbone, batch size, optimizer and augmentation strategy for all methods for fair comparisons. All specific settings, such as more epochs, and hyper-parameters are as suggested in the original papers.

Implementation. All implementation is based on PyTorch(Paszke et al. 2019) and models are trained on a NVIDIA A100 GPU. An 18-layer ResNet (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) is used as backbone for all main experiments. For SoDisam, the only hyper-parameters τ, α, β are fixed as 0.25, 0.3, 1, respectively. Choose PRODEN (Lv et al. 2020) loss as $L(\cdot)$.

Empirical Results

Our methods achieve sota performance. Overall, Tab. 1 provides a thorough comparison of SoDisam against six sota methods across CUB, AWA, and SUN dataset. Notably, at varying levels of ambiguity (q), our SoDisam consistently outperforms competing methods, showcasing its robustness and effectiveness in handling ambiguity. Specifically, SoDisam outperforms sota methods by up to **2.46**%, **3.77**% and **3.35%** within four levels of q for CUB, AWA and SUN respectively. More importantly, SoDisam maintain a high level performance as q increases on AWA, when q increases from 0.05 to 0.2, the performance only decreases by 1.46%. For the largest-scale dataset SUN with the most categories, methods like PiCO perform less effectively. However, our method demonstrates exceptional performance as well. In situations of elevated ambiguity ($q = 0.02, 0.03$), it outperforms the best existing method ParSE by 2.75%, 3.35%, showcasing the strength of our approach.

Our methods achieves attribute localization and focuses on discriminative regions. As shown in Fig. 2, existing methods tend to focus solely on the entire object, and even on irrelevant details such as the background. In contrast, our method, when compared to PRODEN (baseline) and PiCO, exhibits the capacity to emphasize more discriminative attribute features, such as the 'needle bill'. What's more, obtaining the attention heatmaps for specific attributes via a mask through AAM, we can see the capability of our model to accurately focus attributes like 'bill' and 'wing', which holds significant implications in label disambiguation in PLL, providing valuable explanatory insights.

Ablations

Effect of Components of SoDisam. In order to analyze the individual contributions of different components in the

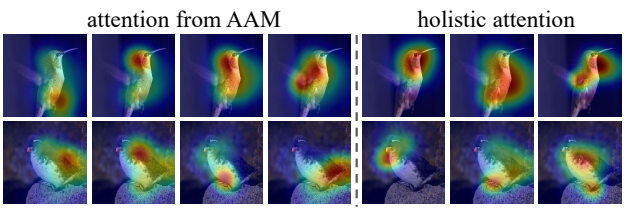

Figure 5: Visualization of DDM and AAM.

SoDisam framework, we conducted ablations as shown in Tab. 2. The results indicate the impact of each component on the performance across different datasets and demonstrate that the AAM, DDM and $\mathcal{L}_{ce}$ improve the performance over baseline consistently, by 3.36% (CUB), 3.02% (AWA), and 6.16% (SUN) gradually. The impact of $\mathcal{L}_{cls}$ *w/* $\mathcal{L}_{att}$ are limited due to their potential disruption of the model's learning objectives. However, by employing AAM, the visual features in the attribute space trained by $\mathcal{L}_{cls}$ can guide the classification of $\mathcal{L}_{cls}$, resulting in performance improvement.

Ablations	$\text{CUB}_{(0.01)}$	$\text{AWA}_{(0.05)}$	$\text{SUN}_{(0.005)}$
$\mathcal{L}_{cls}$	74.16	89.64	54.87
$+\mathcal{L}_{att}$	73.83	89.76	54.56
+ AAM	75.62	91.29	58.45
$+\mathcal{L}_{ce}$	77.02	**92.93**	61.03
+ DDM	**77.47**	92.66	**63.46**

Table 2: Top-1 classification accuracy (%) of ablation study.

Visualization of Dynamic Disambiguation and Attribute Attention. The DDM continually adjusts the candidate set to update shared and distinct attributes, guiding the AAM focus towards more discriminative information. This process holds significant importance. We visualize the attention heatmaps of the attribute attention at different epochs during training with updating from the DDM in Fig. 5. It can been seen that DDM possesses the capability to dynamically adjust its focus on attribute features, thereby guiding the holistic feature extraction process towards more discriminative regions.

Impact of Hyper-parameters. We assess the sensitivity of SoDisam to its parameters, and the findings are presented in Fig. 6 that SoDisam demonstrates robustness to most parameter variations. As α increases, the performance gradually declines, indicating that using attribute annotations for classification is unreliable. Utilizing attributes to guide classification (smaller α) proves to be a reasonable approach.

Related Works

Partial Label Learning (PLL) aims at learning from training samples with a candidate label set which contains latent ground-truth label. Initial works in PLL can be mainly categorized into average-based methods and identification-based methods. The former approaches (Cour, Sapp, and Taskar 2011; Hüllermeier and Beringer 2006; Zhang and Yu 2015) treat all labels in the candidate label set equally and take their average to approximate the true label, while the latter (Jin and Ghahramani 2002; Wang, Li, and Zhang 2019; Xu, Lv, and Geng 2019; Tang and Zhang 2017) strive to identify the real label from the candidate label set. Growing interest focuses on deep PLL for end-to-end training. (Lv et al. 2020) integrates model updating and true label identification with a consistent risk estimator. PiCO (Wang et al. 2021) introduces contrastive learning based on prototypical disambiguation. PaPi (Xia et al. 2023) simplifies PiCO and achieves effective visual representations through the use of prototypical classifiers. However, the performance of the majority of existing methods on more challenging fine-grained data remains less convincing.

Figure 6: Impact of Hyper-parameters τ, α, β.

Fine-Grained Visual Categorization (FGVC) strives to recognize categories characterized by subtle distinctions that often lead to high confusion (Wei et al. 2021). A primary solution involves the localization and analysis of local features for classification, by employing detection or segmentation techniques (Lin et al. 2015; Tang, Yang, and Chen 2023) and leveraging attention mechanisms (Peng, He, and Zhao 2017; Zhu et al. 2022). Others (Min et al. 2020; Zhao et al. 2021) perform high-order feature interactions and (Dubey et al. 2018; Chang et al. 2020; Liu, Chen, and Jia 2022) propose specific loss functions to guide model to learn the subtle difference between categories. Fine-grained zero-shot learning (Huynh and Elhamifar 2020a; Lampert, Nickisch, and Harmeling 2009) represents a significant direction within FGVC, wherein class attributes find extensive utility in establishing connections between seen and unseen classes. Attention-based (Huynh and Elhamifar 2020b; Ji et al. 2018; Huynh and Elhamifar 2020a) and prototype-based (Xu et al. 2020) methods design various attention mechanisms to capture discriminative regions. Nevertheless, these are not suitable for PLL due to the absence of precise class supervision.

Conclusion

In this paper, we consider a novel problem partial label learning on fine-grained images and propose a framework called SoDisam. The core idea is to utilize shared attributes of the ambiguous classes in the candidate set as definite supervision information and distinct attributes as the key for label disambiguation. Empirically, we conducted extensive experiments and show that state-of-the-art performance are established on several fine-grained dataset. Visualization shows SoDisam achieves attribute localization and focusing on discriminative regions. We believe our research can contribute to the advancement of PLL applications in real-world scenarios and inspire new insights in the community.

Acknowledgments

This work has been partially supported by grants from: National Natural Science Foundation of China (Nos. 12071458, 71731009).

References

Chang, D.; Ding, Y.; Xie, J.; Bhunia, A. K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; and Song, Y.-Z. 2020. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29: 4683–4695.

Chen, C.-H.; Patel, V. M.; and Chellappa, R. 2017. Learning from ambiguously labeled face images. *IEEE transactions on pattern analysis and machine intelligence*, 40(7): 1653–1667.

Chen, T.; Lin, L.; Chen, R.; Wu, Y.; and Luo, X. 2018. Knowledge-embedded representation learning for fine-grained image recognition. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 627–634.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *The Journal of Machine Learning Research*, 12: 1501–1536.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dubey, A.; Gupta, O.; Raskar, R.; and Naik, N. 2018. Maximum-entropy fine grained classification. *Advances in neural information processing systems*, 31.

Feng, L.; Lv, J.; Han, B.; Xu, M.; Niu, G.; Geng, X.; An, B.; and Sugiyama, M. 2020. Provably consistent partial-label learning. *Advances in neural information processing systems*, 33: 10948–10960.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, S.; Feng, L.; Lv, F.; Li, W.; and Yang, G. 2022. Partial label learning with semantic label representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 545–553.

Hüllermeier, E.; and Beringer, J. 2006. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5): 419–439.

Huynh, D.; and Elhamifar, E. 2020a. Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems*, 33: 19849–19860.

Huynh, D.; and Elhamifar, E. 2020b. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4483–4493.

Ishida, T.; Niu, G.; Hu, W.; and Sugiyama, M. 2017. Learning from complementary labels. *Advances in neural information processing systems*, 30.

Ishida, T.; Niu, G.; and Sugiyama, M. 2018. Binary classification from positive-confidence data. *Advances in neural information processing systems*, 31.

Ji, Z.; Fu, Y.; Guo, J.; Pang, Y.; Zhang, Z. M.; et al. 2018. Stacked semantics-guided attention model for fine-grained zero-shot learning. *Advances in neural information processing systems*, 31.

Jin, R.; and Ghahramani, Z. 2002. Learning with multiple labels. *Advances in neural information processing systems*, 15.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, 951–958. IEEE.

Li, J.; Jing, M.; Lu, K.; Ding, Z.; Zhu, L.; and Huang, Z. 2019. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7402–7411.

Li, X.; Yang, X.; Wei, K.; Deng, C.; and Yang, M. 2022. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9326–9335.

Li, Y.; Han, H.; Shan, S.; and Chen, X. 2023. DISC: Learning From Noisy Labels via Dynamic Instance-Specific Selection and Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24070–24079.

Lin, D.; Shen, X.; Lu, C.; and Jia, J. 2015. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1666–1674.

Liu, K.; Chen, K.; and Jia, K. 2022. Convolutional fine-grained classification with self-Supervised target relation regularization. *IEEE Transactions on Image Processing*, 31: 5570–5584.

Luo, J.; and Orabona, F. 2010. Learning from candidate labeling sets. *Advances in neural information processing systems*, 23.

Lv, J.; Liu, B.; Feng, L.; Xu, N.; Xu, M.; An, B.; Niu, G.; Geng, X.; and Sugiyama, M. 2023. On the robustness of average losses for partial-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive identification of true labels for partial-label learning. In *international conference on machine learning*, 6500–6510. PMLR.

Lyu, G.; Wu, Y.; and Feng, S. 2022. Deep graph matching for partial label learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3306–3312.

Min, S.; Yao, H.; Xie, H.; Zha, Z.-J.; and Zhang, Y. 2020. Multi-objective matrix normalization for fine-grained visual recognition. *IEEE Transactions on Image Processing*, 29: 4996–5009.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Patterson, G.; and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE conference on computer vision and pattern recognition*, 2751–2758. IEEE.

Peng, Y.; He, X.; and Zhao, J. 2017. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3): 1487–1500.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Song, J.; Liu, H.; Geng, F.; and Zhang, C. 2016. Weakly-supervised classification of pulmonary nodules based on shape characters. In *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, 228–232. IEEE.

Tang, C.-Z.; and Zhang, M.-L. 2017. Confidence-rated discriminative partial label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Tang, Z.; Yang, H.; and Chen, C. Y.-C. 2023. Weakly Supervised Posture Mining for Fine-Grained Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23735–23744.

Tian, Y.; Yu, X.; and Fu, S. 2023. Partial label learning: Taxonomy, analysis and outlook. *Neural Networks*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, D.-B.; Li, L.; and Zhang, M.-L. 2019. Adaptive graph guided disambiguation for partial label learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 83–91.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2021. Pico: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*.

Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2021. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12): 8927–8948.

Wen, H.; Cui, J.; Hang, H.; Liu, J.; Wang, Y.; and Lin, Z. 2021. Leveraged weighted loss for partial label learning. In *International Conference on Machine Learning*, 11091–11100. PMLR.

Wu, D.-D.; Wang, D.-B.; and Zhang, M.-L. 2022. Revisiting consistency regularization for deep partial label learning. In *International Conference on Machine Learning*, 24212–24225. PMLR.

Xia, S.; Lv, J.; Xu, N.; Niu, G.; and Geng, X. 2023. Towards Effective Visual Representations for Partial-Label Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15589–15598.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Xu, N.; Lv, J.; and Geng, X. 2019. Partial label learning via label enhancement. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 33, 5557–5564.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33: 21969–21980.

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2022. Vgse: Visually-grounded semantic embeddings for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9316–9325.

Yang, X.; Song, Z.; King, I.; and Xu, Z. 2022. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, F.; Feng, L.; Han, B.; Liu, T.; Niu, G.; Qin, T.; and Sugiyama, M. 2021. Exploiting class activation value for partial-label learning. In *International Conference on Learning Representations*.

Zhang, M.-L.; and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, 4048–4054.

Zhao, Y.; Yan, K.; Huang, F.; and Li, J. 2021. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15079–15088.

Zhou, Z.-H. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1): 44–53.

Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; and Shan, Y. 2022. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4692–4702.

Zhu, X.; and Goldberg, A. B. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1): 1–130.