

Twofold Debiasing Enhances Fine-Grained Learning with Coarse Labels

Xin-yang Zhao¹, Jian Jin¹, Yang-yang Li¹, Yazhou Yao^{1*}

¹School of Computer Science and Engineering, Nanjing University of Science and Technology,
Nanjing, Jiangsu, China
{zhaoxy, jinj, lyylyyi599, yazhou.yao}@njust.edu.cn

Abstract

The Coarse-to-Fine Few-Shot (C2FS) task is designed to train models using only coarse labels, then leverages a limited number of subclass samples to achieve fine-grained recognition capabilities. This task presents two main challenges: coarse-grained supervised pre-training suppresses the extraction of critical fine-grained features for subcategory discrimination, and models suffer from overfitting due to biased distributions caused by limited fine-grained samples. In this paper, we propose the Twofold Debiasing (TFB) method, which addresses these challenges through detailed feature enhancement and distribution calibration. Specifically, we introduce a multi-layer feature fusion reconstruction module and an intermediate layer feature alignment module to combat the model’s tendency to focus on simple predictive features directly related to coarse-grained supervision, while neglecting complex fine-grained level details. Furthermore, we mitigate the biased distributions learned by the fine-grained classifier using readily available coarse-grained sample embeddings enriched with fine-grained information. Extensive experiments conducted on five benchmark datasets demonstrate the efficacy of our approach, achieving state-of-the-art results that surpass competitive methods.

Code — <https://github.com/Faithzh/TFB>

Introduction

Fine-grained recognition is a longstanding and fundamental problem in pattern recognition, underpinned by a diverse set of real-world applications (Wei et al. 2021; Shen et al. 2022; Xu et al. 2023b; Shen et al. 2024; Zhou et al. 2020; Chen et al. 2019; Zhang et al. 2022; Sun et al. 2021b; Yao et al. 2020). However, acquiring fine-grained labels in large quantities is challenging. On the one hand, this difficulty stems from the complexity of annotating with fine-grained knowledge; on the other hand, in many practical applications, the target label set of interest is not static and may change according to evolving needs. In contrast, coarse labels are more readily obtainable and often constitute the majority of dataset annotations. Therefore, Coarse-to-Fine Few-Shot (C2FS) task was proposed to exploit abundant

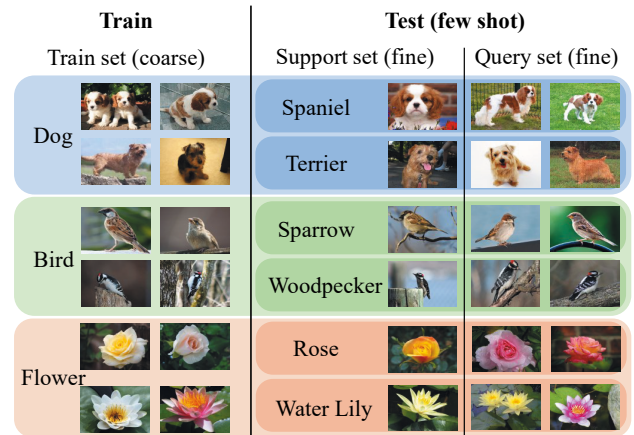


Figure 1: The Coarse-to-Fine Few-Shot (C2FS) task involves training a feature extractor using a training set with coarse labels, then the classifier leverages a support set with fine labels from a limited number of samples to achieve fine-grained recognition capabilities during testing.

coarse labels to enhance fine-grained recognition (Bukchin et al. 2021). The C2FS framework trains models on large-scale coarse-grained samples and then rapidly adapts them to finer recognition tasks using only a few fine-grained samples. As shown in Figure 1, unlike most previous few-shot tasks (Hariharan and Girshick 2017; Wei et al. 2019), the new classes in the C2FS task are fine-grained subclasses of pre-trained categories, contained within known coarse categories.

This task is challenging because coarse-grained labels usually represent broader, more generalized categories. Therefore, the model may learn features that lack intra-class differentiability in pre-training, making it ineffective at distinguishing subtle differences between fine-grained categories. This mismatch in feature representation makes it challenging for the classifier to achieve satisfactory classification performance, even when trained on fine-grained labels. The predominant approach in C2FS effectively utilizes coarse-grained supervision to distinguish different coarse categories, and then combines this with contrastive self-supervised methods to learn distinctive intra-class fea-

*Corresponding author.

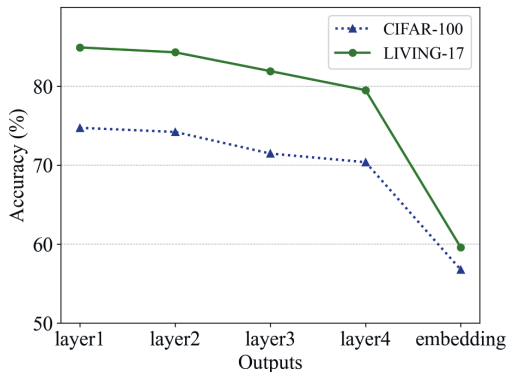


Figure 2: We froze the model trained under coarse-grained supervision and added a classification model to the output of each different layer. Then, these classification models were trained under fine-grained supervision to test their capability for fine-grained classification based on these different layer outputs.

tures (Bukchin et al. 2021; Yang, Yang, and Chen 2021; Touvron et al. 2021; Feng and Patras 2023; Xu et al. 2023a). In this paper, we follow this framework and propose a targeted approach for the C2FS task, termed **TwoFold deBiasing** enhances fine-grained learning (TFB).

As shown in Figure 2, we tested the fine-grained classification potential of outputs from different layers of a coarse-grained pre-trained model. The experiments reveal that under the influence of coarse-grained supervision, fine-grained semantic information gradually diminishes as the network deepens, with the most severe loss occurring in the deepest layer. The model tends to extract and emphasize semantic information directly related to the supervisory signals as the network depth increases while neglecting details that seem less directly relevant to the final task. In fact, although more fine-grained information could help distinguish between coarse categories, the model tends to focus on attributes that can most easily and directly differentiate between coarse categories. We refer to this phenomenon as the simplicity bias caused by coarse-grained supervision (Shah et al. 2020), i.e., neural networks simply ignore several complex predictive features in the presence of few simple predictive features.

To mitigate the simplicity bias induced by coarse-grained supervision and improve the fine-grained feature recognition capabilities, we enhance the model’s self-consistency. The model indirectly learns fine-grained feature representations by reconstructing the original image from the features outputted at the end, which contains all the detailed information (Wei et al. 2023a). Nevertheless, this approach might be suboptimal as the original image encompasses a vast amount of information irrelevant to fine-grained semantics, such as location, background, and pose. These redundant details could potentially be introduced into the model output. We address this by integrating features from different layers, compensating for the semantically irrelevant information during the reconstruction process. Based on our

findings presented in Figure 2, intermediate layer representations retain fine-grained discriminative features, even when directly influenced by coarse-grained supervisory constraints. More intuitively, the intermediate layer representations achieve a balance between preserving fine-grained discriminative attributes and excluding information irrelevant for coarse-grained classification. Building on this insight, we utilize these well-balanced intermediate representations to enhance the fine-grained representational capacity of the final embeddings.

More importantly, due to the limited availability of fine-grained samples, a good model finds it difficult to learn sufficient details from these sparse samples to accurately differentiate between highly similar categories. Notably, models pre-trained on coarse-grained supervision have already outperformed models pre-trained with fine-grained supervision. Especially on CIFAR-100, the previously best method (Xu et al. 2023a) even significantly surpasses the fine upper-bound baseline, exceeding it by 5.89% and 4.93% under 5-way 1-shot and all-way 1-shot conditions respectively. Intuitively, the feature extractor, either under fine-grained label supervision or with the aid of pretext tasks, has developed effective fine-grained discriminative abilities. But classifier trained with sparse data tend to learn a biased distribution, which does not fully reflect the true capabilities of the feature extractor. Previous works have focused on the feature extraction capabilities of backbones, neglecting the equally crucial role of the classifier. By exploiting the intrinsic relationships between base and target class representations, we aim to mitigate the distributional bias of target classes with limited samples by transferring statistical knowledge from richly-sampled base classes. By using the information from the base class samples, we enable the classifier to more accurately reveal the ground-truth data distribution, thereby learning a well-generalized model.

The primary contributions of this work can be summarized as follows:

- We effectively enhance the fine-grained representational capability of features through multi-layer feature fusion reconstruction and intermediate layer feature alignment, thereby mitigating the simplicity bias in features under coarse-grained supervision.
- Our analysis reveals that the distribution bias stemming from limited samples substantially constrains the model’s discriminative capability. To mitigate this issue, we exploit the intrinsic relationships between base and target classes to rectify the distribution bias inherent in fine-grained few-shot scenarios.
- Extensive experiments have demonstrated the effectiveness of TFB across multiple datasets, achieving state-of-the-art accuracy. Our approach offers a new solution for the C2FS task by simultaneously targeting the design of both the feature extractor and classifier.

Related Work

Coarse-to-Fine Learning

Coarse-to-Fine learning (Ristin et al. 2015; Taherkhani et al. 2019; Wei et al. 2023b) has become a significant focus in

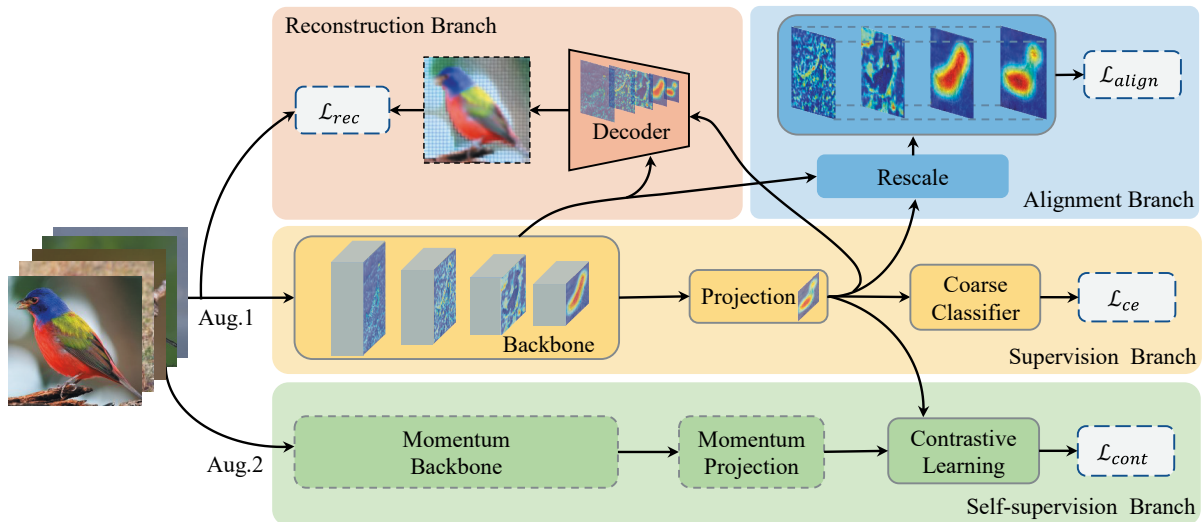


Figure 3: Overview of the TFB model training framework. The training objective consists of four components: coarse-grained cross-entropy loss \mathcal{L}_{CE} for classifying coarse categories, reconstruction loss \mathcal{L}_{rec} and alignment loss \mathcal{L}_{align} to enhance the fine-grained representation capabilities of embeddings, and contrastive learning loss \mathcal{L}_{cont} to learn better feature distributions within coarse category spaces. During testing, only the backbone is used as the feature extractor.

computer vision and machine learning, aiming to leverage coarse-grained labeled data to enhance fine-grained recognition. ANCOR (Bukchin et al. 2021) introduced a fine-grained angular contrastive learning method that uses coarse labels to guide the angular loss function and generate sample pairs. Sun et al. (Sun et al. 2021a) proposed a dynamic metric learning approach that adapts the metric space to different semantic scales. Yang (Yang, Yang, and Chen 2021) address the cross-granularity gap by clustering coarse classes into pseudo-fine classes and introducing a meta-embedder that jointly optimizes both visual and semantic discrimination for effective pseudo-labeling. Graftit (Touvron et al. 2021) implements a joint learning scheme that integrates instance and coarse label supervision losses to learn fine-grained image representations. MaskCon (Feng and Patras 2023) proposes a contrastive learning method that uses coarse labels to generate masked soft labels, leveraging both inter-sample relations and coarse label information. HCM (Xu et al. 2023a) embeds visual representations into a hyperbolic space and enhances their discriminative power using hierarchical cosine margins. FALCON (Grcic, Gadgetsky, and Brbic 2024) enables fine-grained class discovery from coarsely labeled data without requiring fine-grained level supervision. Some studies (Zhao et al. 2021; Xiang et al. 2022) have also explored incremental few-shot learning with mixing of coarse and fine labels.

Self-supervised Learning

As a branch of unsupervised learning, self-supervised learning (SSL) focuses on extracting discriminative features from large-scale unlabeled data, bypassing the need for human annotations (Gui et al. 2024). Context-based SSL methods exploit inherent contextual relationships, such as spatial structures and texture consistency within intact samples, using

domain-specific pretext tasks (Hu et al. 2024) like image rotation prediction (Gidaris, Singh, and Komodakis 2018) and jigsaw puzzles (Noroozi and Favaro 2016). Contrastive learning has progressed from explicitly using negative examples (He et al. 2020; Chen et al. 2020) to ideas like self-distillation and feature decorrelation, all of which adhere to the principle of promoting positive example invariance while pushing apart different examples (Wang and Isola 2020). Masked image modeling (Assran et al. 2022) reconstructs pixels (He et al. 2022) or local features (Baevski et al. 2022), leveraging co-occurrence relationships among image patches as supervision signals.

Few-shot Learning

Few-shot learning (FSL) is designed to mitigate the model’s heavy reliance on large-scale training data, enabling models to quickly generalize to new tasks with limited annotated training data (Wei et al. 2024). This recognition paradigm is ideally suited for fine-grained recognition, as collecting fine-grained samples is often costly. The methodologies for FSL can be taxonomically divided into three primary categories: data augmentation-based methods (Hariharan and Girshick 2017), optimization-based methods (Ravi and Larochelle 2017), and metric-based methods (Jin et al. 2024; Wei et al. 2022a). Recent advances in metric-based methods have incorporated attention mechanisms (Vaswani et al. 2017) to model dependencies between elements in input sequences, utilized dense feature vectors to extract richer and finer image-to-image correlations (Liu et al. 2022). While the aforementioned works have shown success in traditional FSL tasks, they still require a significant number of labels with the same granularity as the test data during training.

Main Approach

Problem Definition

In the study framework as detailed by ANCOR (Bukchin et al. 2021), define $\mathcal{Y}_{coarse} = \{y_1, \dots, y_N\}$ as a collection of N broad training categories. We let $\mathcal{S}_{train}^{coarse} = \{(I_i, y_i) | y_i \in \mathcal{Y}_{coarse}\}_{i=1}^M$ represent a compilation of M training images, each marked solely with labels from \mathcal{Y}_{coarse} . Consider $\mathcal{Y}_{fine} = \{y'_{1,1}, \dots, y'_{1,k_1}, \dots, y'_{N,1}, \dots, y'_{N,k_N}\}$ as a collection of fine-grained subclasses under the broad categories \mathcal{Y}_{coarse} . The encoder E trained by $\mathcal{S}_{train}^{coarse}$ maps images I into a d -dimensional feature space \mathcal{F} . At the testing phase, given a k -shot support set $\mathcal{S}_{sup}^{fine} = \{(I_r, y'_r) | y'_r \in \mathcal{Y}_{fine}^w\}_{r=1}^{k \cdot w}$ for a subset $\mathcal{Y}_{fine}^w \subseteq \mathcal{Y}_{fine}$ comprising w fine categories, our objective is to train a classifier $C_{fine} : \mathcal{F} \rightarrow \mathcal{Y}_{fine}^w$ to achieve optimal accuracy on the query set of \mathcal{Y}_{fine}^w fine categories.

Embedding Learning Debias

The objective of this work is to learn an embedding function $f_{embed}(\cdot)$ that leverages coarsely-labeled training data to project input samples into a feature space with enhanced fine-grained discriminability. As shown in Figure 3, our architecture is built upon the self-supervised learning framework MoCo (He et al. 2020), consisting of a feature encoder E with global average pooling, an projector P , a second pair of momentum encoder E_m and momentum projector P_m for encoding the positive keys in the contrastive objective that are momentum-updated from E and P and a coarse linear classifier C_{coarse} (Bukchin et al. 2021). Our proposed framework incorporates two essential components: a decoder D for input reconstruction through multi-layer feature fusion, and an alignment module for bridging the intermediate representations and final embeddings.

Firstly, we use coarse-grained labels y for supervised learning, enabling the model to initially acquire coarse-grained recognition capabilities. Using I_q to represent the input after data augmentation, we can obtain:

$$\mathcal{L}_{CE} = CrossEntropy(C_{coarse}(P(E(I_q))), y). \quad (1)$$

Multi-layer Feature Fusion for Reconstruction Reconstruction branch aims to alleviate the simplicity bias caused by coarse-grained supervision by enhancing the model’s self-consistency (Ma, Tsao, and Shum 2022). Self-consistency compels the model to learn a variety of features ranging from basic textures to complex objects, thereby achieving a more comprehensive feature representation (Wei et al. 2023a). Specifically, the model indirectly learns fine-grained feature representations by reconstructing the original image I_q , which contains all the detail information. However, during the reconstruction process, it is detrimental for the model to focus on task-irrelevant information and structures within the input image I_q . Therefore, we reconstruct the image by merging features from different layers, utilizing shallow layer features to provide low-level features such as background and positional information, allowing the deeper layers to focus on semantic reconstruction.

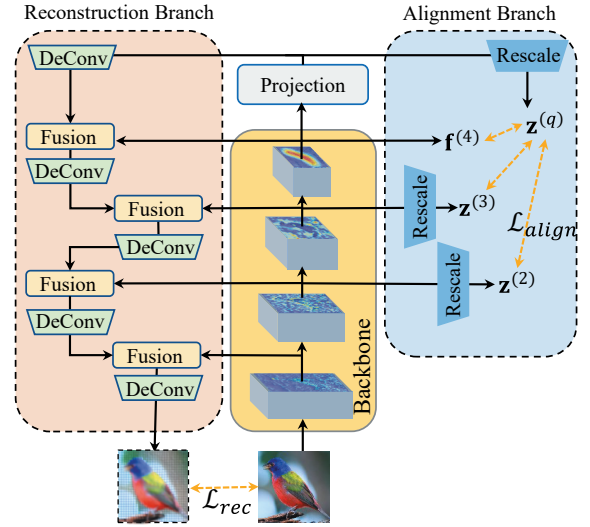


Figure 4: Multi-layer feature fusion for reconstruction branch and Intermediate layer feature alignment branch. The proposed fusion module employs cascaded concatenation operations and convolution layers to facilitate adaptive feature aggregation across different scales. $z^{(i)}$ represents the rescaled $f^{(i)}$.

More specifically, the encoder E transforms an input image I_q into a set of latent representations across multiple levels of abstraction in different layers. Let $f^{(l)}$ represent the feature map at layer l , where l ranges from the shallow to the deep layers within the encoder, capturing varying degrees of semantic and structural details:

$$f^{(l)} = E^{(l)}(I_q). \quad (2)$$

Then, a deconvolutional network D (Zeiler et al. 2010) is applied upon $f^{(l)}$ and f^q , which is expected to reconstruct the input raw image I_q . As detailedly illustrated in Figure 4, this reconstruction is achieved by integrating feature maps $f^{(l)}$ from multiple layers:

$$\hat{I} = D(f^{(1)}, f^{(2)}, \dots, f^q), \quad (3)$$

where f^q is the embedding of I_q . We used the MSE loss as the reconstruction loss \mathcal{L}_{rec} to guide the learning process in faithfully regenerating the original image.

Intermediate Layer Feature Alignment In the experiments shown in Figure 2, we found that the information of fine-grained categories under coarse-grained supervision is mainly lost in the embeddings of the deepest layer outputs, while the model’s intermediate layer outputs still perform well in fine-grained recognition. As shown in Figure 4, a feature alignment module is designed to utilize intermediate layer features for self-knowledge distillation. Each intermediate feature map $f^{(i)}$ and the final embedding f^q are rescaled to match the spatial dimensions of the reference layer feature map $f^{(4)}$. This rescaling is denoted by a function $Rescale(f, s)$, where s represents the target scale size

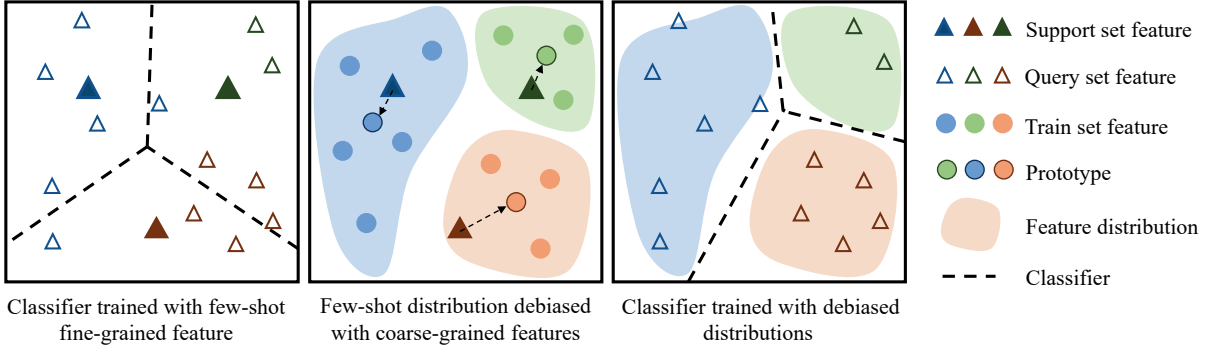


Figure 5: Training a classifier with limited fine-grained samples leads to overfitting (left). Correcting the distribution of few-shot samples in the feature space using training set features (middle). The classifier trained from the calibrated distribution features has better generalization capabilities (right).

derived from $\mathbf{f}^{(4)}$. The function $Rescale(\mathbf{f}, s)$ is actually composed of no more than two layers of convolution or deconvolution. The rescaled feature maps are then aligned through MSE loss minimization. The alignment process effectively acts as a form of regularization, encouraging the network to maintain a consistent representation of details across its depth. The overall align loss \mathcal{L}_{align} is

$$\mathcal{L}_{align} = \sum_{i=2}^4 \|\text{Rescale}(\mathbf{f}^{(i)}, s) - \text{Rescale}(\mathbf{f}^q, s)\|_2^2. \quad (4)$$

Although fine-grained information is enriched through image reconstruction and feature alignment, features may still exhibit arbitrary distributions within coarse category regions, potentially hindering discrimination between fine-grained categories. Therefore, we adopt the contrastive learning strategy \mathcal{L}_{cont} proposed in ANCOR (Bukchin et al. 2021) to enhance the self-organization and separability of fine-grained features.

The final training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{cont} + \alpha \mathcal{L}_{rec} + \beta \mathcal{L}_{align}, \quad (5)$$

where α and β are tuning parameters that balance the importance of different losses against the primary task performance. To simplify the tuning process, we set $\alpha = \beta$.

Fine-grained Classifier Debias

When trained on sparsely available fine-grained data, well-pretrained models also suffers from distribution bias and overfitting issues. Considering this, we aim to utilize the extensive information from coarse-grained training datasets to address these issues. If we can identify the relationships between training set samples and support set samples, we can calibrate the distribution of fine-grained classes using training set samples. Calibrating the distribution in the original data space is challenging, so we attempt to calibrate the distribution in the feature space. The feature space has far fewer dimensions than the original data space and is also easier to calibrate (Xian et al. 2018; Yang et al. 2021). Therefore, our approach is based on the premise that two semantically similar samples, once processed through a feature extractor, will yield proximate features in the feature space.

Following the feature extractor training, we establish a feature repository \mathcal{T} by extracting and storing feature representations from the entire training set. Using $\mathcal{S}_{sup}^{y'}$ to represent the support set features corresponding to the fine-grained class y' , the prototype $P^{y'}$ (Wei et al. 2022b) is initialized as the centroid of $\mathcal{S}_{sup}^{y'}$. We first determine the coarse category y to which y' belongs. The y is determined through a k-Nearest Neighbors (kNN) classification scheme operating on the training set feature repository \mathcal{T} . Specifically, for the prototype $P^{y'}$, its coarse label y is assigned based on majority voting among its k nearest neighbors:

$$y = \arg \max_c \sum_{i=1}^k \mathbf{1}(y_i = c), \quad (6)$$

where y_i denotes the label of the i -th nearest neighbor in feature space, and $\mathbf{1}(\cdot)$ is the indicator function. Upon identifying the coarse label y , A feature subset \mathcal{T}_y is extracted from \mathcal{T} , consisting of embeddings labeled as y . This subset provides a potential addition feature pool for $\mathcal{S}_{sup}^{y'}$:

$$\mathcal{T}_y = \{x \in \mathcal{T} : \text{label}(x) = y\}. \quad (7)$$

Then, an additional support set $\mathcal{S}_{add}^{y'}$ is constructed by selecting the m nearest neighbors in feature space from \mathcal{T}_y , where the distance is computed with $P^{y'}$. Subsequently, the average features of $\mathcal{S}_{add}^{y'}$ are used to correct the original prototype $P^{y'}$:

$$P^{y'} = \frac{1}{|\mathcal{S}_{sup}^{y'}|} \sum_{\mathbf{x} \in \mathcal{S}_{sup}^{y'}} \mathbf{x} + \frac{1}{|\mathcal{S}_{add}^{y'}|} \sum_{\mathbf{x} \in \mathcal{S}_{add}^{y'}} \mathbf{x}. \quad (8)$$

This prototype calibration process needs to be repeated multiple times. The corrected prototype is then used to continue finding new m features to add to $\mathcal{S}_{add}^{y'}$ from \mathcal{T}_y until the feature count in $\mathcal{S}_{add}^{y'}$ meets the requirement of n items. Finally, the additional support set $\mathcal{S}_{add}^{y'}$ along with the original support set $\mathcal{S}_{sup}^{y'}$ are then served as the training data for fine-grained classifier.

Methods	LIVING-17		NONLIVING-26		ENTITY-13		ENTITY-30	
	5-way	all-way	5-way	all-way	5-way	all-way	5-way	all-way
Fine upper-bound	90.75±0.48	62.65±0.18	90.33±0.47	60.68±0.14	94.72±0.33	65.18±0.09	94.02±0.36	63.72±0.10
MoCo-v2	56.66±0.70	18.57±0.11	63.51±0.75	21.07±0.11	82.00±0.67	33.06±0.07	80.37±0.62	28.62±0.06
MoCo-v2-ImageNet	82.21±0.73	40.29±0.14	77.07±0.78	34.78±0.13	85.24±0.60	35.62±0.08	83.06±0.62	31.73±0.08
SWAV-ImageNet	79.83±0.65	38.79±0.15	76.26±0.71	33.94±0.11	81.15±0.65	33.57±0.07	79.91±0.54	31.15±0.07
ANCOR	89.23±0.55	45.14±0.12	86.23±0.54	43.10±0.11	90.58±0.54	42.29±0.08	88.12±0.54	41.79±0.08
SCGM-G	89.72±0.54	48.74±0.15	89.87±0.51	49.25±0.13	90.15±0.51	40.00±0.08	92.90±0.46	42.17±0.08
SCGM-A	90.97±0.55	49.31±0.16	88.78±0.55	46.93±0.13	88.48±0.59	41.07±0.09	91.22±0.51	44.14±0.09
PE-HCM	90.94±0.43	53.09±0.11	89.97±0.42	50.12±0.11	91.24±0.38	41.64±0.09	92.95±0.40	44.53±0.09
Ours	91.89±0.54	53.42±0.12	90.35±0.53	49.74±0.11	91.56±0.53	43.23±0.08	91.73±0.52	45.03±0.08

Table 1: Accuracy comparisons of different methods on the BREEDS dataset under 5-way 1-shot and all-way 1-shot settings. Bold numbers are the best results.

Experiments

Benchmark Datasets

Our experiments were performed on BREEDS (Santurkar, Tsipras, and Madry 2020) and CIFAR-100 (Krizhevsky 2009) datasets. BREEDS is a collection of four datasets derived from ImageNet and recalibrated to ensure that classes at the same hierarchy level exhibit similar visual granularity. The CIFAR-100 dataset includes 20 coarse-grained classes and a total of 100 fine-grained classes.

Baselines

We compare our method with the most relevant state-of-the-art models on Coarse-to-Fine Few-Shot task: (1) MoCo-v2 (Chen et al. 2020), trained on respective training sets. (2) MoCo-v2-ImageNet (Chen et al. 2020), represents the complete ImageNet pre-trained version of the model. (3) SWAV-ImageNet, the official model described in (Caron et al. 2020). (4) ANCOR (Bukchin et al. 2021), integrates supervised inter-class and self-supervised intra-class learning with an innovative angular normalization component. (5) SCGM (Ni et al. 2022), simulates the sample generation process using hierarchical class structures and explicitly incorporates latent variables to account for unobserved subclasses. (6) PE-HCM (Xu et al. 2023a), embeds visual representations into hyperbolic space, enhancing their discriminative capacity through hierarchical cosine margins. (7) Fine upper-bound (Bukchin et al. 2021), naturally trained on fine-grained labels.

Additional Implementation Details

For fair comparisons with other methods, we use ResNet-12 (He et al. 2016) as the backbone network for CIFAR-100 (Krizhevsky 2009) and ResNet-50 (He et al. 2016) for BREEDS (Santurkar, Tsipras, and Madry 2020). The output dimension of these backbone networks is $d = 2048$ or $d = 640$. In all experiments, the projector is implemented as a three-layer MLP with an output dimension of 128. The model is trained using the SGD optimizer on 4 GeForce RTX 3090 GPUs for 200 epochs. For CIFAR-100 and BREEDS, the batch sizes are 512 and 256 respectively;

Methods	5-way	all-way
Fine upper-bound	75.53±0.68	31.35±0.11
ANCOR	74.56±0.70	29.84±0.11
SCGM-G	76.19±0.73	29.92±0.11
SCGM-A	77.37±0.77	25.91±0.10
PE-HCM	81.42±0.69	36.28±0.12
Ours	84.15±0.69	39.03±0.11

Table 2: Comparisons on CIFAR-100 (Krizhevsky 2009). Bold numbers are the best results.

the initial learning rates are 0.12 and 0.03; the hyperparameters α are set at 1 and 10 respectively. The learning rates decrease by tenfold at the 140th and 180th epochs. We implement random data augmentation techniques including random resized crop, random horizontal flipping, and random color jitter during training. Other hyperparameter settings remain the same as those used in ANCOR. During the test phase, we evaluate the performance using 5-way and all-way 1-shot settings. The hyperparameters k , m and n are set to 10, 20 and 100 respectively. The evaluation is conducted on 1000 random episodes, and we report the mean accuracy along with a 95% confidence interval.

Main Result

Following common experimental setups (Tian et al. 2020), we present results for both 5-way 1-shot and all-way 1-shot configurations, with 15 queries per test episode. Table 1 and table 2 report the average accuracy for the core use-case of the task evaluation. It is evident that our approach delivers superior performance across all datasets. Especially on the CIFAR-100 dataset, we observed an approximate 3% significant advantage in accuracy. Traditional methods have focused on designing fine-grained pretext tasks to enable feature extractors to learn fine-grained feature distributions from coarse-grained datasets, subsequently verifying these distributions through fine-grained classifier.

Our method introduces an alternative novel solution that not only preserves the fine-grained modeling capabilities of the feature extractor but also guides the fine-grained clas-

Methods			LIVING-17		CIFAR-100	
REC	ALIGN	CD	5-way	all-way	5-way	all-way
			89.45	45.59	77.67	33.83
✓			90.74	47.32	80.05	36.80
✓	✓		91.06	48.73	81.79	37.41
✓	✓	✓	91.89	53.42	84.15	39.03

Table 3: Comparisons of our proposals on LIVING-17 and CIFAR-100. REC: Multi-layer feature fusion for reconstruction. ALIGN: Intermediate layer feature alignment. CD :Fine-grained classifier debias.

Methods	LIVING-17		CIFAR-100	
	5-way	all-way	5-way	all-way
All features	91.02	51.71	83.19	37.61
kNN-based	91.89	53.42	84.15	39.03
Coarse labeled	92.91	54.95	86.92	43.22

Table 4: Comparative experiment of different coarse-grained supervisions in the debiasing process of fine-grained classifiers. All features: completely disregards coarse granularity, obtaining $\mathcal{S}_{add}^{y'}$ from all training features \mathcal{T} . kNN-based: uses a kNN classifier to obtain coarse categories in our method. Coarse labeled: the support set is annotated with coarse-grained labels.

sifier to learn the correct distributions as much as possible. Compared to pretext-task-based approaches, our method offers two advantages: (1) Although only a limited amount of fine-grained data is used as the support set, this set provides accurate and authentic fine-grained category information. For instance, as demonstrated in the PE-HCM method (Xu et al. 2023a), the number of clusters significantly influences the performance of fine-grained recognition. (2) Features learned under coarse-grained labels calibrate the distribution bias in fine-grained classifiers for few-shot scenarios, effectively leveraging the relationship between the sparse fine-grained data in the support set and the abundant coarse-grained data in the training set, thereby bridging the gap between few-shot and many-shot conditions.

Ablation Study

The ablation analysis presented in table 3 quantitatively demonstrates the contribution of each model component to the overall performance across two benchmark datasets. The results demonstrate that incorporating the REC component significantly improves accuracy, while the subsequent integration of the ALIGN component brings further enhancements. These two methods effectively maintain the fine-grained information preservation capability of the feature extractor. As there is no explicit modeling of fine-grained classification tasks, our model does not show a significant advantage over other methods before the use of CD. However, the CD strategy noticeably enhances the overall performance. This debiasing of the fine-grained classifier is crucial. In fact, our strategy has the potential to be applicable to previous methods. Overall, experimental results demonstrate that designing both the feature extractor and classifier

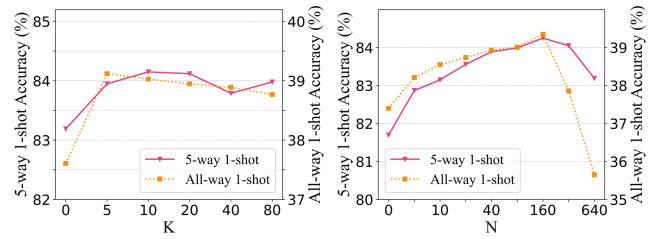


Figure 6: Ablation study of selected hyperparameters on the CIFAR-100 dataset. The Impact of Different kNN Parameters on Results(left). Influence of the Size n of the Additional Support Set $\mathcal{S}_{add}^{y'}$ on Results(right).

specifically for the C2FS is effective and necessary.

Table 4 shows the effectiveness of incorporating coarse-grained labels in the debiasing process of fine-grained classifiers. The results demonstrate that coarse-grained label of the support set plays a crucial role in optimizing classifier performance. With coarse-grained labels available, the model more accurately identifies the corresponding feature subset \mathcal{T}_y for each category, resulting in substantial performance gains, notably a 6% improvement in all-way accuracy on LIVING-17. When coarse-grained categories are disregarded, the CD strategy shows limited improvement. Evidently, utilizing the training set to additionally train a better coarse-grained classifier, instead of relying on simple kNN, can effectively guide the debiasing process.

Additionally, as illustrated in the left side of the Figure 6, we conducted an ablation study on the kNN parameter k , finding that our method is not highly sensitive to k . On the right side of the Figure 6, we experimented with different numbers of coarse granularity features n in $\mathcal{S}_{add}^{y'}$. The results indicate that when n is within a reasonable range, it can significantly improve model performance. However, when n is sufficiently large, $\mathcal{S}_{add}^{y'}$ introduces a large number of false positive samples, leading to incorrect adjustments in the distribution.

Conclusion

In this paper, we introduce a novel methodology called TwoFold deBiasing (TFB), which enhances fine-Grained learning for the C2FS task. Our approach mitigates the simplicity bias of feature extractors under coarse-grained supervision and tackles the biased distributions learned by fine-grained classifiers in few-shot scenarios. Specifically, TFB employs multi-layer feature fusion and intermediate layer feature alignment to enrich the detailed information required for fine-grained classification. Furthermore, by leveraging the hierarchical relationship between coarse and fine labels, TFB incorporates training set features to calibrate biased distributions in few-shot learning, enabling classifiers to better approximate the true data distribution. Our comprehensive experiments across benchmark datasets show the benefits of designing integrated training and testing strategies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62472222), and Natural Science Foundation of Jiangsu Province (No. BK20240080).

References

- Assran, M.; Caron, M.; Misra, I.; Bojanowski, P.; Bordes, F.; Vincent, P.; Joulin, A.; Rabbat, M.; and Ballas, N. 2022. Masked siamese networks for label-efficient learning. In *Proceedings of the European Conference on Computer Vision*, 456–473.
- Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, 1298–1312.
- Bukchin, G.; Schwartz, E.; Saenko, K.; Shahar, O.; Feris, R.; Giryes, R.; and Karlinsky, L. 2021. Fine-grained angular contrastive learning with coarse labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8730–8740.
- Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 9912–9924.
- Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5177–5186.
- Feng, C.; and Patras, I. 2023. MaskCon: Masked contrastive learning for coarse-labelled dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 19913–19922.
- Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*.
- Grcic, M.; Gadetsky, A.; and Brbic, M. 2024. Fine-grained Classes and How to Find Them. In *International conference on machine learning*.
- Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; and Tao, D. 2024. A Survey on Self-supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Hariharan, B.; and Girshick, R. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE international conference on computer vision*, 3018–3027.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 16000–16009.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9729–9738.
- 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.
- Hu, F.; Zhang, C.; Guo, J.; Wei, X.-S.; Zhao, L.; Xu, A.; and Gao, L. 2024. An Asymmetric Augmented Self-Supervised Learning Method for Unsupervised Fine-Grained Image Hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 17648–17657.
- Jin, J.; Shen, Y.; Fu, Z.; and Yang, J. 2024. Few-shot open-set recognition via pairwise discriminant aggregation. *Neurocomputing*, 602: 128214.
- Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Technical report. Science Department, University of Toronto, Tech.*, 1–60.
- Liu, Y.; Zhang, W.; Xiang, C.; Zheng, T.; Cai, D.; and He, X. 2022. Learning to affiliate: Mutual centralized learning for few-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 14411–14420.
- Ma, Y.; Tsao, D.; and Shum, H.-Y. 2022. On the principles of parsimony and self-consistency for the emergence of intelligence. *Frontiers of Information Technology & Electronic Engineering*, 23(9): 1298–1323.
- Ni, J.; Cheng, W.; Chen, Z.; Asakura, T.; Soma, T.; Kato, S.; and Chen, H. 2022. Superclass-Conditional Gaussian Mixture Model For Learning Fine-Grained Embeddings. In *International Conference on Learning Representations*, 1–12.
- Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, 69–84.
- Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.
- Ristin, M.; Gall, J.; Guillaumin, M.; and Van Gool, L. 2015. From categories to subcategories: large-scale image classification with partial class label refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 231–239.
- Santurkar, S.; Tsipras, D.; and Madry, A. 2020. BREEDS: Benchmarks for Subpopulation Shift. *arXiv preprint arXiv:2008.04859*.
- Shah, H.; Tamuly, K.; Raghunathan, A.; Jain, P.; and Ne-trapalli, P. 2020. The Pitfalls of Simplicity Bias in Neural Networks. In *Advances in Neural Information Processing Systems*, 9573–9585.
- Shen, Y.; Sun, X.; Wei, X.-S.; Jiang, Q.-Y.; and Yang, J. 2022. SEMICON: A learning-to-hash solution for large-scale fine-grained image retrieval. In *Proceedings of the European Conference on Computer Vision*, 531–548.

- Shen, Y.; Sun, X.; Wei, X.-S.; Xu, A.; and Gao, L. 2024. Equiangular Basis Vectors: A Novel Paradigm for Classification Tasks. *International Journal of Computer Vision*, 1–26.
- Sun, Y.; Zhu, Y.; Zhang, Y.; Zheng, P.; Qiu, X.; Zhang, C.; and Wei, Y. 2021a. Dynamic metric learning: Towards a scalable metric space to accommodate multiple semantic scales. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5393–5402.
- Sun, Z.; Yao, Y.; Wei, X.-S.; Zhang, Y.; Shen, F.; Wu, J.; Zhang, J.; and Shen, H. T. 2021b. Webly supervised fine-grained recognition: Benchmark datasets and an approach. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10602–10611.
- Taherkhani, F.; Kazemi, H.; Dabouei, A.; Dawson, J.; and Nasrabadi, N. M. 2019. A weakly supervised fine label classifier enhanced by coarse supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6459–6468.
- Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need? In *Proceedings of the European Conference on Computer Vision*, 266–282.
- Touvron, H.; Sablayrolles, A.; Douze, M.; Cord, M.; and Jégou, H. 2021. Graft: Learning fine-grained image representations with coarse labels. In *Proceedings of the IEEE international conference on computer vision*, 874–884.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 5998–6008.
- Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, 9929–9939.
- Wei, X.-S.; Shen, Y.; Sun, X.; Wang, P.; and Peng, Y. 2023a. Attribute-aware deep hashing with self-consistency for large-scale fine-grained image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13904–13920.
- 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.
- Wei, X.-S.; Wang, P.; Liu, L.; Shen, C.; and Wu, J. 2019. Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE Transactions on Image Processing*, 28(12): 6116–6125.
- Wei, X.-S.; Xu, H.-Y.; Yang, Z.; Duan, C.-L.; and Peng, Y. 2024. Negatives Make a Positive: An Embarrassingly Simple Approach to Semi-Supervised Few-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2091–2103.
- Wei, X.-S.; Xu, H.-Y.; Zhang, F.; Peng, Y.; and Zhou, W. 2022a. An embarrassingly simple approach to semi-supervised few-shot learning. *NeurIPS*, 35: 14489–14500.
- Wei, X.-S.; Xu, S.-L.; Chen, H.; Xiao, L.; and Peng, Y. 2022b. Prototype-based classifier learning for long-tailed visual recognition. *Science China Information Sciences*, 65(6): 160105.
- Wei, X.-S.; Xu, Y.-Y.; Zhang, C.-L.; Xia, G.-S.; and Peng, Y.-X. 2023b. CAT: A coarse-to-fine attention tree for semantic change detection. *Visual Intelligence*, 1(1): 3.
- Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5542–5551.
- Xiang, X.; Tan, Y.; Wan, Q.; Ma, J.; Yuille, A.; and Hager, G. D. 2022. Coarse-to-fine incremental few-shot learning. In *Proceedings of the European Conference on Computer Vision*, 205–222.
- Xu, S.-L.; Sun, Y.; Zhang, F.; Xu, A.; Wei, X.-S.; and Yang, Y. 2023a. Hyperbolic space with hierarchical margin boosts fine-grained learning from coarse labels. In *Advances in Neural Information Processing Systems*, volume 36, 71263–71274.
- Xu, Z.; Shang, H.; Yang, S.; Xu, R.; Yan, Y.; Li, Y.; Huang, J.; Yang, H. C.; and Zhou, J. 2023b. Hierarchical painter: Chinese landscape painting restoration with fine-grained styles. *Visual Intelligence*, 1(1): 19.
- Yang, J.; Yang, H.; and Chen, L. 2021. Towards cross-granularity few-shot learning: Coarse-to-fine pseudo-labeling with visual-semantic meta-embedding. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3005–3014.
- Yang, S.; Wu, S.; Liu, T.; and Xu, M. 2021. Bridging the gap between few-shot and many-shot learning via distribution calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9830–9843.
- Yao, Y.; Hua, X.; Gao, G.; Sun, Z.; Li, Z.; and Zhang, J. 2020. Bridging the web data and fine-grained visual recognition via alleviating label noise and domain mismatch. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1735–1744.
- Zeiler, M. D.; Krishnan, D.; Taylor, G. W.; and Fergus, R. 2010. Deconvolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2528–2535.
- Zhang, C.; Lin, G.; Wang, Q.; Shen, F.; Yao, Y.; and Tang, Z. 2022. Guided by meta-set: a data-driven method for fine-grained visual recognition. *IEEE Transactions on Multimedia*, 25: 4691–4703.
- Zhao, H.; Fu, Y.; Kang, M.; Tian, Q.; Wu, F.; and Li, X. 2021. Mgsvf: Multi-grained slow versus fast framework for few-shot class-incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1576–1588.
- Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9719–9728.