

CausalFSFG: Rethinking Few-Shot Fine-Grained Visual Categorization from Causal Perspective

Zhiwen Yang, Jinglin Xu and Yuxin Peng

Abstract—Few-shot fine-grained visual categorization (FS-FGVC) focuses on identifying various subcategories within a common superclass given just one or few support examples. Most existing methods aim to boost classification accuracy by enriching the extracted features with discriminative part-level details. However, they often overlook the fact that the set of support samples acts as a confounding variable, which hampers the FS-FGVC performance by introducing biased data distribution and misguiding the extraction of discriminative features. To address this issue, we propose a new causal FS-FGVC (CausalFSFG) approach inspired by causal inference for addressing biased data distributions through causal intervention. Specifically, based on the structural causal model (SCM), we argue that FS-FGVC infers the subcategories (i.e., effect) from the inputs (i.e., cause), whereas both the few-shot condition disturbance and the inherent fine-grained nature (i.e., large intra-class variance and small inter-class variance) lead to unobservable variables that bring spurious correlations, compromising the final classification performance. To further eliminate the spurious correlations, our CausalFSFG approach incorporates two key components: (1) Interventional multi-scale encoder (IMSE) conducts sample-level interventions, (2) Interventional masked feature reconstruction (IMFR) conducts feature-level interventions, which together reveal real causalities from inputs to subcategories. Extensive experiments and thorough analyses on the widely-used public datasets, including CUB-200-2011, Stanford Dogs, and Stanford Cars, demonstrate that our CausalFSFG achieves new state-of-the-art performance. The code is available at https://github.com/PKU-ICST-MIPL/CausalFSFG_TMM.

Index Terms—Few-shot fine-grained visual categorization, causal intervention, structural causal model, inherent fine-grained nature

I. INTRODUCTION

Fine-grained visual categorization (FGVC) is a lasting and crucial problem in computer vision [1], which endeavors to discern various subcategories that belong to a shared superclass. The FGVC task poses significant difficulties stemming from the inherent fine-grained nature of large intra-class variance and small inter-class variance. Since deep learning has demonstrated great potential in computer vision tasks, existing fine-grained visual categorization techniques leverage deep learning models extensively, which heavily rely on a substantial amount of labeled images [2]–[8]. However, different from the traditional image classification task, the labels

This work was supported by the grants from the National Natural Science Foundation of China (62525201, 62132001, 62432001) and Beijing Natural Science Foundation (L247006, L257005).

Zhiwen Yang and Yuxin Peng are with the Wangxuan Institute of Computer Technology, Peking University, Beijing, 100871, China.

Jinglin Xu is with the School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China.

Corresponding author: Yuxin Peng (e-mail: pengyuxin@pku.edu.cn).

of fine-grained images in the FGVC task are impractical and expensive due to the scarcity of available samples and the requirement for fine-grained expertise. Therefore, achieving accurate categorization with minimal labeled samples remains challenging.

In order to emulate the human capacity of acquiring novel knowledge from a limited set of samples [9], few-shot fine-grained visual categorization (FS-FGVC) [10]–[12] has garnered considerable attention from academic researchers, whose primary objective is to discern novel subcategories given limited support samples. The key challenges of FS-FGVC are twofold: (1) The inherent fine-grained nature, where images within the same subcategory can exhibit significant variations and different subcategories often share a high degree of visual similarity. (2) The few-shot condition, the selected images cannot comprehensively represent the entire subcategory's variance, even introducing spurious correlations. To tackle these challenges, existing methods adopt several techniques, such as meta-learning [13]–[15], data augmentation [16]–[18], transfer learning [19]–[21], and metric learning [22]–[24]. These methods primarily emphasize the recognition of discriminative part-level details to address the first challenge for better classification accuracy, but overlook the fact that the set of support samples acts as a confounder, impeding the classification performance with biased data distribution and misguidance in extracting discriminative features. For instance, when selecting categories exhibiting large variance, such as “Fish Cow” and “Yellow Billed Cuckoo”, models are prone to focusing on conspicuous coarse-grained features like shapes and outlines, while overlooking discriminative fine-grained features like local textures. Therefore, varied class selections lead to distinct feature distributions that deviate significantly from the joint distribution of the entire dataset, as illustrated in Fig. 1 (a).

To address the aforementioned issues, we reformulate FS-FGVC with a structural causal model (SCM), mitigating the biased data distribution from the few-shot condition via causal intervention. Beginning with a causal interpretation of FS-FGVC, our CausalFSFG approach illustrates the causalities among the input, features, prediction, fine-grained dataset, observable data, and inherent fine-grained nature, as depicted in Fig. 1 (b). In the context of FGVC, the input samples are directly drawn i.i.d. from the fine-grained dataset, allowing the entire fine-grained dataset to be observed during training. Consequently, the classification model is able to learn from the joint distribution of all subcategories in the fine-grained dataset, and thereby address the challenge of inherent fine-grained nature. While in the FS-FGVC task, a subset of

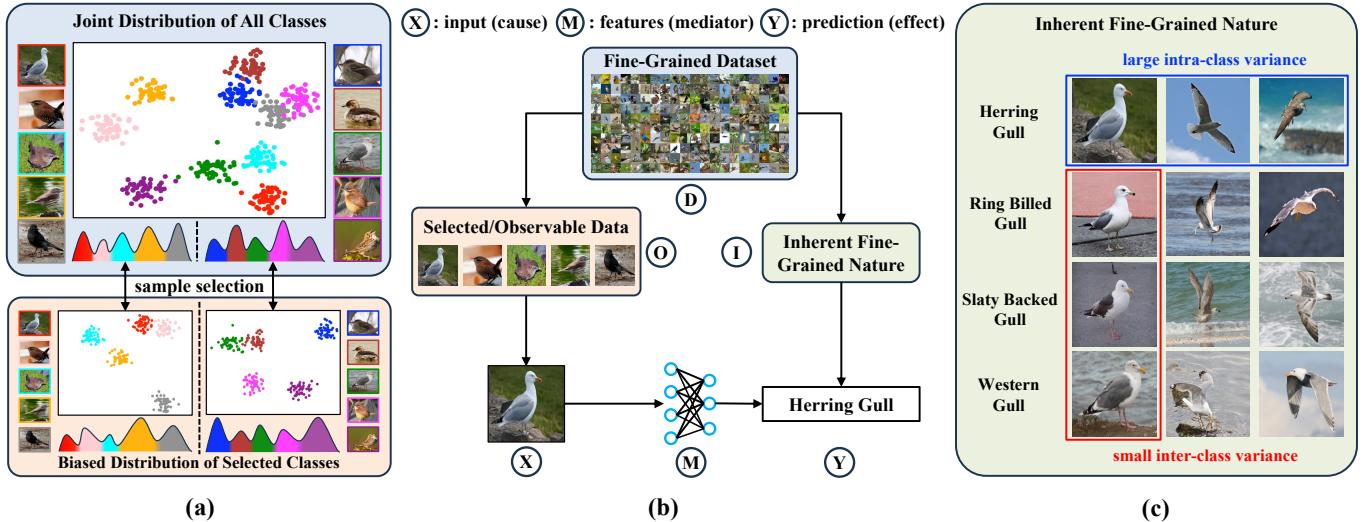


Fig. 1. (a) Illustration of the distribution bias caused by the selection operation under the few-shot condition. (b) Illustration of aligning the FS-FGVC problem with the Structural Causal Model assumption, aiming to infer the accurate prediction (effect) from the input (cause) through extracted features (mediator). (c) Illustration of the inherent fine-grained nature of the fine-grained data.

specific classes and samples are selected from the fine-grained dataset to construct the observable data comprising certain support and query samples, serving as the input for performing few-shot classification. As a consequence, the selected subset of observable data hinders the causalities from the fine-grained dataset and the inherent fine-grained nature, introducing biased data distribution to the training process. In our SCM assumption, this subset acts as a confounding variable, and the accompanying biased data distribution further makes the set of confounding variables (D, O, I) unobservable, causing spurious correlations that limit the classification performance.

To eliminate the spurious correlations, we propose a new CausalFSFG approach that learns discriminative features concerning the entire fine-grained dataset on two levels. On the sample level, the interventional multi-scale encoder (IMSE) is proposed to mitigate the biased distributions of selected samples. On the feature level, the interventional masked feature reconstruction (IMFR) is proposed to extract more discriminative features across the query sets. The above complementary modules collectively conduct the causal intervention on the input samples (cause) and the extracted features (mediator), thereby revealing the real causalities from inputs to subcategories for better FS-FGVC performance.

The main contributions can be summarized as follows:

- We reformulate the FS-FGVC task from the causal perspective to alleviate the biased data distribution caused by the few-shot condition.
- We propose the CausalFSFG approach that reveals real causalities from inputs to subcategories by conducting the sample-level and feature-level interventions with an interventional multi-scale encoder (IMSE) and an interventional masked feature reconstruction (IMFR) modules.
- Extensive comparison experiments on the widely-used public datasets, including CUB-200-2011, Stanford Dogs, and Stanford Cars, demonstrate that our CausalFSFG achieves new state-of-the-art.

The rest of the paper is organized as follows: Section II provides a brief review of related work on fine-grained visual

categorization, few-shot visual categorization, and causal inference. Section III states the detailed definition and parameterization of the FS-FGVC problem, as well as the proposed SCM assumption as its causal reformulation. Section IV describes the implementation of the proposed CausalFSFG approach based on the frontdoor adjustment rule. Section V presents the comparison experiments, analyses, and ablation studies. Finally, Section VI concludes the paper.

II. RELATED WORK

This section briefly reviews related works about fine-grained visual categorization, few-shot visual categorization, and causal inference.

A. Fine-Grained Visual Categorization

Recent FGVC methods predominantly focus on identifying discriminative regions and extracting features for fine-grained visual classification. Among them, some methods locate distinctive semantic parts and build a mid-level representation for precise classifications, such as utilizing multi-level attention to localize multiple discriminative regions and encode their features in the meanwhile [25], and adopting reinforcement learning paradigm to determine the location and number of discriminative regions [7]. Some methods target at modeling subtle differences between fine-grained subcategories in an end-to-end feature encoding manner, such as integrating structure and appearance information to enhance fine-grained representation [6], extracting part-level information and enhancing multi-granularity feature representation [26]. Apart from that, some methods leverage external information to aid the fine-grained classification, such as web data [27], and multi-modal data [28]–[30]. The aforementioned methods depend heavily on large-scale annotated datasets, while in reality, a substantial amount of data is hard to acquire and costly to label. To mitigate this problem, a more challenging few-shot visual categorization setting is proposed where the model is asked to distinguish fine-grained subcategories with only one or few supporting examples.

B. Few-Shot Visual Categorization

Given the limited availability of support samples, the meta-learning paradigm has gained popularity for few-shot visual classification. In this framework, the training stage comprises multiple N -way K -shot episodes designed to simulate the testing stage. Within the meta-learning paradigm, few-shot visual classification methods are generally categorized into two mainstreams: optimization-based and metric-based.

Optimization-based methods: The idea of optimization-based methods were initially introduced in MAML [31], which aims to acquire an optimal initialization of model parameters that facilitates smooth fine-tuning. In general, optimization-based methods [32]–[36] start with training the model with auxiliary data, followed by fine-tuning the network with additional supporting data sampled from unseen classes. MetaOptNet [37] harnesses the differentiation conditions and dual formulation of convex optimization problems to boost generalization with high-dimensional embeddings. MattML [38] utilizes a multi-attention mechanism for both base and task learners to locate discriminative part-level details. C2-Net [39] integrates outputs from multiple layers with channel activation and position matching operations. However, one drawback of optimization-based methods is their requirement for online training for novel classes.

Metric-based methods: Metric-based methods [40]–[46] embed both support and query images into a vector space, and perform classification by distance or similarity metrics. ProtoNet [11] calculates the average embedding vector of support images as the prototype of each class, and the classification is carried out through distances between a query image and prototypes. RelationNet [47] builds a network to learn the suitable distance metric instead of relying on predefined metrics. BSNet [48] develops a bi-similarity network to merge two similarity metrics for learning fewer but more discriminative regions. DUAL ATT-Net [49] leverages dual attention streams to model relations among object parts and capture discriminative details. FRN [13] constructs a feature map reconstruction network that directly regresses from support features to query features in a closed form. TDM [14] designs a task discrepancy maximization module for leveraging class-wise channel importance for improved classification. Bi-FRN [15] proposes a bi-directional feature reconstruction framework among support and query samples to address both inter-class and intra-class variance. BTG-Net [50] filters noise on mid-level features and retains cross-task general knowledge through prompting mechanisms. ATR-Net [51] adaptively selects task-specific information by interacting with local feature patches for better integration of task-level and instance-level information. The above methods primarily concentrate on enhancing the features of selected samples, but are constrained by the biased distributions inherent in this subset. Therefore, we attempt to rectify the confusion arising from the biased distributions from a causal perspective.

C. Causal Inference

In the realm of statistics and data science, causal inference serves as the foundation of uncovering the cause-and-effect

relationships underlying observed phenomena. At its essence, causal inference targets at elucidating the mechanisms concerning correlated variables and discerning real causalities from spurious correlations. Causal inference made its initial attempt into machine learning through the works of [52], [53] and has been adopted in various fields of computer vision since then, such as long-tailed recognition [54], semantic segmentation [55], and image classification [56]. It is worth noting that the interventional few-shot learning paradigm [57], [58] proposes analyzing the few-shot visual categorization problem with a Structural Causal Model assumption which views the pretrained knowledge as the confounder hindering the classification performance. Despite its effectiveness in the general few-shot classification, IFSL is not compatible with the FS-FGVC problem where the train-from-scratch paradigm eliminates the confounder within the pre-trained knowledge, but instead, the inherent fine-grained nature confuses the subcategory predictions as new confounding variables.

In this paper, we propose a CausalFSFG approach to causally reformulate the FS-FGVC problem with an SCM assumption and address the biased distribution of the limited support samples through the causal intervention.

III. PROBLEM REFORMULATIONS

This section states the definition of the FS-FGVC problem and introduces the proposed causal reformulation including the SCM assumption and the casual intervention through the frontdoor adjustment.

A. Few-Shot Fine-Grained Visual Categorization

In few-shot fine-grained visual categorization scenarios, given a dataset $\mathcal{D} = \{(x_i, y_i), y_i \in C_{total}\}$, we divide it into three subsets: the base training set $\mathcal{D}_{train} = \{(x_i, y_i), y_i \in C_{train}\}$, the validation set $\mathcal{D}_{val} = \{(x_i, y_i), y_i \in C_{val}\}$, and the novel testing set $\mathcal{D}_{test} = \{(x_i, y_i), y_i \in C_{test}\}$, where x_i and y_i denote the i^{th} image and corresponding class label, respectively. The training, validation, and testing classes are disjoint, i.e., $C_{train} \cap C_{val} \cap C_{test} = \emptyset$. Generally, few-shot visual classification aims to boost an N -way K -shot classification performance on the testing set \mathcal{N} , where N classes are randomly chosen from novel testing classes. Each selected class comprises K labeled images and U unlabelled images. We refer to the labelled images as the support set $S = \{(x_j, y_j)\}_{j=1}^{N \times K}$, and the unlabelled images as the query set $Q = \{(x_j, y_j)\}_{j=1}^{N \times U}$.

Our approach adopts the meta-learning paradigm, where the training stage is designed to imitate the N -way K -shot episodes of the testing stage. Concretely, for each episode of the training stage, a meta-training set is randomly sampled from the training set \mathcal{B} . Similarly, each meta-training set is composed of N classes from the base train classes C_{base} , and each class contains K labeled images and U unlabelled images, forming the support set S and the query set Q .

B. Problem Parameterization

To further investigate the FS-FGVC problem, we parameterize the problem with the elements depicted in Fig. 1 (b):

- D : the full fine-grained dataset.
- O : the observable data for the model during training.
- I : the inherent fine-grained nature.
- X : the input samples (cause), which are usually drawn i.i.d. from the observable data O .
- Y : the subcategories (effect) corresponding to X .
- M : the features extracted from models (mediator).

First, we can decompose and parameterize the FS-FGVC problem as two successive feature extraction, $P_\phi(M)$, and class prediction, $P_\theta(Y)$, stages. Specifically, given input samples, the feature maps are extracted conditionally:

$$P_\phi(M) = P(M|X) = P(M|X)P_x(X), \quad (1)$$

where $P_x(X)$ denotes the observed distribution of the input samples. Then, the extracted feature maps are adopted to perform the classification concerning the observed expression of the inherent fine-grained nature:

$$P_\theta(Y) = P(Y|M, I) = P(Y|M, I)P_\phi(M)P_D(I), \quad (2)$$

where $P_D(I)$ denotes the expression of the inherent fine-grained nature conditioned in the full fine-grained dataset. Notice that in the FGVC context, the observable data coincides with the full fine-grained dataset, from which the input samples are drawn i.i.d. :

$$P_x(X) \stackrel{\text{i.i.d.}}{=} P(D). \quad (3)$$

Therefore, by bringing $P_x(X)$ into Eqn. (1) and (2), we can observe that the model has access to the distribution of the full dataset $P(D)$ and thereby the expression of the inherent fine-grained nature $P_D(I)$.

While in the FS-FGVC context, the input samples are drawn i.i.d. from the observable data, which is a selected subset of the full fine-grained dataset:

$$P_x(X) \stackrel{\text{i.i.d.}}{=} P(O) = P(O|D)P(D). \quad (4)$$

It can be observed that the few-shot condition disturbs the observed data distribution with the selection operation $P(O|D)$ which can be decomposed as the class and sample selections:

$$\begin{aligned} P(\mathbf{x}, \mathbf{y}|O) &= \sum_{x \in \mathcal{X}} P_{ss}(\mathcal{X}|\mathcal{Y}) \sum_{y \in \mathcal{Y}} P_{cs}(\mathcal{Y}|D) \\ &\neq P(\mathbf{x}, \mathbf{y}|D), \end{aligned} \quad (5)$$

where x, y denotes the input sample and corresponding class label, P_{cs} and P_{ss} denote the class and sample selection operations, \mathcal{Y} is the set of the selected classes, and \mathcal{X} is the set of selected samples. Since the class and sample selections are both random operations, the biased distribution makes the distribution of the full fine-grained dataset together with the inherent fine-grained nature unobservable during training.

C. Structural Causal Model

From the above discussion, we can see that the FS-FGVC problem can align seamlessly with a Structural Causal Model (SCM), which considers a set of variables associated with the vertices of a directed acyclic graph and depending on their parents in the graph, as illustrated in Fig. 2.

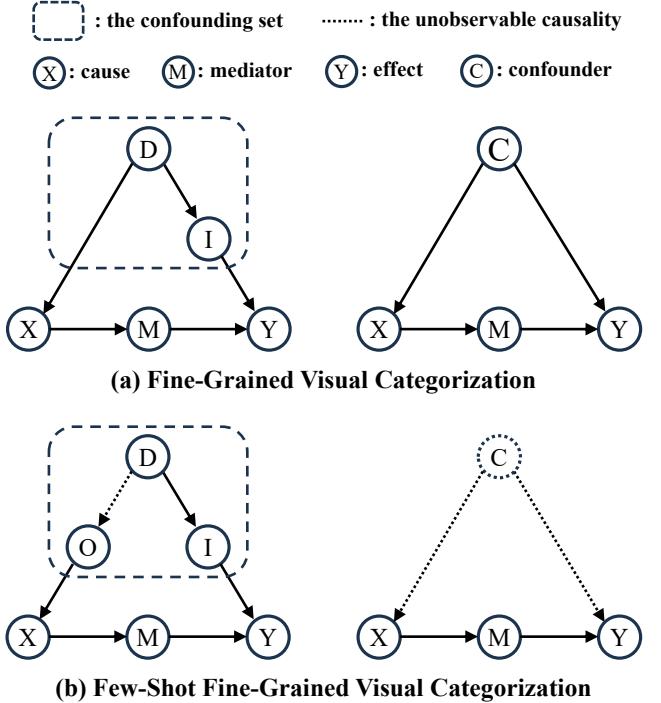


Fig. 2. The Structural Causal Model assumptions for the FS-FGVC problem, dashed lines mean the causality is unobservable during training. (a) In the FGVC context, the confounder is observable since the observable data coincides with the full dataset. (b) In the FS-FGVC context, the biased distribution of the selected observable data makes the confounder unobservable.

- $D \rightarrow I$: this link represents that the inherent fine-grained nature is derived from the distribution of the full dataset.
- $D \rightarrow O$: this connection represents the causalities from the full fine-grained dataset to the observable data during training. Notice that in the FGVC context, the observable data D coincides with the full fine-grained dataset D , as shown in the left part of Fig. 2 (a). While in the FS-FGVC context, the observable data is a selected subset of the full dataset, as shown in the left part of Fig. 2 (b).
- $O \rightarrow X$: this link represents the causalities that the input samples are drawn i.i.d. from the observable data.
- $X \rightarrow M$: this link represents the causalities that the feature maps are extracted depending on the input samples, as formulated in Eqn (1).
- $M \rightarrow Y \leftarrow I$: this assumption models the class prediction stage described in Eqn (2) and can be interpreted as follows. a) $M \rightarrow Y$: the predicted classification results are directly obtained via the extracted features. b) $I \rightarrow Y$: the inherent fine-grained nature is abstracted from the underlying distribution of the full fine-grained dataset and hence affects the classification implicitly.

So far, we have demonstrated the alignment between the SCM assumption and the FS-FGVC problem from the causal perspective. In the following analysis, we refer to the set of the full dataset, observable data, and inherent fine-grained nature as the confounder in the SCM, denoted as C , for simplicity. Then, the FS-FGVC problem can be reformulated as follows:

- The main target is to generate correct classification re-

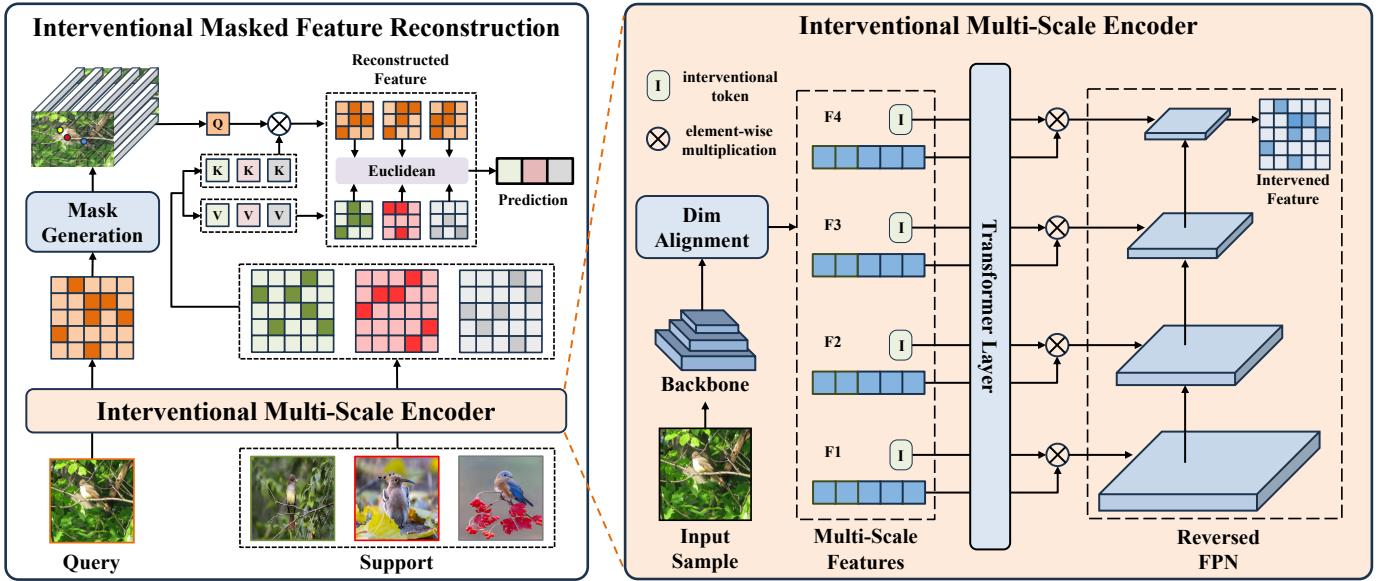


Fig. 3. The overall framework of our proposed CausalFSFG approach. The Interventional Multi-Scale Encoder module implements the sample-level intervention to extract the intervened features. Then, the Interventional Masked Feature Reconstruction module further implements the feature-level intervention to improve the classification performance.

sults (effect) corresponding to the input samples (cause) through the extracted features (mediator), while the few-shot condition and inherent fine-grained nature (confounder) together confuse the classification results with spurious correlations.

- As formulated in Eqn. 5, in the FS-FGVC problem, the biased distribution is unobservable during training (marked with dashed lines), making the summarized confounder unobservable as well.

D. Causal Intervention via Frontdoor Adjustment

An ideal classification pipeline is to capture the direct causality from X to Y , eliminating the confusion from confounder C . However, the direct probabilistic likelihood $P(Y|X)$ in the SCM fails to do so since the likelihood Y given X is not only determined by the observable causal path $X \rightarrow M \rightarrow Y$, but also affected by the spurious correlation path $X \leftarrow C \rightarrow Y$. Therefore, to obtain the true causality between input samples X and classification results Y , we need to compute the causal intervention [59] likelihood $P(Y|do(X))$ instead of the direct likelihood $P(Y|X)$.

Since the confounder C is unobservable under the few-shot condition, which causes the spurious correlations of $X \leftarrow C \rightarrow Y$ cannot be directly computed and eliminated either. To tackle this issue, we adopt the frontdoor adjustment [60] as the causal intervention towards $P(Y|do(X))$. Frontdoor adjustment estimates the causal intervention likelihood $P(T|do(X))$ via intervention on the mediator M :

$$\begin{aligned} P(Y|do(X)) &= \sum_m P(Y|do(X), M)P(M|do(X)) \\ &= \sum_m P(Y|do(M))P(M|do(X)) \\ &= \sum_m P(M|X) \sum_x P(Y|M, X)P(X) \end{aligned} \quad (6)$$

It can be observed that frontdoor adjustment conducts two intervention operations based on the mediator M , i.e., the sample-level intervention $P(M|do(X))$ and the feature-level intervention $P(Y|do(M))$. In the following section, we will present our implementation of the frontdoor adjustment, CausalFSFG, based on the above two intervention operations.

IV. METHODOLOGY

This section introduces the overall pipeline of the proposed CausalFSFG approach and elaborates on each component.

A. Overview

The whole pipeline of the proposed CausalFSFG approach is illustrated in Fig. 3. Inspired by the frontdoor adjustment, our CausalFSFG approach eliminates the spurious correlations caused by the unobserved inherent fine-grained nature (confounder) through the causal intervention. Specifically, an Interventional Multi-Scale Encoder (IMSE) module is proposed as an implementation for $P(M|X)$ in Eqn. (6) and executes the sample-level intervention. Furthermore, an Interventional Masked Feature Reconstruction (IMFR) module is proposed as an implementation for $P(Y|M, X)$ in Eqn. (6) and executes the feature-level intervention.

B. Interventional Multi-Scale Encoder

We propose to integrate multi-scale features conditionally, based on the discovery that multi-scale features extracted from different convolution layers of the backbone contain complementary information [61].

Suppose that given an input sample X , the extracted multi-scale features consist of feature maps of 4 scales (take the Conv-4 and ResNet-12 backbone as examples), i.e., $M = \{M_1, M_2, M_3, M_4\}$. In most previous methods, only the last

feature map M_4 is further processed to derive the classification results. In other words, a one-hot distribution $[0, 0, 0, 1]$ is applied to the extracted multi-scale features, where all feature maps but the last one are overlooked.

To make full use of the multi-scale features and execute sample-level intervention, we implement the target conditional distribution $P(M|X) = P([M_1, M_2, M_3, M_4]|X)$ by introducing interventional tokens which:

- promotes the extraction of discriminative features of each scale as general representations,
- guides the conditional integration of the extracted multi-scale discriminative features.

To begin with, we conduct the dim alignment with 1×1 convolution layers to resize multi-scale features with a unified embedding dim $\Gamma + 1$, for each feature map M_i :

$$[F_i, I_i] = DA(M_i), \quad (7)$$

where $DA(\cdot)$ denotes the dim alignment operation, F_i is the resized feature map taking up the first Γ embedding dim, and I_i is the corresponding interventional token at the last one embedding dim.

Then, the resized multi-scale features and the corresponding interventional tokens are flattened and went through a standard Transformer layer to capture discriminative information of each scale. In the meanwhile, the interventional tokens work as general representations which generalize the discriminative features of each scale into one dimension:

$$[F_i, I_i] = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{\Gamma}}\right) V_i, \quad (8)$$

where Q_i, K_i, V_i are a set of learnable weight parameters projected from the feature map and corresponding interventional token of scale i .

Notice that the interventional tokens can be considered as the general representations of the multi-scale features, we implement the target distribution $P(M|X)$ by applying the softmax function to the interventional tokens:

$$[I'_1, \dots, I'_4] = \text{Softmax}([I_1, \dots, I_4]). \quad (9)$$

So far, we have implemented the target distribution over as $P(M|X) = [I'_1, \dots, I'_4]$, which are employed to guide the integration of the multi-scale features $[F_1, \dots, F_4]$. The element-wise multiplication is conducted to get weighted multi-scale features $F'_i = F_i \otimes I'_i$, which are then integrated by a reversed Feature Pyramid Network (FPN) [62]:

$$F_I = F'_4 + \text{Maxpool}(F'_3 + \dots + \text{Maxpool}(F'_1)), \quad (10)$$

where F_I is the output intervened feature.

C. Interventional Masked Feature Reconstruction

Notice that the following decomposition accords with the few-shot classification paradigm:

$$\begin{aligned} & \sum_x P(Y|M, X)P(X) \\ &= \frac{1}{|S|} \sum_{s \in S} P(Y|M, s) + \frac{1}{|Q|} \sum_{q \in Q} P(Y|M, q) \end{aligned} \quad (11)$$

where S and Q denote the support and query set respectively.

It can be observed that given an input sample and its intervened feature M , the execution of the feature-level intervention requires two implementations:

- $P(Y|M, q)$: This query term indicates the information interaction within the query set, which is implemented with a shared mask generation block for all query samples.
- $P(Y|M, s)$: This support term indicates the information exchange across the query set and the support set, which is implemented with a masked feature reconstruction operation between the query samples and the support samples.

More concretely, for the N -way K -shot classification task, we first apply the IMSE module to extract support features of the n^{th} class, i.e. $S_i = [s_k^n] \in \mathbb{R}^{C \times H \times W}$, where $n \in [1, 2, \dots, N]$ and $k \in [1, 2, \dots, K]$, and query features $q_i \in \mathbb{R}^{C \times H \times W}$, where $i \in [1, 2, \dots, |Q|]$, where $|Q|$ is the number of query samples.

To facilitate the information interaction within the query set, we perform channel-wise max-pooling and average-pooling for each query feature $q_i \in C \times H \times W$, integrating global signals as two $1 \times H \times W$ tensors. Subsequently, these two tensors are concatenated to form a tensor of size $2 \times H \times W$, and a convolution block followed by a sigmoid function is then utilized to produce a normalized global matrix $G_i \in \mathbb{R}^{H \times W}$.

Then we select k elements of G_i with the highest activation values as the indicative mask of general discriminative regions within the query set:

$$\mathcal{G}_i(h, w) = \begin{cases} 1, & G_i(h, w) \in \text{top-}k(G_i) \\ 0, & \text{otherwise} \end{cases}, \quad (12)$$

where $\text{top-}k(G_i)$ denotes the set of top- k highest values among the elements of G_i . Notice that the convolution block is shared by all query samples, enabling \mathcal{G}_i to implicitly indicate the information interaction within the query set. Therefore, a residual connection is conducted:

$$\hat{q}_i = q_i + q_i \times \mathcal{G}_i, \quad (13)$$

which enhances the query features implements the query term as $P(Y|M, q) = P(Y|M + M \times \mathcal{G}(Q))$.

Based on the enhanced query features, we develop a direct implementation of the support term $P(Y|M, s)$ with a feature reconstruction operation. The support features of the same class are first averaged into a prototype of the class:

$$s_n = \frac{1}{K} \sum_{k=1}^K s_k^n, \quad n = [1, \dots, N]. \quad (14)$$

Then for each prototype s_j , the feature reconstruction is conducted as follows:

$$q_i^j = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{\Gamma}}\right) V_j, \quad (15)$$

where Q_i is the query matrix projected from \hat{q}_i and K_j, V_j are the key and value matrices projected from s_j .

D. Learning Objectives

Regarding the query features \mathbf{q}_i , we compute the Euclidean distance between the reconstructed query feature \mathbf{q}_i^j and the projected value matrix of support the feature V_j :

$$d_{ij} = \|\mathbf{q}_i^j - V_j\|_2, \quad (16)$$

and the predicted probability of the query sample belonging to class j is:

$$p_{ij} = \frac{e^{-d_{ij}}}{\sum_{n=1}^N e^{-d_{in}}}. \quad (17)$$

The cross-entropy loss is adopted as the final optimization objective to train the model:

$$\mathcal{L} = CE([p_{i1}, \dots, p_{iN}], \mathbf{y}_i), \quad (18)$$

where y_i is the correct subcategory label.

V. EXPERIMENTS

In this section, we assess the efficacy of the proposed CausalFSFG approach on three widely used fine-grained benchmark datasets. We begin by providing a concise overview of these three datasets and how data pre-processing and separation are conducted. Subsequently, we present our implementation details for all experiments. Following this, we compare our approach with state-of-the-art methods to showcase the effectiveness of CauFSFG. Finally, extensive experimental analyses are presented including ablation studies and parameter analysis to validate the contribution of each component in our approach.

A. Datasets and Evaluation Metric

Three widely used fine-grained datasets, i.e., the CUB-200-2011, Stanford Dogs, and Stanford Cars datasets, are adopted in the experiments. Detailed descriptions of the three datasets are as follows:

- CUB-200-2011 (CUB) [63] consists 11,788 images representing 200 bird species. Following [15], the input images are cropped using human-annotated bounding boxes.
- Stanford Dogs (Dogs) [64] comprises 20,580 images across 120 dog breeds.
- Stanford Cars (Cars) [65] includes 16,185 images depicting 120 different car types.

For each dataset, we follow [15], [38] to split the original images into three disjoint subsets: D_{train} , D_{val} , D_{test} for training, validation, and testing respectively. Details are shown in Table I.

We adopt the widely used classification accuracy as the metric to validate the performance of the proposed CausalFSFG approach and other comparison methods.

B. Implementation Details

For a fair comparison with state-of-the-art methods, we adopt two widely used backbone networks: Conv-4 and ResNet-12. The input images are resized to 84×84 , and the output is a feature map with $64 \times 5 \times 5$ elements for the

TABLE I

CATEGORY SPLIT FOR THREE DATASETS. C_{total} , C_{train} , C_{val} , C_{test} REPRESENTS THE NUMBER OF SUBCATEGORIES IN THE WHOLE DATASET, TRAINING SET, VALIDATION TEST, AND TESTING SET, RESPECTIVELY.

Dataset	CUB	Cars	Dogs
C_{total}	200	196	120
C_{train}	130	130	70
C_{val}	20	17	20
C_{test}	50	49	30

Conv-4 backbone and $640 \times 5 \times 5$ elements for the ResNet-12 backbone. We apply data augmentation, which includes random crops, random horizontal flips, color jitter at the meta-training stage, and center crops at the testing stage, in all implemented experiments.

As for the meta-learning paradigm, we adopt two different settings for more comprehensive comparisons with the state-of-the-art methods:

- Following [15], we adopt 30-way 5-shot episodes for the Conv-4 backbone and 15-way 5-shot episodes for the ResNet-12 backbone.
- Following [14], [66], we adopt 5-way 5-shot episodes for both the Conv-4 and ResNet-12 backbones.

In both scenarios, we conduct tests for 5-way 1-shot and 5-way 5-shot episodes, selecting 15 query images for each class. The unified embedding dimension Γ is configured as 128 for the Conv-4 backbone and 256 for the ResNet-12 backbone. The top- k threshold is designated as $k = 5$ for Conv-4 and $k = 3$ for ResNet-12. The results are presented as mean accuracy (MA) with 95% confidence intervals across 10,000 sampled testing episodes. During meta-training, all models are trained from scratch in an end-to-end fashion. Training spans 800 epochs for both Conv-4 and ResNet-12 models, utilizing the SGD optimizer with Nesterov momentum set to 0.9. The initial learning rate is 0.1 with a weight decay of 3e-4. Throughout training, the learning rate diminishes by a scaling factor of 20 every 400 epochs. Experiments are conducted using PyTorch on a single NVIDIA GeForce RTX 4090 GPU.

C. Experimental Results

We compare the proposed CausalFSFG approach with state-of-the-art methods on the three fine-grained datasets with both meta-learning paradigms, and the results are shown in Table II, III, IV, and V.

- Table II presents the comparison results with the Conv-4 backbone following the meta-learning paradigm in [15]. The proposed CausalFSFG approach achieves superior performance on most experimental settings, achieving **81.94%**, **93.33%**, **67.56%**, **82.83%**, **79.96%**, **93.07%** for the 1-shot and 5-shot settings on the CUB, Dogs, and Cars datasets respectively. Compared with the representative SOTA methods BiFRN [15] which accommodates for inter-class and intra-class variations through a bi-directional feature reconstruction, our CauFSFG approach achieves **2.86%**, **1.11%**, **2.82%**, **1.54%**, **4.22%**, **1.49%**

TABLE II

COMPARISON RESULTS MEAN \pm STD ON THE CUB, DOGS, AND CARS DATASETS WITH THE CONV-4 BACKBONE. **BOLD VALUE** INDICATES THE BEST PERFORMANCE AND UNDERLINE VALUE INDICATES THE SUBOPTIMAL PERFORMANCE.

Method	Published In	CUB		Dogs		Cars	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [11]	NeurIPS 2017	64.82 \pm 0.23	85.74 \pm 0.14	46.66 \pm 0.21	70.77 \pm 0.16	50.88 \pm 0.23	74.89 \pm 0.18
Relation [47]	CVPR 2018	63.94 \pm 0.92	77.87 \pm 0.64	47.35 \pm 0.88	66.20 \pm 0.74	46.04 \pm 0.91	68.52 \pm 0.78
DN4 [67]	CVPR 2019	57.45 \pm 0.89	84.41 \pm 0.58	39.08 \pm 0.76	69.81 \pm 0.69	34.12 \pm 0.68	87.47 \pm 0.47
PARN [68]	ICCV 2019	74.43 \pm 0.95	83.11 \pm 0.67	55.86 \pm 0.97	68.06 \pm 0.72	66.01 \pm 0.94	73.74 \pm 0.70
SAML [69]	ICCV 2019	65.35 \pm 0.65	78.47 \pm 0.41	45.46 \pm 0.36	59.65 \pm 0.51	61.07 \pm 0.47	88.73 \pm 0.49
DeepEMD [22]	CVPR 2020	64.08 \pm 0.50	80.55 \pm 0.71	46.73 \pm 0.49	65.74 \pm 0.63	61.63 \pm 0.27	72.95 \pm 0.38
LRPABN [70]	TMM 2021	63.63 \pm 0.77	76.06 \pm 0.58	45.72 \pm 0.75	60.94 \pm 0.66	60.28 \pm 0.76	73.29 \pm 0.58
BSNet(D&C) [48]	TIP 2021	62.84 \pm 0.95	85.39 \pm 0.56	43.42 \pm 0.86	71.90 \pm 0.68	40.89 \pm 0.77	86.88 \pm 0.50
CTX [71]	NeurIPS 2020	72.61 \pm 0.21	86.23 \pm 0.14	57.86 \pm 0.21	73.59 \pm 0.16	66.35 \pm 0.21	82.25 \pm 0.14
FRN [13]	CVPR 2021	74.90 \pm 0.21	89.39 \pm 0.12	60.41 \pm 0.21	79.26 \pm 0.15	67.48 \pm 0.22	87.97 \pm 0.11
FRN+TDM [14]	CVPR 2022	72.01 \pm 0.22	89.05 \pm 0.12	51.57 \pm 0.23	75.25 \pm 0.16	65.67 \pm 0.22	86.44 \pm 0.12
Bi-FRN [15]	AAAI 2023	79.08 \pm 0.20	92.22 \pm 0.10	64.74 \pm 0.22	81.29 \pm 0.14	75.74 \pm 0.20	91.58 \pm 0.09
C2-Net [39]	AAAI 2024	-	-	66.42 \pm 0.50	81.23 \pm 0.34	81.29 \pm 0.45	91.08 \pm 0.26
Our CausalFSFG	-	81.94 \pm 0.19	93.33 \pm 0.10	67.56 \pm 0.22	82.83 \pm 0.14	79.96 \pm 0.19	93.07 \pm 0.09

TABLE III

COMPARISON RESULTS MEAN \pm STD ON THE CUB, DOGS, AND CARS DATASETS WITH THE RESNET-12 BACKBONE. **BOLD VALUE** INDICATES THE BEST PERFORMANCE AND UNDERLINE VALUE INDICATES THE SUBOPTIMAL PERFORMANCE.

Method	Published In	CUB		Dogs		Cars	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ProtoNet [11]	NeurIPS 2017	81.02 \pm 0.20	91.93 \pm 0.11	73.81 \pm 0.21	87.39 \pm 0.12	85.46 \pm 0.19	95.08 \pm 0.08
CTX [71]	NeurIPS 2020	80.39 \pm 0.20	91.01 \pm 0.11	73.22 \pm 0.22	85.90 \pm 0.13	85.03 \pm 0.19	92.63 \pm 0.11
DeepEMD [22]	CVPR 2020	75.59 \pm 0.30	88.23 \pm 0.18	70.38 \pm 0.30	85.24 \pm 0.18	80.62 \pm 0.26	92.63 \pm 0.13
FRN [13]	CVPR 2021	84.30 \pm 0.18	93.34 \pm 0.10	76.76 \pm 0.21	88.74 \pm 0.12	88.01 \pm 0.17	95.75 \pm 0.07
FRN+TDM [14]	CVPR 2022	85.15 \pm 0.18	93.99 \pm 0.09	78.02 \pm 0.20	89.85 \pm 0.11	88.92 \pm 0.16	96.88 \pm 0.06
Bi-FRN [15]	AAAI 2023	85.44 \pm 0.18	94.73 \pm 0.09	76.89 \pm 0.21	88.27 \pm 0.12	90.44 \pm 0.15	97.49 \pm 0.05
RAsD [44]	TMM 2024	-	-	73.75 \pm 0.93	86.65 \pm 0.54	87.27 \pm 0.70	95.01 \pm 0.49
C2-Net [39]	AAAI 2024	-	-	75.50 \pm 0.49	87.65 \pm 0.28	88.96 \pm 0.37	95.16 \pm 0.20
Our CausalFSFG	-	87.05 \pm 0.17	95.26 \pm 0.08	78.79 \pm 0.20	90.07 \pm 0.11	90.71 \pm 0.14	97.60 \pm 0.05

performance improvements respectively. We attribute the improvements brought by our CausalFSFG approach to the utilizing of the causal intervention to address the biased distribution of the limited support samples, which breaks the restriction of the selected samples and enables the model to learn a more general distribution of the entire fine-grained dataset.

- Table III presents the comparison results with the ResNet-12 backbone following the meta-learning paradigm in [15]. Notice that the ResNet-12 backbone is a more complicated and powerful network than the Conv-4 backbone, which naturally learns a better data distribution and weakens the effectiveness of our proposed approach to some extent, and leads to less significant performance gains compared with the Conv-4 backbone. In spite of

this, the proposed CausalFSFG approach still achieves the best performance of **87.05%**, **95.26%**, **78.79%**, **90.07%**, **90.71%**, **97.60%** for the 1-shot and 5-shot settings on the CUB, Dogs, and Cars datasets respectively.

- Table IV and V present the comparison results on the CUB dataset for both the Conv-4 and the ResNet-12 backbones following the meta-learning paradigm in [14]. Our CausalFSFG approach maintains the best performance of **79.12%**, **92.01%**, **85.01%**, **94.56%** for the 1-shot and 5-shot settings, respectively. Compared with the representative SOTA methods LCCRN [66] which learns local content-enriched features, our CausalFSFG approach also achieves **0.46%**, **2.58%**, **1.65%**, **0.93%** performance gains, which verifies the effectiveness of our CausalFSFG approach under different training paradigms.

TABLE IV

COMPARISON RESULTS ON THE CUB DATASET WITH THE CONV-4 BACKBONE. **BOLD VALUE** INDICATES THE BEST PERFORMANCE AND UNDERLINE VALUE INDICATES THE SUBOPTIMAL PERFORMANCE.

Method	Published In	1-shot	5-shot
ProtoNet [11]	NeurIPS 2017	61.82 ± 0.23	83.37 ± 0.75
FRN [13]	CVPR 2021	73.46 ± 0.21	88.13 ± 0.13
Dual Att-Net [49]	AAAI 2022	72.89 ± 0.50	86.60 ± 0.31
FRN+TDM [14]	CVPR 2022	74.39 ± 0.21	88.89 ± 0.13
LCCRN [66]	TCSVT 2023	76.22 ± 0.21	89.39 ± 0.13
RSaD [44]	TMM 2024	71.15 ± 0.92	84.03 ± 0.62
FicNet [72]	TMM 2024	75.27 ± 0.61	88.48 ± 0.37
C2-Net [39]	AAAI 2024	$\underline{78.66} \pm 0.46$	89.43 ± 0.28
Our CausalFSFG	-	79.12 ± 0.20	92.01 ± 0.11

TABLE V

COMPARISON RESULTS ON THE CUB DATASET WITH THE RESNET-12 BACKBONE. **BOLD VALUE** INDICATES THE BEST PERFORMANCE AND UNDERLINE VALUE INDICATES THE SUBOPTIMAL PERFORMANCE.

Method	Published In	1-shot	5-shot
ProtoNet [11]	NeurIPS 2017	79.64 ± 0.20	91.15 ± 0.11
FRN [13]	CVPR 2021	83.11 ± 0.19	92.49 ± 0.11
FRN+TDM [14]	CVPR 2022	83.36 ± 0.19	92.80 ± 0.10
LAGPF [73]	PR 2023	78.73 ± 0.84	89.77 ± 0.47
BSFA [74]	TCSVT 2023	82.27 ± 0.46	90.76 ± 0.26
LCCRN [66]	TCSVT 2023	82.97 ± 0.19	93.63 ± 0.10
FicNet [72]	TMM 2024	80.97 ± 0.57	93.17 ± 0.32
RSaD [44]	TMM 2024	82.45 ± 0.79	92.02 ± 0.44
Our CausalFSFG	-	85.01 ± 0.18	94.56 ± 0.09

D. Ablation Study

To further demonstrate the effectiveness of the proposed CausalFSFG approach, we evaluate the key components in our CausalFSFG framework on the CUB dataset, adopting the Conv-4 and ResNet-12 backbones respectively. The results of ablation experiments are presented in Table VI. We can observe that: (1) Based on the Conv-4 backbone, IMSE aggregates multi-scale features to achieve the few-shot fine-grained classification accuracy of 77.13% and 88.24% for the 1-shot and 5-shot configurations, which outperforms the baseline method by margins of 12.31% and 2.5% and verifies the effectiveness of conducting sample-level intervention for learning less biased feature distribution. The same trend is observed on the ResNet-12 backbone, where IMSE achieves 4.6% and 1.18% performance improvements. (2) IMF reconstructs the features of query images to achieve the few-shot fine-grained classification accuracy of 73.51% and 88.75% for the 1-shot and 5-shot configurations, which outperforms the baseline method by margins of 8.69% and 3.01% and verifies the effectiveness of conducting feature-level intervention for extracting more discriminative features. The same trend is observed on the ResNet-12 backbone, where our IMF module achieves 3.44% and 2.45% performance improvements. (3) By combining the IMSE and IMF modules, our CausalFSFG approach first learns less biased feature distributions and then extracts more discriminative features, which eliminates spurious correlations caused by the few-shot condition and further achieves performance improvements of 17.12%, 7.59%, 6.03%, 3.33% on both backbones, respectively.

E. Parameter Analysis

To further investigate the effectiveness of our proposed approach, we conduct parameter experiments about the embedding dim Γ , and the top- k threshold on the CUB dataset to explore different designs in our framework.

a) *Embedding Dim*: The experimental results on the embedding dim are presented in Table VII, where the value

TABLE VI
ABLATION STUDY ON THE CUB DATASET OF DIFFERENT COMPONENTS.
BOLD VALUE INDICATES THE BEST PERFORMANCE.

IMSE	IMFR	Conv-4		ResNet-12	
		1-shot	5-shot	1-shot	5-shot
		64.82 ± 0.23	85.74 ± 0.14	81.02 ± 0.20	91.93 ± 0.11
✓		77.13 ± 0.21	88.24 ± 0.13	85.62 ± 0.18	93.11 ± 0.10
	✓	73.51 ± 0.21	88.75 ± 0.12	84.46 ± 0.18	94.38 ± 0.09
✓	✓	81.94 ± 0.19	93.33 ± 0.10	87.05 ± 0.17	95.26 ± 0.08

of embedding dim ranges in $\{62, 128, 256\}$. The best performances are achieved when the embedding dim is set as 128 and 256 for the Conv-4 and ResNet-12 backbones respectively. The different best values can be attributed to the different designs of the backbones. The Conv-4 backbone, with its straightforward architecture, generates feature maps of size $64 \times 5 \times 5$, but struggles with overfitting when the embedding dimension is raised to 256. In contrast, the ResNet-12 backbone, with its more intricate architecture, captures richer semantic information and generates larger feature maps of size $640 \times 5 \times 5$. Consequently, classification accuracy using the ResNet-12 backbone continues to improve as the embedding dimension increases from 64 to 256.

b) *Top-k Thresholds*: In this section, we investigate the effect of top- k thresholds on model performance, as shown in Table VIII. We can observe that slight performance fluctuation happens with different thresholds, which indicates that our proposed approach is relatively insensitive to the value of thresholds. In particular, we set the value of thresholds as 5 with Conv-4 and 3 with ResNet-12 for the best performance.

F. Visualization Experiments

Fig. 4 illustrates the visualization results of the proposed CausalFSFG framework. Specifically, the input images are presented in the first row. The visualization results of the multi-scale features extracted from different layers are present

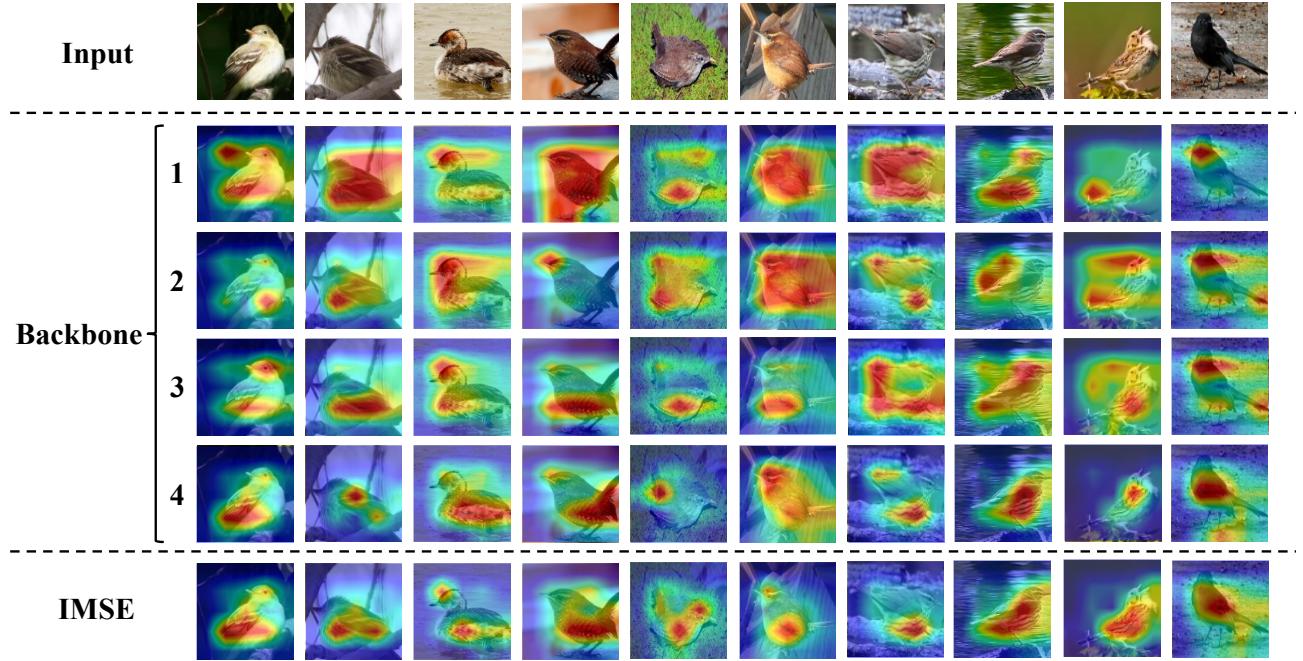


Fig. 4. The visualization examples of the proposed CausalFSFG framework on the CUB dataset with the ResNet-12 backbone. The top block presents the input samples. The middle block presents the visualizations of the multi-scale features extracted by the first to fourth layers of the backbone, respectively. The bottom block presents the visualizations of the intervened feature generated by our IMSE module.

TABLE VII
PARAMETER ANALYSIS ON THE CUB DATASET OF DIFFERENT EMBEDDING DIMS. **BOLD VALUE** INDICATES THE BEST PERFORMANCE.

Γ	Conv-4		ResNet-12	
	1-shot	5-shot	1-shot	5-shot
64	80.88 ± 0.20	92.88 ± 0.10	86.44 ± 0.17	95.04 ± 0.08
128	81.94 ± 0.19	93.33 ± 0.10	86.22 ± 0.17	95.02 ± 0.08
256	81.71 ± 0.19	93.08 ± 0.10	87.05 ± 0.17	95.26 ± 0.08

TABLE VIII
PARAMETER ANALYSIS ON THE CUB DATASET OF DIFFERENT SELECTIVE TOP-K NUMBERS. **BOLD VALUE** INDICATES THE BEST PERFORMANCE.

k	Conv-4		ResNet-12	
	1-shot	5-shot	1-shot	5-shot
3	81.77 ± 0.19	93.23 ± 0.10	87.05 ± 0.17	95.26 ± 0.08
5	81.94 ± 0.19	93.33 ± 0.10	86.83 ± 0.17	95.25 ± 0.08
7	81.53 ± 0.19	93.01 ± 0.10	86.82 ± 0.17	95.19 ± 0.08
10	81.71 ± 0.19	93.23 ± 0.10	86.97 ± 0.17	95.33 ± 0.08

in the second to fifth rows, and the visualization results of our intervened features are shown in the last row. It can be observed that: (1) The multi-scale features contain complementary information concerning each other, which focus on different parts of the input sample. (2) The multi-scale features

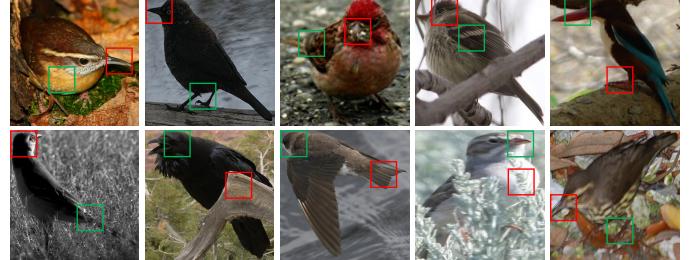


Fig. 5. Visualizations of erroneous results on the CUB test set. Red boxes indicate false image contents identified by the model, and green boxes highlight discriminative image contents for correct classification.

TABLE IX
COMPUTATIONAL COMPLEXITY COMPARISON AGAINST Bi-FRN WITH THE CONV-4 BACKBONE.

Method	Params (K)	Memory (G)	GFLOPs
Bi-FRN [15]	150.53	13.34	60.30
Our CausalFSFG	168.26	14.81	66.07

alone are affected by the biased distribution of the limited samples and fail to precisely locate the salient object and its discriminative parts. (3) Our intervened features alleviate the biased distribution and locate the salient objects and capture corresponding discriminative parts with higher precision.

Failure case analysis. Figure 5 presents the visualization of erroneous results in Figure 5, where red boxes indicate false image contents identified by the model, and green boxes highlight discriminative image contents for correct classifica-

tion. It can be observed that misclassifications often occur when discriminative features are visually compromised due to extreme poses and occlusions, or misled by salient yet spurious background cues when the foreground object is ambiguous.

G. Complexity Analysis

Table IX compares the computational complexity of our CausalFSFG approach against the best comparison method Bi-FRN [15], with the Conv-4 backbone. It can be observed that our CausalFSFG achieves superior FS-FGVC performance with only slightly increased computational burden.

VI. CONCLUSION

In this paper, we propose a novel CausalFSFG approach for few-shot fine-grained visual categorization, which reformulates the FS-FGVC problem from the causal perspective to alleviate the biased data distribution inherent in the few-shot condition through causal intervention. We first align the FS-FGVC problem with a Structural Causal Model (SCM) assumption, where the few-shot condition and the inherent fine-grained nature collectively constitute an unobservable confounder, restricting the classification performance by introducing spurious correlations. To further mitigate this issue, an Interventional Multi-Scale Encoder (IMSE) module and an Interventional Masked Feature Reconstruction (IMFR) module are proposed to conduct the sample and feature level intervention respectively, which eliminates the spurious correlations and reveals real causalities from inputs to subcategories. Extensive experiments and analyses on three public fine-grained datasets validate the superiority and practicality of the proposed CausalFSFG approach.

In our ongoing efforts, we aim to enhance this research by consolidating sample and feature level interventions into a unified network architecture. This integration is expected to offer a more concise and efficient solution for addressing biased distributions and improving overall model efficacy.

REFERENCES

- [1] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, “Fine-grained image analysis with deep learning: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [2] C. Wang, H. Fu, and H. Ma, “Learning mutually exclusive part representations for fine-grained image classification,” *IEEE Transactions on Multimedia*, 2023.
- [3] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, Z. Zhang, and Y.-F. Li, “Transifc: invariant cues-aware feature concentration learning for efficient fine-grained bird image classification,” *IEEE Transactions on Multimedia*, 2023.
- [4] H. Sun, X. He, and Y. Peng, “Hcl: Hierarchical consistency learning for weakly supervised fine-grained recognition,” *IEEE Transactions on Multimedia*, 2023.
- [5] C. Zhang, H. Bai, and Y. Zhao, “Fine-grained image classification by class and image-specific decomposition with multiple views,” *IEEE Transactions on Multimedia*, 2022.
- [6] H. Sun, X. He, and Y. Peng, “Sim-trans: Structure information modeling transformer for fine-grained visual categorization,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5853–5861.
- [7] X. He, Y. Peng, and J. Zhao, “Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization,” *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1235–1255, 2019.
- [8] Y. Peng, X. He, and J. Zhao, “Object-part attention model for fine-grained image classification,” *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2017.
- [9] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.
- [10] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, “Matching networks for one shot learning,” *Advances in neural information processing systems*, vol. 29, 2016.
- [11] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [12] X. He and Y. Peng, “Only learn one sample: Fine-grained visual categorization with one sample training,” in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1372–1380.
- [13] D. Wertheimer, L. Tang, and B. Hariharan, “Few-shot classification with feature map reconstruction networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8012–8021.
- [14] S. Lee, W. Moon, and J.-P. Heo, “Task discrepancy maximization for fine-grained few-shot classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5331–5340.
- [15] J. Wu, D. Chang, A. Sain, X. Li, Z. Ma, J. Cao, J. Guo, and Y.-Z. Song, “Bi-directional feature reconstruction network for fine-grained few-shot image classification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2821–2829.
- [16] S. Huang, M. Zhang, Y. Kang, and D. Wang, “Attributes-guided and pure-visual attention alignment for few-shot recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 9, 2021, pp. 7840–7847.
- [17] Q. Luo, L. Wang, J. Lv, S. Xiang, and C. Pan, “Few-shot learning via feature hallucination with variational inference,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3963–3972.
- [18] C. Wang, S. Song, Q. Yang, X. Li, and G. Huang, “Fine-grained few shot learning with foreground object transformation,” *Neurocomputing*, vol. 466, pp. 16–26, 2021.
- [19] N. Sun and P. Yang, “T2l: Trans-transfer learning for few-shot fine-grained visual categorization with extended adaptation,” *Knowledge-Based Systems*, vol. 264, p. 110329, 2023.
- [20] J. K. Tam, M. Gustineli, and A. Miyaguchi, “Transfer learning and mixup for fine-grained few-shot fungi classification,” *arXiv preprint arXiv:2507.08248*, 2025.
- [21] C. Li, S. Li, H. Wang, F. Gu, and A. D. Ball, “Attention-based deep meta-transfer learning for few-shot fine-grained fault diagnosis,” *Knowledge-Based Systems*, vol. 264, p. 110345, 2023.
- [22] C. Zhang, Y. Cai, G. Lin, and C. Shen, “Deepemd: Few-shot image classification with differentiable earth mover’s distance and structured classifiers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 203–12 213.
- [23] H. Wu, Y. Zhao, and J. Li, “Selective, structural, subtle: Trilinear spatial-awareness for few-shot fine-grained visual recognition,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE Computer Society, 2021, pp. 1–6.
- [24] Y. Liu, L. Zhu, X. Wang, M. Yamada, and Y. Yang, “Bilaterally normalized scale-consistent sinkhorn distance for few-shot image classification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11475–11485, 2023.
- [25] X. He, Y. Peng, and J. Zhao, “Fast fine-grained image classification via weakly supervised discriminative localization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1394–1407, 2018.
- [26] J. Wang, Q. Xu, B. Jiang, B. Luo, and J. Tang, “Multi-granularity part sampling attention for fine-grained visual classification,” *IEEE Transactions on Image Processing*, 2024.
- [27] C. Zhang, Y. Yao, H. Liu, G.-S. Xie, X. Shu, T. Zhou, Z. Zhang, F. Shen, and Z. Tang, “Web-supervised network with softly update-drop training for fine-grained visual classification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 781–12 788.
- [28] K. Song, X.-S. Wei, X. Shu, R.-J. Song, and J. Lu, “Bi-modal progressive mask attention for fine-grained recognition,” *IEEE Transactions on Image Processing*, vol. 29, pp. 7006–7018, 2020.
- [29] X. Jiang, H. Tang, J. Gao, X. Du, S. He, and Z. Li, “Delving into multi-modal prompting for fine-grained visual classification,” in *Proceedings*

- of the AAAI conference on artificial intelligence*, vol. 38, no. 3, 2024, pp. 2570–2578.
- [30] S. Cheng, F. Zhang, H. Zhou, and C. Xu, “Multi-modal knowledge-enhanced fine-grained image classification,” in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2024, pp. 333–346.
- [31] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
- [32] A. Antoniou, H. Edwards, and A. Storkey, “How to train your maml,” *arXiv preprint arXiv:1810.09502*, 2018.
- [33] M. A. Jamal and G.-J. Qi, “Task agnostic meta-learning for few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11719–11727.
- [34] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, “Dense classification and implanting for few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9258–9267.
- [35] J. Sun, T. Chen, T. Luo, H. Gu, Y. Tian, X. Li, J. Zheng, and J. Fu, “Multi-scale cross-modal collaborative reconstruction network,” in *2025 5th International Conference on Computer Vision, Application and Algorithm (CVAA)*. IEEE, 2025, pp. 171–179.
- [36] S. Yang, X. Li, D. Chang, Z. Ma, and J.-H. Xue, “Channel-spatial support-query cross-attention for fine-grained few-shot image classification,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9175–9183.
- [37] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, “Meta-learning with differentiable convex optimization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10657–10665.
- [38] Y. Zhu, C. Liu, and S. Jiang, “Multi-attention meta learning for few-shot fine-grained image recognition,” in *IJCAI*, 2020, pp. 1090–1096.
- [39] Z.-X. Ma, Z.-D. Chen, L.-J. Zhao, Z.-C. Zhang, X. Luo, and X.-S. Xu, “Cross-layer and cross-sample feature optimization network for few-shot fine-grained image classification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4136–4144.
- [40] S. W. Yoon, J. Seo, and J. Moon, “Tapnet: Neural network augmented with task-adaptive projection for few-shot learning,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 7115–7123.
- [41] L. Yang, L. Li, Z. Zhang, X. Zhou, E. Zhou, and Y. Liu, “Dpgn: Distribution propagation graph network for few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13390–13399.
- [42] Y. Guo and N.-M. Cheung, “Attentive weights generation for few shot learning via information maximization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13499–13508.
- [43] J. Wang, J. Lu, J. Yang, M. Wang, and W. Zhang, “An unbiased feature estimation network for few-shot fine-grained image classification,” *Sensors*, vol. 24, no. 23, p. 7737, 2024.
- [44] H. Liu, C. P. Chen, X. Gong, and T. Zhang, “Robust saliency-aware distillation for few-shot fine-grained visual recognition,” *IEEE Transactions on Multimedia*, 2024.
- [45] L.-J. Zhao, Z.-D. Chen, Z.-X. Ma, X. Luo, and X.-S. Xu, “Angular isotonic loss guided multi-layer integration for few-shot fine-grained image classification,” *IEEE Transactions on Image Processing*, vol. 33, pp. 3778–3792, 2024.
- [46] J. Wu, D. Chang, A. Sain, X. Li, Z. Ma, J. Cao, J. Guo, and Y.-Z. Song, “Bi-directional ensemble feature reconstruction network for few-shot fine-grained classification,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 9, pp. 6082–6096, 2024.
- [47] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.
- [48] X. Li, J. Wu, Z. Sun, Z. Ma, J. Cao, and J.-H. Xue, “Bsnet: Bi-similarity network for few-shot fine-grained image classification,” *IEEE Transactions on Image Processing*, vol. 30, pp. 1318–1331, 2020.
- [49] S.-L. Xu, F. Zhang, X.-S. Wei, and J. Wang, “Dual attention networks for few-shot fine-grained recognition,” 2022.
- [50] Z.-X. Ma, Z.-D. Chen, L.-J. Zhao, Z.-C. Zhang, T. Zheng, X. Luo, and X.-S. Xu, “Bi-directional task-guided network for few-shot fine-grained image classification,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8277–8286.
- [51] L. Yu, Z. Guan, W. Zhao, Y. Yang, and J. Tan, “Adaptive task-aware refining network for few-shot fine-grained image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [52] S. Magliacane, T. Van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, “Domain adaptation by using causal inference to predict invariant conditional distributions,” *Advances in neural information processing systems*, vol. 31, 2018.
- [53] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, “A meta-transfer objective for learning to disentangle causal mechanisms,” *arXiv preprint arXiv:1901.10912*, 2019.
- [54] K. Tang, J. Huang, and H. Zhang, “Long-tailed classification by keeping the good and removing the bad momentum causal effect,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1513–1524, 2020.
- [55] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, “Causal intervention for weakly-supervised semantic segmentation,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 655–666, 2020.
- [56] K. Chalupka, P. Perona, and F. Eberhardt, “Visual causal feature learning,” *arXiv preprint arXiv:1412.2309*, 2014.
- [57] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, “Interventional few-shot learning,” *Advances in neural information processing systems*, vol. 33, pp. 2734–2746, 2020.
- [58] S. Wang, J. Lu, H. Ben, Y. Hao, X. Gao, and M. Wang, “Interventional feature generation for few-shot learning,” *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.
- [59] J. Pearl *et al.*, “Models, reasoning and inference,” Cambridge, UK: Cambridge University Press, vol. 19, no. 2, p. 3, 2000.
- [60] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- [61] R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y.-Z. Song, and J. Guo, “Fine-grained visual classification via progressive multi-granularity training of jigsaw patches,” in *European Conference on Computer Vision*. Springer, 2020, pp. 153–168.
- [62] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [63] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.
- [64] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, “Novel dataset for fine-grained image categorization: Stanford dogs,” in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, no. 1. Citeseer, 2011.
- [65] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations for fine-grained categorization,” in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.
- [66] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, and J.-H. Xue, “Locally-enriched cross-reconstruction for few-shot fine-grained image classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [67] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, “Revisiting local descriptor based image-to-class measure for few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7260–7268.
- [68] Z. Wu, Y. Li, L. Guo, and K. Jia, “Parn: Position-aware relation networks for few-shot learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6659–6667.
- [69] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, “Collect and select: Semantic alignment metric learning for few-shot learning,” in *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 2019, pp. 8460–8469.
- [70] H. Huang, J. Zhang, J. Zhang, J. Xu, and Q. Wu, “Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1666–1680, 2020.
- [71] C. Doersch, A. Gupta, and A. Zisserman, “Crosstransformers: spatially-aware few-shot transfer,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 21981–21993, 2020.
- [72] H. Zhu, Z. Gao, J. Wang, Y. Zhou, and C. Li, “Few-shot fine-grained image classification via multi-frequency neighborhood and double-cross modulation,” *IEEE Transactions on Multimedia*, 2024.
- [73] H. Tang, C. Yuan, Z. Li, and J. Tang, “Learning attention-guided pyramidal features for few-shot fine-grained recognition,” *Pattern Recognition*, p. 108792, 2022.
- [74] Z. Zha, H. Tang, Y. Sun, and J. Tang, “Boosting few-shot fine-grained recognition with background suppression and foreground alignment,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.