mathematics

MDPI

Article

Learning High-Order Features for Fine-Grained Visual Categorization with Causal Inference

Yuhang Zhang [1], Yuan Wan [1,*], Jiahui Hao [1], Zaili Yang [2] and Huanhuan Li [2,*]

[1] School of Mathematics and Statistics, Wuhan University of Technology, 122 Luoshi Road, Wuhan 430070, China; zhangyuhang@whut.edu.cn (Y.Z.); haojiahui@whut.edu.cn (J.H.)
[2] Liverpool Logistics, Offshore and Marine Research Institute, Liverpool John Moores University, Liverpool L3 3AF, UK; z.yang@ljmu.ac.uk
* Correspondence: wanyuan@whut.edu.cn (Y.W.); h.li2@ljmu.ac.uk (H.L.)

Abstract: Recently, causal models have gained significant attention in natural language processing (NLP) and computer vision (CV) due to their capability of capturing features with causal relationships. This study addresses Fine-Grained Visual Categorization (FGVC) by incorporating high-order feature fusions to improve the representation of feature interactions while mitigating the influence of confounding factors through causal inference. A novel high-order feature learning framework with causal inference is developed to enhance FGVC. A causal graph tailored to FGVC is constructed, and the causal assumptions of baseline models are analyzed to identify confounding factors. A reconstructed causal structure establishes meaningful interactions between individual images and image pairs. Causal interventions are applied by severing specific causal links, effectively reducing confounding effects and enhancing model robustness. The framework combines high-order feature fusion with interventional fine-grained learning by performing causal interventions on both classifiers and categories. The experimental results demonstrate that the proposed method achieves accuracies of 90.7% on CUB-200, 92.0% on FGVC-Aircraft, and 94.8% on Stanford Cars, highlighting its effectiveness and robustness across these widely used fine-grained recognition datasets. Comprehensive evaluations of these three widely used fine-grained recognition datasets demonstrate the proposed framework's effectiveness and robustness.

Keywords: causal models; causal inference; fine-grained visual categorization; feature fusion; causal intervention

MSC: 68T07

Academic Editors: Carlos Soubervielle-Montalvo and Cesar Puente

Received: 28 February 2025
Revised: 14 April 2025
Accepted: 17 April 2025
Published: 19 April 2025

Citation: Zhang, Y.; Wan, Y.; Hao, J.; Yang, Z.; Li, H. Learning High-Order Features for Fine-Grained Visual Categorization with Causal Inference. *Mathematics* **2025**, *13*, 1340. https://doi.org/10.3390/math13081340

1. Introduction

Fine-grained Visual Categorization (FGVC) presents significant challenges due to subtle inter-class differences and large intra-class variations [1]. Across different categories, objects often share similarities in shape, color, and texture [2], making classification difficult. At the same time, variations within a single category arise due to factors such as pose, viewpoint, and background differences, further complicating the recognition process. For example, in bird species classification, distinguishing between two visually similar species may rely on minute differences in their feather patterns or beak structure. Additionally, variations in viewing angles can result in vastly different visual representations of the same bird. This discrepancy occurs because an image is a two-dimensional projection of

a three-dimensional object, where changes in perspective naturally yield distinct patterns and textures.

Due to the high cost of manual labeling [3], weakly supervised learning has become a widely adopted approach for FGVC. Deep learning-based methods [4] have significantly advanced this field, and among them, high-order feature pooling techniques have been gained. These methods apply higher-order transformations to convolutional neural network (CNN) features before classification, with notable examples including bilinear models [5–14] and kernel pooling [15,16]. High-order feature fusion enhances contextual semantic information by capturing complex feature interactions, leading to notable accuracy improvements across different FGVC tasks [17,18].

A key challenge in high-order feature fusion is the presence of confounding factors in images. When features are extracted and merged without considering these confounders, the model may incorporate misleading information, leading to incorrect classifications. Objects in FGVC tasks often appear in distinct environments that may influence classification decisions. As illustrated in Figure 1, Great Crested Flycatchers are commonly seen in forested areas (green backgrounds), while Olive-Sided Flycatchers are typically observed against a sky backdrop (blue). Consequently, a model may misclassify a bird solely based on its background rather than its intrinsic attributes.

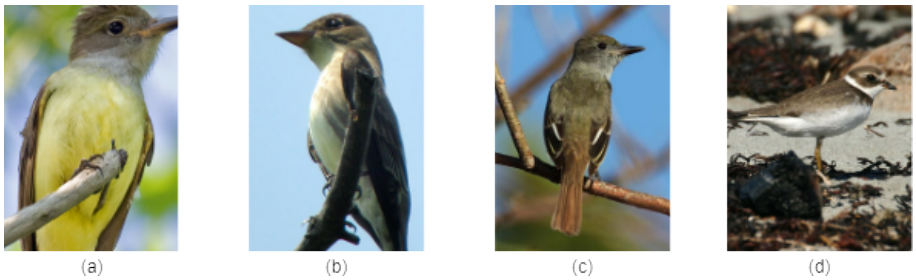


Figure 1. Visual comparison of misclassified samples: Great Crested Flycatcher vs. Olive-Sided Flycatcher. (**a**,**c**) show Great Crested Flycatchers; (**b**,**d**) show Olive-Sided Flycatchers. These two species are visually similar, which contributes to classification challenges. In (**c**), the absence of distinctive abdominal feathers causes background features to dominate, leading to misclassification. In (**d**), the bird's posture causes its feather patterns to blend with the background, further complicating accurate identification.

Given the limitations of conventional background noise removal techniques, distinguishing meaningful background cues from misleading ones is crucial. Unlike standard denoising methods that indiscriminately suppress background information, a more effective approach should be able to selectively preserve useful features while mitigating the impact of confounding factors.

To address this, causal intervention is introduced to improve fine-grained recognition by disentangling relevant and irrelevant information. In recent years, causal models have been successfully applied in computer vision [19–23], particularly through causal representation learning [24,25]. Unlike traditional machine learning models that assume independent and identically distributed (i.i.d.) data, causal representation learning leverages stable causal mechanisms across datasets, making it robust to challenges such as limited samples, imbalanced data, and biased observations.

Inspired by the work of Rao et al. [26] on counterfactual attention learning for fine-grained image classification, this study extends causal reasoning into a comprehensive high-order feature learning framework. To better utilize high-order features for fine-grained categorization, Interventional High-Order Feature Fusion (IHFF) is proposed, integrating causal reasoning with high-order feature learning. The approach is built upon structural

causal models (SCMs) [24,25], aiming to reconstruct latent causal relationships that capture meaningful semantic interactions.

Previous methods relied on high-order feature fusion to extract fine-grained details. However, when encountering highly similar objects across different classes, conventional feature fusion is often insufficient for accurate classification. To address this limitation, high-order features are decomposed into two components:

(1) Intra-object relationships—capturing structural dependencies within the same object.

(2) Inter-object relationships—modeling semantic differences among distinct objects within the same category.

By fusing these two levels of semantic relationships, a more discriminative and context-aware representation of fine-grained details is obtained. Furthermore, analysis of the structural causal model for FGVC enables the identification of causal links through which confounders affect classification performance. Causal interventions are then applied to sever these links, reducing confounding effects and increasing the proportion of effective features in high-order representations. Causal intervention enables the selective extraction of high-order features while preserving essential patterns without overemphasizing misleading information. The contributions of this paper are summarized below:

(1) Constructing a general SCM for FGVC and introducing high-order feature fusion via tensor product spaces to describe the relationships between different semantic information spaces.

(2) Analyzing the reconstructed SCM to identify causal links through which confounders influence classification outcomes and applying causal interventions to mitigate their impact.

(3) Proposing IHFF, which simultaneously applies high-order feature fusion and causal intervention.

(4) Conducting comprehensive experiments on three widely used fine-grained public datasets (CUB-200-2011, FGVC-Aircraft, and Stanford Cars) to demonstrate the effectiveness of the proposed method.

The subsequent sections of this paper are structured as follows: Section 2 reviews the relevant literature. Section 3 introduces the structural causal model and its application to FGVC. Section 4 details a proposed methodology. Section 5 presents the experimental setup, results, and analysis. Finally, Section 6 concludes the study.

2. Related Work

This section presents a review of relevant research, focusing on two key areas: high-order feature pooling techniques in fine-grained image analysis and the application of causal interventions in computer vision.

2.1. Fine-Grained High-Order Feature Fusion

In fine-grained image analysis, high-order feature pooling plays a pivotal role in enhancing the discriminative power of features, crucial for distinguishing subtle variations in texture, shape, and color [2] among closely resembling categories. This method integrates complex interactions of features, crucial for identifying minor differences that conventional pooling methods often overlook. Research has demonstrated that high-order pooling techniques, such as bilinear pooling and cross-layer pooling, are particularly effective in capturing complex patterns and fine-grained structural details.

For instance, in fine-grained recognition tasks, the Fisher vector encoding of SIFT features has the ability to outperform outputs from fully connected layers [27,28], prompting further exploration into advanced feature fusion techniques. A notable development in this area is the bilinear CNN model, which computes an image representation as the outer prod-

uct of features from two separate deep convolutional neural networks [5]. This approach captures the second-order statistics of the features, significantly enhancing fine-grained recognition accuracy. When two identical networks are utilized, this method essentially forms a covariance matrix of the features. However, it inherently leads to high-dimensional feature spaces and an increased parameter count.

To mitigate these challenges, Gao et al. [6] introduced tensor sketching techniques that approximate these second-order statistics, effectively reducing the dimensionality of the features. Similarly, Kong et al. [7] implemented a low-rank approximation of the covariance matrix and developed a low-rank bilinear classifier that could avoid the direct computation of the bilinear feature matrix, thus significantly cutting down on the number of parameters. Further innovations by Li et al. [29] involved modeling pairwise feature interactions through quadratic transformations under low-rank constraints. Yu et al. [11] addressed the issue of dimensionality explosion by applying dimensionality reduction projections prior to bilinear pooling and introduced cross-layer bilinear pooling to harness inter-layer feature interactions across different layers of a convolutional neural network. Additionally, Zheng et al. [13] focused on simplifying bilinear transformations by calculating pairwise interactions within each channel group.

Expanding beyond bilinear methods, other methods attempt to capture higher-order interactions of features. Cui et al. [15] presented a kernel pooling approach that could capture arbitrary order and nonlinear features through compact feature mappings. Moreover, Cai et al. [9] proposed a polynomial kernel-based predictor for modeling higher-order feature interactions across multiple layers, which facilitated capturing detailed part interactions, further advancing the field of fine-grained image analysis.

2.2. Causal Inference

The integration of causal reasoning into computer vision represents a significant shift toward addressing confounding biases and enhancing the robustness of models. By employing causal interventions, researchers can isolate the effects of specific variables on visual phenomena, enabling more precise interpretations and predictions. This approach utilizes methods such as counterfactual reasoning and structural causal models, which are instrumental in deciphering the underlying mechanisms of visual representations and their influences on AI decision-making processes. Such methods are particularly effective in scenarios where traditional correlation-based techniques falter, like in conditions of occlusion, variable illumination, and dynamic environments.

Historically, causal inference [25] has roots in diverse fields like psychology, political science, and epidemiology [30–34] and has recently made significant inroads into deep learning [35,36], especially noted in natural language processing [37–39]. Inspired by these successes, numerous studies have embarked on integrating causal representation learning into computer vision, achieving notable advancements across various applications. These include image classification [19,20], few-shot learning [21], long-tail recognition [22], and semantic segmentation [23].

Specifically, causal interventions are employed to remove confounding factors, thus refining model performance [40]. For instance, Tang et al. [41] utilized causal graphs to represent specific tasks and calculated true causal effects based on these graphs, effectively eliminating biases. Qi et al. [42] differentiated visual dialog (VisDial) from Visual Question Answering (VQA) by incorporating historical information and tackled the challenges using a causal inference framework, enhancing all VisDial baseline models to state-of-the-art performance. Niu et al. [43] addressed the issue of language bias in visual and language integrated tasks by isolating the direct influence of language from the combined effects, thereby reducing biases significantly.

Moreover, Zhang et al. [23] applied causal graphs to analyze element relationships in weakly supervised semantic segmentation, pinpointing context priors in datasets as confounding factors. Their strategy of implementing causal interventions severed the correlation between context priors and image data, improving result accuracy. Yue et al. [21] discussed the dual role of pre-training as both a source of rich prior knowledge and a potential confounder. By applying causal interventions, they effectively eliminated these confounding factors to boost performance in few-shot learning scenarios.

Additionally, some approaches have embraced counterfactual learning [26], integrating it with attention mechanisms [44,45] in fine-grained image classification to help networks focus more intensely on primary classification targets. This blending of causal representation with advanced learning techniques underscores the growing importance and utility of causal methods in computer vision, paving the way for more accurate and unbiased models.

3. Structural Causal Model for FGVC

This section outlines the construction of a causal graph for fine-grained image classification, illustrated in Figure 2 [24,25]. It begins by analyzing causal assumptions and identifying confounding factors within a baseline model. Subsequently, the section outlines methods to implement causal interventions by severing causal links effectively. This approach clarifies the mechanisms underpinning the baseline model and pinpoints strategic interventions to mitigate biases and enhance accuracy.

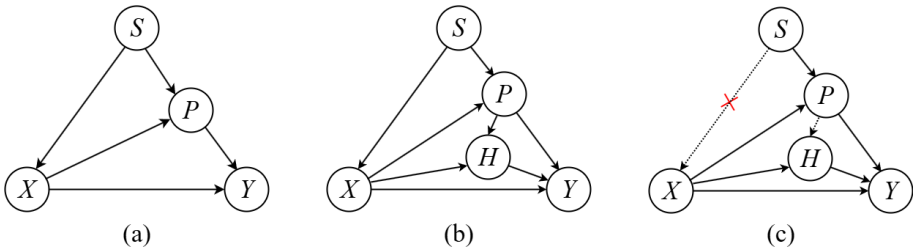


Figure 2. *S*: input image; *X*: feature map; *P*: basis of feature space; *Y*: classification; *H*: high-order feature space. (**a**) Causal graph for FGVC. (**b**) A new collider *H* is added to describe feature interactions. (**c**) After analysis, *P* can be fully represented by *H*, so the only causal link influencing the final classification *Y* is $S \rightarrow X \rightarrow H \rightarrow Y$. Therefore, it is sufficient to sever $S \rightarrow X$ to block the entire causal link, thereby preventing any confounding factors from affecting the classification result.

3.1. General Structural Causal Model for FGVC

***S*→*X*:** In this causal link, *X* represents the feature representation, while *S* refers to the input image, which inherently contains both natural semantic information and confounding factors. For instance, consider a dataset *S* and its corresponding feature extraction network, Ω. This causal link implies that the feature map *X* is derived from the input image S through the transformation applied by the network Ω.

***S*→*P*←*X*:** *P* represents a transformed version of *X*, with its foundation stemming from *S*. This link consists of two key components:

S*→*P: The space *P* is spanned by the basis of the feature space and serves as the projection of the input image *S* onto this feature space. This projection is typically realized through linear transformations in neural networks. Consequently, *P* encapsulates not only essential semantic information but also features influenced by confounding factors.

X*→*P: Feature map *X* undergoes linear or nonlinear transformations, resulting in the formation of the feature space *P*, which subsequently feeds into the fully connected layers of the model.

This paper differentiates the foundation of the feature space P into two distinct levels: the classifier level and the class label level. To mitigate the influence of confounding factors, causal interventions are applied through backdoor adjustments at each of these levels separately.

***X*→*Y*←*P*:** Let Y represent the classification outcome (e.g., logits), which is influenced by the feature map X through two distinct pathways: (1) a direct projection from X to Y, and (2) an intermediary projection via P. Typically, the direct path can be disregarded if X is fully encapsulated by P, especially when taking into account the dimensionality reduction in features. For example, in a conventional neural network architecture composed of convolutional and pooling layers, the feature map X can be completely described in a linear manner using the basis of the feature space. Consequently, the structural causal model simplifies to $S \rightarrow P \rightarrow Y$. However, in FGVC, this simplification toward causal links overlooks the role of contextual factors, which are critical in high-level feature fusion, as these high-level features cannot be linearly represented by their bases. Regarding the pathway through the intermediary P, this mechanism naturally arises because the variables forming the basis of the final classification function are derived from the basis of P, suggesting that the classification function can always be expressed as an implicit function of the feature space P.

3.2. Reconstructed Structural Causal Model for FGVC

In FGVC, capturing higher-order semantic information is crucial. Advanced feature extractors that can handle high-order features have shown superior effectiveness in achieving precise classification. Their success is largely due to their ability to discern subtle differences between highly similar sub-categories. Additionally, these extractors are pivotal in identifying invariant features across different poses, scales, and rotations. Traditional causal interventions typically focus on an SCM with low-dimensional features, which are inadequate for FGVC. Therefore, we propose reconstructing the causal link to amplify the impact of higher-order semantic information through the following model:

***X*→*H*←*P*:** H represents the pairwise features derived from two distinct sets of features extracted by networks, such as outputs from different convolutional neural networks (CNNs). This link helps reconstruct the relationships between high-order features (pairwise features) and their representation in lower dimensions. For example, let P denote a linear combination of $k + m$ base vectors, along with a residual noise component. This approach enhances our understanding and manipulation of complex feature interactions within FGVC:

$$\begin{aligned} x =& c_1 x_1 + \cdots + c_k x_k + c_{k+1} x_{k+1} + \cdots + c_{k+m} x_{k+m} + e, \\ classifier f(x) =& f(c_1 x_1 + \cdots + c_k x_k + c_{k+1} x_{k+1} + \cdots + c_{k+m} x_{k+m} + e), \end{aligned} \tag{1}$$

where e is the residual noise, $\{x_1, x_2, \ldots, x_k\}$ and $\{x_{k+1}, x_{k+2}, \ldots, x_{k+m}\}$ are from two different feature extractors. The relationship between them can be described by the tensor product:

Definition 1. *Tensor product of multilinear functions.*

Given a k-linear function $f \in L(V_1, \ldots, V_k; \mathbb{R})$, the set of all multilinear functions L, and an m-linear function $g \in L(W_1, \ldots, W_m; \mathbb{R})$, define the tensor product of both as a (k+m)-linear function $f \otimes g \in L(V_1, \ldots, V_k, W_1, \ldots, W_m; \mathbb{R})$; it satisfies the following:

$$(f \otimes g)(v_1 \ldots, v_k, w_1, \ldots, w_m) := f(v_1 \ldots, v_k) g(w_1, \ldots, w_m). \forall v_i \in V_i, w_i \in W_i. \tag{2}$$

It is evident that the basis of H is derived from the basis of P, enabling the causal relationship of H to inherit the causal structure of P. In other words, three causal links can be identified: $X \to H$, $P \to H$, and $H \to Y$. Let the last convolutional layers of two feature extractors be represented by two vector spaces, V and W. Thus, P is the direct sum of V and W: $P = V \oplus W$, while H is the tensor product space of V and W (as discussed in Section 4.1): $H = V \otimes W = span\{vs. \otimes w | v \in V, w \in W\}$. There are fundamental differences between H and P, as $\dim(V \oplus W) = \dim V + \dim W$ and $\dim(V \otimes W) = \dim V * \dim W$. Therefore, due to the nature of H as an increased dimensionality representation of the feature space P (as proven in Appendix A), P can be fully represented by H. Consequently, similar to the previous analysis, the connection $P \to H$ can be omitted. Thus, only the causal link $X \to H \to Y$ needs to be considered. This simplification allows for focusing on the backdoor adjustment of the $S \to X$ path in causal interventions.

4. Method

The structural causal model for FGVC in Section 3 consists of a single causal link from S to Y: $S \to X \to H \to Y$. To mitigate the influence of confounding factors on the classification outcome, the backdoor adjustments [46] are applied.

This section constructs mathematical models for causal interventions in fine-grained visual classification FGVC, consisting of two parts. As shown in Figure 3, first, the high-order feature space H is defined. Second, backdoor adjustments are applied by severing the causal link where confounding factors influence Y, specifically the path $S \to X$.

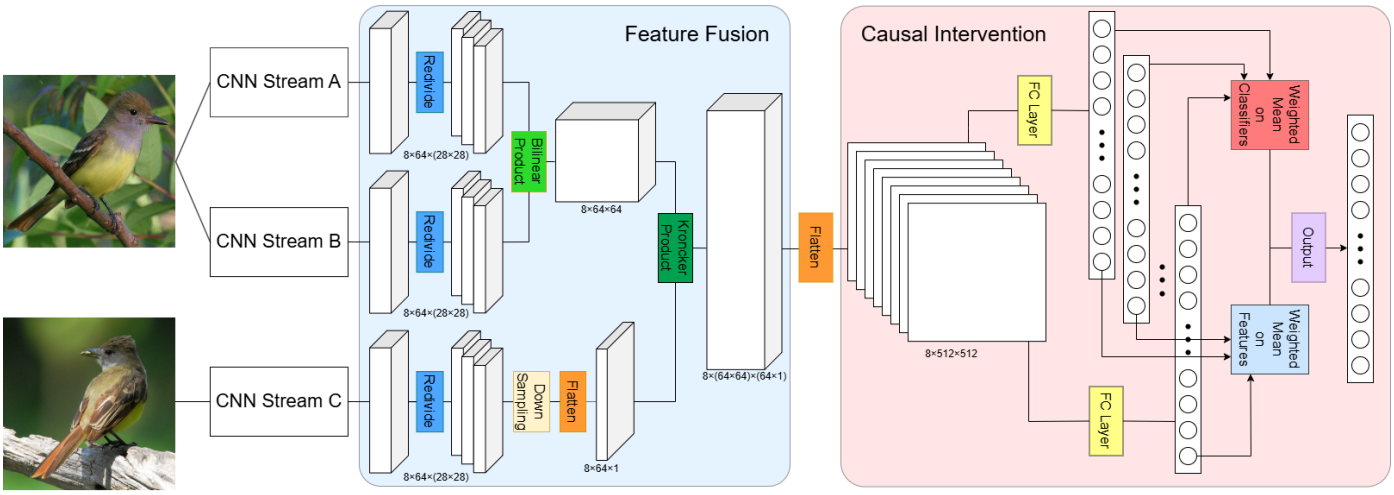


Figure 3. The architecture of the proposed network. The network first computes the bilinear product of the final convolutional layer outputs from two neural networks processing the same image. Then, the convolutional output of an image with the same label is downsampled and flattened. A tensor product between these two feature spaces is performed. For causal intervention, the tensor product space is divided into blocks, which are sequentially processed by fully connected layers. Finally, backdoor adjustments are applied by multiplying the tensor product output with a corresponding weight matrix.

4.1. High-Order Feature Space

Given the last convolutional layers V of a network, V is manually divided into k parts: $\{V_1, V_2, \ldots, V_k\}$. Similarly, the last convolutional layer W of another network is partitioned into m parts: $\{W_1, W_2, \ldots, W_m\}$. Each part is treated as a group representing a subset of semantic information. Furthermore, each part is subdivided into smaller vector spaces. For instance, $V_1 = \{v_1, v_2, \ldots, v_i\}$ and $W_1 = \{w_1, w_2, \ldots, w_j\}$, with each v_i corresponding positionally to w_j. It is essential that k and m are equal. The classification effects are then represented by two multilinear functions:

$$\begin{aligned} &f \in L(V_1, \ldots, V_k; \mathbb{R}), \\ &g \in L(W_1, \ldots, W_k; \mathbb{R}), \\ &f \otimes g \in L(V_1, \ldots, V_k, W_1, \ldots, W_k; \mathbb{R}), \end{aligned} \tag{3}$$

where $\mathbb{R}$ is the classification labels.

In fact, $v \in V$, $w \in W$ can be treated as linear functions on V^* and W^*, respectively. Thus, Equation (4) is obtained:

$$vs. \otimes w \in L(V^*, W^*; \mathbb{R}). \tag{4}$$

Assume that V has a basis $a_\mu = \{a_1, a_2, \ldots, a_k\}$ and W has a basis $b_\nu = \{b_1, b_2, \ldots, b_k\}$. Thus, the tensor product space $V \otimes W$ has a basis $\{a_\mu \otimes b_\nu\} = \{a_i \otimes b_j : 1 \leq i, j \leq k\}$. Moreover, based on the tensor product of functions defined in Equation (2), each element $a_i \otimes b_j$ belongs to the space $L(V^*, W^*; \mathbb{R})$, thus $V \otimes W$ is a subspace of $L(V^*, W^*; \mathbb{R})$. $\{a_\mu \otimes b_\nu\}$ is the basis of $L(V^*, W^*; \mathbb{R})$. It follows that $V \otimes W = L(V^*, W^*; \mathbb{R})$.

The derivation above demonstrates that the constructed bilinear mapping satisfies a universal property. This implies that, in an isomorphic sense, the bilinear mapping from V and W to L is unique. Hence, the tenser product $V \otimes W$ is rigorously proved unique as well.

4.2. Feature Extraction with Contextual Semantic Information

To extract high-order features with contextual semantic information, it is essential to pairwisely train one single image while considering the inter-relationships within it. Suppose V is the last convolutional layer of image X_1: the inter-relationship feature can be extracted by $V \otimes V$. Typically, high-order features are extracted using two different networks on a single image. However, this does not render the tensor product on a feature space itself meaningless. Actually, experiments on a bilinear CNN (B-CNN) have indicated that utilizing the same two neural networks to extract features can also enhance the accuracy of classification. The discussion in Appendix A.2 offers a more rational mathematical explanation for these experimental outcomes: Taking the tensor product of a feature space with itself essentially results in dimensionality expansion.

Next, consider another image X_2 that shares the same label as X_1, along with its corresponding final convolutional layer representation W. Pairwise relationships between X_1 and X_2 are captured using the tensor product $V \otimes W$. With both inter-relationships ($V \otimes V$) and outer-relationships ($V \otimes W$) established, feature fusion is then applied to integrate them. In tensor product operations, the sequence in which inter- or outer-relational features are extracted does not affect the final outcome of feature fusion. This property is consistent with real-world scenarios. A formal proof of this property is presented below.

The element f within $V \otimes W$ can be unfolded as follows:

$$f = f^{\mu\nu} a_\mu \otimes b_\nu, \tag{5}$$

Similarly, function g within $(V \otimes W) \otimes Z$ can be unfolded as follows:

$$g^{\mu\nu\sigma} \left(f^{\mu\nu} a_\mu \otimes b_\nu\right) \otimes c_\sigma = g^{\mu\nu\sigma} f^{\mu\nu} \left(a_\mu \otimes b_\nu \otimes c_\sigma\right), \tag{6}$$

where c_σ is a basis of Z, $\left\{a_\mu \otimes b_\nu \otimes c_\sigma\right\}$ is a basis of $(V \otimes W) \otimes Z$. Similarly, $\left\{a_\mu \otimes b_\nu \otimes c_\sigma\right\}$ is also a basis of $V \otimes (W \otimes Z)$. Thus, Equation (7) is obtained:

$$V \otimes W \otimes Z = (V \otimes W) \otimes Z = V \otimes (W \otimes Z). \tag{7}$$

Then, the high-order feature space H is revisited, comprising the inter-relationship, which is composed of the inter-relationship $V \otimes V$ and outer-relationship $V \otimes W$:

$$H = V \otimes V \otimes W = (V \otimes V) \otimes W = V \otimes (V \otimes W). \tag{8}$$

From Equation (8), it is clear that the calculation order does not alter the result. Whether we calculate $V \otimes V$ first or calculate $V \otimes W$ first, they both ultimately equal Equation (8). This equation confirms the uniqueness of the constructed higher-order feature space H from another perspective.

4.3. Causal Intervention with Rebuilt Causal Link

Let Y be the classification effect, X be the input feature, and z be the semantic information set containing confounding factors. Then the probability output formula in the general network is as follows:

$$P(Y|X) = \sum_{z} P(Y|X,z)P(z|X) = \frac{P(Y,X)}{P(X)}. \tag{9}$$

Causal intervention is essentially an adaptive weighted probability involving the traversal of all objects in the Z set and the calculation of the conditional probability after intervention. Normally, in an SCM with only one collider (In an SCM, the junction $S \rightarrow P \leftarrow X$ is called a "collider", making S and X independent even though S and X are linked via P), the intervention is as follows:

$$\begin{aligned} P(Y|do(X)) &= \sum_{Z} P(Y|X,z)P(z) \\ &= \sum_{z} \frac{P(Y,X,z)P(z)}{P(X,z)}. \end{aligned} \tag{10}$$

Furthermore, by taking into account the internal relationship of a single image, Equation (10) is transformed into the following:

$$\begin{aligned} P(Y|do(X)) &= \sum_{Z} P(Y|X,z_1,z_2)P(z_1,z_2) \\ &= \sum_{z} \frac{P(Y,X,z_1,z_2)P(z_1,z_2)}{P(X,z_1,z_2)}, \end{aligned} \tag{11}$$

where z_1 and z_2 are two events that occur within the same scenario Z.

Assume that z_1 and z_2 are causally influential features co-determining label Y. For example, let Y = 008. Rhinoceros_Auklet, z_1 represents double white stripes on the eyes and z_2 represents a red bird beak. Then, under Equation (11), $P(Y,X,z_1,z_2) = P(X,z_1,z_2)$, hence we have

$$P(Y|do(X)) = \sum_{Z} P(z_1,z_2) \tag{12}$$

Obviously, this probability approaches 1 in theoretical computation, indicating that applying backdoor adjustment after high-order feature fusion is effective. Equation (10) can correctly adjust its probability.

Similarly to the previous derivation, to perform backdoor adjustment with pairwise features from two images, the features of the two images can be averaged, as they cannot appear simultaneously:

$$\begin{aligned} P(Y|do(X)) &= \sum_{Z} P(Y|X_1, X_2, z_1, z_2)P(z_1, z_2) \\ &= \frac{1}{2}\sum_{z}\left(\frac{P(Y, X_1, z_1, z_2)P(z_1, z_2)}{P(X_1, z_1, z_2)}\right. \\ &\quad \left. + \frac{P(Y, X_2, z_1, z_2)P(z_1, z_2)}{P(X_2, z_1, z_2)}\right), \end{aligned} \tag{13}$$

where X_1 and X_2 are different images with the same category labels.

4.4. Interventional High-Order Feature Learning

Suppose that V is the last convolutional layer of image X_1 through network Ω_1. V is divided into k equal-sized, disjoint subsets in order. Similarly, let W denote the last convolutional layer of image X_2 through network Ω_2. Thus, Equation (14) is obtained:

$$\begin{aligned} V &= \{V_1, V_2, V_3, V_4\}, \\ W &= \{W_1, W_2, W_3, W_4\}, \end{aligned} \tag{14}$$

Since the output layer is divided in order, the semantic information within each individual sub-feature space should be considered positionally similar. Therefore, the focus is on extracting the relationships between them.

For each subspace, it is further divided into p parts in order. If $p = 8$, Equation (15) is obtained:

$$\begin{aligned} V &= \left\{V_i | V_i = \left\{v_{ij} | 1 \le j \le 8\right\}, 1 \le i \le 4\right\}, \\ W &= \left\{W_i | W_i = \left\{w_{ij} | 1 \le j \le 8\right\}, 1 \le i \le 4\right\}. \end{aligned} \tag{15}$$

The tensor product is then applied to each pair of (v_{ij}, v_{ij}):

$$\bar{v}_{ij} = v_{ij} \otimes v_{ij}, \tag{16}$$

where $1 \le i \le 4, 1 \le j \le 8$.

Thus, the tensor product space $\bar{V} = V \otimes V$ is obtained. Next, the same operation is applied on $\bar{V}$ and W to obtain the following:

$$H = \bar{V} \otimes W = V \otimes V \otimes W. \tag{17}$$

As there is no prior knowledge while training, it is difficult to determine the number of features that have a causal effect toward classification. In other words, the number of $Z = z_1, z_2$ is unknown and infinite to some extent. Thus, it is prohibitive to achieve the above backdoor adjustment through Equation (13). However, the probability can be approximated using the inverse probability weighting formulation in Equation (18):

$$P(Y = i|do(X)) \approx \frac{1}{K}\sum_{k=1}^{K} \tilde{P}(Y = i, X = x|Z = z_1, z_2). \tag{18}$$

Thus, a multi-head strategy is naturally applied [47]. For every $\bar{v}_{ij} = v_{ij} \otimes v_{ij}$, where $1 \le i \le 4, 1 \le j \le 8$, it can be considered a fine-grained sampling. Hence, the logit calculation with the classifiers' backdoor adjustments for $P(Y = i|do(X))$ can be formulated as follows:

$$P(Y = i|do(X)) = \frac{1}{K}\sum_{K=1}^{K} P\left(Y|\left(w_i^k\right)^T x^k\right), \tag{19}$$

where w_i^k is the weight.

Next, feature backdoor adjustments are applied. Class adjustments are quantized into weights, which are multiplied by $P(Y = i|do(X))$, and subsequently normalized:

$$P(Y = i|do(X)) = \frac{1}{n} \frac{P(Y|x)}{\sum_{n=1}^{n} P(Y|x)P(Y|P(y_i|x)\bar{x}_i)}. \tag{20}$$

where $P(y_i|x)$ represents the probability that x belongs to the i-th label and $\bar{x}_i$ is the mean feature of the i-th class.

To make the causal intervention more fine-grained, two backdoor adjustments are applied simultaneously. By combining and organizing Equations (19) and (20), the following is obtained:

$$P(Y = i|do(X)) = \frac{1}{K} \sum_{K=1}^{K} P\left(Y|\left(w_i^k\right)^T x^k \oplus \frac{1}{n} \sum_{n=1}^{n} \left(w_i^k\right)^T P\left(y_i|x^k\right)\bar{x}_i^k\right), \tag{21}$$

where $\oplus$ denotes vector concatenation. This combination is straightforward: vector concatenation treats classifier backdoor adjustments as equally important as feature backdoor adjustments.

5. Experiments

This section evaluates the experimental results from three key perspectives: (1) a comparative analysis of traditional accuracy metrics, (2) ablation studies to assess the effectiveness of the proposed method, and (3) an investigation of the impact of different hyperparameter values through comparative experiments.

5.1. Datasets

The effectiveness of the proposed interventional multilinear learning method is assessed on three widely used datasets for Fine-grained Visual Categorization, including Caltech-UCSD Birds (CUB-200-2011) [48], FGVC-Aircraft [49], and Stanford Cars [50]. The datasets' details are shown in Table 1: (1) Caltech-UCSD Birds-200-2011 (CUB) is an extension of CUB-200, which includes 200 classes, and each class has around 60 samples. (2) The FGVC-Aircraft dataset contains 10,200 aircraft images, with each of the 100 different aircraft model variants having 102 images. (3) The Stanford Cars dataset consists of 196 classes of cars with a total of 16,185 images. It is important to note that only category labels are used in experiments.

Table 1. Descriptions of the three datasets used in the experiments.

Dataset	Category	Training Set	Testing Set
CUB-200-2011	200	5994	5794
Stanford Cars	196	8144	8041
FGVC-Aircraft	100	6667	3333

5.2. Implementation Details

Overall framework. The 16-layer Visual Geometry Group (VGG-16) and the 18-layer ResNet (ResNet-18) [51] were pre-trained and used as backbones. When VGG-16 was employed as the backbone, the input image was first resized to 448 × 448 pixels, which is the required input size for VGG-16. For fine-tuning the fully connected layers, the 1000-way classification layer pre-trained on the ImageNet dataset was replaced with a k-way softmax layer, where k corresponds to the number of classes in the fine-grained dataset. The final pooling layer was then replaced with a high-order feature pooling layer.

It is important to note that the network's previous parameters were frozen to allow for training of only the last layer. For high-order feature pooling, bilinear feature fusion was used as the inter-relationship feature extractor on a single image. To capture outer-relationship features, this bilinear feature was fused with the output layer of another image sharing the same label but differing from the first. The parameters of the softmax layer were randomly initialized. The fully connected layer was trained with a higher learning rate while monitoring the accuracy of the validation set. After training the fully connected layer, its parameters were incorporated into the overall network training, with previously frozen parameters unfrozen.

The classification layer during initial training can be interpreted as a prior probability. Due to the difficulty and labor-intensive nature of obtaining part-level annotations in fine-grained datasets, backdoor adjustment could not be performed as part-level prior probabilities were unavailable. Therefore, using an adaptive network to acquire part-level prior probabilities proved both suitable and efficient.

Feature fusion details. Taking VGG-16 as an example, the output feature size is 512×28. Through bilinear fusion with itself, the inter-relationship is represented as a 512×512 feature map. For causal intervention, the feature space is partitioned into k parts (with k set to 8 in VGG-16), resulting in eight smaller subspaces, each of size 64×64. Similarly, the output layer of another image is segmented into eight parts, each of size 64.

For outer-relationship feature fusion, each 64-dimensional vector is first expanded into a 64×1 matrix. The Kronecker product is then computed between each 64×64 matrix and the corresponding 64×1 matrix, yielding a high-order feature subspace of size $64 \times 64 \times 64$, which can be reshaped into a 512×512 structure. By introducing causal intervention into high-order feature fusion, the final feature space is reduced from $512 \times 512 \times 512$ to $8 \times 512 \times 512$, while preserving the rank structure of the Kronecker product.

The operation of reshaping the fused feature into an $8 \times 512 \times 512$ structure is intentional. Here, the "8" represents the use of eight parallel classifiers, each corresponding to a distinct layer of backdoor adjustment within the causal intervention framework. These classifiers are designed to process the feature maps at different levels, allowing the model to simulate multiple intervention scenarios and refine the feature representations accordingly.

This design enables the model to learn not just from observed data, but also from intervention-based reasoning, which strengthens its ability to generalise and identify causally relevant patterns. While the causal reasoning module may appear abstract, it plays a vital role in guiding the model toward more discriminative and reliable decision-making in fine-grained classification tasks.

Configuration details. During the preprocessing of the training set, data augmentation is performed using RandomHorizontalFlip, followed by random cropping to a size of 448 and normalization. To ensure consistency with real-world image classification tasks, RandomHorizontalFlip is not applied to the validation or test sets. Initially, only the classifiers are trained using logistic regression, with a batch size of 16, a weight decay of 1×10^{-8}, and a learning rate of 1. Subsequently, the entire network is fine-tuned using stochastic gradient descent, with a batch size of 64, a weight decay of 1×10^{-6}, and a learning rate of 1×10^{-2}.

5.3. Results and Analysis

5.3.1. Comparative Analysis and Efficacy of IHFF in FGVC

Table 2 and Figure 4 present the comparative experiments between the proposed method and several classical methods, all of which were fine-tuned. The key findings are as follows:

(1) Accuracy improvements: The results demonstrate that the new method, IHFF, leads to significant accuracy enhancements across a variety of datasets and backbone networks, particularly showing notable improvements over the baseline model (B-CNN with VGG-16 as the backbone). This indicates that IHFF is effective across different datasets and backbone architectures.

(2) Effectiveness of feature fusion and causal reasoning: The data in Table 2 clearly show that methods incorporating feature fusion are generally more effective. For instance, IHFF achieves an accuracy of 90.7% on the CUB dataset using a ResNet backbone, compared to DBTNet's accuracy at 88.1%. This underscores the utility of causal reasoning in enhancing Fine-grained Visual Categorization, with IHFF showing an average improvement of 3.40% over B-CNN. Furthermore, among the newer models, IHFF—except for DCAL—has achieved state-of-the-art performance. Models such as GBP, SFSCF-Net, and I2-HOFI are all enhancements based on high-order feature fusion, which underscores the effectiveness and significance of the causal intervention strategy employed in IHFF.

(3) Comparison with B-CNN: The B-CNN initially introduced high-order feature fusion into Fine-grained Visual Categorization. Our proposed IHFF method outperforms the B-CNN, demonstrating an average improvement of 3.43% across three datasets and an even higher accuracy gain of 6.07% with the ResNet backbone. These findings validate the application of causal reasoning in this domain and illustrate that integrating causal interventions into high-order feature fusion boosts performance rather than causing antagonism.

(4) Comparison with CAL: CAL pioneers the use of counterfactual causal reasoning in Fine-grained Visual Categorization but shows a lower accuracy on the CUB dataset compared to IHFF. The variance may be linked to the differing depths of the backbone networks utilized. Nonetheless, IHFF's superior accuracy supports the effectiveness of our causal intervention approach and its underlying mathematical principles, showcasing the potential of causal reasoning in fine-grained image analysis.

(5) Comparison with DCAL: DCAL, currently the top-performing network in FGVC, achieves a higher overall accuracy than IHFF. Both methods employ high-order feature fusion, but DCAL may have an edge due to its integration of self-attention mechanisms, which likely improves its capability of capturing contextual information.

(6) Dataset-Specific Performance: IHFF exhibits notably higher performance improvement on the CUB dataset compared to the Aircraft and Cars datasets. This may be attributed to the CUB dataset having a larger number of categories and more training images per category. The data suggest that causal intervention learning particularly enhances network performance on datasets with smaller samples by mitigating confounding factors through backdoor adjustments, thereby focusing on truly impactful features. Conversely, the lesser improvement on the Aircraft dataset may be due to the classification task relying less on feature interactions to extract contextual semantic information, as identifying aircraft types might often depend more on recognizing distinct physical features, such as the number of windows.

Table 2. Top-1 classification accuracy.

Method	Backbone	CUB	Cars	Aircraft	Feature Fusion
ResNet-50 [51]	ResNet-50	84.5	-	-	×
B-CNN [5]	VGGD+VGGM	84.1	91.3	83.9	✓
DBTNet [13]	ResNet-101	88.1	94.5	91.6	×
Improved B-CNN [52]	VGGD+VGGM	85.8	92.0	88.5	✓
LRBP [7]	VGG-16	84.2	90.9	87.3	✓

Table 2. *Cont.*

Method	Backbone	CUB	Cars	Aircraft	Feature Fusion
HBP [11]	VGG-16	87.1	93.7	90.3	✓
GBP [53]	GCNN	87.8	93.5	89.6	✓
SFSCF-Net [54]	ResNet-50	89.6	94.5	-	✓
I2-HOFI [54]	ResNet-50	90.1	94.3	92.3	✓
CAL [26]	ResNet-101	90.6	95.5	94.2	×
DCAL [55]	R50-ViT-Base	92.0	95.3	93.3	✓
IHFF (ours)	VGG-16	87.4	93.9	88.2	✓
	ResNet-50	90.7	94.8	92.0	✓

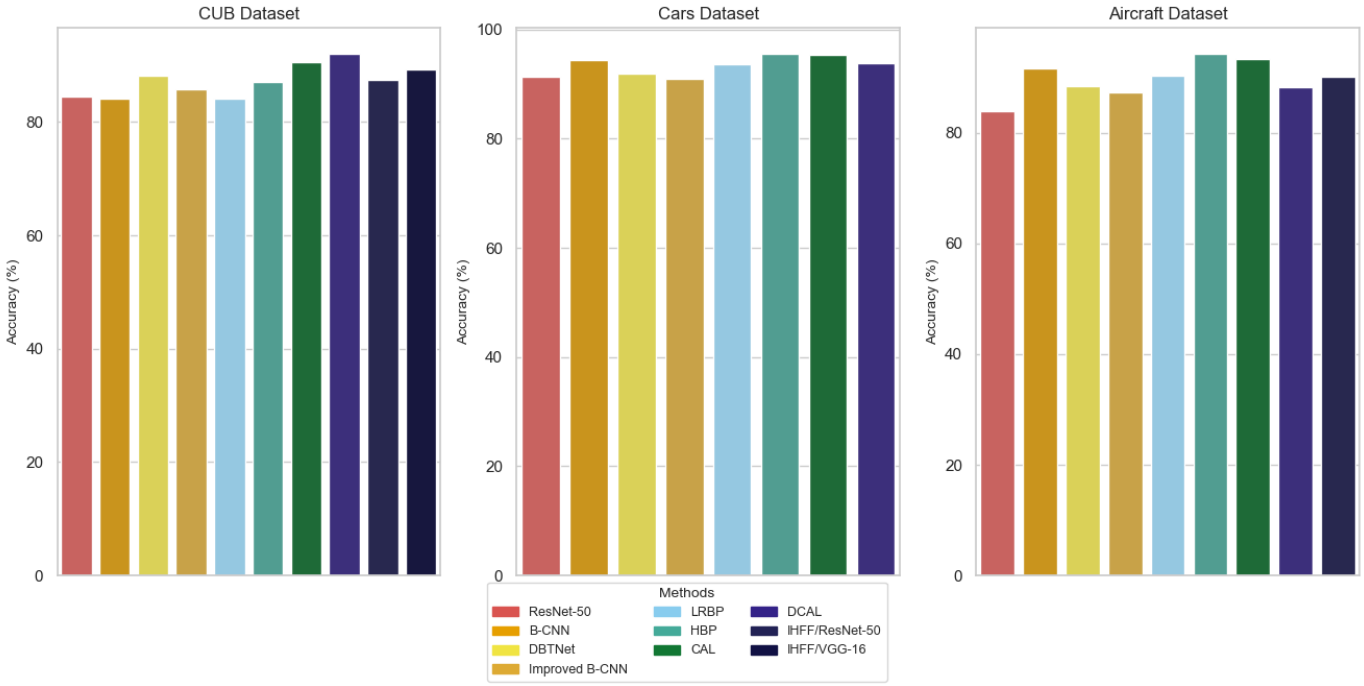


Figure 4. Bar chart of Top-1 classification accuracy.

5.3.2. Ablation Studies

The ablation experiment freezes the parameters of the backbone network and only trains the fully connected layer with a learning rate of 1, without fine-tuning. It is important to note that the baseline model removes the final pooling and fully connected layers from the backbone networks, replacing them with bilinear pooling layers [5]:

$$\begin{aligned} x &= \text{vec}(\sum_i V_i^T W_i), \\ f(x) &= \frac{\text{sign}(x)\sqrt{|x|}}{\left\|\text{sign}(x)\sqrt{|x|}\right\|_2}, \end{aligned} \tag{22}$$

where V_i and W_i represent the output layer parts of different networks that correspond to the same position.

Table 3 and Figure 5 reveal that prior research has highlighted the value of feature fusion in Fine-grained Visual Categorization, and our findings substantiate this further:

(1) Enhanced bilinear pooling: Utilizing higher-order feature fusion via the Kronecker product has significantly enhanced two-dimensional bilinear pooling. This approach resulted in accuracy increases of 4.96% for VGG-16 and 4.81% for ResNet-18.

(2) Backdoor adjustment efficacy: Applying backdoor adjustments to the class has boosted performance by an average of 6.28%, whereas adjustments to the classifier have shown a slightly higher average improvement of 6.44%. However, applying backdoor adjustments across all methods only yielded a modest average increase of 1.22%, indicating that the overall impact of backdoor adjustments may be limited.

(3) Impact of post-adjustment: Implementing post-adjustments without incorporating high-order feature fusion led to a performance boost of about 1%, which is lower than when using high-order feature fusion alone. Nevertheless, the results were less effective compared to scenarios where adjustments were applied after high-order feature fusion. This finding underscores the potential of causal interventions in managing high-dimensional data by eliminating confounding factors and retaining more impactful features.

(4) Overall methodology impact: The collective application of all methods led to an improvement of approximately 7.58% over the baseline model and about 1.20% over using high-order feature fusion alone. Although the proposed method markedly enhances accuracy compared to the baseline, the combined benefit of all methods is not additive, likely due to the diminishing returns associated with increased complexity.

Table 4 demonstrates the effectiveness of the proposed IHFF module after fine-tuning the entire network with a batch size of 64, a weight decay of 1×10^{-6}, and a learning rate of 1×10^{-2}. Even after end-to-end training, the high-order feature fusion and causal intervention module continued to yield an improvement in classification accuracy. However, the overall accuracy gain was lower compared to the scenario where only the proposed module was trained while keeping the backbone frozen. This may be attributed to the increased complexity and number of trainable parameters during full fine-tuning, which potentially introduces model instability and partial overfitting to the training data.

Table 3. Ablation experiment results with only high-order feature fusion and causal intervention modules trained.

Backbone	Method	Accuracy	Comparison	High-Order Features
VGG-16	Baseline	74.86	-	×
	With feature fusion	79.81	**+4.95**	✓
	With class BDA	81.05	**+6.19**	✓
	With classifier BDA	81.24	**+6.38**	✓
	With both BDA	80.17	**+5.31**	×
	With all	82.22	**+7.36**	✓
ResNet-18	Baseline	77.92	-	×
	With feature fusion	82.73	**+4.81**	✓
	With class BDA	84.29	**+6.37**	✓
	With classifier BDA	84.42	**+6.50**	✓
	With both BDA	84.06	**+6.14**	×
	With all	85.71	**+7.79**	✓

Figure 6 shows the results of an ablation study on pairwise learning, utilizing a VGG-16 backbone network. It should be noted that this ablation experiment froze the parameters of the backbone network and only trained high-order feature fusion and causal intervention modules. The figure clearly demonstrates that feature fusion strategies that incorporate outer relationships generally outperform those limited to inter-relationships. With few exceptions, the accuracy gained by integrating both inter- and outer-relationships consistently surpasses that achieved through inter-relationship alone. Since traditional feature fusion methods primarily leverage inter-relationship information derived from features within the same image, these findings are significant. They point to a promising new direction for enhancing feature fusion techniques in Fine-grained Visual Categorization.

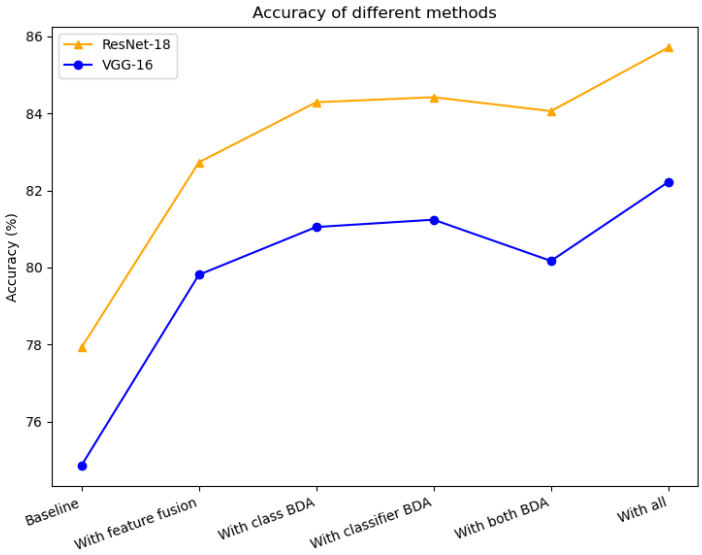


Figure 5. Results for ablation studies with different backbones on CUB-200-2011. The blue line refers to backbone VGG-16, the orange line refers to ResNet-18.

Table 4. Ablation experiment results with fine tuned network.

Backbone	Method	Accuracy	Comparison	High-Order Features
VGG-16	Baseline	84.12	-	×
	With feature fusion	85.39	**+1.17**	✓
	With class BDA	86.13	**+2.01**	✓
	With classifier BDA	86.38	**+2.26**	✓
	With both BDA	86.00	**+1.82**	×
	With all	87.42	**+3.30**	✓

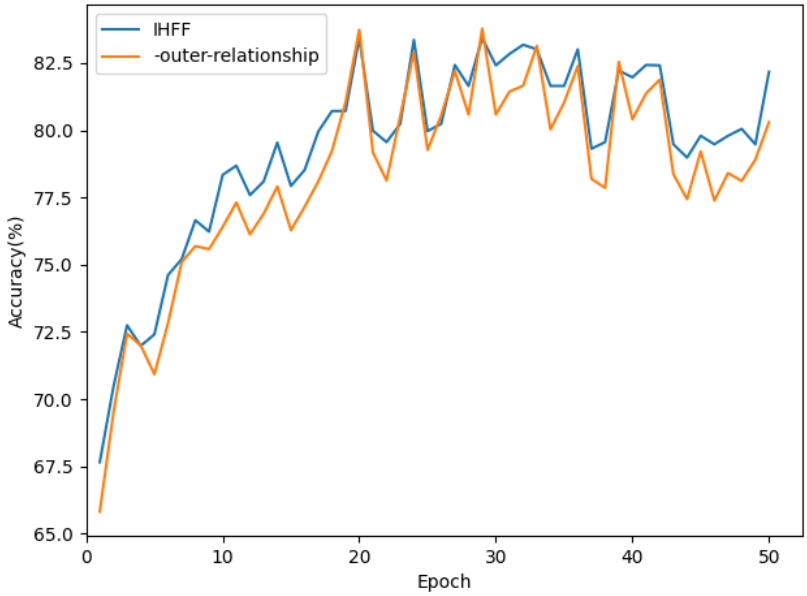
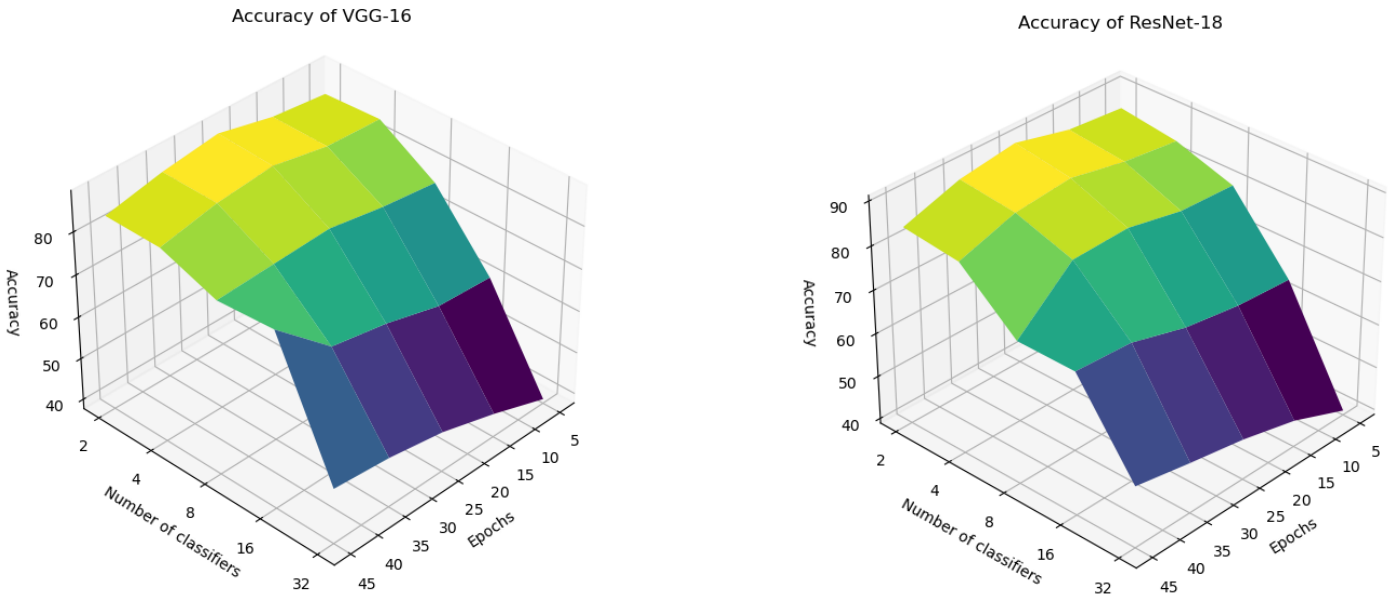


Figure 6. Ablation study results using VGG-16 backbone on pairwise learning. The blue line represents IHFF, while the orange line denotes the variant where outer-relationships are excluded during feature fusion, illustrating the impact of these relationships on performance.

5.3.3. Research on Varying Numbers of Classifiers

Figure 7 illustrates the accuracy variation with respect to different values of n_c (the number of classifiers) and epochs for the backbone models VGG-16 and ResNet-18. It

can be observed that as n_c increases, the initial accuracy improves more rapidly. This improvement may be due to the multiple classifier layers, intervened by the backdoor, pre-training the fully connected layers, which leads to better initial performance. However, as training progresses, an excessive number of classifiers results in a decrease in accuracy.



(**a**) (**b**)

Figure 7. Accuracy of different n_c (number of classifiers) and epochs within backbone VGG-16 (as shown in (**a**)) and ResNet-18 (as shown in (**b**)).

Specifically, Figure 8 shows accuracy variations for different values of the number of classifiers. Generally, performance is better when the hyperparameter n_c is set to 8 or 16. Beyond this range, a sharp decline in performance occurs. Contrary to the analysis in ablation study point 4, this decline is not caused by overfitting. This can be explained by the total dimension at the final classification layer, which is given by

$$\dim f(x) = n_c \left(\frac{\dim x}{n_c} \right)^3, \tag{23}$$

where f is the classification function.

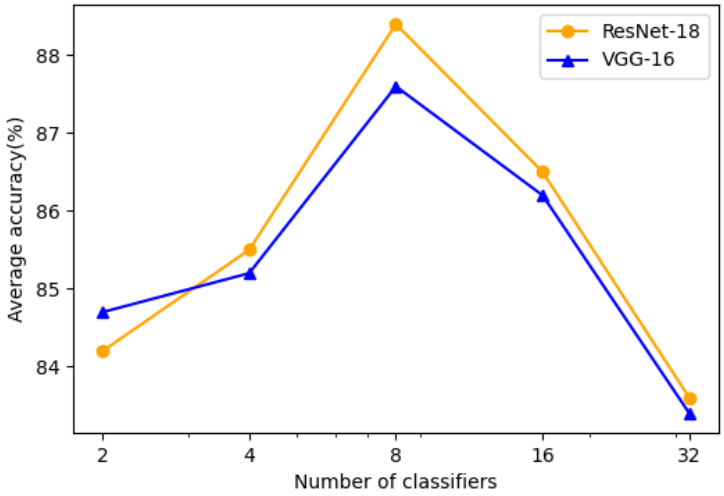


Figure 8. Results for two backbone networks with varying numbers of classifiers, where the x-axis is plotted on a logarithmic scale.

From Equation (23), it is evident that when the number of classifiers doubles, the overall dimensionality is reduced by a factor of 4. This suggests that an excessive number of classifiers will lead to a loss of semantic information. By dividing the feature space too much, the same semantic features could be split into separate parts, resulting in the loss of fine-grained part-level details. This finding indirectly supports the effectiveness of the proposed approach in extracting contextual semantic information, as discussed in Section 4.2.

5.3.4. Convergence Speed Analysis

Figure 9 illustrates the convergence speed of the loss function under different values of the hyperparameter n_c. The experiment compares model convergence performance across varying n_c settings. Results show that increasing n_c significantly accelerates convergence and reduces the final loss. The blue curve, representing the baseline model without causal reasoning, exhibits the slowest convergence, with the loss stabilizing around 2 after 30 epochs. This indicates that the absence of causal reasoning hinders convergence efficiency. With $n_c = 4$ (orange curve), the initial loss reduction is faster than the baseline, but the overall decline remains moderate, ending with a final loss around 1. When $n_c = 8$ (red curve), the convergence further improves, showing a sharp decline in the first 10 epochs and reaching a final loss near 0.6. The best performance is observed with $n_c = 16$ (green curve), where the loss drops rapidly in the early epochs and stabilizes after 20 epochs at approximately 0.3. These findings suggest that higher n_c values enhance both convergence speed and model performance. The inclusion of causal reasoning contributes significantly to more efficient training and faster inference.

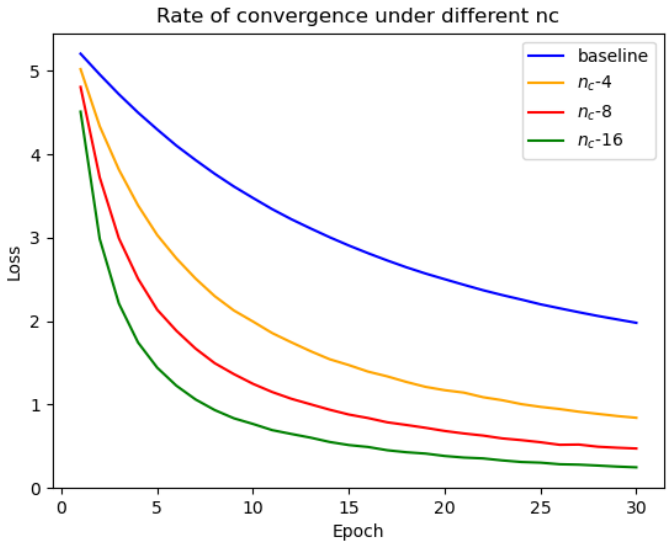


Figure 9. Convergence speed analysis (learning rate: 0.1; batch size: 16).

5.3.5. Visualization

Figure 10 presents the generated heat map, highlighting the selected high-response regions. The results indicate that IHFF consistently identifies the most discriminative areas within an image. Specifically, in the CUB dataset, it effectively emphasizes feature-rich regions, such as the bird's beak and feather texture.

In DCAL, self-attention is modified by replacing the keys (K) and values (V) with representations from two separate images, allowing the model to capture interactions between different instances of the same class. Inspired by this, to extract more effective features by using paired learning IHFF introduces an auxiliary feature extraction path (Stream C) that processes a second image of the same category. The resulting features are

then downsampled to reduce their impact on the main feature stream, allowing the model to benefit from inter-class guidance while preserving the stability and fine-grained nature of the bilinearly pooled features from Streams A and B.

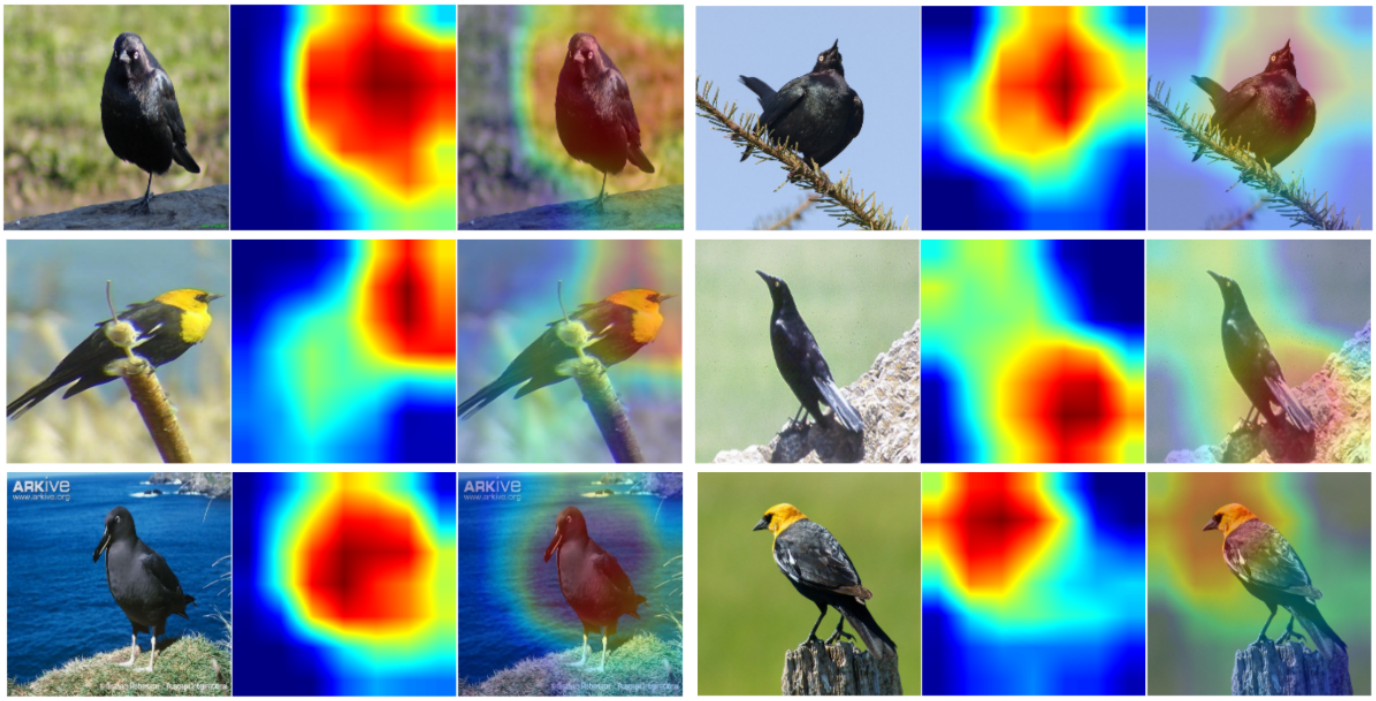


Figure 10. Visualization of the heat map for IHFF in the CUB dataset. Each group of the images in order is the CUB dataset, the heat map of IHFF, and their superposition. Specifically, (**a**) represents the effectiveness of feature extraction in IHFF, and (**b**) represents that after causal intervention, the network focuses on features with causal correlations in IHFF.

The goal is not to alter or dilute the main feature representation, but to provide additional semantic cues that can help the model better identify discriminative regions by learning from another instance of the same class. This mechanism effectively encourages the model to focus attention on the most informative parts of the object. The visualization experiment supports the effectiveness of this auxiliary feature extraction path. In Figure 10a, it can be seen that the high response attention area of IHFF is concentrated in areas where fine-grained targets have identifiable features, such as bird heads and beaks. This proves that IHFF, like most models, can effectively guide model attention to effective areas.

In Figure 10b, it can be observed that causal intervention effectively reduces misleading attention, guiding the network toward more relevant discriminative cues. For instance, the first image in Figure 10b primarily focuses on the bird itself, excluding tree branches from the high-response features. This suggests that causal intervention helps disentangle features with statistical correlation but no causal relationship. Specifically, since this bird species frequently perches on tree branches, the model may mistakenly learn branch features as intrinsic to the bird. However, tree branches are merely environmental elements and not part of the bird itself. The causal intervention successfully mitigates this confounding factor.

Similarly, the third image in Figure 10b demonstrates that causal intervention prevents the misinterpretation of tree stump textures as part of the bird's feather pattern, ensuring that high-response areas are concentrated in the correct regions. Furthermore, the second image in Figure 10b demonstrates the ability of causal representation learning to reinforce essential discriminative features. Since the bird's tail feathers are its most distinctive characteristic, causal representation learning effectively captures this feature even in the

presence of a complex background. Rather than dispersing attention across the entire bird, it focuses on the most discriminative region, enhancing classification accuracy. This discriminative and targeted focus prevents misclassification, which could otherwise arise due to the bird's predominantly black body blending with the background.

6. Conclusions

This paper introduces a novel high-order feature learning framework with causal inference for fine-grained categorization. By leveraging a tensor product space, the framework extracts high-order feature representations while mitigating the influence of confounding factors through causal interventions, specifically backdoor adjustments. To the best of the author's knowledge, this is the first comprehensive application of causal representation learning in fine-grained image analysis tasks.

The proposed method does not require bounding boxes or part annotations and can be trained end-to-end, making it flexible and widely applicable. Extensive experiments on three benchmark datasets—CUB-200, FGVC-Aircraft, and Stanford Cars—demonstrate the effectiveness and robustness of the IHFF approach.

By leveraging causal inference and high-order feature learning, this method enhances the influence robustness and interpretability, making it beneficial in real-world scenarios where fine-grained distinctions are crucial. Future work will inspire further exploration of causal interventions in fine-grained visual classification and other CV tasks. The integration of causal reasoning into computer vision presents a promising direction, and the success of this method will encourage the adoption of causal models across diverse deep learning domains. This could further drive the development of multimodal model fusion, akin to the advancements seen with transformer architectures.

Author Contributions: Conceptualization, Y.Z., Y.W. and H.L.; methodology, Y.Z., Y.W. and H.L.; software, Y.Z.; validation, Y.Z. and Z.Y.; formal analysis, Y.Z., Y.W. and J.H.; investigation, Y.Z., Y.W. and J.H.; resources, Y.Z., Y.W. and Z.Y.; data curation, Y.Z.; writing—original draft preparation, Y.Z. and J.H.; writing—review and editing, Y.W., H.L. and Z.Y.; visualization, Y.Z., Y.W., H.L. and J.H.; supervision, Y.W. and H.L.; project administration, H.L. and Z.Y.; funding acquisition, Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 864724).

Data Availability Statement: The data presented in this study are available on request from the first author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1. Proof—The Semantic Information of V and W Is Both Included in $V \otimes W$

As discussed in Section 3, it is clear that $(V \otimes W, \otimes)$ is a group. Then let

$$\begin{cases} V' = \{(v, w) \in V \otimes W \mid w = e_W\} \\ W' = \{(v, w) \in V \otimes W \mid v = e_V\}. \end{cases} \tag{A1}$$

Obviously, $(e_V, e_W) \in V'$, then for any $(v_1, e_W), (v_2, e_W) \in V'$, Equation (A2) is obtained:

$$(v_1, e_W) \otimes (v_2, e_W) = (v_1 v_2, e_W) \in V'. \tag{A2}$$

Therefore, V' is closed under multiplication. Ultimately, for each $(v, e_W) \in V'$, it follows that $(v^{-1}, e_W) \in V'$. Taken together, V' is a subgroup of $V \otimes W$. Similarly, W' is also a subgroup of $V \otimes W$.

In conclusion, the semantic information in space P is included in H, indicating that P is completely represented by H. As a result, the causal link $P \rightarrow H$ can be disregarded.

Appendix A.2. Proof—The Tensor Product of Feature Space and Its Own Is Essentially the Dimensionality Increase in Features

If the basis in $(v, e_W) \in V'$ is replaced with (v, e_V), let

$$V' = \{(v, v') \in V \otimes V \mid v' = e_v\} \tag{A3}$$

then $(v, e_V) \in V'$. This indicates that, in a certain sense, V' is a subspace of $V \otimes V$, as any basis vector in V combined with a unit vector forms a basis vector of V'. Therefore, taking the tensor product within the same feature space increases the dimension of this feature space, which aligns with the goal of high-order feature learning.

References

1. Wei, X.; Song, Y.; Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; Belongie, S. Fine-Grained Image Analysis with Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8927–8948. [CrossRef] [PubMed]
2. Ge, Y.; Xiao, Y.; Xu, Z.; Wang, X.; Itti, L. Contributions of Shape, Texture, and Color in Visual Recognition. In *European Conference on Computer Vision*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; pp. 369–386.
3. Gebru, T.; Krause, J.; Deng, J.; Fei-Fei, L. Scalable Annotation of Fine-Grained Categories Without Experts. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; CHI '17, pp. 1877–1881. [CrossRef]
4. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
5. Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN Models for Fine-Grained Visual Recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1449–1457. . [CrossRef]
6. Gao, Y.; Beijbom, O.; Zhang, N.; Darrell, T. Compact Bilinear Pooling. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 317–326. [CrossRef]
7. Kong, S.; Fowlkes, C. Low-Rank Bilinear Pooling for Fine-Grained Classification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7025–7034. [CrossRef]
8. Wang, Q.; Li, P.; Zhang, L. G2DeNet: Global Gaussian Distribution Embedding Network and Its Application to Visual Recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6507–6516. [CrossRef]
9. Cai, S.; Zuo, W.; Zhang, L. Higher-Order Integration of Hierarchical Convolutional Activations for Fine-Grained Visual Categorization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 21–26 July 2017; pp. 511–520. [CrossRef]
10. Li, P.; Xie, J.; Wang, Q.; Gao, Z. Towards Faster Training of Global Covariance Pooling Networks by Iterative Matrix Square Root Normalization. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 947–955. [CrossRef]
11. Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; You, X. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. In Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 8–14 September 2018; pp. 595–610.
12. Wei, X.; Zhang, Y.; Gong, Y.; Zhang, J.; Zheng, N. Grassmann Pooling as Compact Homogeneous Bilinear Pooling for Fine-Grained Visual Classification. In Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; pp. 365–380.
13. Zheng, H.; Fu, J.; Zha, Z.J.; Luo, J. Learning deep bilinear transformation for fine-grained image representation. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 8–14 December 2019.
14. Min, S.; Yao, H.; Xie, H.; Zha, Z.J.; Zhang, Y. Multi-Objective Matrix Normalization for Fine-Grained Visual Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4996–5009. [CrossRef] [PubMed]
15. Cui, Y.; Zhou, F.; Wang, J.; Liu, X.; Lin, Y.; Belongie, S. Kernel Pooling for Convolutional Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3049–3058. [CrossRef]
16. Engin, M.; Wang, L.; Zhou, L.; Liu, X. DeepKSPD: Learning Kernel-Matrix-Based SPD Representation For Fine-Grained Image Recognition. In Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; pp. 629–645.

17. Gao, Z.; Wu, Y.; Zhang, X.; Dai, J.; Jia, Y.; Harandi, M. Revisiting Bilinear Pooling: A Coding Perspective. *Proc. Aaai Conf. Artif. Intell.* **2020**, *34*, 3954–3961. [CrossRef]
18. Feng, F.; Zhang, Y.; Zhang, J.; Liu, B. Small Sample Hyperspectral Image Classification Based on Cascade Fusion of Mixed Spatial-Spectral Features and Second-Order Pooling. *Remote Sens.* **2022**, *14*, 505. [CrossRef]
19. Chalupka, K.; Perona, P.; Eberhardt, F. Visual causal feature learning. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, Arlington, VA, USA, 12–16 July 2015; UAI'15, pp. 181–190.
20. Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Schölkopf, B.; Bottou, L. Discovering Causal Signals in Images. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 58–66. [CrossRef]
21. Yue, Z.; Zhang, H.; Sun, Q.; Hua, X.S. Interventional Few-Shot Learning. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Dutchess County, NY, USA, 2020; Volume 33, pp. 2734–2746.
22. Tang, K.; Huang, J.; Zhang, H. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 6–12 December 2020; NIPS '20.
23. Zhang, D.; Zhang, H.; Tang, J.; Hua, X.S.; Sun, Q. Causal Intervention for Weakly-Supervised Semantic Segmentation. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Dutchess County, NY, USA, 2020; Volume 33, pp. 655–666.
24. Pearl, J. Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, USA, 2–5 August 2001; UAI'01; pp. 411–420.
25. Pearl, J.; Glymour, M.; Jewell, N.P. *Causal Inference in Statistics: A Primer*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
26. Rao, Y.; Chen, G.; Lu, J.; Zhou, J. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 1005–1014. [CrossRef]
27. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the Fisher Kernel for Large-Scale Image Classification. In Proceedings of the Computer Vision—ECCV 2010, Heraklion, Crete, 5–11 September 2010; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 143–156.
28. Lowe, D. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157. [CrossRef]
29. Li, Y.; Wang, N.; Liu, J.; Hou, X. Factorized Bilinear Models for Image Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2098–2106.. [CrossRef]
30. MacKinnon, D.P.; Fairchild, A.J.; Fritz, M.S. Mediation Analysis. *Annu. Rev. Psychol.* **2007**, *58*, 593–614. [CrossRef] [PubMed]
31. Gomez, J.P.; Akleman, D.; Akleman, E.; Pavlidis, I. Causality Effects of Interventions and Stressors on Driving Behaviors under Typical Conditions. *Mathematics* **2018**, *6*, 139. [CrossRef]
32. Keele, L. The Statistics of Causal Inference: A View from Political Methodology. *Political Anal.* **2015**, *23*, 313–335. [CrossRef]
33. Richiardi, L.; Bellocco, R.; Zugna, D. Mediation Analysis in Epidemiology: Methods, Interpretation and Bias. *Int. J. Epidemiol.* **2013**, *42*, 1511–1519. [CrossRef] [PubMed]
34. Li, H.; Hai, M.; Tang, W. Prior Knowledge-Based Causal Inference Algorithms and Their Applications for China COVID-19 Analysis. *Mathematics* **2022**, *10*, 3568. [CrossRef]
35. Pearl, J.; Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*; Basic Books: New York, NY, USA, 2018.
36. Su, H.; Wang, W. Invariant Feature Learning Based on Causal Inference from Heterogeneous Environments. *Mathematics* **2024**, *12*, 696. [CrossRef]
37. Veitch, V.; Sridhar, D.; Blei, D. Adapting Text Embeddings for Causal Inference. In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), Virtual, 3–6 August 2020; Peters, J., Sontag, D., Eds.; PMLR: New York, NY, USA, 2020; Volume 124, pp. 919–928.
38. Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E.H.; Beutel, A. Counterfactual Fairness in Text Classification through Robustness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 27–28 January 2019; AIES '19, pp. 219–226. [CrossRef]
39. Feder, A.; Keith, K.A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M.E.; et al. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1138–1158. [CrossRef]
40. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2009.
41. Tang, K.; Niu, Y.; Huang, J.; Shi, J.; Zhang, H. Unbiased Scene Graph Generation from Biased Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.

42. Qi, J.; Niu, Y.; Huang, J.; Zhang, H. Two Causal Principles for Improving Visual Dialog. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
43. Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.S.; Wen, J.R. Counterfactual VQA: A Cause-Effect Look at Language Bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
44. Zheng, H.; Fu, J.; Mei, T.; Luo, J. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5219–5227. [CrossRef]
45. Sun, M.; Yuan, Y.; Zhou, F.; Ding, E. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. In Proceedings of the Computer Vision—ECCV 2018, Cham, Switzerland, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; pp. 834–850.
46. Pearl, J. Causal Diagrams for Empirical Research. *Biometrika* **1995**, *82*, 669–688. [CrossRef]
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; NIPS'17, pp. 6000–6010.
48. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. 2011. Available online: https://authors.library.caltech.edu/records/cvm3y-5hh21/files/CUB_200_2011.pdf?download=1 (accessed on 16 April 2025).
49. Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; Vedaldi, A. Fine-grained Visual Classification of Aircraft. *arXiv* **2013**, arXiv:1306.5151.
50. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3D Object Representations for Fine-Grained Categorization. In Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, Sydney, NSW, Australia, 2–8 December 2013; pp. 554–561. [CrossRef]
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
52. Lin, T.Y.; Maji, S. Improved Bilinear Pooling with CNNs. *arXiv* **2017**, arXiv:1707.06772.
53. Cheung, M.; Shi, J.; Jiang, L.; Wright, O.; Moura, J.M.F. Pooling in Graph Convolutional Neural Networks. In Proceedings of the 2019 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 3–6 November 2019; pp. 462–466. [CrossRef]
54. Sikdar, A.; Liu, Y.; Kedarisetty, S.; Zhao, Y.; Ahmed, A.; Behera, A. Interweaving Insights: High-Order Feature Interaction for Fine-Grained Visual Recognition. *Int. J. Comput. Vis.* **2025**, *133*, 1755–1779. [CrossRef] [PubMed]
55. Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; Shan, Y. Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 4682–4692.