Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Deconfounded hierarchical multi-granularity classification

Ziyu Zhao, Leilei Gan, Tao Shen, Kun Kuang *, Fei Wu

Zhejiang University, Hangzhou, Zhejiang 310027, China

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Hierarchical multi-granularity classification
Causal inference

ABSTRACT

Hierarchical multi-granularity classification (HMC) assigns labels at varying levels of detail to images using a structured hierarchy that categorizes labels from coarse to fine, such as [“Suliformes”, “Fregatidae”, “Frigatebird”]. Traditional HMC methods typically integrate hierarchical label information into either the model’s architecture or its loss function. However, these approaches often overlook the spurious correlations between coarse-level semantic information and fine-grained labels, which can lead models to rely on these non-causal relationships for making predictions. In this paper, we adopt a causal perspective to address the challenges in HMC, demonstrating how coarse-grained semantics can serve as confounders in fine-grained classification. To comprehensively mitigate confounding bias in HMC, we introduce a novel framework, Deconf-HMC, which consists of three main components: (1) a causal-inspired label prediction module that combines fine-level features with coarse-level prediction outcomes to determine the appropriate labels at each hierarchical level; (2) a representation disentanglement module that minimizes the mutual information between representations of different granularities; and (3) an adversarial training module that restricts the predictive influence of coarse-level representations on fine-level labels, thereby aiming to eliminate confounding bias. Extensive experiments on three widely used datasets demonstrate the superiority of our approach over existing state-of-the-art HMC methods.

1. Introduction

Traditional image classification tasks, like fine-grained vision classification (FGVC), typically concentrate on single-granularity classification. However, real-world scenarios often present labels with a hierarchical structure (Zhang et al., 2022; Li et al., 2022, 2023). For example, as depicted in Fig. 1, a bird can be categorized into various levels of granularity, such as [“Suliformes”, “Fregatidae”, “Frigatebird”]. This hierarchical structure allows for more precise classifications by considering attributes across different levels. Consequently, there is increasing interest in hierarchical multi-granularity classification (HMC) within the field of computer vision. Cerri et al. (2016), Wehrmann et al. (2018), Giunchiglia and Lukasiewicz (2020), Chen et al. (2018), Chang et al. (2021), Chen et al. (2022).

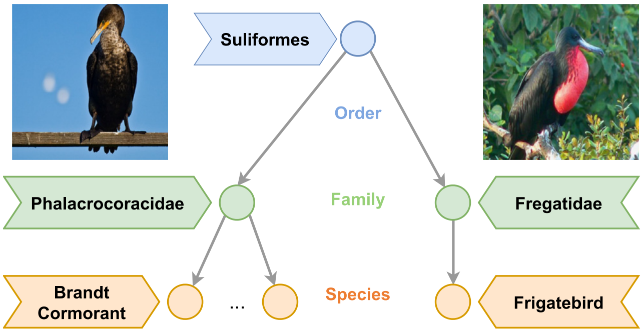


Fig. 1. Label hierarchy of birds according to the biological taxonomy.

Previous studies on Hierarchical Multi-granularity Classification (HMC) have primarily focused on embedding hierarchical label information into the network structure or loss function to improve model accuracy (Cerri et al., 2016; Wehrmann et al., 2018; Giunchiglia and Lukasiewicz, 2020; Chen et al., 2018). However, these methods fall short in exploring the interactions between different levels of granularity, as they do not explicitly model or disentangle information across these levels, which can lead to suboptimal performance in HMC. Recent research (Chang et al., 2021; Chen et al., 2022) has started to bridge this gap by more thoroughly examining the interactions across granularities. Specifically, Chen et al. (2022) propose that fine-level classes should inherit attributes from their coarse-level superclasses to enrich the classification context, thereby enhancing the predictive capability. Meanwhile, Chang et al. (2021) demonstrate that an over-reliance on

* Corresponding author.
E-mail addresses: kunkuang@zju.edu.cn (K. Kuang), wufei@zju.edu.cn (F. Wu).

https://doi.org/10.1016/j.cviu.2024.104108
Received 29 May 2024; Received in revised form 17 July 2024; Accepted 6 August 2024
Available online 20 August 2024

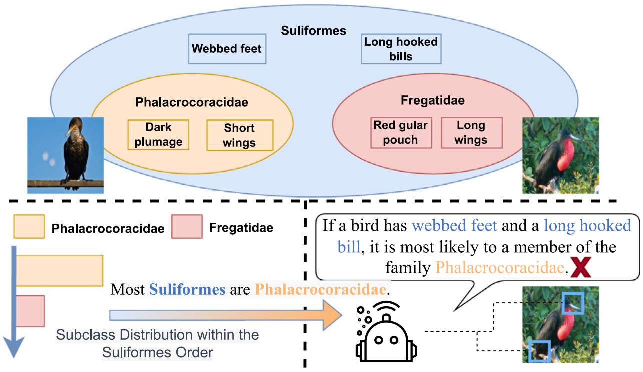

Suliformes
Webbed feet
Long hooked bills
Phalacrocoracidae
Dark plumage
Short wings
Fregatidae
Red gular pouch
Long wings
Phalacrocoracidae
Fregatidae
If a bird has webbed feet and a long hooked bill, it is most likely to a member of the family Phalacrocoracidae. ✗
Most Suliformes are Phalacrocoracidae.
Subclass Distribution within the Suliformes Order

Fig. 2. Illustration of how an imbalanced subclass distribution can create spurious correlations between coarse-level features and fine-level label predictions.

coarse-level features can impair the performance of fine-level classifiers. Despite these advancements, these approaches overlook the issue of spurious correlations between coarse-grained semantic information and fine-grained labels. This oversight can cause models to rely on these non-causal relationships for predictions, thereby compromising the reliability and accuracy of the classification outcomes.

Here, we present a straightforward example to demonstrate how spurious correlations between coarse-grained semantic information and fine-grained labels can arise in HMC. As shown in Fig. 2, the Suliformes order primarily encompasses two families: Phalacrocoracidae and Fregatidae. While each subfamily is marked by distinct characteristics, the Suliformes order generally exhibits common traits such as webbed feet and long hooked bills. In the CUB200 dataset (Wah et al., 2011), a majority of the birds classified within the Suliformes order are from the Phalacrocoracidae family. This predominance leads to spurious correlations between the general features of the Suliformes and the specific label of Phalacrocoracidae, triggered by a skewed label distribution. Consequently, when models use these broad Suliformes traits for prediction, they are prone to incorrectly identifying birds as Phalacrocoracidae, neglecting the unique characteristics of other subfamilies. Relying on such spurious correlations can therefore result in erroneous predictions, such as inaccurately identifying a bird with webbed feet and long hooked bills as belonging to the Phalacrocoracidae family, when it may actually not.

To address this gap, we approached the problem from a causal perspective to thoroughly analyze the mechanisms driving these spurious correlations. In the causal graph for HMC, we identified two backdoor paths that introduce confounding biases: one path arises from the correlation between semantic information at various granularities, and the other results from the influence of coarse-grained features on fine-grained label predictions. To counteract these biases, we developed a novel framework, Deconf-HMC, which comprises three integral components: (1) a label prediction module that utilizes fine-level features and coarse-level predictions to accurately forecast labels at each hierarchical level; (2) a representation disentanglement module designed to reduce the mutual information between representations at different granularities, thus enabling the model to focus on relevant semantic information for accurate label prediction; and (3) an adversarial training module specifically tailored to limit the influence of coarse-level representations on fine-grained label predictions. Our extensive experiments demonstrate that the Deconf-HMC framework not only effectively mitigates confounding biases but also consistently outperforms existing state-of-the-art methods.

The main contributions of this paper are summarized as follows. (1) We investigate a significant yet often overlooked challenge in hierarchical multi-granularity classification (HMC), which is the spurious correlations between coarse-level features and fine-grained labels that arise from imbalanced subclass distributions. (2) We analyze the HMC problem from a causal perspective and introduce a novel framework, Deconf-HMC, designed to eliminate confounding biases inherent in HMC. (3) Through experiments on three widely used HMC datasets, we demonstrate that our proposed method not only outperforms existing state-of-the-art approaches but also effectively resolves the confounding bias prevalent in HMC.

2. Related works

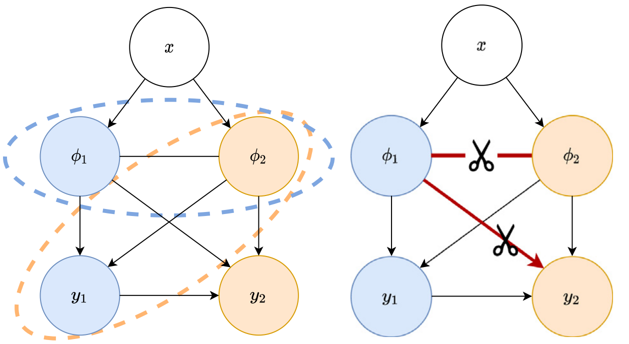
2.1. Leveraging the hierarchical multi-label

Apart from traditional vision tasks focused on singular granularity (Guo et al., 2024; Wei et al., 2021), hierarchical multi-label classification (HMC) organizes labels into a hierarchy based on prior knowledge and performs multi-label classification accordingly. Previous works have typically integrated hierarchical label information into network structures or adapted loss functions to enhance HMC performance. For instance, HMC-LMLP enhances subclass representation by using the predicted labels of the parent class as feature inputs for the subclass classifier (Cerri et al., 2016). Similarly, HSE introduces a hierarchical semantic embedding with a tree structure, leveraging the predicted score vector from the upper level as prior knowledge to refine feature representations (Chen et al., 2018). Furthermore, C-HMCNN modifies the binary cross-entropy loss to conform to the parent–child constraint, ensuring that if a sample is classified within a subclass, it is also recognized under all corresponding parent classes (Giunchiglia and Lukasiewicz, 2020). However, these methods do not adequately address the reinforcing effects between representations at different granularities, which can result in suboptimal model performance.

Recent studies in fine-grained vision classification (FGVC) have leveraged the inherent hierarchical structure of fine-grained labels to organize them into hierarchies for multi-label classification. Chang et al. (2021) developed a technique employing classification heads specific to each granularity level, which isolates coarse-level features from fine-grained ones. This approach allows for the integration of fine-grained features into coarser-grained label predictions while restricting gradient flow to update only the parameters within each classification head. Meanwhile, Chen et al. (2022) introduced the Hierarchical Representation Network (HRN), inspired by the notion that labeling objects at multiple levels should facilitate the transfer of hierarchical knowledge. In HRN, lower-level classes inherit attributes from their upper-level superclasses, enhancing the model's understanding of inter-level dynamics. While these methods effectively address interactions between different granularities, they often overlook the spurious correlations that arise from imbalanced subclass distributions, which can hinder the model's ability to rely on accurate causal relationships for classification.

2.2. Causal representation learning

Causal representation learning aims to develop balanced representations that mitigate confounding bias, enabling models to focus on genuine causal relationships. Previous work in this area has emphasized the importance of reducing correlations between confounders and treatments by imposing specific distance constraints. Common strategies include using the Wasserstein distance to control the distributional distance between intervention and control groups, which has been shown to be particularly effective for binary treatments (Shalit et al., 2017; Guo et al., 2020; Wu et al., 2022). For scenarios with high-dimensional or continuous treatments, other studies have employed distance constraints such as the Hilbert–Schmidt Independence Criterion (HSIC) to address confounding (Bahng et al., 2020; Ma and Tresp, 2021; Zhao et al., 2022). Furthermore, some approaches focus on learning de-biased representations by applying sample weighting techniques post-training to balance the distributions (Zou et al., 2020; Hassanpour and Greiner, 2019). Recent advancements have explored methods to learn distribution weights or refine representations through adversarial training, which improves model calibration (Kallus, 2020; Ozery-Flato et al., 2018).



(a) Causal graph for HMC (b) Deconfounded HMC

Fig. 3. The causal graph of HMC. In this notation, we use x to represent the input image, ϕ_1 and ϕ_2 to represent the semantics of the superclass and subclass, respectively, and y_1 and y_2 to denote the labels of the superclass and subclass, respectively.

3. A causal view of HMC

In this work, we approach the problem of hierarchical multi-granularity classification (HMC) using a Structural Causal Model (SCM) (Pearl, 2009)[1]. To construct the SCM, we examine the causal relationships among key elements, including the image x, coarse-level representation ϕ_1, fine-level representation ϕ_2, coarse-level label y_1, and fine-level label y_2. Fig. 3(a) illustrates these relationships, with each directed link indicating a causal connection between two nodes.

Given an input image x, we derive two distinct representations: ϕ_1 and ϕ_2, coarse-grained and fine-grained levels, respectively. Due to the nuanced distinction between these semantic levels, each representation influences both the coarse-grained label y_1 and the fine-grained label y_2. Additionally, in alignment with the hierarchical structure of the labels in our Structural Causal Model (SCM), the coarse-grained labels y_1 exert a causal influence on the corresponding fine-grained labels y_2. It is crucial to acknowledge that while ϕ_1 and ϕ_2 are correlated, the directionality of causality between these representations remains ambiguous. Consequently, we depict their relationship using an undirected edge, as recommended by Tchetgen Tchetgen et al. (2021).

3.1. Causal analysis of hierarchical label prediction

Here, we discuss how to utilize information from different granularity levels, as depicted in the causal graph (Fig. 3(a)), to predict labels across various levels of granularity.

Takeaway #1: Fine-level features enhance coarse-level prediction. **For coarse-level prediction**, we focus on the causal effect from $\{\phi_1, \phi_2\}$ to the coarse-level label y_1,as highlighted by the variables enclosed by blue dotted lines in Fig. 3(a). Ideally, by utilizing both coarse-level and fine-level features, we can enhance the accuracy of predicting the coarse-level label. The integration of fine-grained information, ϕ_2, is justified because it often includes discriminative attributes that help identify broader categories. For example, the presence of a red gular pouch is a distinctive trait of frigatebirds, suggesting classification within the Suliformes order. As there is no backdoor path[2] from $\{\phi_1, \phi_2\} \rightarrow y_1$, the coarse-level label y_1 can be accurately predicted with $\{\phi_1, \phi_2\}$ (Greenland and Pearl, 2007).

Takeaway #2: Coarse-level features can impair fine-level prediction accuracy. **For fine-level prediction**, we concentrate on the causal effect from $\{y_1, \phi_2\}$ to the fine-level label y_2, as indicated by the variables within orange dotted lines in Fig. 3(a). Ideally, fine-grained classification should utilize both the coarse-grained label and the fine-grained representation. For example, recognizing that a bird belongs to the Suliformes order and exhibits specific traits such as a red gular pouch and long wings, enables precise classification as Fregatidae. However, reliance on coarse-grained features alone may foster spurious associations. However, this process can be complicated by the coarse-grained feature ϕ_1, which may act as a confounder due to its correlations with both y_2 and ϕ_2, thereby introducing confounding bias. This interplay can lead to spurious associations and affect the accuracy of fine-grained label predictions.

3.2. Deconfounded representation learning

Given our previous analysis, which confirmed the absence of confounding bias in predicting coarse-grained labels, Fig. 3(b) delves into the sources of confounding bias in fine-level prediction. The bias primarily stems from two specific relationships within the graph, highlighted in red. Focusing on the causal link from $\{\phi_2, y_1\}$ to y_2, we identify confounding bias arising from two backdoor paths: $\phi_2 \leftrightarrow \phi_1 \rightarrow y_1 \rightarrow y_2$, and $y_1 \leftarrow \phi_1 \rightarrow y_2$. To mitigate these biases, it is crucial to develop balanced representations that neutralize the effects of these backdoor paths. Strategies to achieve balanced representations across different granularity levels as follows.

- **Eliminate Correlations Between Different Granularity Levels:** This involves disrupting the existing correlation between ϕ_1 and ϕ_2, aiming to sever the indirect path $\phi_2 \not\leftrightarrow \phi_1 \rightarrow y_1 \rightarrow y_2$ (Section 4.2.1).
- **Restrict Predictive Ability of Coarse-Level Representations on Fine-Level Labels:**, This strategy reduces the impact of ϕ_1 on y_2, thereby modifying the causal path to $y_1 \leftarrow \phi_1 \not\rightarrow y_2$ (Section 4.2.2).

4. Methodology

This section introduces *Deconf-HMC*, a causal-inspired framework designed to address confounding bias in HMC. The model structure is depicted in Fig. 4. Assuming the label hierarchy comprises K levels, the process begins with an input image x being fed through a trunk network, $\mathcal{F}$, to extract image features. The output, $\phi = \mathcal{F}(x)$, is segmented into K portions, represented as $\phi_1, \ldots, \phi_k$, which serve as inputs for the prediction heads at different granularity levels, denoted as $\mathcal{G}_1, \ldots, \mathcal{G}_k$. Each prediction head is tasked with predicting a specific label from the set $y_1, \ldots, y_k$, spanning from the coarsest to the finest granularity. The framework incorporates three main modules for addressing confounding bias in HMC: the label prediction module, the representation disentanglement module, and the adversarial training module.

4.1. Causally-inspired hierarchical label prediction

To account for varying levels of granularity $\{\phi_1, \ldots, \phi_K\}$, we adopt K separate prediction heads for labeling each hierarchical level. We employ a straightforward architecture for each layer i, denoted $\mathcal{G}_i$, which consists of a fully connected layer followed by a softmax activation function.

In alignment with the hierarchical labeling structure and the principles of our proposed causal graph, our label prediction module adheres to the following guidelines:

[1] Although the causal graph presented here includes label structures with only two levels, the principles outlined are applicable to structures with additional levels.

[2] A backdoor path from X to Y is a path that starts with a parent of X and ends at Y, potentially confounding the direct effect of X on Y (Greenland and Pearl, 2007).

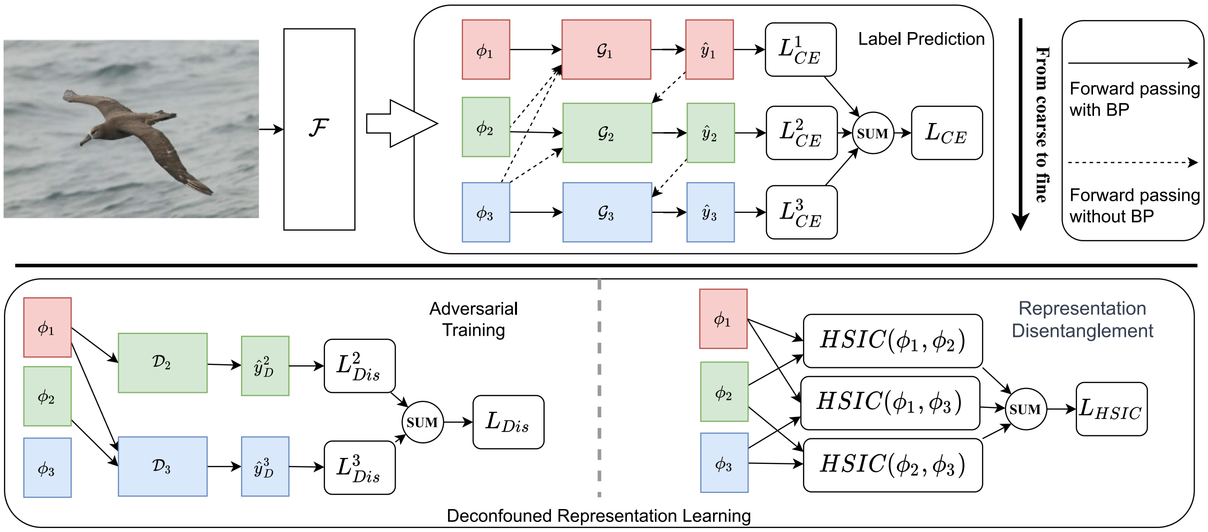


Fig. 4. The network structure of the proposed Deconf-HMC. The model comprises three main components: (1) Label prediction, (2) Representation disentanglement, and (3) Adversarial training. BP: Backpropagation.

1. Avoidance of using the representation of ancestor nodes to prevent dependencies on spurious associations in label predictions.
2. Utilization of the representation of descendant nodes to enhance label predictions.
3. Forwarding of the output from parent-level classifiers to child-level classifiers to facilitate informed predictions.

The output of the ith layer classifier $\mathcal{G}_i$, is calculated using the formula:

$$\hat{y}_i = \mathcal{G}_i(\hat{y}_{i-1}, \phi_i, \Gamma(\phi_{i+1}), \dots \Gamma(\phi_k)), \tag{1}$$

where $\Gamma(\cdot)$ represents the gradient controller used to regulate the gradient of the representation of descendant nodes, following Chang et al. (2021). During the training of the classifier $\mathcal{G}_i$, gradient flow is restricted to the features ϕ_i of the classifier itself, using $\Gamma(\cdot)$ to halt the gradients of other features. This strategy helps to disentangle information across different levels and prevents fine-grained representations from inheriting biases from coarse-grained data.

The loss function of each level is calculated as $\mathcal{L}^i_{CE} = CrossEntropy(\hat{y}_i, y_i)$, with the overall cross-entropy loss for all levels expressed as:

$$\mathcal{L}_{CE} = \sum_{i=1}^{k} \mathcal{L}^i_{CE} = \sum_{i=1}^{k} CrossEntropy(\hat{y}_i, y_i). \tag{2}$$

4.2. Deconfounded representation learning

4.2.1. Representation disentanglement between different granularities

Although we differentiate representations based on labels of different granularities, we cannot guarantee the complete elimination of correlations between these representations. In other words, $\phi_2 \leftrightarrow \phi_1 \rightarrow y_1 \rightarrow y_2$ remains unblocked. Consequently, it is necessary to quantify the degree of independence between representations at different granularity levels. To this end, we employ the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al., 2007) to measure and minimize the correlation between these representations. The HSIC is defined as:

$$\mathrm{HSIC}(\mu, \nu) = (N-1)^{-2} \operatorname{tr}(\mathrm{K}^{\mu}\mathrm{H}\mathrm{K}^{\nu}\mathrm{H}), \tag{3}$$

where μ and ν represents the sample sets of two random variables, K^{μ} and K^{ν} are their respective kernel matrices, and H is the centering matrix defined as $\mathrm{H} = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$, with $\operatorname{tr}(\cdot)$ denoting the trace of a matrix. As the value of HSIC approaches zero, it indicates that the two random variables are more independent. Therefore, HSIC serves as a crucial criterion for supervising feature decorrelation (Bahng et al., 2020; Zhang et al., 2021). We incorporate the following loss function to enforce the HSIC constraint across different granular levels of representations:

$$\mathcal{L}_{HSIC} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \mathrm{HSIC}(\phi_i, \phi_j). \tag{4}$$

By imposing this constraint, we aim to eliminate correlations between representations at different granularities. This allows each representation to focus on distinct information, thereby reducing confounding bias and effectively blocking the path $\phi_2 \nleftrightarrow \phi_1 \rightarrow y_1 \rightarrow y_2$.

4.2.2. Eliminating spurious correlations through adversarial training

Through the representation disentanglement module, we have effectively eliminated the confounding bias caused by the correlation between representations of different granularities. However, the confounding bias due to the correlation between coarse-level representations and fine-level labels ($y_1 \leftarrow \phi_1 \rightarrow y_2$) remains a challenge. To mitigate this bias, we have integrated an adversarial training component designed to restrict the predictive capabilities of coarse-level features for fine-level labels. Specifically, we introduce a discriminator $\mathcal{D}_i$, at each level i, except for the top level, to limit the influence of coarse-level representations on fine-grained classification. Prior to each training epoch, we calculate the output of the discriminator as follows:

$$\hat{y}^i_D = \mathcal{D}_i(\phi_1, \dots, \phi_{i-1}), \tag{5}$$

where $\hat{y}^i_D$ represents the aggregated coarse-level features. We then use this output to compute the adversarial training loss for enhancing fine-grained label prediction accuracy:

$$\mathcal{L}_{Dis} = \sum_{i=2}^{k} CrossEntropy(\hat{y}^i_D, y_i). \tag{6}$$

Discriminators are updated at the beginning of each training epoch, ensuring that this adjustment occurs without altering the main network architecture or the features during this phase.

During the primary training phase, our objective is to maximize $\mathcal{L}_{Dis}$ in order to reduce the influence of coarse-level features on fine-level label predictions. To achieve this, we update all network layers except for the discriminators $\{\mathcal{D}_2, \dots, \mathcal{D}_k\}$. The overall loss function can be formulated as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{CE} + \mathcal{L}_{HSIC} - \alpha\mathcal{L}_{Dis}, \tag{7}$$

where α is a hyper-parameter that balances the different components of the loss. By minimizing this function, we optimize network performance while excluding the discriminators from updates. This method allows the model to learn balanced representations and effectively eliminate confounding bias in HMC.

4.3. *Overall framework of deconf-HMC*

The Deconf-HMC model comprises three primary components: the backbone network $\mathcal{F}$, the classifiers $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$, and the discriminators $\{\mathcal{D}_2, \ldots, \mathcal{D}_k\}$. The training process is structured as follows:

- Update $\mathcal{F}$ and $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$ with $\mathcal{L}_{overall} = \mathcal{L}_{CE} + \mathcal{L}_{HSIC} - \alpha\mathcal{L}_{Dis}$.
- Update the discriminators $\{\mathcal{D}_2, \ldots, \mathcal{D}_k\}$ with $\mathcal{L}_{Dis}$.

By iteratively executing these updates, the model learns to effectively reduce confounding biases, enhancing its performance in hierarchical multi-granularity classification tasks. The detailed pseudocode for implementing these processes is provided in Algorithm 1.

Algorithm 1 Pseudocode for Deconf-HMC

```
Initialize the backbone network $\mathcal{F}$
Initialize classifiers $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$
Initialize discriminators $\{\mathcal{D}_2, \ldots, \mathcal{D}_k\}$
while not converged do
    for each batch $x$ do
        $[\phi_1, \cdots, \phi_K] \leftarrow \mathcal{F}(x)$        ▷ Extract features
        for $i = 1$ to $k$ do
            if $i == 1$ then
                $\hat{y}_i \leftarrow \mathcal{G}_i(\phi_i)$
            else
                $\hat{y}_i \leftarrow \mathcal{G}_i(\hat{y}_{i-1}, \phi_i)$
            end if
        end for
        $\mathcal{L}_{CE} \leftarrow \sum_{i=1}^{k}$ CrossEntropy$(\hat{y}_i, y_i)$
        $\mathcal{L}_{HSIC} \leftarrow$ HSIC$(\phi_1, \cdots, \phi_K)$
        $\mathcal{L}_{overall} \leftarrow \mathcal{L}_{CE} + \mathcal{L}_{HSIC} - \alpha \times \mathcal{L}_{Dis}$
        Update $\mathcal{F}, \{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$ with $\mathcal{L}_{overall}$
        for $i = 2$ to $k$ do
            $\hat{y}_D^i \leftarrow \mathcal{D}_i(\phi_1, \ldots, \phi_{i-1})$
            $\mathcal{L}_{Dis}^i \leftarrow$ CrossEntropy$(\hat{y}_D^i, y_i)$
        end for
        $\mathcal{L}_{Dis} \leftarrow \sum_{i=2}^{k} \mathcal{L}_{Dis}^i$
        Update $\{\mathcal{D}_2, \ldots, \mathcal{D}_k\}$ with $\mathcal{L}_{Dis}$
    end for
end while
```

4.4. *Further discussion*

4.4.1. *The main difference between Deconf-HMC and base model*

Given that our model builds upon the framework proposed by Chang et al. (2021), it is crucial to delineate the specific enhancements made:

Incorporating the prediction of the coarse-level label. Unlike the base model, our approach integrates the prediction results of parent class labels, which were previously unconsidered. This inclusion addresses the base model's limitation due to the absence of explicit parent class label information, which hindered the proper disentanglement of fine-grained semantic information. Consequently, the base model's fine-grained representations were often conflated with coarse-level information.

Explicitly constrained the mutual information of different granularity representations. We have implemented constraints on the HSIC between representations of different granularities, enabling more focused attention to distinct information. The base model lacks such constraints, which has traditionally prevented it from effectively decoupling information within the learned representations.

Limit the ability of coarse-grained features to predict fine-grained labels. Our analysis reveals that coarse-grained information often acts as a confounder in the fine-grained classification process. By restricting the predictive influence of coarse-grained representations on fine-grained labels, we effectively address the issue of confounding bias—a consideration absent in previous works on HMC including those by Cerri et al. (2016), Wehrmann et al. (2018), Giunchiglia and Lukasiewicz (2020), Chen et al. (2018), Chang et al. (2021), Chen et al. (2022).

4.4.2. *Scalability of Deconf-HMC*

The Deconf-HMC model incurs extra computational overhead primarily due to its adversarial training and representation disentanglement modules, which are crucial for addressing the challenges inherent in hierarchical multi-granularity classification tasks. These modules, operational during the training phase, effectively eliminate spurious correlations and decouple representations at varying granularity levels, thus removing biases and unnecessary correlations from the training data to enhance the model's ability to generalize to real-world data. Importantly, these modules do not entail additional computational demands during the inference phase, thereby preserving the operational efficiency of the model in practical applications. Furthermore, the majority of hierarchical label structures typically involve three to five levels (Chang et al., 2020; Giunchiglia and Lukasiewicz, 2020; Chang et al., 2021), confirming that the Deconf-HMC model's architecture is aptly designed to meet the demands of most real-world scenarios.

5. Experiments

5.1. *Datasets*

We evaluated the proposed methods using three widely recognized fine-grained visual classification (FGVC) datasets. For datasets that only provided fine-grained labels, we adopted hierarchical labels by tracking parent nodes across Wikipedia, following the method described in Chang et al. (2021). A detailed description of the datasets is as follows: (i) **CUB-200-2011** (Wah et al., 2011) is a comprehensive dataset aimed at recognizing 200 bird species. We organized it into a three-level label hierarchy consisting of 13 orders, 38 families, and 200 species. (ii) **FGVC-Aircraft** (Maji et al., 2013) is specifically designed for aircraft recognition. This dataset includes 10,000 images categorized into a three-tier hierarchy comprising 30 makers, 70 families, and 100 plane models. (iii) **Stanford Cars** (Krause et al., 2013) focuses on car recognition and includes 16,185 images across 196 different car classes. We restructured this dataset into a two-level hierarchy by introducing 9 superordinate car types. In all experiments, following the precedent set by Chang et al. (2021), Chen et al. (2022), we do not utilize bounding box or part annotations and adhere to the official train/test splits provided.

5.2. *Dataset analysis*

The primary cause of spurious correlations in classification tasks is the imbalanced distribution of subclasses. As shown in Fig. 5, we examined the subclass distribution within the Passeriformes order, the largest sample group in the CUB200 dataset (Wah et al., 2011). Our analysis identified a significant imbalance among the subclasses. This disparity can lead to spurious correlations between the features of the parent class and the labels of the subclasses.

Table 1
Comparisons with the state-of-the-art methods on both HMC task and FGVC task with hierarchical multi-labels. We report the overall accuracy for different granularity and the $AU(\overline{PRC})$ for each dataset. The granularity indicated by the underline corresponds to the performance that traditional FGVC tasks focus on.

	CUB-200–2011				FGVC-Aircraft				Stanford Cars		
	Order	Family	$\underline{\text{Species}}$	$AU(\overline{PRC})$	Maker	Family	$\underline{\text{Model}}$	$AU(\overline{PRC})$	Type	$\underline{\text{Maker}}$	$AU(\overline{PRC})$
HMC-LMLP (Cerri et al., 2016)	98.45	94.24	79.60	0.945	97.09	94.39	90.25	0.968	96.98	87.65	0.953
HMCN (Wehrmann et al., 2018)	97.29	93.15	79.75	0.934	96.07	92.56	87.19	0.959	95.21	88.71	0.953
C-HMCNN (Giunchiglia and Lukasiewicz, 2020)	98.48	94.63	81.58	0.960	97.45	95.41	91.69	0.979	96.75	90.64	0.971
HSE (Chen et al., 2018)	98.17	94.35	83.46	0.965	96.6	94.53	91.43	0.979	97.23	92.24	0.980
Chang et al. (2021)	97.76	94.17	85.56	0.968	96.88	95.28	91.92	0.981	96.40	93.65	0.977
HRN (Chen et al., 2022)	98.67	95.51	86.60	0.969	97.45	95.79	92.58	0.976	97.41	94.03	**0.981**
Ours	**99.15**	**96.11**	**87.50**	**0.975**	**97.48**	**95.85**	**93.15**	**0.982**	**97.49**	**94.34**	0.980

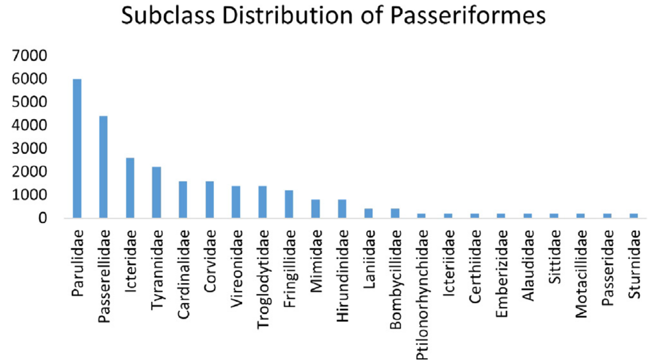


Fig. 5. An illustration of in the CUB200 dataset (Wah et al., 2011), there is a significant imbalance in the data distribution of the subclass (family level) in the Passerimormes order, which can lead to bias in fine-level prediction.

5.3. Evaluation metrics

To evaluate the performance of our proposed method in the HMC setting, we utilize overall accuracy and the area under the average precision and the recall curve $AU(\overline{PRC})$ as metrics, following (Chang et al., 2021; Chen et al., 2022). $AU(\overline{PRC})$ calculates the average precision–recall curve across all classes in the hierarchy, providing a comprehensive measure of the model's probability output vector. For each threshold value, one point $(\overline{Prec}, \overline{Rec})$ on the average precision–recall curve is computed as follows:

$$\overline{Prec} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}, \quad (8)$$

$$\overline{Rec} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}, \quad (9)$$

where i ranges over all classes, and TP_i, FP_i, and FN_i represent the counts of true positives, false positives, and false negatives for each class label i, respectively. By varying the threshold, we derive the average PRC curve. The $AU(\overline{PRC})$ is the area under this curve, which offers a significant advantage as it is independent of the class prediction threshold that varies widely with specific applications. This metric is also one of the most commonly used in HMC literature (Bi and Kwok, 2011; Wehrmann et al., 2018; Giunchiglia and Lukasiewicz, 2020; Chen et al., 2022)".

5.4. Implementation details

For a fair comparison, we adopt the base model as (Chang et al., 2021), which uses ResNet50 pre-trained on ImageNet as the backbone and splits the resulting representation into three parts for predicting different granularity classes. In all of our experiments, we resize the input images to 448 × 448 and train every experiment for 180 epochs. Standard data augmentation techniques like random horizontal flipping and random cropping are applied. We use stochastic gradient descent (SGD) with a momentum of 0.9, and a weight decay of 5e-4 to optimize our model. The learning rate of newly added FC layers is initialized as 2e-3, and ImageNet pre-trained convolutional layers are initialized as 2e-4. The learning rate is adjusted by the cosine annealing strategy. The batch size is set to 8.

5.5. Comparison with state-of-the-art methods

In this section, we compared our method with the state-of-the-art methods across three datasets. We report the overall accuracy of each level and the $AU(\overline{PRC})$ on the test sets of three FGVC datasets: CUB-200-2011, FGVC-Aircraft, and Stanford Cars on Table 1. We compare with the two groups of HMC methods: (1) those not considering the interaction between representations of different granularities, including HMC-LMP (Cerri et al., 2016), HMCN (Wehrmann et al., 2018), C-HMCNN (Giunchiglia and Lukasiewicz, 2020), and HSE (Chen et al., 2018); (2) those that account for cooperation between granularities, such as Chang et al. (2021) and HRN (Chen et al., 2022). All of these methods did not take into account the confounding bias that exists in HMC.

Our experimental results reveal the following: (1) Methods that disregard the interaction between hierarchical representations (HMC-LMLP, HMCN, C-HMCNN, HSE) fail to make accurate predictions and do not mitigate the confounding bias; (2) While the approaches by Chang et al. (2021) and HRN consider the mutual influence between different levels of representations, they still overlook the confounding bias, leading to suboptimal model performance; (3) Our method effectively resolves the confounding bias and consistently achieves superior predictive performance for all granularity levels across the datasets. Additionally, we evaluated these methods on traditional FGVC tasks, where the underlined granularity indicates the primary focus of these tasks. Our results demonstrate that our method outperforms the baseline methods in FGVC tasks by addressing the confounding bias in hierarchical labels. Although some techniques (Du et al., 2020; Yang et al., 2018; Chang et al., 2020) can further enhance FGVC task performance, they are not the primary focus of this paper. (4) Notably, our model shows the most significant performance improvement at the finest granularity level. In coarse-grained classification, the impact of confounding bias is relatively minor. However, finer-grained classification requires a more effective disentanglement of information at various levels, making deconfounding particularly advantageous in this context.

5.6. Ablation studies

In this section, we perform ablation studies to validate the effectiveness of disentanglement loss $\mathcal{L}_{HSIC}$ and adversarial loss $\mathcal{L}_{Dis}$. We report the overall accuracy of each level. The results of the three datasets are presented in Table 2. From the experimental results, the following observations can be made: (1) Utilizing $\mathcal{L}_{HSIC}$ alone does

Table 2
Ablation studies of $\mathcal{L}_{HSIC}$ and $\mathcal{L}_{Dis}$ on three HMC datasets.

Ablation of Deconf-HMC			CUB-200–2011			FGVC-Aircraft			Stanford Cars	
$\mathcal{L}_{CE}$	$\mathcal{L}_{HSIC}$	$\mathcal{L}_{Dis}$	Order	Family	Species	Maker	Family	Model	Type	Maker
✓			98.74	95.47	85.99	97.12	95.73	91.74	97.11	93.65
✓	✓		98.83	95.84	86.52	97.41	95.70	91.65	96.99	93.84
✓		✓	99.05	95.77	87.22	97.48	95.88	92.88	97.30	94.17
✓	✓	✓	99.15	96.11	87.50	97.48	96.00	93.15	97.49	94.34

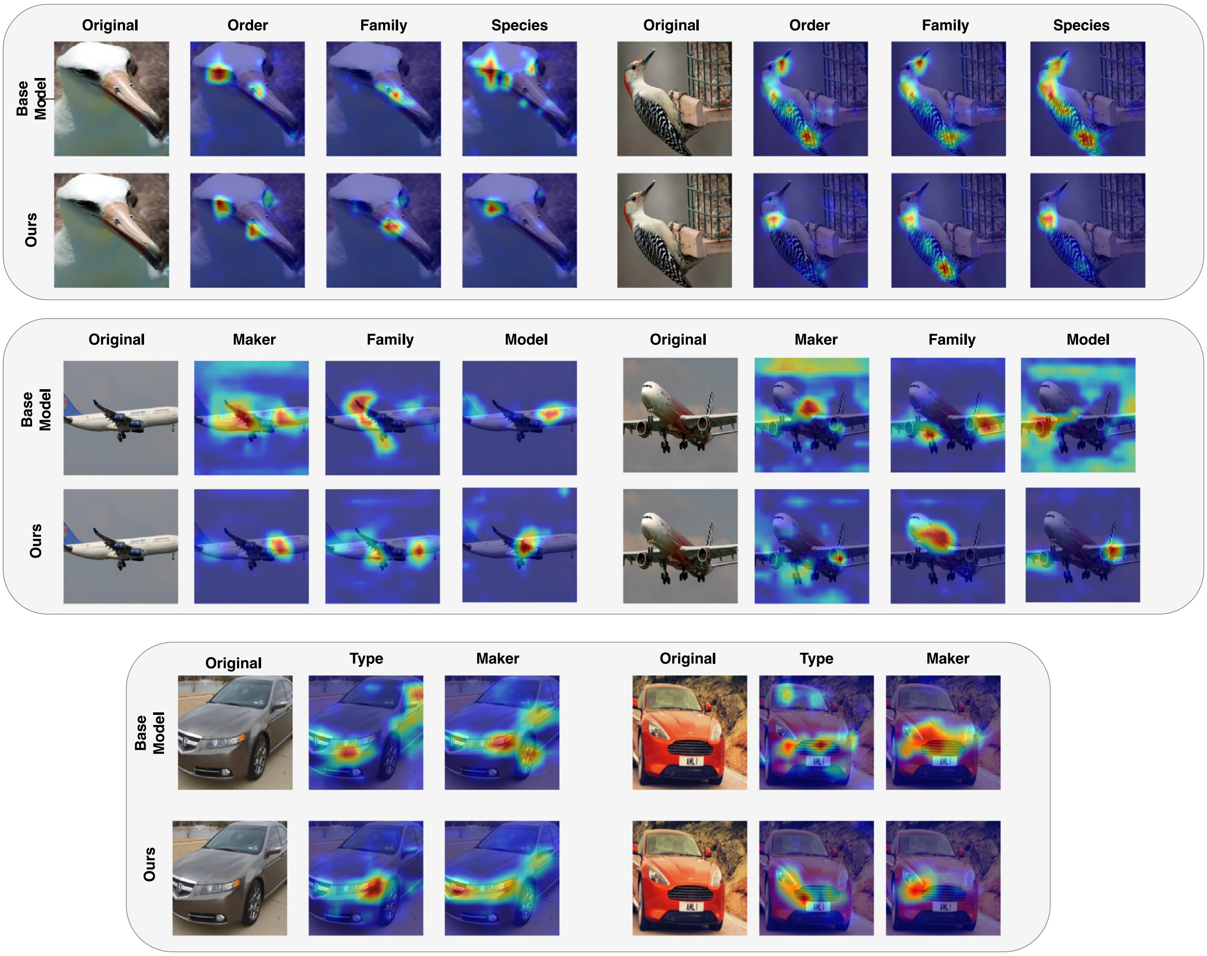


Fig. 6. The supporting visual regions for classifiers at the different granularity of the base model (Chang et al., 2021) compared with our method.

not substantially enhance the model performance. This outcome suggests that, despite its ability to constrain mutual information between representations of different granularities, $\mathcal{L}_{HSIC}$ alone is insufficient to effectively decouple these representations. The persistence of confounding bias during training limits its effectiveness. (2) $\mathcal{L}_{Dis}$ is crucial for improving the model's performance. As the spurious correlation between coarse-grained representation information and fine-grained labels is the main source of confounding bias, restricting the prediction ability of coarse-level representation to fine-level labels significantly improves the model's performance. (3) There is a mutually reinforcing effect between $\mathcal{L}_{HSIC}$ and $\mathcal{L}_{Dis}$, which makes the model's performance significantly improve when the two losses are combined. As $\mathcal{L}_{Dis}$ restricts the main source of confounding bias and determines which information representations at different granularities should attend to, decoupling the representations at different granularities enables them to more accurately focus on the corresponding information, resulting in optimal performance.

5.7. Further analysis

Do the coarse-level features harm the fine-level prediction? We made some modifications to the well-trained model of Deconf-HMC to demonstrate the harmful effects of coarse-level representations on fine-grained classifiers. We kept the trunk network and the feature layer of the model fixed and retrained the fine-grained classifier by introducing coarse-level representations for training. We evaluated the results of family-level and species-level for the CUB dataset, family-level and model-level for the Aircraft dataset, and maker-level for the Cars dataset, since there is no coarser representation for the top level. As shown in Fig. 7, after training, we can find a significant decrease in the model's performance. This indicates that the decoupled coarse-level representation has actually significantly negative effects on the classification of fine-grained labels. Furthermore, this highlights the fact that having more information does not necessarily lead to better classification results, and it is important to be aware of and eliminate confounding bias during the process.

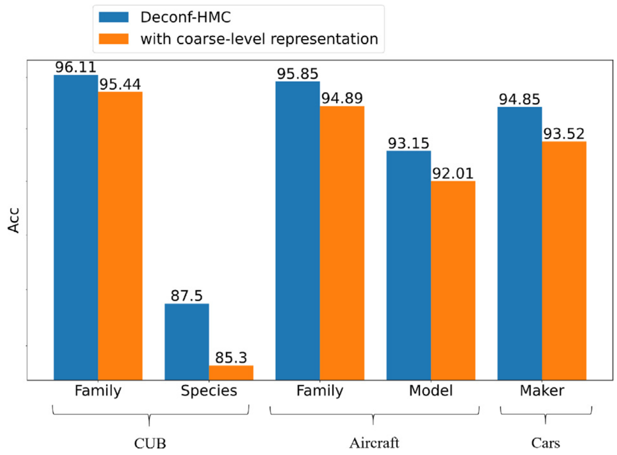
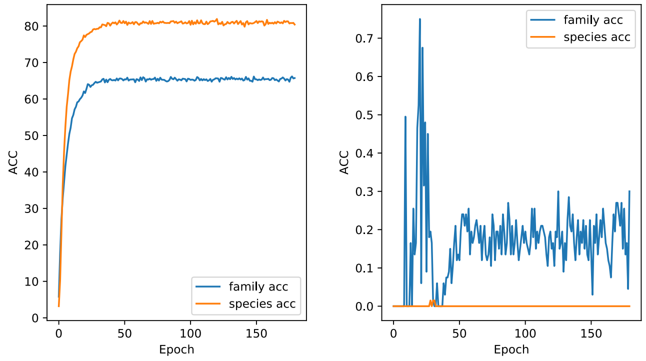


Fig. 7. Training classifiers without coarse-level representations vs. with coarse-level representations.



(a) without adversarial training

(b) with adversarial training

Fig. 8. Accuracy of the discriminator of different granularities on dataset CUB-200-2011.

Does the proposed Deconf-HMC limit the prediction ability of coarse-level features to fine-level labels? We monitored the discriminator's precision during the training process in Fig. 8. Fig. 8(a) shows the performance if we do not limit the prediction ability of coarse-level representation to fine-level labels ($\alpha = 0$ in Eq. (7)), the learned coarse-level representation has a strong prediction ability for the fine-level label, even if we do not introduce coarse-grained representations when training the fine-grained classifier. This indicates that without restrictions, the representations cannot be effectively disentangled and the confounding bias cannot be well eliminated. Fig. 8(b) shows that through adversarial training strategies, we are able to effectively limit the predictive power of coarse-level representations for fine-grained labels, thereby reducing the confounding bias (see Fig. 9).

Does the proposed Deconf-HMC disentangle the representations for different granularities? To evaluate whether our proposed method can effectively decouple representations at different granularities, we adopted Grad-Cam (Selvaraju et al., 2017) to visualize the different image regions attended by each classifier by backpropagating gradients to the original input x. Through the heatmaps, we hoped to show that classifiers at different granularities focus on different discriminative regions for classification. We compared our method with the base model (Chang et al., 2021), and from the results in Fig. 6, it is evident that Deconf-HMC effectively disentangles representations and the classifier attends different discriminative regions for classification. In contrast, the baseline model (Chang et al., 2021) often focuses on large repeated regions, indicating that their methods do not effectively disentangle representation information at different granularities. Due to some features being crucial for distinguishing categories of varying granularity, there might be some overlap on the heatmaps of different granularities. Overall, our method focuses on more detailed parts compared to the baseline model, and the heatmaps of different granularities also pay attention to more reasonable areas for discrimination.

5.8. Hyper-parameter sensitivity analysis

We investigated the impact of varying the hyperparameter α across a range of values $\{0, 1, 2, 3, 5, 8, 10, 20, 50\}$ in Fig. 9 to understand its influence on the model's performance across three datasets. Our observations indicate that a small α value leads to inadequate constraint by the adversarial training loss on the coarse-grained representations, preventing them from predicting fine-grained information effectively and thus yielding suboptimal performance. In contrast, a large α value excessively emphasizes the adversarial loss during label prediction, potentially overshadowing other important model dynamics. Optimal model performance was achieved when α was set between 1 and 10, with the most significant improvements noted. Consequently, we selected $\alpha = 5$ as the default setting for all three datasets, balancing the influence of adversarial training while maintaining robust overall model performance.

6. Conclusion

This paper explores the hierarchical multi-granularity classification problem from a causal perspective. We identify that confounding bias, stemming from skewed label distributions, impairs accurate prediction and fosters spurious correlations between coarse-grained features and fine-grained labels. To address these issues, we introduce the Deconf-HMC framework, which incorporates three critical modules designed to mitigate confounding bias. The label prediction module utilizes both coarse-level label information and fine-level representation information to improve predictive accuracy across hierarchical levels. Furthermore, we apply Hilbert–Schmidt Independence Criterion to enhance the disentanglement of representations at different granularities by constraining their mutual information. To further counteract confounding bias, our framework employs adversarial training, which restricts the impact of coarse-level representations on fine-level labels. Extensive experiments conducted on three commonly used datasets confirm the superiority of our proposed method over existing state-of-the-art HMC methods.

CRediT authorship contribution statement

Ziyu Zhao: Writing – original draft, Visualization, Validation, Methodology. **Leilei Gan:** Writing – review & editing. **Tao Shen:** Writing – review & editing. **Kun Kuang:** Writing – review & editing. **Fei Wu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376243, 62441605, U20A20387, 62037001), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study, China (SN-ZJU-SIAS-0010). All opinions in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

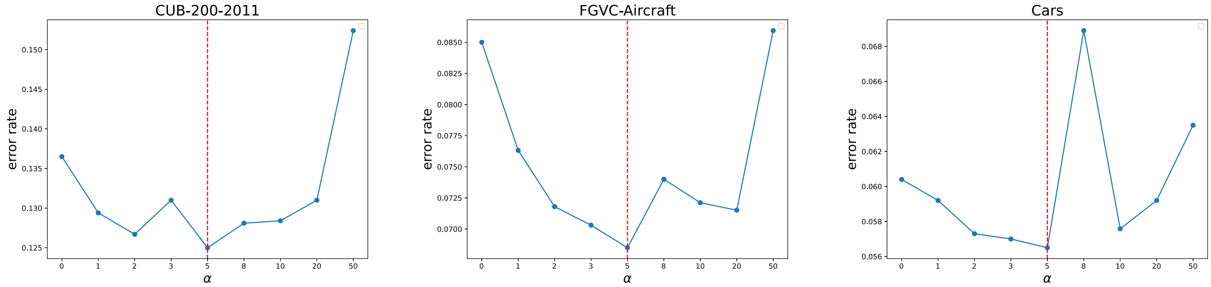


Fig. 9. Hyper-parameter sensitivity analysis on three datasets. The blue lines show the error rate of the hyper-parameter α within the specified range $\{0, 1, 5, 10, 20, 50\}$. The red line indicates the parameters chosen by Deconf-HMC.

References

Bahng, H., Chun, S., Yun, S., Choo, J., Oh, S.J., 2020. Learning de-biased representations with biased representations. In: International Conference on Machine Learning. PMLR, pp. 528–539.

Bi, W., Kwok, J.T., 2011. Multi-label classification on tree-and dag-structured hierarchies. In: Proceedings of the 28th International Conference on Machine Learning. ICML-11, pp. 17–24.

Cerri, R., Barros, R.C., PLF de Carvalho, A.C., Jin, Y., 2016. Reduction strategies for hierarchical multi-label classification in protein function prediction. BMC Bioinform. 17 (1), 1–24.

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

Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.Z., Guo, J., 2021. Your" Flamingo" is my" Bird": fine-grained, or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11476–11485.

Chen, J., Wang, P., Liu, J., Qian, Y., 2022. Label relation graphs enhanced hierarchical residual network for hierarchical multi-granularity classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4858–4867.

Chen, T., Wu, W., Gao, Y., Dong, L., Luo, X., Lin, L., 2018. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In: Proceedings of the 26th ACM International Conference on Multimedia. pp. 2023–2031.

Du, R., Chang, D., Bhunia, A.K., Xie, J., Ma, Z., Song, Y.Z., Guo, J., 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX. Springer, pp. 153–168.

Giunchiglia, E., Lukasiewicz, T., 2020. Coherent hierarchical multi-label classification networks. Adv. Neural Inf. Process. Syst. 33, 9662–9673.

Greenland, S., Pearl, J., 2007. Causal diagrams.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., Smola, A., 2007. A kernel statistical test of independence. Adv. Neural Inf. Process. Syst. 20.

Guo, D., Li, K., Hu, B., Zhang, Y., Wang, M., 2024. Benchmarking micro-action recognition: Dataset, method, and application. IEEE Trans. Circuits Syst. Video Technol. 34 (7), 6238–6252.

Guo, R., Li, J., Liu, H., 2020. Learning individual causal effects from networked observational data. In: Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 232–240.

Hassanpour, N., Greiner, R., 2019. CounterFactual regression with importance sampling weights. In: IJCAI. pp. 5880–5887.

Kallus, N., 2020. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In: International Conference on Machine Learning. PMLR, pp. 5067–5077.

Krause, J., Stark, M., Deng, J., Fei-Fei, L., 2013. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561.

Li, L., Wang, W., Zhou, T., Quan, R., Yang, Y., 2023. Semantic hierarchy-aware segmentation. IEEE Trans. Pattern Anal. Mach. Intell.

Li, L., Zhou, T., Wang, W., Li, J., Yang, Y., 2022. Deep hierarchical semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1246–1257.

Ma, Y., Tresp, V., 2021. Causal inference under networked interference and intervention policy enhancement. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 3700–3708.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A., 2013. Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151.

Ozery-Flato, M., Thodoroff, P., Ninio, M., Rosen-Zvi, M., El-Hay, T., 2018. Adversarial balancing for causal inference. arXiv preprint arXiv:1810.07406.

Pearl, J., 2009. Causality. Cambridge University Press.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Shalit, U., Johansson, F.D., Sontag, D., 2017. Estimating individual treatment effect: generalization bounds and algorithms. In: International Conference on Machine Learning. PMLR, pp. 3076–3085.

Tchetgen Tchetgen, E.J., Fulcher, I.R., Shpitser, I., 2021. Auto-g-computation of causal effects on a network. J. Amer. Statist. Assoc. 116 (534), 833–844.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-Ucsd Birds-200–2011 Dataset. California Institute of Technology.

Wehrmann, J., Cerri, R., Barros, R., 2018. Hierarchical multi-label classification networks. In: International Conference on Machine Learning. PMLR, pp. 5075–5084.

Wei, X.-S., Song, Y.-Z., Mac Aodha, O., Wu, J., Peng, Y., Tang, J., Yang, J., Belongie, S., 2021. Fine-grained image analysis with deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 44 (12), 8927–8948.

Wu, A., Yuan, J., Kuang, K., Li, B., Wu, R., Zhu, Q., Zhuang, Y.T., Wu, F., 2022. Learning decomposed representations for treatment effect estimation. IEEE Trans. Knowl. Data Eng..

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L., 2018. Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 420–435.

Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z., 2021. Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5372–5382.

Zhang, S., Xu, R., Xiong, C., Ramaiah, C., 2022. Use all the labels: A hierarchical multi-label contrastive learning framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16660–16669.

Zhao, Z., Bai, Y., Xiong, R., Cao, Q., Ma, C., Jiang, N., Wu, F., Kuang, K., 2022. Learning individual treatment effects under heterogeneous interference in networks. ACM Trans. Knowl. Discov. Data.

Zou, H., Cui, P., Li, B., Shen, Z., Ma, J., Yang, H., He, Y., 2020. Counterfactual prediction for bundle treatment. Adv. Neural Inf. Process. Syst. 33, 19705–19715.