Empowering Vision Transformers with Multi-Scale Causal Intervention for Long-Tailed Image Classification

Xiaoshuo Yan[1], **Zhaochuan Li**[2], **Lei Meng**[1*], **Zhuang Qi**[1], **Wei Wu**[1], **Zixuan Li**[1], **Xiangxu Meng**[1]

[1]School of Software, Shandong University, Jinan, China

[2]Inspur, China

{yanxiaoshuo, z_qi, wu_wei, lizixuan0707}@mail.sdu.edu.cn, lizhaoch@inspur.com, {lmeng, mxx}@sdu.edu.cn

Abstract

Causal inference has emerged as a promising approach to mitigate long-tail classification by handling the biases introduced by class imbalance. However, along with the change of advanced backbone models from Convolutional Neural Networks (CNNs) to Visual Transformers (ViT), existing causal models may not achieve an expected performance gain. This paper investigates the influence of existing causal models on CNNs and ViT variants, highlighting that ViT's global feature representation makes it hard for causal methods to model associations between fine-grained features and predictions, which leads to difficulties in classifying tail classes with similar visual appearance. To address these issues, this paper proposes TSCNet, a two-stage causal modeling method to discover fine-grained causal associations through multi-scale causal interventions. Specifically, in the hierarchical causal representation learning stage (HCRL), it decouples the background and objects, applying backdoor interventions at both the patch and feature level to prevent model from using class-irrelevant areas to infer labels which enhances fine-grained causal representation. In the counterfactual logits bias calibration stage (CLBC), it refines the optimization of model's decision boundary by adaptive constructing counterfactual balanced data distribution to remove the spurious associations in the logits caused by data distribution. Extensive experiments conducted on various long-tail benchmarks demonstrate that the proposed TSCNet can eliminate multiple biases introduced by data imbalance, which outperforms existing methods.

1 Introduction

Real-world data typically follows long-tailed distributions, resulting in models that primarily optimize for head classes and demonstrate limited generalization to tail classes [Zhang *et al.*, 2023]. Existing CNN-based long-tailed algorithms including class balancing methods [Cui *et al.*, 2019; Ren *et al.*, 2020], data augmentation [Wang *et al.*, 2024a; Ahn *et al.*, 2022], enhanced training strategies [Wang *et al.*, 2021; Du *et al.*, 2023]. With the development of Transformers, ViT employs an attention-based global feature extraction approach, facilitating the capture of finer-grained features relative to CNN architectures. However, this does not change the essence of the model's reliance on statistical information from the data, leading to an overall performance gain but a persistent gap between head and tail class performance, leaving the long-tail problem unresolved.

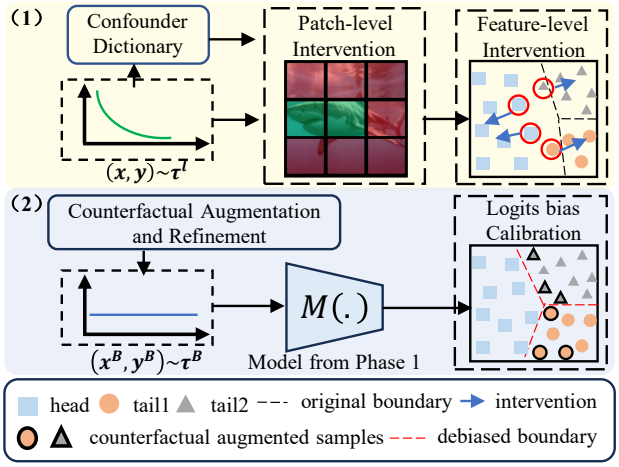

Figure 1: The illustration of the proposed TSCNet. It removes semantic bias through hierarchical causal intervention to enhance the causal representation of tail classes. In the second stage, it adaptively calibrates logit bias through counterfactual intervention.

Long-tail image classification in ViT can be improved using two main approaches: parameter-efficient fine-tuning strategies and information enhancement. The former method [Li *et al.*, 2024a] mainly leverages pre-trained knowledge to enhance the generalization of tail classes. LPT [Dong *et al.*, 2023] designs visual prompts for group-wise categories to improve the learning of unique representations for tail classes, while LIFT [Shi *et al.*, 2023] adopts a partial parameter fine-tuning approach to enhance the discriminative ability for tail classes. However, they struggle with tail classes that have high intra-class complexity by fine-tuning with a lim-

*Corresponding author

ited number of parameters. Information enhancement methods aim to augment the information for tail classes. VL-LTR [Tian *et al.*, 2022] leverages textual features to enhance the learning of image features, and DeiT [Rangwani *et al.*, 2024] extracts information from pre-trained CNNs through knowledge distillation. These methods often struggle to obtain accurate knowledge of tail classes, leading to semantic confusion. The key to solving these problems lies in learning highly relevant visual features and mitigating spurious correlations caused by long-tailed distribution, which can be achieved through causal inference methods. However, directly applying existing causal methods [Tang *et al.*, 2020; Zhu *et al.*, 2022] to ViT fails to yield performance gains akin to CNN-based models due to they are difficult to model the spurious association between the fine-grained features and the predictions by calibrating the logits with the estimated category consistency bias. This leads to the problem that existing causal methods struggle to eliminate the spurious associations between a woman's image and related categories like "girl" or "table," with most of the misclassified tail categories being confused with similar categories, as shown in Figure 2.

To address these issues, this paper proposes a two-stage causal modeling framework by multi-scale causal intervention termed TSCNet, as shown in Figure 1. To enhance the model's fine-grained causal representation and mitigate the spurious associations on logits, we design two stages: hierarchical causal representation learning (HCRL) and counterfactual logits bias calibration (CLBC). HCRL enhances the model's fine-grained causal representation for tail classes by introducing class-independent semantic information such as background at both the patch-level and global feature-level. This enables the model to focus on class-relevant regions through hierarchical interventions. CLBC calibrates the spurious associations in label predictions caused by domain distribution from counterfactual perspective. By counterfactual generation and adaptively refining the intensity of counterfactual augmentation to construct different distributions, we effectively model category relationships and calibrate logits' bias caused by long-tailed distribution. The two-stage causal modeling method enables independent interventions on multiple biases, allowing TSCNet to maintain head class performance while improving tail class accuracy.

Experiments were conducted on two datasets, including performance comparison, ablation study, case study, and other in-depth analyses. The results confirm that TSCNet effectively enhances causal representations for tail data, while mitigating the logits bias caused by long-tailed distributions. The main contributions of this paper are:

- This paper points out that due to the different feature extraction, existing long-tailed causal methods face challenges when applied to transformer architectures.
- This paper proposes a two-stage causal framework with multi-scale interventions. To the best of our knowledge, it's the first causal framework that uses backdoor adjustment to remove various biases and is applicable to ViT.
- Experiments show that TSCNet effectively mitigates semantic and distributional biases, reducing errors in both head and tail class predictions.

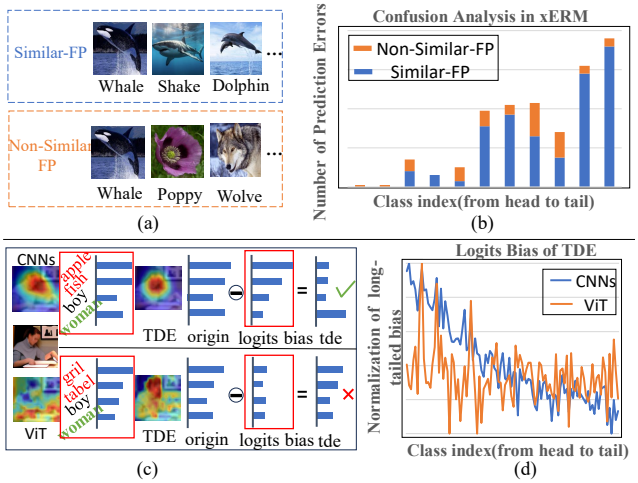

Figure 2: The example of Similar-FP and Non-similar-FP on (a), the error confusion analysis of the existing causal method xERM on ViT on (b), the bias correction mechanism of causal methods TDE on (c), the difference in counterfactual estimation logits class consistency long-tail bias between the method TDE on ViT and CNNs. on (d).

2 Related Work

2.1 Long-tailed Image Classification

Previous studies addressing the negative impacts of long-tail distributions have focused on three distinct aspects: class balancing methods, which enhance optimization for tail classes by designing resampling strategies [Cui *et al.*, 2019; Dang *et al.*, 2023], reweighting loss functions [Ren *et al.*, 2020; Zhou *et al.*, 2024], and adjusting logits [Hong *et al.*, 2021; Liao *et al.*, 2024]; data augmentation, which improve the information scarcity of tail classes through curriculum learning for image augmentation[Ahn *et al.*, 2022; Wang *et al.*, 2024a], transfer learning from head classes to tail classes [Chen and Su, 2023; Dang *et al.*, 2024a], and feature enhancement for tail classes [Qi *et al.*, 2023; Li *et al.*, 2024b]; improving training strategies [Du *et al.*, 2023; Fu *et al.*, 2025], which typically involve decoupling representation learning from classifiers [Kang *et al.*, 2019] and employing ensemble learning strategies [Zhou *et al.*, 2020; Cui *et al.*, 2022; Dang *et al.*, 2024b; Dang *et al.*, 2025] to further optimize both head and tail classes. Causal methods [Tang *et al.*, 2020; Zhu *et al.*, 2022] have shown remarkable performance improvements in CNNs by calibrating the logits bias induced by data distribution through backdoor adjustments. Transformer-based long-tailed methodologies[Xu *et al.*, 2023; Zhu *et al.*, 2024] primarily focus on fine-tuning strategies for ViT, including prompt tuning to enhance shared prompts for tail classes[Dong *et al.*, 2023; Li *et al.*, 2024a], parameter-efficient fine-tuning techniques[Shi *et al.*, 2023] to facilitate the learning of tail classes. Some methods further enhance the representation of tail classes by incorporating external knowledge through visual-language contrastive learning[Tian *et al.*, 2022] and knowledge distillation techniques[Rangwani *et al.*, 2024].

2.2 Causal Inference in Image Classification

Causal inference and counterfactual reasoning have received increasing attention in a variety of tasks in computer vision, including scene graph generation[Sun *et al.*, 2023], image recognition[Meng *et al.*, 2025; Guan *et al.*, 2023], and video analysis[Wang *et al.*, 2024d; Dang *et al.*, 2024c]. Causal methods in the field of image classification demonstrated significant performance improvements. Existing research has implemented backdoor adjustment strategies by designing causal classifiers[Liu *et al.*, 2022], using attention mechanisms[Yang *et al.*, 2023] to identify and mitigate the interference of confounding factors[Zhang *et al.*, 2024]. Moreover, prevailing front-door adjustment strategies involve designing local-global feature attention mechanisms[Wang *et al.*, 2024c] to extract distinguishable causal features. Additionally, causal invariant representation learning methods [Mao *et al.*, 2022; Liu *et al.*, 2024] utilize style generation models or Fourier transform techniques in conjunction with invariant loss functions[Lv *et al.*, 2022] to improve the identification of causal factors. However, existing causal methods in image classification are inadequate for addressing the long-tailed distribution problem, as they overlook the influence of long-tailed bias on causal graph construction and the limited effectiveness of interventions in tail classes with sparse data.

3 Problem Formulation

The long-tailed dataset is represented as $\mathcal{D} = \{x, y\}$. Let n_j denote the number of training sample for class j, and let $n = \sum_{j=1}^{C} n_j$ be the total number of training sample and $n_1 \gg n_C$. Conventional methods extract visual features : $F_v = M_v(x)$, where $M_v(\cdot)$ denotes the feature extractor. Then, predicting the category of the sample, i.e. $P = classifier(F_v)$. In the first stage, we initially extract confounder $S = [s_1, s_2, ...s_n]$ and perform causal interventions $P(Y|do(X))$ at both the token level and the global feature level, thereby obtaining a causally enhanced model $M(\cdot)$. In the second stage, we construct a counterfactual balanced distribution $\overline{x}$ through counterfactual data augmentation $F(\overline{x}, L_c^e)$ while adaptively adjusting the intensity L_c^e to perform causal interventions $P(Y|do(D))$. This process can get predictions after calibration $P_c = M(\overline{x})$.

4 Method

This study proposes a two-stage debiasing method for long-tail learning, called TSCNet. The method constructs a structured causal graph to analyze the interfering factors in the inference path. Specifically, TSCNet consists of two main stages, as shown in Figure 4. The Hierarchical Causal Representation Learning (HCRL) stage enhances the model's fine-grained causal representation through debiasing at the patch and global feature levels eliminating semantic confusion. The Counterfactual Bias Calibration (CLBC) stage utilizes counterfactual data augmentation and refinemention strategies to reduce the logits' bias caused by data distribution.

4.1 Causal View at Long-tailed Classification

We use the structural causal model to model the variable relationships of complex spatiotemporal data in long-tailed image classification tasks. As illustrated in Figure 3, it is a directed acyclic graph $\mathcal{G} = \{N, E\}$ in which the nodes N denote variables and edges E denote association between variables. The SCM $\mathcal{G}$ includes four variables: image X, semantic confounder S, data distribution confounder D and prediction Y. The correlations in graph $\mathcal{G}$ are as follows:

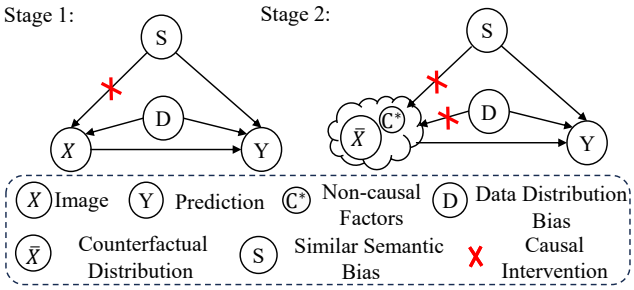

Figure 3: The Causal view of Long-tailed Image Classification.

$\boldsymbol{S \rightarrow X}$. This path indicates that similar semantic factors such as the background influence the composition of image content and tend to affect the tail classes where data is sparse.

$\boldsymbol{S \rightarrow Y}$. This path indicates that the predicted label distributions(logits) follow their own training domain prior.

$\boldsymbol{D \rightarrow X}$. This path indicates that an image is sampled according to the selected data distribution,e.g., imbalanced data distribution is prone to head classes.

$\boldsymbol{D \rightarrow Y}$. This path indicates that the predicted label distributions follow their own training domain prior.

$\boldsymbol{X \leftarrow S \rightarrow Y}, \boldsymbol{X \leftarrow D \rightarrow Y}$. This two back-door path contribute spurious correlation between X and Y, where S and B acts as confounder.

4.2 Hierarchical Causal Representation Learning

To eliminate the semantic confusion between tail classes and head classes, we propose a hierarchical causal representation learning method. By extracting class-agnostic information and performing causal interventions at both the patch and global feature levels, the method enhances the model's fine-grained representation learning for tail classes.

Mitigating the bias caused by S is to intervene on X, ensuring that class-agnostic semantic information contributes equally to the image classification. We extract class-agnostic semantic information and intervene on X:

$$P(Y \mid do(X)) = \sum_S P(Y \mid X, S) P(S) \qquad (1)$$

where $do(X)$ denotes intervene on X, the path in Fig. 3 from S to X is cut-off. Due to the inability to combine all class-agnostic information with the image, only approximate interventions are possible. We propose a hierarchical intervention strategy to strengthen the intervention for tail class and improve the fine-grained causal representation of tail classes.

Patch-level Intervention

At the patch-level intervention, we introduce class-agnostic patch information alongside the original image patches, leveraging the encoder's attention mechanism to finely uncover the causal regions within the image. Specifically, we can apply this method on CNNs to merge the original image with

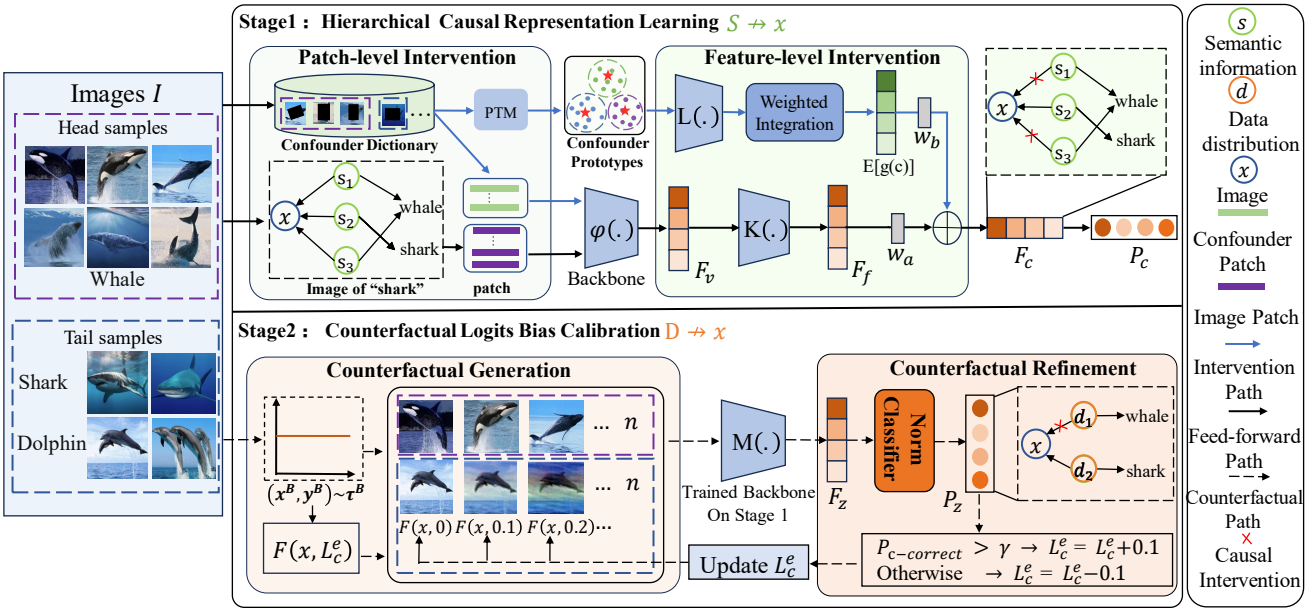

Figure 4: Illustration of the proposed TSCNet. It contains two main stages: HCRL and CLBC. The former introduces class-independent semantic information and performs backdoor adjustments to enhance the model's fine-grained causal representation for tail classes $S \nrightarrow X$. The latter generates counterfactual distribution to calibrate logits bias and model category relationships $D \nrightarrow X$.

theclass-agnostic information.

Confounder Dictionary S We extract class-irrelevant semantic information from the training set and construct a confounder dictionary. Given an image x_i, it uses off-the-shelf methods such as Grad-CAM [Selvaraju *et al.*, 2020] to detect the main subject of the image. Then, we apply an inverse transformation to obtain the class-agnostic information mask M_i. Then we paste the mask onto the original image by $M_i \odot x_i$ to obtain a class-irrelevant image s_i.Then, TSCNet constructs the confounder dictionary $S = [s_1, s_2, ..., s_n]$.

Patch-level intervention requires multiple backpropagations over all irrelevant information. To reduce training cost, we assume a uniform distribution of confounding patches and apply a random sampling strategy:

$$P(Y \mid \mathrm{do}(X)) \approx \frac{1}{N}\sum_{j=1}^{N} P(Y \mid f(X, s_j)) \approx P(Y \mid f(E(X), E(s_k))) \tag{2}$$

where s_k is randomly sampled from the confounder dictionary S, $E(\cdot)$ is the patch embeddings of the image, and $f(x, s_k)$ is the function that indicates stacking and concatenation on the sequence length dimension, the length of the sequence can be changed by random sampling from $E(s_k)$.

Then,we can get the causal-enhanced representation:

$$F_v = \varphi(f(E(X), E(s_k))) \tag{3}$$

where $\varphi(.)$ is the visual backbone.

Feature-level Intervention

To further mitigate the effects caused by class-agnostic semantics on the feature distribution and enhance the model's generalization ability for tail classes, TSCNet introduces a global feature causal intervention module.

We construct a confounder prototype dictionary $S_p = [c1, c2, ..., c_l]$ structured as an $l * d$ matrix to systematically address these feature-level factors. Where l is the dictionary size and d is the feature dimension using the pre-trained backbone. We apply k-means++ to derive c_i from the confounder dictionary S, each c_i is the average feature of its cluster. To implement the theoretical interventions in Eq 2 and reduce computation, we use Normalised Weighted Geometric Mean [Xu, 2015] to approximate the results expected from the above feature layers:

$$P(Y \mid \mathrm{do}(X)) \overset{\mathrm{NWGM}}{\approx} \sum_S P(Y \mid X, S) P(S) \tag{4}$$

We parameterize the network model to approximate the conditional probability of Eq.6, inspired by[48], as follows:

$$P(Y \mid do(X)) = W_a F_f + W_b \mathbb{E}_c[g(c)] \tag{5}$$

where $W_a \in \mathbb{R}^{d_m \times d_a}$ and $W_b \in \mathbb{R}^{d_m \times d}$ are learnable parameters. We approximate $\mathbb{E}_c[g(c)]$ as a weighted integration of all background prototypes:

$$\mathbb{E}_c[g(c)] = 1/N_i \sum_{i=1}^{N} \mu_i c_i \tag{6}$$

where μ_i represents the important weight coefficient measuring the interaction between each c_i and the feature F_v.

4.3 Counterfactual Logits Bias Calibration

Although we obtained features of causal enhancement in the first stage, the model relies on this long-tail distribution prior, leading to bias in label prediction. Therefore, we propose a model-agnostic counterfactual intervention method, which generates a balanced distribution through counterfactual aug-

mentation and refinement to adaptively calibrate the logits:

$$P(Y|do(X)) = \sum_D P(Y|X,D)P(D) \tag{7}$$

where $d = 0$ denotes the imbalanced data distribution, and $d = 1$ denotes the balanced data distribution.

Counterfactual Generation

Simple balanced sampling is not applicable as it leads to the model's overfitting to tail classes. We use Fourier transformation [Lv *et al.*, 2022] to perform counterfactual augmentation by disturbing non-causal factors to construct a counterfactual balanced distribution:

$$\mathcal{F}(x) = \mathcal{A}(x) \times e^{-j \times \mathcal{P}(x)} \tag{8}$$

where $\mathcal{A}(x)$, $\mathcal{P}(x)$ denote the amplitude and phase components respectively. We then perturb the amplitude information via linearly interpolating between the amplitude spectrums of the original image x and an image x' sampled randomly:

$$\hat{\mathcal{A}}(x^o) = (1-\lambda)\mathcal{A}(x^o) + \lambda\mathcal{A}((x')^o) \tag{9}$$

where $\lambda \sim U(0, L_c^e)$ and L_c^e controls the strength of perturbation, which adjusts the perturbation strength for class c during an epoch e. Then we can obtain the counterfactual augmented image. We can obtain the counterfactual augmented image:

$$\mathcal{F}(x^a) = \hat{\mathcal{A}}(x^o) \times e^{-j \times \mathcal{P}(x^o)}, x^a = \mathcal{F}^{-1}(\mathcal{F}(x^a)) \tag{10}$$

Counterfactual Refinement

The counterfactual refinement module adaptively adjust the strength L_c^e of counterfactual data augmentation to enable the model to progressively adjust the logits bias from easy to difficult. At epoch e, we define a computation function P_l for each class to adaptively update the perturbation strength:

$$L_c^e = P_l(\mathcal{D}_c, L_c^{e-1}, M(\cdot), \gamma) \tag{11}$$

where γ is threshold hyperparameter, $M(\cdot)$ is the model from stage 1. We can update V_{LoL} as follows:

$$P_l = L_c^{e-1} + 0.1 \quad \text{if } Acc(\mathcal{D}_c, M(\cdot)) \geq \gamma \tag{12}$$

$$V_{\text{LoL}} = L_c^{e-1} - 0.1 \quad \text{otherwise} \tag{13}$$

where Acc is a function which outputs the number of correctly predicted examples by the model f_θ.

After updating L_c^e, TSCNet control the intensity of generating the non-causal factor $\hat{\mathcal{A}}(x^o)$ and conterfactual augmented image x^a to construct a balanced dataset $\overline{x} = (x^B, y^B) \sim \tau^B$. Then, We can get $P_z = M(\overline{x})$.

4.4 Training Strategies

The training of TSCNet follows two steps: the first step is de-confounded training for HCRL. After obtaining a causal representation-enhanced model, counterfactual fine-tuning is subsequently applied for CLBC. The details are as follows:

The HCRL stage uses causal intervention modules at the patch and feature levels. The process is constrained by:

$$\mathcal{L}_{cls} = -(\sum_{i=1}^{C} y_i \log(\hat{y}_i)) \tag{14}$$

The CLBC construct and refine counterfactual distribution $\overline{x}$ to mitigate the long-tail bias. The process is constrained by:

$$\mathcal{L}_f = \mathcal{L}_{cls} + \alpha_{gf} \frac{1}{N} \sum_{i=1}^{N} \|M(x) - M(x')\|_2^2 \tag{15}$$

where α_{gf} is the weight factor, $M(.)$ is the model from step 1, x' is a counterfactually augmented sample of x.

5 Experiments

5.1 Experiment Settings

Datasets

Experiments are conducted on two datasets: CIFAR-100LT and the more challenging VireoFood-172 [Chen and Ngo, 2016] of 66,071 training and 33,154 test images.

Evaluation Protocol

For CIFAR-100LT dataset, we evaluated Top-1 accuracy under three different imbalance ratios: 100/50/10. For VireoFood-172, we evaluated Top-1 accuracy under an imbalance ratio of 50. We followed TDE [Tang *et al.*, 2020] and CMLTNet [Li *et al.*, 2024c] to test the performance of the head, middle, and tail classes in the CIFAR100 dataset and VireoFood-172 dataset.

Implementation Details

For CIFAR100-LT, we use warm-up scheduler for fair comparisons. All models were trained by using SGD optimizer with momentum μ = 0.9 and batch size 64. The learning rate was decayed by a cosine scheduler from 0.01 to 0.0 over 200 epochs for the ResNet50 and 40 epochs for the ViT and VPT . For VireoFood-172-LT, all models were trained by using Adam optimizer with momentum μ = 0.1 and batch size 64. The learning rate is chosen in the range of 1e-4 to 5e-5. The learning rate decays every 4 epochs, with each model decaying 3 times by a factor of 0.1.

5.2 Performance Comparison

We conducted a comprehensive comparison involving 3 visual modal backbones, 3 causal methods and 8 long-tailed methods: ResNet50 [He *et al.*, 2016], ViT [Alexey, 2020], VPT[Jia *et al.*, 2022], CCIM [Yang *et al.*, 2023], GOAT [Wang *et al.*, 2024b], CaDeT [Pourkeshavarz *et al.*, 2024], TDE [Tang *et al.*, 2020], xERM [Zhu *et al.*, 2022], PLOT [Zhou *et al.*, 2024], LiVT [Xu *et al.*, 2023], Gpaco [Cui *et al.*, 2023], H2T [Li *et al.*, 2024b], LPT [Dong *et al.*, 2023]. To make a fair comparison, the hyper-parameters of all models are chosen in above section. The following observations are drawn from Table 1:

- **TSCNet achieved significant improvements across different vision networks especially in the Transformers**. This is due to multi-level causal interventions that enhance the model's fine-grained causal representations, along with the introduction of a model-agnostic counterfactual bias calibration strategy.
- **TSCNet generally achieved better performance than other algorithms in both datasets.** The two-stage causal debiasing framework has been validated, demonstrating significant improvements in tail class performance while maintaining head class performance.
- **Causal methods for image classification enhance head class performance but offer limited gains for tail classes**. They enhance causal representation for data-rich head classes, but fail to provide fine-grained

Algorithms	Backbone	CIFAR100-ratio0.01			CIFAR100-ratio0.02			CIFAR100-ratio0.1			VireoFood172-ratio0.02		
		Acc@all	Acc@h	Acc@t	Acc@all	Acc@h	Acc@t	Acc@all	Acc@h	Acc@t	Acc@all	Acc@h	Acc@t
ResNet50	ResNet50	0.404	0.661	0.128	0.454	0.690	0.219	0.557	0.662	0.559	0.748	0.850	0.530
ViT	ViT	0.795	0.929	0.613	0.817	0.932	0.712	0.885	0.926	0.836	0.811	0.884	0.642
VPT	VPT	0.812	0.930	0.657	0.840	0.941	0.748	0.895	0.929	0.862	0.826	0.895	0.658
CCIM	ViT	0.779	0.934	0.567	0.829	**0.945**	0.711	0.888	**0.931**	0.841	0.826	**0.896**	0.657
CaDeT	ViT	0.788	0.933	0.597	0.823	0.936	0.703	0.887	0.928	0.842	0.812	0.887	0.649
GOAT	ViT	0.819	0.930	0.671	0.837	0.941	0.736	0.890	0.926	0.845	0.829	0.890	0.672
TDE	ResNet50	0.450	0.644	0.202	0.490	0.635	0.328	0.561	0.665	0.432	0.751	0.814	0.567
xERM	ResNet50	0.455	0.680	0.174	0.492	0.670	0.296	0.575	0.691	0.426	0.770	0.835	0.572
CUDA	ResNet50	0.431	0.639	0.197	0.472	0.612	0.255	0.565	0.671	0.431	0.764	0.823	0.550
PLOT	ResNet50	0.445	0.640	0.219	0.487	0.607	0.330	0.573	0.680	0.443	0.769	0.830	0.573
TDE	ViT	0.803	0.937	0.624	0.836	0.930	0.737	0.887	0.924	0.840	0.810	0.880	0.651
xERM	ViT	0.799	0.930	0.615	0.834	0.946	0.720	0.888	0.926	0.841	0.813	0.883	0.650
LiVT	ViT	0.807	0.921	0.674	0.823	0.923	0.732	0.885	0.924	0.847	0.834	0.873	0.752
Gpaco	ViT	0.832	0.913	0.717	0.858	0.934	0.789	0.907	0.915	0.899	0.831	0.875	0.746
H2T	ViT	0.840	0.915	0.740	0.832	0.919	0.731	0.887	0.916	0.853	0.798	0.630	0.876
LPT	VPT	0.861	0.933	0.778	0.884	0.931	0.853	0.908	0.916	0.899	0.830	0.888	0.690
TSCNet	ResNet50	0.472	0.640	0.258	0.510	0.674	0.331	0.590	0.657	0.487	0.790	0.832	0.598
TSCNet	ViT	0.860	0.932	0.778	0.877	0.937	0.819	0.905	0.927	0.885	0.847	0.885	0.750
TSCNet	VPT	**0.887**	**0.934**	**0.830**	**0.901**	0.937	**0.877**	**0.915**	0.924	**0.917**	**0.875**	0.890	**0.819**

Table 1: Performance comparison of algorithms on CIFAR100 and VireoFood-172

Models	CIFAR100-ratio0.02			VireoFood172-ratio0.02		
	Acc@all	Acc@h	Acc@t	Acc@all	Acc@h	Acc@t
ViT	0.817	0.932	0.712	0.811	0.884	0.642
+I	0.839	**0.944**	0.728	0.823	0.893	0.666
+F	0.828	0.942	0.713	0.820	0.892	0.665
+I+F	0.845	0.943	0.748	0.826	**0.893**	0.667
+I+F+C	0.864	0.933	0.805	0.839	0.878	0.741
+I+F+C+R	**0.877**	0.937	**0.819**	**0.847**	0.885	**0.750**

Table 2: Ablation study of TSCNet with ViT backbone.

causal representations for tail classes and do not address decision boundary bias in long-tail distributions.

- **Long-tail algorithms boost tail class performance but degrade head class.** Existing long-tail algorithms perform poorly on the imbalanced VireoFood-172 test set due to spurious associations, which cause a trade-off between head and tail class performance.

5.3 Ablation Study

In this section, we further studied the working mechanism of each module of TSCNet, as shown in Table 2:

- **Hierarchical causal interventions(+I+F) effectively improve the feature representation for tail classes.** Incorporating patch-level intervention (+I) and feature-level intervention (+F) can significantly improve ViT's performance for tail classes, but there is still a significant gap compared to the head classes.
- **Counterfactual generation (+C) helps further optimize the decision boundary.** Training the model with a counterfactual balanced distribution (+C) leads to a significant improvement in tail class performance. However, the uncontrollable counterfactual strength leads to performance decline in the head classes.

Setting	CIFAR100-ratio0.02			VireoFood172-ratio0.02		
	Acc@all	Acc@h	Acc@t	Acc@all	Acc@h	Acc@t
Base	0.817	0.932	0.712	0.811	0.884	0.642
Random	0.821	0.938	0.702	0.806	0.881	0.642
Zero	0.816	0.937	0.688	0.810	0.881	0.642
Average	0.820	0.938	0.703	0.810	0.887	0.637
Confounder	**0.828**	**0.942**	**0.713**	**0.826**	**0.896**	**0.657**

Table 3: The results on different versions of dictionary in Feature-level Invention model. Random, Zero, Average, and Confounder Dictionary use random, zero, average features of the image, and the average features of the class-independent factors as confounders.

- **Counterfactual refinement (+R) enhances learning for both head and tail classes.** By adopting a difficulty-progressive refining strategy (+R), it simulates different data distributions to finely model the spurious correlations between head and tail logits.

5.4 In-depth Analyses

Effectiveness Analyses of the Dictionary S_p

As delineated in Table 3, we scrutinized the efficacy of the class-non confounder dictionary S_p within the BD module. Experimental results show that replacing S_p with a random dictionary or a zero dictionary significantly deteriorates performance. Using average Image features as a confounder dictionary is less effective than class-agnostic confounder features, indicating that random dictionaries and class averages, among others, are insufficient as confounder.

Effect of the counterfactual enhancement parameter L_c^e

As shown in Figure 5, compared to the multiple fixed parameters (0.36, 0.64), the adaptive adjustment of the enhancement parameter L_c^e achieves performance improvements on most categories of the VireoFood-172 dataset. This effectively demonstrates that the designed adaptive enhancement parameter L_c^e can progressively increase the counterfactual

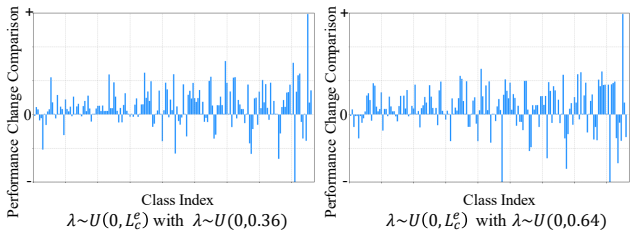

Figure 5: Comparison of our adaptive adjustment of the enhancement parameter L_c^e with using fixed parameters 0.36 and 0.64 in terms of performance, positive values indicate that our parameter adjustment performs better in this category.

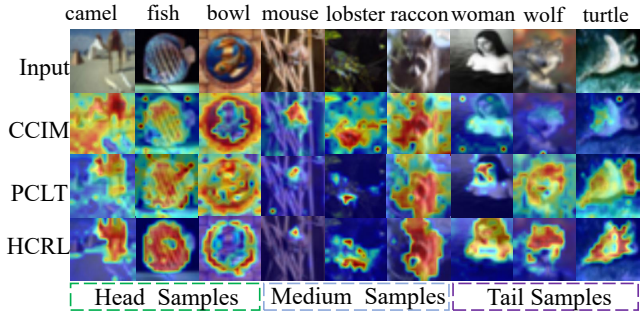

Figure 6: Visualization of attention on sampled image from the CIFAR100-LT. Four lines respectively represent the input, the CCIM, the Patch-level Intervention (PCLT) and HCRL.

data strength from easy to difficult and assign appropriate enhancement intensity to each class.

5.5 Case Study

Visualization of the Causal Representation by HCRL

To assess the efficacy of HCRL, we conducted a comparative evaluation against CCIM [Liu *et al.*, 2022] on the CIFAR100 validation set, emphasizing causal visual information in long-tail data using GradCAM [Selvaraju *et al.*, 2020] heatmaps as shown in Figure 6. CCIM demonstrated weak causal perception for tail-class images such as "wolf" and "woman." In contrast, the proposed patch-level intervention method (PCLT) enriched fine-grained causal representations, enabling the model to capture intricate features, such as facial details in "woman" and "mouse," as well as edge features in "flowerpot," effectively mitigating the interference of irrelevant information. Furthermore, HCRL improved the model's attention, enhancing its intervention on tail-class.

Visualization of the Decision Boundary by CLBC

We utilized tSNE visualization to illustrate the feature distribution and decision boundaries of three commonly confused head-tail categories in the CIFAR100 dataset, as shown in Figure 7. We observed a clearly separable boundary for the three categories in the feature space through HCRL. However, due to the confounding effect of the data distribution, a significant number of tail-class samples were misclassified as head-class samples. By incorporating the CLBC module and constructing a counterfactual balanced distribution, the model's bias towards tail classes at the decision boundary was

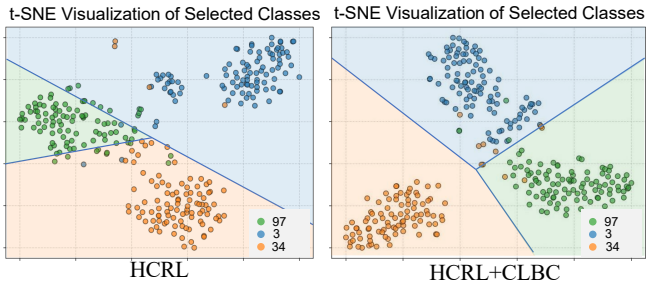

Figure 7: Visualization results of HCRL and HCRL+CLBC on the CIFAR100 dataset. The regions in different colors represent the predictions for the corresponding color categories.

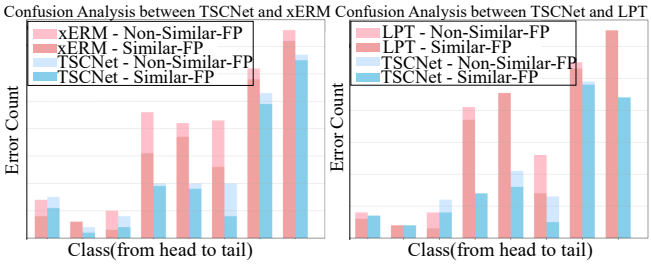

Figure 8: The error confusion analysis of TSCNet, and the comparison of xERM and LPT in terms of similar-FP and non-similar-FP in CIFAR100.

effectively mitigated. Notably, the CLBC module further enhanced the model's representation of long-tail data.

Error confusion analysis of TSCNet

We compared the error confusions of different models with our method, as shown in Figure 8. The error confusions were categorized into two types: similar-class confusion and non-similar-class confusion. The results demonstrate that TSCNet effectively alleviates error confusions in similar classes by leveraging fine-grained causal representations and bias calibration, particularly in the middle and tail classes, with a more significant improvement observed in the middle classes. In contrast, xERM and LPT faces challenges in modeling the relationship between feature regions in ViT and model predictions, leading to a concentration of errors in similar classes.

6 Conclusion

To effectively mitigate the biases induced by long-tail distributions and tackle the challenges associated with applying existing causal methods to ViT, this paper introduces TSCNet. It strengthens the model's fine-grained causal representation through hierarchical causal representation learning. Furthermore, TSCNet employs a model-agnostic counterfactual logits bias calibration stage to adaptively eliminate the prediction biases caused by long-tail distributions. Experimental results indicate that the synergistic interaction of the two stages significantly enhances long-tailed image classification performance across various backbones. Future work will focus on exploring the use of causal methods in large models to further improve long-tail performance.

Acknowledgments

This work is supported in part by the Shandong Province Excellent Young Scientists Fund Program (Overseas) (Grant no. 2022HWYQ-048).

References

[Ahn *et al.*, 2022] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. Cuda: Curriculum of data augmentation for long-tailed recognition. In *ICLR*, 2022.

[Alexey, 2020] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[Chen and Ngo, 2016] Jingjing Chen and Chong-Wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *MM*, pages 32–41, 2016.

[Chen and Su, 2023] Jiahao Chen and Bing Su. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In *CVPR*, pages 19978–19987, 2023.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019.

[Cui *et al.*, 2022] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *TPAMI*, 45(3):3695–3706, 2022.

[Cui *et al.*, 2023] Jiequan Cui, Zhisheng Zhong, Zhuotao Tian, Shu Liu, Bei Yu, and Jiaya Jia. Generalized parametric contrastive learning. *TPAMI*, 2023.

[Dang *et al.*, 2023] Jisheng Dang, Huicheng Zheng, Jinming Lai, Xu Yan, and Yulan Guo. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *TIP*, 32:3924–3938, 2023.

[Dang *et al.*, 2024a] Jisheng Dang, Huicheng Zheng, Bimei Wang, Longguang Wang, and Yulan Guo. Temporo-spatial parallel sparse memory networks for efficient video object segmentation. *T-ITS*, 25(11):17291–17304, 2024.

[Dang *et al.*, 2024b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, and Yulan Guo. Unified spatio-temporal dynamic routing for efficient video object segmentation. *T-ITS*, 25(5):4512–4526, 2024.

[Dang *et al.*, 2024c] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, and Yulan Guo. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *TIP*, 33:4853–4866, 2024.

[Dang *et al.*, 2025] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, Qingyong Hu, and Yulan Guo. Adaptive sparse memory networks for efficient and robust video object segmentation. *TNNLS*, 36(2):3820–3833, 2025.

[Dong *et al.*, 2023] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: Long-tailed prompt tuning for image classification. In *ICLR*, 2023.

[Du *et al.*, 2023] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *CVPR*, pages 15814–15823, 2023.

[Fu *et al.*, 2025] Lele Fu, Sheng Huang, Yanyi Lai, Chuanfu Zhang, Hong-Ning Dai, Zibin Zheng, and Chuan Chen. Federated domain-independent prototype learning with alignments of representation and parameter spaces for feature shift. *TMC*, pages 1–16, 2025.

[Guan *et al.*, 2023] Qing-Ling Guan, Yuze Zheng, Lei Meng, Li-Quan Dong, and Qun Hao. Improving the generalization of visual classification models across iot cameras via cross-modal inference and fusion. *IoT*, 10(18):15835–15846, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hong *et al.*, 2021] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, pages 6626–6636, 2021.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022.

[Kang *et al.*, 2019] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2019.

[Li *et al.*, 2024a] Mengke Li, Ye Liu, Yang Lu, Yiqun Zhang, Yiu-ming Cheung, and Hui Huang. Improving visual prompt tuning by gaussian neighborhood minimization for long-tailed visual recognition. *NeurIPS*, 37:103985–104009, 2024.

[Li *et al.*, 2024b] Mengke Li, HU Zhikai, Yang Lu, Weichao Lan, Yiu-ming Cheung, and Hui Huang. Feature fusion from head to tail for long-tailed visual recognition. In *AAAI*, volume 38, pages 13581–13589, 2024.

[Li *et al.*, 2024c] Xiangxian Li, Yuze Zheng, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. Cross-modal learning using privileged information for long-tailed image classification. *CVM*, pages 1–12, 2024.

[Liao *et al.*, 2024] Tianchi Liao, Lele Fu, Jialong Chen, Zhen Wang, Zibin Zheng, and Chuan Chen. A swiss army knife for heterogeneous federated learning: Flexible coupling via trace norm. In *NeurIPS*, volume 37, pages 139886–139911, 2024.

[Liu *et al.*, 2022] Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, and Tao Yang. Contextual debiasing for visual recognition with causal mechanisms. In *CVPR*, pages 12755–12765, 2022.

[Liu *et al.*, 2024] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. Causality-inspired invariant

representation learning for text-based person retrieval. In *AAAI*, volume 38, pages 14052–14060, 2024.

[Lv *et al.*, 2022] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, pages 8046–8056, 2022.

[Mao *et al.*, 2022] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *CVPR*, pages 7521–7531, 2022.

[Meng *et al.*, 2025] Lei Meng, Xiangxian Li, Xiaoshuo Yan, Haokai Ma, Zhuang Qi, Wei Wu, and Xiangxu Meng. Causal inference over visual-semantic-aligned graph for image classification. In *AAAI*, volume 39, pages 19449–19457, 2025.

[Pourkeshavarz *et al.*, 2024] Mozhgan Pourkeshavarz, Junrui Zhang, and Amir Rasouli. Cadet: a causal disentanglement approach for robust trajectory prediction in autonomous driving. In *CVPR*, pages 14874–14884, 2024.

[Qi *et al.*, 2023] Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. Cross-silo prototypical calibration for federated learning with non-iid data. In *MM*, pages 3099–3107, 2023.

[Rangwani *et al.*, 2024] Harsh Rangwani, Pradipto Mondal, Mayank Mishra, Ashish Ramayee Asokan, and R Venkatesh Babu. Deit-lt: Distillation strikes back for vision transformer training on long-tailed datasets. In *CVPR*, pages 23396–23406, 2024.

[Ren *et al.*, 2020] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 33:4175–4186, 2020.

[Selvaraju *et al.*, 2020] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *IJCV*, 128:336–359, 2020.

[Shi *et al.*, 2023] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Xin-Yan Han, Jie-Jing Shao, and Yu-Feng Li. Parameter-efficient long-tailed recognition. *arXiv preprint arXiv:2309.10019*, 2023.

[Sun *et al.*, 2023] Shuzhou Sun, Shuaifeng Zhi, Qing Liao, Janne Heikkilä, and Li Liu. Unbiased scene graph generation via two-stage causal modeling. *TPAMI*, 45(10):12562–12580, 2023.

[Tang *et al.*, 2020] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 33:1513–1524, 2020.

[Tian *et al.*, 2022] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *ECCV*, pages 73–91. Springer, 2022.

[Wang *et al.*, 2021] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *CVPR*, pages 943–952, 2021.

[Wang *et al.*, 2024a] Binwu Wang, Pengkun Wang, Wei Xu, Xu Wang, Yudong Zhang, Kun Wang, and Yang Wang. Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning. In *ICLR*, 2024.

[Wang *et al.*, 2024b] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *CVPR*, pages 13139–13150, 2024.

[Wang *et al.*, 2024c] Yuqing Wang, Xiangxian Li, Yannan Liu, Xiao Cao, Xiangxu Meng, and Lei Meng. Causal inference for out-of-distribution recognition via sample balancing. *CAAI Transactions on Intelligence Technology*, 9(5):1172–1184, 2024.

[Wang *et al.*, 2024d] Yuqing Wang, Lei Meng, Haokai Ma, Yuqing Wang, Haibei Huang, and Xiangxu Meng. Modeling event-level causal representation for video classification. In *MM*, pages 3936–3944, 2024.

[Xu *et al.*, 2023] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning imbalanced data with vision transformers. In *CVPR*, pages 15793–15803, 2023.

[Xu, 2015] Kelvin Xu. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[Yang *et al.*, 2023] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. Context de-confounded emotion recognition. In *CVPR*, pages 19005–19015, 2023.

[Zhang *et al.*, 2023] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *TPAMI*, 45(9):10795–10816, 2023.

[Zhang *et al.*, 2024] Youliang Zhang, Wenxuan Liu, Danni Xu, Zhuo Zhou, and Zheng Wang. Bi-causal: Group activity recognition via bidirectional causality. In *CVPR*, pages 1450–1459, 2024.

[Zhou *et al.*, 2020] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020.

[Zhou *et al.*, 2024] Zhipeng Zhou, Liu Liu, Peilin Zhao, and Wei Gong. Pareto deep long-tailed recognition: A conflict-averse solution. In *ICLR*, 2024.

[Zhu *et al.*, 2022] Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *AAAI*, volume 36, pages 3589–3597, 2022.

[Zhu *et al.*, 2024] Beier Zhu, Kaihua Tang, Qianru Sun, and Hanwang Zhang. Generalized logit adjustment: Calibrating fine-tuned models by removing label bias in foundation models. *NeurIPS*, 36, 2024.