Computational Visual Media
https://doi.org/10.1007/s41095-023-0382-0 Vol. 10, No. 5, October 2024, 981–992

Research Article

Cross-modal learning using privileged information for long-tailed image classification

Xiangxian Li[1], Yuze Zheng[1], Haokai Ma[1], Zhuang Qi[1], Xiangxu Meng[1], and Lei Meng[1] (✉)

Abstract The prevalence of long-tailed distributions in real-world data often results in classification models favoring the dominant classes, neglecting the less frequent ones. Current approaches address the issues in long-tailed image classification by rebalancing data, optimizing weights, and augmenting information. However, these methods often struggle to balance the performance between dominant and minority classes because of inadequate representation learning of the latter. To address these problems, we introduce descriptional words into images as cross-modal privileged information and propose a cross-modal enhanced method for long-tailed image classification, referred to as CMLTNet. CMLTNet improves the learning of intra-class similarity of tail-class representations by cross-modal alignment and captures the difference between the head and tail classes in semantic space by cross-modal inference. After fusing the above information, CMLTNet achieved an overall performance that was better than those of benchmark long-tailed and cross-modal learning methods on the long-tailed cross-modal datasets, NUS-WIDE and VireoFood-172. The effectiveness of the proposed modules was further studied through ablation experiments. In a case study of feature distribution, the proposed model was better in learning representations of tail classes, and in the experiments on model attention, CMLTNet has the potential to help learn some rare concepts in the tail class through mapping to the semantic space.

Keywords long-tailed classification; cross-modal learning; representation learning; privileged information

1 School of Software, Shandong University, Jinan 250101, China. E-mail: X. Li, xiangxian_lee@mail.sdu.edu.cn; Y. Zheng, zhengyuze@mail.sdu.edu.cn; H. Ma, mahaokai@mail.sdu.edu.cn; Z. Qi, z_qi@mail.sdu.edu.cn; X. Meng, mxx@sdu.edu.cn; L. Meng, lmeng@sdu.edu.cn (✉).

Manuscript received: 2023-01-11; accepted: 2023-09-29

1 Introduction

The long-tailed phenomenon in a data distribution means that most samples belong to a small number of head classes, with many tail classes occupying only a small part of the samples. Learning image classification from long-tailed data tends to cause the model to be dominated by the head class, resulting in poor accuracy on the tail. Therefore, existing studies mainly attempt to rebalance the data distribution [1, 2] and reassign the optimization weights to compensate for the tail [3–5]; however, the lack of diversity in tail-class information may necessitate a trade-off between head and tail performance. Therefore, recent studies have proposed the application of data augmentation [6–8], adversarial training [9, 10], and transfer learning [11] to supplement the information on tail classes. However, these methods in visual modality face the dilemma of adversely affecting the head or exacerbating the imbalanced situation, as they still play a role in rebalancing. Therefore, when representation learning is not sufficiently improved, problems like interference from background noise may be more serious in tail classes and can hinder solutions.

Due to the popularity of multi-modal data [12], images are usually accompanied by semantic information such as tags or description words, which make it easier for the classification model to distinguish confusing classes, as shown in Fig. 1. Therefore, cross-modal semantic information is introduced to supplement the training process, such as in learning using privileged information (LUPI) paradigm [13, 14], which has potential for improving the representation learning of the model. Studies in this area are primarily divided into cross-modal constraint and cross-modal alignment

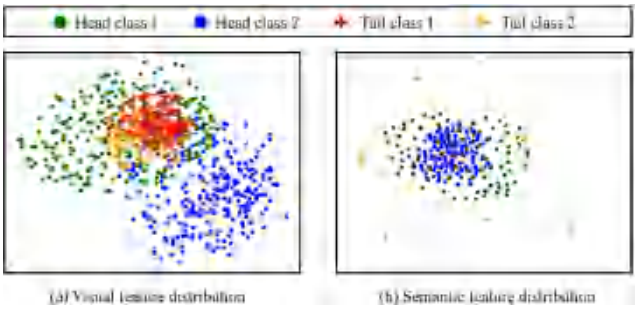

Fig. 1 Visualization of distributions. (a) In visual space, distribution of visual features among classes is messy, with tail features intermixed with the head class. (b) In semantic space, the intra-class distribution of the head is more concentrated, and the tail features are distributed in specific regions that can be clearly distinguished.

methods. Cross-modal constraint methods utilize semantic information as an additional constraint in local [15, 16] or global feature extraction [17, 18], whereas cross-modal alignment methods make the range [19, 20] or distribution [21, 22] of visual and semantic features more similar. However, existing studies achieve limited performance gains because of uncontrollable constraints and modal heterogeneity. In addition, because the long-tailed distribution can also exist in semantic space, the bias may be further exacerbated in cross-modal learning [23].

To address the aforementioned problems, we propose a cross-modal learning method, termed CMLTNet, to improve the learning of visual representations in long-tailed image classification. Through the introduction of cross-modal semantic information, the visual representations are enhanced for both the head and the tail classes. The overall idea of CMLTNet is shown in Fig. 2, which consists of three main processes, **alignment** between cross-modal information, **inference** from visual to semantic space, and the cross-modal information **fusion**. To make full use of the information in the semantic modality during training, we first propose feature-level alignment to make the cluttered visual features more similar to the focused and distinguishable semantic features. Since the alignment has limited effects due to the modal heterogeneity, in another aspect, we encourage the model to map from visual to semantic space, that is, visual–semantic inference, finding the meaningful semantic information from visual features to achieve communication between modalities. Finally, the representation learning of the model is enhanced from the fusion of distribution alignment and visual–semantic inference, which improves the intra-class similarity and inter-class discrimination learning to achieve better performance on long-tailed image classification.

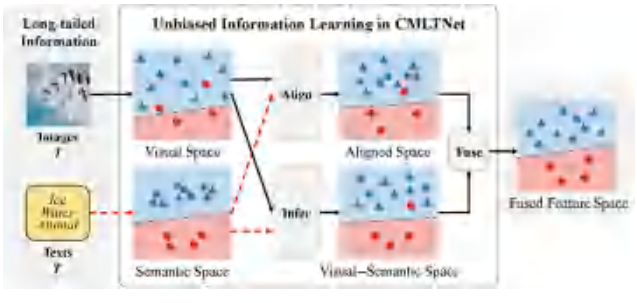

Fig. 2 CMLTNet improves representation learning in long-tailed image classification for both head (blue triangles) and tail (red circles) classes. In visual space, features are scattered, and decision-making is dominated by the head class, whereas representations learned from description words are clear in semantic space. CMLTNet encourages the model to learn from semantic information during training through the alignment of feature distribution and mapping from visual to semantic space.

In experiments, the effectiveness of CMLTNet was demonstrated on two cross-modal long-tailed datasets, NUS-WIDE and VireoFood-172. The experimental results show that the proposed method can effectively enhance the prediction of the entire class, especially the tail classes, without head loss. In the ablation study, we analyzed the effectiveness of cross-modal alignment and inference, and the effects of different fusion strategies. The enhancement effect of CMLTNet on representation learning and the improvement of model attention to long-tailed data in cross-modal learning were further demonstrated through case studies.

The main contributions of this study are as follows:

- Incorporating cross-modal privilege information into long-tailed image classification was explored, thereby proposing CMLTNet to effectively improve representation learning and alleviate the problem of long-tailed image classification in vision, which is pioneering work.
- The effectiveness of cross-modal learning methods on long-tailed image classification was analyzed, thereby proposing a model-agnostic "alignment–inference–fusion" framework with the advantages demonstrated in representation learning and filtering visual noise.

2 Related works

2.1 Long-tailed image classification

Existing approaches for long-tailed image classification typically focus on addressing imbalanced data distributions and biased optimization weights.

Concerning data distribution rebalancing, various resampling techniques such as reverse sampling [1] and square-root sampling [24] are employed to redistribute the weight of the class samples. Nevertheless, it is worth noting that resampling can sometimes lead to overfitting of tail classes. Consequently, resampling is typically utilized as a strategy for classifier retraining in decoupling training [2] or as a method to balance sampling branches in dual-branch learning [1]. By contrast, the loss weight adjustment method aims to mitigate bias by compensating for tail classes, which involves adjusting loss weights based on the training sample distribution [3, 4, 25] or model predictions [26, 27].

Although these approaches have alleviated bias in model decision-making to a certain extent, they often fall short in providing sufficient information to enhance model learning of visual representations. This limitation makes it challenging to achieve overall performance improvement. Consequently, methods that leverage data augmentation to diversify tail samples [6] have been proposed. These techniques may involve the use of head samples to augment tail samples at either the sample or feature level [28, 29]. Methods rooted in contrastive learning enhance long-tailed classification performance by improving the selection of positive and negative samples and optimizing contrastive loss [30, 31]. Furthermore, adversarial training can effectively distinguish between head and tail samples by introducing perturbations [9, 10]. Apart from these methods, there are also approaches that leverage multi-expert mixture [32], causal inference [33], and biased optimization in long-tailed image classification.

2.2 Cross-modal learning for image classification

Incorporating cross-modal semantics into the pre-training of visual models has been demonstrated to significantly improve model generalization capabilities for downstream tasks [34]. However, a noteworthy limitation is the inability to effectively address modal heterogeneity [35], which leads to an extensive reliance on substantial training data to build visual–semantic relationships.

In real-world scenarios, the availability of cross-modal text data for images is often limited. Therefore, it is essential to address the challenge of modal heterogeneity within the context of LUPI [13, 14]. Two primary approaches have emerged to address this issue: implicit cross-modal constraints and explicit cross-modal alignment. The cross-modal constraint method leverages semantic information as a predictable label to constrain the semantic information prediction of the local visual area [15, 16] or the entire image [17, 18, 36]. The constraint introduces an additional regularization component along with image classification loss. By contrast, the cross-modal alignment approach explicitly brings visual and semantic features closer to each other within a shared space. This is primarily achieved through the use of similarity loss, such as between visual and semantic features [19] or the covariance matrices of the features [20]. These techniques guide the model to effectively filter the noise in visual feature space [21, 22].

3 Method

3.1 Overview

To fully utilize cross-modal information in the training phase and improve the learning of long-tailed images, we constructed an alignment–inference–fusion learning framework in CMLTNet, as shown in Fig. 3.

In this process, first, in the visual representation enhancement module, cross-modal alignment is used to bring the visual feature distribution close to the semantic distribution, to improve the learning of intra-class representation as modal heterogeneity limits the effects of alignment. In the cross-modal representation inference module, semantic information is used as a constraint to guide the mapping of visual features to semantic space, to promote effective learning of semantically meaningful visual–semantic knowledge, and reduce the inter-class confusion caused by visual noise. Finally, in the cross-modal information fusion module, through the fusion of the above features learned from different channels, CMLTNet obtains debiased information and thus improves the overall effect of long-tailed image classification.

3.2 Visual representation enhancement module

As mentioned in Section 3.1, the main objective of the visual representation enhancement module is to make the extracted visual features closer to the semantic features at the distribution level during classification. For input images $\mathcal{V} = \{v_i | i = 1, 2, \cdots, N\}$ and the corresponding description words $\mathcal{S} = \{s_i | i =$

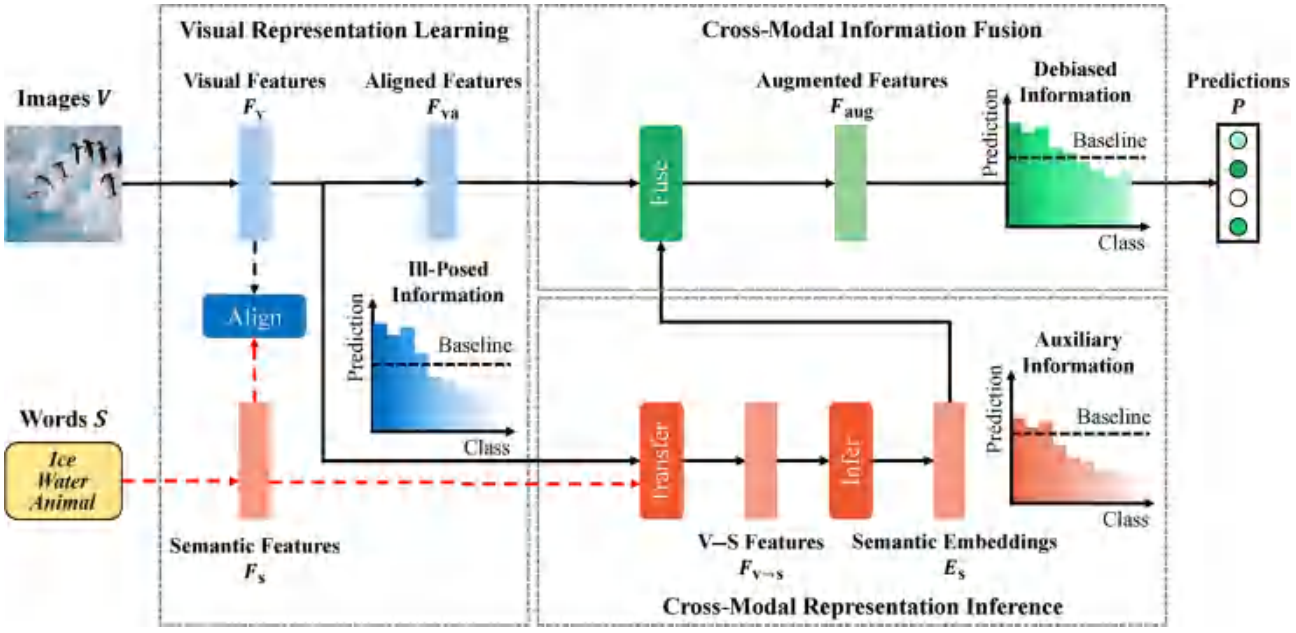

Fig. 3 Schematic diagram of CMLTNet. Description words $\mathcal{S}$ are introduced for images $\mathcal{V}$ in the training phase as cross-modal semantic features $\boldsymbol{F}_{\mathrm{s}}$ to help (i) alignment learning from visual features $\boldsymbol{F}_{\mathrm{v}}$ to visually aligned features $\boldsymbol{F}_{\mathrm{va}}$ in a shared space; (ii) better transfer visual features $\boldsymbol{F}_{\mathrm{v}}$ to semantic space and infer semantic embeddings $\boldsymbol{E}_{\mathrm{s}}$ through words $\mathcal{S}$. Finally the learned features $\boldsymbol{F}_{\mathrm{va}}$ and embeddings are fused as augmented features $\boldsymbol{F}_{\mathrm{aug}}$. Through the training of the CMLTNet framework, the bias information in learning long-tailed images is alleviated, thereby improving the image classification ability in the test.

$1, 2, \cdots, N\}$, the model first extracts visual features $\boldsymbol{F}_{\mathrm{v}} = \rho_{\mathrm{v}}(\mathcal{V})$ and semantic features $\boldsymbol{F}_{\mathrm{s}} = \rho_{\mathrm{s}}(\mathcal{S})$ through the visual encoder $\rho_{\mathrm{v}}(.)$ and semantic encoder $\rho_{\mathrm{s}}(.)$, respectively. The model then attempts to find a shared space that minimizes the distance of the distribution of $\boldsymbol{F}_{\mathrm{v}}$ and $\boldsymbol{F}_{\mathrm{s}}$.

$$\min\{\mathrm{Distance}(\alpha_{\mathrm{v}}(\boldsymbol{F}_{\mathrm{v}}), \alpha_{\mathrm{s}}(\boldsymbol{F}_{\mathrm{s}}))\} \tag{1}$$

where Distance(.) denotes measurement of distance, such as L_p Norm; α_{v} and α_{s} are shared space mappings for visual and semantic features, respectively. In CMLTNet, one-layer linear projection is followed by ReLU activation.

In CMLTNet, the aligned features are mapped by shared space $\boldsymbol{F}_{\mathrm{va}} = \alpha_{\mathrm{v}}(\boldsymbol{F}_{\mathrm{v}})$ and $\boldsymbol{F}_{\mathrm{sa}} = \alpha_{\mathrm{s}}(\boldsymbol{F}_{\mathrm{s}})$ achieving the goal of Eq. (1) through KL-divergence, by making visual features closer to semantic features in shared space.

$$\mathcal{L}_{\mathrm{exp}} = \mathrm{KLD}(\mathrm{Softmax}(\boldsymbol{F}_{\mathrm{va}}), \mathrm{Softmax}(\boldsymbol{F}_{\mathrm{sa}})) \tag{2}$$

In the above process, the visual and semantic features are mapped into a shared space to generate alignment features $\boldsymbol{F}_{\mathrm{va}} = \alpha_{\mathrm{v}}(\boldsymbol{F}_{\mathrm{v}})$ and $\boldsymbol{F}_{\mathrm{sa}} = \alpha_{\mathrm{s}}(\boldsymbol{F}_{\mathrm{s}})$. By imposing classification constraints, the features of the two modalities are further optimized to improve classification, thereby forming implicit constraints as Eq. (3):

$$\mathcal{L}_{\mathrm{imp}} = \mathcal{L}_{\mathrm{cls}}(f(\boldsymbol{F}_{\mathrm{va}}), \mathcal{C}) + \mathcal{L}_{\mathrm{cls}}(f(\boldsymbol{F}_{\mathrm{sa}}), \mathcal{C}) \tag{3}$$

where $\mathcal{L}_{\mathrm{cls}}$ is the classification loss, which can be cross-entropy in the single-label classification task or binary cross-entropy in multi-label classification; $\mathcal{C}$ denotes the labels of the samples; and $f(.)$ is the linear classifier for both visual and semantic features.

3.3 Cross-modal representation learning module

Visual representation is enhanced by alignment; however, modal heterogeneity limits the effects of the alignment. Thus, visual noise and error propagation remain serious. Therefore, we designed a cross-modal representation learning method to infer from the visual modality, semantic modality features.

To extract semantically meaningful visual information, a cross-modal transfer mapping is required to convert visual features into visual-semantic features, that is, $\boldsymbol{F}_{\mathrm{v}\to\mathrm{s}} = \mathrm{Trans}(\boldsymbol{F}_{\mathrm{v}})$ which closely corresponds to the description words $\mathcal{S}$. The target of cross-modal inference is

$$\min\{\mathrm{Error}(\mathrm{g}(\boldsymbol{F}_{\mathrm{v}\to\mathrm{s}}), \mathcal{S})\} \tag{4}$$

where Error(.) denotes the error in word predictions, and g(.) is the predicted mapping of words. In CMLTNet, we applied two blocks of linear projections followed by LeakyReLU activation as the modal-transfer mapping Trans(.); g(.) denotes the one-layer

linear projection.

To achieve the goal of Eq. (4), semantic words $\mathcal{S}$ are used as targets for word predictions, and semantic features $\boldsymbol{F}_{\mathrm{s}}$ are also used to improve the cross-modal transfer mapping:

$$\mathcal{L}_{\mathrm{tran}} = \mathrm{BCE}(\mathrm{g}(\boldsymbol{F}_{\mathrm{v}\to\mathrm{s}}), \mathcal{S}) + \beta_{\mathrm{t}} \cdot \mathrm{MSE}(\boldsymbol{F}_{\mathrm{v}\to\mathrm{s}}, \boldsymbol{F}_{\mathrm{s}}) \quad (5)$$

where $\mathrm{BCE}(.,.)$ is the binary cross-entropy loss; $\mathrm{MSE}(.,.)$ is the mean squared error loss; and β_{t} is the coefficient of transfer loss with the range specified in Section 4.2.2.

The semantic predictions $\boldsymbol{P}_{\mathrm{v}\to\mathrm{s}} = \mathrm{g}(\boldsymbol{F}_{\mathrm{v}\to\mathrm{s}})$ contain the word probabilities for given images, whereby this information is used to enhance the representations learned in visual space. CMLTNet encodes $\boldsymbol{P}_{\mathrm{v}\to\mathrm{s}}$ as embeddings:

$$\boldsymbol{E}_{\mathrm{s}} = \theta(\mathrm{Emb}(\mathrm{Topk}(\boldsymbol{P}_{\mathrm{v}\to\mathrm{s}}))) \quad (6)$$

where $\mathrm{Topk}(.)$ is the operation to choose the top-k predicted words and $\mathrm{Emb}(.)$ is the word-embedding layer; $\theta(.)$ denotes the word-embedding fusion operation, which is expressed as an average of embeddings or linear fusion of embeddings. The effects of different $\theta(.)$ are shown in Section 4.4.

The process of learning embeddings is constrained by the class prediction loss:

$$\mathcal{L}_{\mathrm{embed}} = \mathcal{L}_{\mathrm{cls}}(f_{\mathrm{e}}(\boldsymbol{E}_{\mathrm{s}}), \mathcal{C}) \quad (7)$$

where $f_{\mathrm{e}}(.)$ is the class mapping of semantic embeddings.

Therefore, the overall cross-modal inference loss is

$$\mathcal{L}_{\mathrm{infer}} = \mathcal{L}_{\mathrm{tran}} + \mathcal{L}_{\mathrm{embed}} \quad (8)$$

3.4 Cross-modal information fusion module

Representations of visual features are strengthened by alignment; however, there is still an ill-posed gap between the head and tail classes. After cross-modal inference, visual noise is filtered, with the loss of information resulting in a drop in performance. Therefore, we propose fusing the two parts of the features to combine the advantages of the two modalities.

$$\boldsymbol{F}_{\mathrm{aug}} = \mathrm{Fusion}(\phi(\boldsymbol{F}_{\mathrm{va}}), \phi(\boldsymbol{E}_{\mathrm{s}})) \quad (9)$$

where $\mathrm{Fusion}(.,.)$ is a feature-level operation, such as feature concatenation, add, min, and max operations, and $\phi(.)$ is a linear layer followed by LeakyReLU activation.

A classification constraint is applied to the fusion:

$$\mathcal{L}_{\mathrm{fusion}} = \mathcal{L}_{\mathrm{cls}}(f_{\mathrm{f}}(\boldsymbol{F}_{\mathrm{aug}}), \mathcal{C}) \quad (10)$$

where $\mathcal{L}_{\mathrm{cls}}$ is the CE loss for single-label classification or BCE loss for multi-label classification, and $f_{\mathrm{f}}(.)$ is a feature fused to the class mapping.

3.5 Training strategy

3.5.1 Multi-stage training

To improve the training efficiency, the training of CMLTNet can be divided into the following stages according to the aforementioned process:

- **Stage 1**: Training the visual encoder $\rho_{\mathrm{v}}(.)$, semantic encoder $\rho_{\mathrm{s}}(.)$, and shared space mapping net in visual and semantic modality, i.e., $\alpha_{\mathrm{v}}(.)$ and $\alpha_{\mathrm{s}}(.)$. During training, the explicit alignment constraint $\mathcal{L}_{\mathrm{exp}}$ is combined with implicit classification constraint $\mathcal{L}_{\mathrm{imp}}$, using an adjustable factor γ_{imp} on $\mathcal{L}_{\mathrm{imp}}$. Thus, the loss function of training Stage 1 is $\mathcal{L}_{\mathrm{s1}} = \mathcal{L}_{\mathrm{exp}} + \gamma_{\mathrm{imp}} \cdot \mathcal{L}_{\mathrm{imp}}$.
- **Stage 2**: Freezing the parameters of the networks trained in Stage 1, and training the transfer network $\mathrm{Trans}(.)$, semantic prediction mapping $\mathrm{g}(.)$, and embedding $\mathrm{Emb}(.)$. The above process is constrained by the loss of Stage 2: $\mathcal{L}_{\mathrm{s2}} = \mathcal{L}_{\mathrm{infer}}$.
- **Stage 3**: Freezing the networks of Stage 1 and Stage 2, in addition to training the linear net $\phi(.)$ and class mapping $f_{\mathrm{f}}(.)$ during fusion under the constraints of $\mathcal{L}_{\mathrm{s3}} = \mathcal{L}_{\mathrm{fusion}}$.

3.5.2 One-stage training

CMLTNet can also be used end-to-end in one stage by combining the above losses; however, in this case, the parameter adjustment needs to be fine-tuned, as discussed further in Section 4.2.2:

$$\mathcal{L} = \gamma_{\mathrm{imp}} \cdot \mathcal{L}_{\mathrm{imp}} + \mathcal{L}_{\mathrm{exp}} + \gamma_{\mathrm{t}} \cdot \mathcal{L}_{\mathrm{infer}} + \mathcal{L}_{\mathrm{fusion}} \quad (11)$$

where γ_{imp} and γ_{t} are the weight factors of losses.

4 Experiments

4.1 Datasets

Experiments were conducted on two cross-modal long-tailed datasets, as shown in Table 1, where the imbalance ratio (IR) [3, 4] measures the degree of imbalance in datasets, that is, $\mathrm{IR} = \max n_c / \min n_c$, where n_c means the number of samples in the class c, and the IR illustrates the ratio of the sample amount in the most and least sampled classes.

NUS-WIDE [37]: a multi-label classification dataset containing images in 81 classes. Each image

Table 1 Statistical details of NUS-WIDE and VireoFood-172 with # indicating the categories

Dataset	#Classes	#Words	IR (train)	IR (test)
NUS-WIDE	81	1000	1083.62	1465.70
VireoFood-172	172	353	5.57	5.50

corresponds to several texts, and the total number of word classes is 1000. Following previous studies [37–39], we split the training and test sets and removed samples with missing labels or text. Finally, 203,598 samples remained, including 121,962 training samples and 81,636 testing samples, with an IR of 1083.62 in the training set and 1465.70 in the test set.

VireoFood-172 [17]: a single-label classification dataset with a total of 99,225 images corresponding to 172 categories. Each image corresponds to multiple texts, and the total number of classes is 353. Following a procedure similar to that of NUS-WIDE, there were 66,071 samples in the training set and 33,154 samples in the test set. The IRs of the training and testing sets were 5.57 and 5.50, respectively.

4.2 Experimental settings

4.2.1 Evaluation protocol

Following previous studies on multi-label long-tailed classification [40, 41], the mean average precision (mAP) was adapted to the multi-label NUS-WIDE dataset to evaluate the performance of the algorithms. We report the performances in three disjoint class subsets, divided by the frequency of occurrences in the training set, as in Ref. [42]: head classes (classes each with over 5000 occurrences), medium classes (classes each with 2000–5000 occurrences), and tail classes (classes each with less than 2000 occurrences).

The accuracy score was used to evaluate the classification performance of the algorithms on the single-label dataset VireoFood-172, as in previous studies [9, 42]. As mentioned in the protocol settings of NUS-WIDE, we also divided the classes of VireoFood-172 into three disjoint subsets: head classes (classes each with over 500 occurrences), medium classes (classes each with 300–500 occurrences), and tail classes (classes under 300 occurrences).

4.2.2 Implementation details

For general hyperparameters, the settings were as follows: batch size was fixed at 64; learning rate decay occurred at intervals of four epochs; and each model underwent three decay steps, each reducing the learning rate by 0.1, followed by an additional training epoch. The Adam optimizer was employed with weight decay options of $[1 \times 10^{-3}, 5 \times 10^{-4}, 2 \times 10^{-4}, 1 \times 10^{-4}]$. The learning rate was selected to be within the range of 5×10^{-5} to 5×10^{-3}.

For the hyperparameters used in comparative experiments of long-tailed learning methods, we selected the tunable focusing parameter γ in Focal [26] from [0.1, 0.2, 0.5, 1.0, 2.0, 5.0], the independent constant in LDAM-DRW [4] from [0.1, 0.2, 0.5, 1.0, 2.0], and the hyperparameter β in class-balanced (CB) [3] resample, reweight, and LDAM-DRW [4] from [0.9, 0.99, 0.999, 0.9999].

For the hyperparameters used in comparative experiments of cross-modal learning methods, the weights of alignment loss were selected from [0.1, 0.2, 0.5, 1.0, 1.5, 2.0] in ATNet [19], and the weights of semantic loss were selected from [0.5, 1.0, 1.5, 2.0] in ARCH-D [17], CMRR [15], and CMFL [18]. We fine-tuned CLIP [34] on the aforementioned datasets until convergence was reached by loading the pretrained model with a learning rate of 1×10^{-5} to 1×10^{-4}. The dimension of the semantic latent space was set to 2048.

As for the parameters in CMLTNet and its variants, the coefficient of losses β_{align}, β_{transfer} were selected from [0.1, 0.2, 0.5, 1.0, 1.5, 2.0]. For one-stage training, the weights of β_{t}, γ_{imp} and γ_{t} losses were chosen from [0.1, 0.5, 1.0, 2.0], and the dimension of semantic latent space was 300.

4.3 Performance comparison

Comparative experiments were conducted to verify the effectiveness of the proposed CMLTNet. Visual backbone methods included pretrained basic networks ResNet-18, ResNet-50 [43], VGG [44], two improved networks WRN [45] and WISeR [46] based on ResNet-50, and the recent Transformer-based backbone ViT [47]. Long-tailed learning methods included focal loss [26], CB [3] resample and reweight, and LDAM-DRW [4]. Cross-modal learning methods included our in-house implementation of constraint-based methods ARCH-D [17], CMRR [15], CMFL [18], and alignment-based methods ATNet [19]. The methods above use the pretrained ResNet-50 as the backbone. In addition, the CLIP [34] model pretrained on 400 million image–text pairs was included in the comparisons. From the results in Table 2, we make the following observations:

- The performance of CMLTNet is comparable or better among the comparison methods. Compared with the benchmark long-tailed learning and cross-modal learning methods, CMLTNet simultaneously improves the per-

Table 2 Comparative results on multi-label NUS-WIDE, with mAP and accuracy scores reported for single-label VireoFood-172. In the table, All, Head, Med, and Tail represent the results on the entire, head, medium, and tail classes, respectively

Method	Model	NUS-WIDE				VireoFood-172			
		All	Head	Med	Tail	All	Head	Med	Tail
Visual backbone	ResNet-18	0.421	0.681	0.508	0.332	0.782	0.784	0.785	0.767
	ResNet-50	0.444	0.692	0.536	0.357	0.817	0.817	0.824	0.798
	VGG	0.436	0.694	0.531	0.346	0.811	0.805	0.820	0.801
	WRN	0.451	0.711	0.546	0.361	0.825	0.817	0.830	0.823
	WISeR	0.451	0.711	0.544	0.362	0.828	0.832	0.829	0.819
	ViT	0.455	0.709	0.544	0.367	0.836	0.829	0.846	0.830
Long-tailed learning	Focal (ResNet-50)	0.452	0.714	**0.569**	0.356	0.821	0.821	0.827	0.801
	CB Resample (ResNet-50)	0.467	0.691	0.518	0.397	0.812	0.802	0.821	0.811
	CB Reweight (ResNet-50)	0.459	0.695	0.534	0.379	0.820	0.817	0.826	0.810
	LDAM-DRW (ResNet-50)	0.470	0.701	0.548	0.392	0.833	0.826	0.840	0.825
Cross-modal learning	ATNet (ResNet-50)	0.458	0.693	0.531	0.380	0.829	0.824	0.837	0.814
	ARCH-D (ResNet-50)	0.450	0.695	0.532	0.366	0.825	0.824	0.833	0.804
	CMRR (ResNet-50)	0.450	0.686	0.508	0.375	0.819	0.815	0.824	0.812
	CMFL (ResNet-50)	0.478	0.706	0.564	0.398	0.831	0.829	0.833	0.816
	CLIP (ResNet-50)	0.490	**0.717**	0.567	0.412	0.834	0.827	0.842	0.830
	CMLTNet (ResNet-18)	0.478	0.702	0.539	0.405	0.792	0.785	0.800	0.781
	CMLTNet (ResNet-50)	0.486	0.707	0.548	0.413	0.833	0.825	0.842	0.823
	CMLTNet (ViT)	**0.494**	0.715	0.557	**0.420**	**0.843**	**0.837**	**0.850**	**0.832**

formance of head, medium, and tail classes under the same visual backbone.

- CMLTNet achieves more stable performance gains across datasets in different domains compared with other methods. Note that on NUS-WIDE, CMLTNet using ResNet-18 outperforms most cross-modal learning methods using ResNet-50, which indicates that CMLTNet can effectively combine valuable information in different modalities.
- There is an obvious head-to-tail deviation in the visual backbones, including the conventional convolutional and transform-based networks. The improvements introduced by the enhanced backbones on the tail classes are limited to NUS-WIDE with higher IR. The performance of medium classes is higher than the head on VireoFood-172 because there are more confusing classes in the head.
- To enhance performance, long-tailed learning methods prioritize head class optimization to prevent inadvertent weakening of the head while simultaneously improving the tail. CB resample on VireoFood-172 is such an example, in that, as the tail is enhanced by 1.2%, the medium and head are weakened by 0.3% and 1.8%, respectively, resulting in an overall reduction of 0.6%. Focal loss encounters the problem of an increased head-to-tail gap on NUS-WIDE.
- Cross-modal learning methods generally improve the tail predictions, with the effect varying for different datasets. For example, since deception words form a diverse inner class on NUS-WIDE, the effectiveness of align-based ATNet is limited. On VireoFood-172, less improvement is observed on the tail with the cross-modal constraint-based methods ARCH-D and CMRR.
- CLIP demonstrates performance on par with CMLTNet, outperforming other cross-modal and long-tailed learning methods. This highlights the effectiveness of CLIP's extensive pre-training and fine-tuning. Moreover, considering CLIP's substantial data requirements, CMLTNet results indicate that addressing modal heterogeneity efficiently enhances data utilization.

4.4 Ablation study

To analyze the mechanisms behind the performance improvement of CMLTNet, we gradually added modules to the base model in the ablation studies, as shown in Table 3. The visual representation learning module incorporated feature alignment (+A), whereas the cross-modal representation inference module included cross-modal inference (+I) and

Table 3 Ablation study on CMLTNet using ResNet-50 as the base model. In the table, +A denotes cross-modal alignment; +I denotes cross-modal inference comprising two word-embedding combination methods, mean of features (M) and linear projection (L); and F denotes feature fusion

Model	NUS-WIDE				VireoFood-172			
	All	Head	Med	Tail	All	Head	Med	Tail
Base	0.444	0.692	0.536	0.357	0.817	0.817	0.824	0.798
+A	0.466	0.698	0.540	0.388	0.821	0.821	0.834	0.822
+I(M)	0.355	0.594	0.411	0.279	0.780	0.790	0.797	0.708
+I(L)	0.401	0.628	0.446	0.332	0.801	0.802	0.812	0.771
+A+I(M)+F	0.473	0.703	0.543	0.397	0.829	0.823	0.837	0.816
+A+I(L)+F	**0.486**	**0.707**	**0.548**	**0.413**	**0.833**	**0.825**	**0.842**	**0.823**

experiments in two modes: embedding averaging (+I(M)) and linear mapping fusion of embeddings (+I(L)). Finally, feature fusion (+F) was performed by combining +A and +I.

- Following alignment (+A), the model demonstrates better performance across all classes in comparison to the base model. Notably, the increase in performance for the tail class surpasses that of both the head and middle classes. For instance, in the case of NUS-WIDE, the improvements in the head, middle, and tail classes are 4%, 0.8%, and 8%, respectively. This illustrates that the incorporation of cross-modal information enhances the information augmentation for the tail class, effectively mitigating the issue of class imbalance.
- Regarding cross-modal inference (+I), this procedure encompasses the filtration of visual noise during the inference phase. While this diminishes the gap between the head and tail classes, it also carries the potential risk of discarding valuable information that may be relevant for classification. As a result, the direct prediction accuracy for cross-modal inference remains relatively modest, irrespective of whether mean +I(M) or linear projection +I(L) are employed to construct embeddings. Nevertheless, the significant semantic insights acquired can effectively complement the aligned features.
- The supplementary effects are evident when considering the impact of cross-modal fusion (+F). In comparison to aligned predictions, the performances of the head, middle, and tail classes experience additional enhancements.

4.5 In-depth analysis of fusion strategy

In this section, the fusion strategy employed in CMLTNet is discussed. Table 4 lists the performance outcomes resulting from the various fusion strategies used to augment features. Notably, we found that employing the element-wise maximum (Max) method yields the highest overall performance. Employing element-wise feature addition (Add), feature concatenation (Con), and maximum (Max) or minimum (Min), the predictions for the head, middle, and tail classes are consistently enhanced by CMLTNet. This suggests that CMLTNet effectively extracts valuable information from both visual alignment and cross-modal inferences.

4.6 Case study

4.6.1 Representation learning in feature spaces

In the above analysis, each module of CMLTNet played a positive role in alleviating the imbalance problem. In this section, we further delve into the feature space to understand how it improves representation learning, randomly choosing two confusing head and tail classes from VireoFood-172 (with an imbalance ratio of 4.8) and using t-SNE to observe the distributions in feature space.

The results are shown in Fig. 4. The features are heavily mixed in visual space because of model bias during optimization. However, in semantic space, the feature dimension is relatively low, so distinguishing the head and tail classes is easier, and the distribution

Table 4 Performance of CMLTNet using different fusion strategies. Align, Inference, and Cross-modal fusion represent the performance using aligned visual features, semantic embeddings, and fused augmented features, respectively

Class	Align	Inference	Cross-modal fusion			
			Add	Con	Max	Min
All	0.466	0.401	0.482	0.483	**0.486**	0.484
Head	0.698	0.628	**0.709**	0.707	0.707	0.707
Med	0.540	0.446	**0.550**	0.549	0.548	0.549
Tail	0.388	0.332	0.407	0.410	**0.413**	0.411

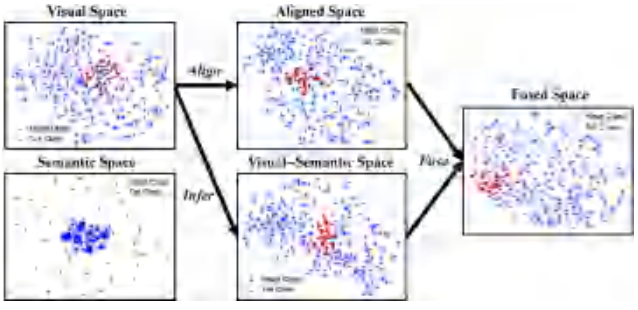

Fig. 4 t-SNE visualization of the feature distribution in the latent embedding spaces, where blue crosses represent head class samples, and red dots represent tail class samples.

of features is the aggregation of multiple clusters with distinct semantic features.

Through the alignment operation of CMLTNet, we found that the features of both the head and tail classes in semantic space tend to form small clusters, making mixed head and tail features more distinguishable. By contrast, after semantic inference, the head and tail are gathered in their respective spaces, and there is a clear demarcation between the classes. Finally, in fusion space, the features combine the characteristics of the above two spaces, using intra-class aggregation and inter-class separation simultaneously so that both the head and tail achieve better representation learning.

4.6.2 Visual attention of different features

Previous experiments demonstrated that CMLTNet improves representation learning. In this section, we further analyze whether CMLTNet learns semantically meaningful information from features in the head, middle, and tail classes using GradCAM [48] visualization, as shown in Fig. 5.

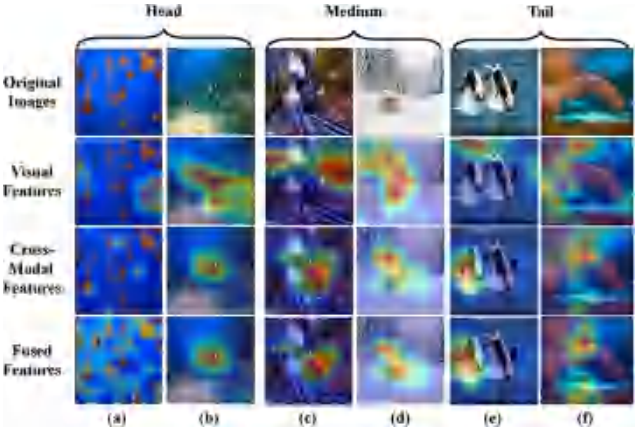

Fig. 5 Visualization of model attention. Visual, cross-modal, and fused features represent model attention in different feature spaces.

In the experiments, it was observed that visual models are easily disrupted by noise, particularly when dealing with images from underrepresented categories with limited information diversity. In terms of visual features, in the cases (c), (e), and (f), attention focuses on the background of the images. After cross-modal inference, the attention of the model is more focused on the visual modality, significantly reducing the issue of background focus in the cases (c) and (e). However, as demonstrated in the cases (a) and (d), there is also the risk of allocating more attention to the background. Therefore, in cross-modal fusion, CMLTNet combines the attention of both modalities, expanding the receptive field, thereby reducing errors caused by visual and cross-modal inference noise. Furthermore, it is observed that the visual modality pays less attention to rare concepts in tail categories, such as *rainbow* and *whale*, whereas the semantic modality learns them better. This explains the effectiveness of CMLTNet in mitigating long-tailed problems.

5 Conclusions

This study introduced CMLTNet, which enhances long-tailed classification based on cross-modal privilege information. Through heterogeneous feature alignment, cross-modal transfer, and fusion enhancement representation learning, CMLTNet strengthens the focus on minority classes, improves overall prediction ability, and provides an "alignment–inference–fusion" framework for enhancing classification using cross-modal information.

This study is a preliminary step in cross-modal long-tailed classification. In the future, we will consider enriching the diversity at the sample level using methods such as contrastive learning [49] and by introducing causal inference [33, 50] into feature learning to improve the extraction of key information and further enhance the learning of tail features.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62006141), the National Key R&D Program of China (2021YFC3300203), the Overseas Innovation Team Project of the "20 Regulations for New Universities" Funding Program of Jinan (2021GXRC073), and

the Excellent Youth Scholars Program of Shandong Province (2022HWYQ-048).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Zhou, B.; Cui, Q.; Wei, X. S.; Chen, Z. M. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9716–9725, 2020.

[2] Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In: Proceedings of the International Conference on Learning Representations, 2019.

[3] Cui, Y.; Jia, M.; Lin, T. Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9268–9277, 2019.

[4] Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In: Proceedings of the Advances in Neural Information Processing Systems, 1567–1578, 2019.

[5] Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; Jia, J. Parametric contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 715–724, 2021.

[6] Chou, H. P.; Chang, S. C.; Pan, J. Y.; Wei, W.; Juan, D. C. Remix: Rebalanced mixup. In: *Computer Vision – ECCV 2020 Workshops. Lecture Notes in Computer Science, Vol. 12540*. Bartoli, A.; Fusiello, A. Eds. Springer Cham, 95–110, 2021.

[7] Zhang, Y.; Wei, X. S.; Zhou, B.; Wu, J. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, No. 4, 3447–3455, 2021.

[8] Park, S.; Hong, Y.; Heo, B.; Yun, S.; Choi, J. Y. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6877–6886, 2022.

[9] Li, X.; Ma, H.; Meng, L.; Meng, X. Comparative study of adversarial training methods for long-tailed classification. In: Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia, 1–7, 2021.

[10] Kim, J.; Jeong, J.; Shin, J. M2m: Imbalanced classification via major-to-minor translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13893–13902, 2020.

[11] Liu, J.; Sun, Y.; Han, C.; Dou, Z.; Li, W. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2970–2979, 2020.

[12] Ma, H.; Qi, Z.; Dong, X.; Li, X.; Zheng, Y.; Meng, X.; Meng, L. Cross-modal content inference and feature enrichment for cold-start recommendation. In: Proceedings of the International Joint Conference on Neural Networks, 1–8, 2023.

[13] Vapnik, V.; Vashist, A. A new learning paradigm: Learning using privileged information. *Neural Networks* Vol. 22, Nos. 5–6, 544–557, 2009.

[14] Vapnik, V.; Izmailov, R. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research* Vol. 16, No. 61, 2023–2049, 2015.

[15] Chen, J. J.; Ngo, C. W.; Chua, T. S. Cross-modal recipe retrieval with rich food attributes. In: Proceedings of the 25th ACM International Conference on Multimedia, 1771–1779, 2017.

[16] Min, W.; Liu, L.; Luo, Z.; Jiang, S. Ingredient-guided cascaded multi-attention network for food recognition. In: Proceedings of the 27th ACM International Conference on Multimedia, 1331–1339, 2019.

[17] Chen, J.; Ngo, C. W. Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 24th ACM International Conference on Multimedia, 32–41, 2016.

[18] George, A.; Marcel, S. Cross modal focal loss for RGBD face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7882–7891, 2021.

[19] Meng, L.; Chen, L.; Yang, X.; Tao, D.; Zhang, H.; Miao, C.; Chua, T. S. Learning using privileged information for food recognition. In: Proceedings of the 27th ACM International Conference on Multimedia, 557–565, 2019.

[20] Sun, B.; Saenko, K. Deep CORAL: Correlation

alignment for deep domain adaptation. In: Proceedings of the European Conference on Computer Vision, 443–450, 2016.

[21] Li, S.; Xie, B.; Wu, J.; Zhao, Y.; Liu, C. H.; Ding, Z. Simultaneous semantic alignment network for heterogeneous domain adaptation. In: Proceedings of the 28th ACM International Conference on Multimedia, 3866–3874, 2020.

[22] Li, X.; Xu, Z.; Wei, K.; Deng, C. Generalized zero-shot learning via disentangled representation. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, No. 3, 1966–1974, 2021.

[23] Gao, J.; Chen, J.; Fu, H.; Jiang, Y. G. Dynamic mixup for multi-label long-tailed food ingredient recognition. *IEEE Transactions on Multimedia* Vol. 25, 4764–4773, 2023.

[24] Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; van der Maaten, L. Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision, 181–196, 2018.

[25] Ren, J.; Yu, C.; Sheng, S.; Ma, X.; Zhao, H.; Yi, S.; Li, H. Balanced meta-softmax for long-tailed visual recognition. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, Article No. 351, 2020.

[26] Lin, T. Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2980–2988, 2017.

[27] Wang, Y.; Gan, W.; Yang, J.; Wu, W.; Yan, J. Dynamic curriculum learning for imbalanced data classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 5017–5026, 2019.

[28] Chu, P.; Bian, X.; Liu, S.; Ling, H. Feature space augmentation for long-tailed data. In: Proceedings of the 17th European Conference on Computer Vision, 694–710, 2020.

[29] Hong, Y.; Zhang, J.; Sun, Z.; Yan, K. SAFA: Sample-adaptive feature augmentation for long-tailed image classification. In: Proceedings of the 17th European Conference on Computer Vision, 587–603, 2022.

[30] Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; Feng, J. Exploring balanced feature spaces for representation learning. In: Proceedings of the International Conference on Learning Representations, 2021.

[31] Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R.; Indyk, P.; Katabi, D. Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6918–6928, 2022.

[32] Xiang, L.; Ding, G.; Han, J. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12350*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 247–263, 2020.

[33] Tang, K.; Huang, J.; Zhang, H. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: Proceedings of the 34th Conference on Neural Information Processing Systems, 1513–1524, 2020.

[34] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 8748–8763, 2021.

[35] Meng, L.; Feng, F.; He, X.; Gao, X.; Chua, T. S. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In: Proceedings of the 28th ACM International Conference on Multimedia, 3460–3468, 2020.

[36] Jiang, S.; Min, W.; Liu, L.; Luo, Z. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing* Vol. 29, 265–276, 2020.

[37] Chua, T. S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. NUS-WIDE: A real-world web image database from National University of Singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval, Article No. 48, 2009.

[38] Tang, J.; Shu, X.; Li, Z.; Qi, G. J.; Wang, J. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 12, No. 4s, Article No. 68, 2016.

[39] Tang, J.; Shu, X.; Qi, G. J.; Li, Z.; Wang, M.; Yan, S.; Jain, R. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 8, 1662–1674, 2017.

[40] Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; Lin, D. Distribution-balanced loss for multi-label classification in long-tailed datasets. In: Proceedings of the 16th European Conference on Computer Vision, 162–178, 2020.

[41] Guo, H.; Wang, S. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15089–15098, 2021.

[42] Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S. X. Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2537–2546, 2019.

[43] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.

[44] Simonyan, K.; Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556, 2014.

[45] Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv preprint* arXiv:1605.07146, 2017.

[46] Martinel, N.; Foresti, G. L.; Micheloni, C. Wide-slice residual networks for food recognition. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 567–576, 2018.

[47] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations, 2021.

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

[49] Chen, Z.; Qi, Z.; Cao, X.; Li, X.; Meng, X.; Meng, L. Class-level structural relation modelling and smoothing for visual representation learning. *arXiv preprint* arXiv:2308.04142, 2023.

[50] Wang, Y.; Li, X.; Qi, Z.; Li, J.; Li, X.; Meng, X.; Meng, L. Meta-causal feature learning for out-of-distribution generalization. In: *Computer Vision – ECCV 2022 Workshops. Lecture Notes in Computer Science, Vol. 13806*. Karlinsky, L.; Michaeli, T.; Nishino, K. Eds. Springer Cham, 530–545, 2023.

Xiangxian Li is a Ph.D. student supervised by Prof. Xiangxu Meng and Prof. Lei Meng in the School of Software, Shandong University, China. His research interests include long-tailed classification and cross-modal learning. He received an ACM MM Student Travel Grant in 2021.

Lei Meng is professor with Shandong University, China. His research interests cover multimedia computing and its application in smart family and social governance. He has published a monograph of social media computing and more than 60 academic papers in multimedia and artificial intelligence journals and conferences. His research lab, i.e., the Lab of Multimedia Mining, Reasoning, and Application (MMRC), has been selected into the "20 New Universities" Innovation Team Program of Jinan City. He presided several national and provincial projects, including the National Key R&D Program of China and the NSFC Young Scholar Project. He is an associate editor of *Applied Soft Computing*, the executive member of CCF Multimedia Committee, and the Chairman of CCF YOCSEF Jinan.