Class-level Structural Relation Modelling and Smoothing for Visual Representation Learning

Zitan Chen
chenzt@mail.sdu.edu.cn
Shandong University

Zhuang Qi
z_qi@mail.sdu.edu.cn
Shandong University

Xiao Cao
xiaocao@u.nus.edu
National University of Singapore

Xiangxian Li
xiangxian_lee@mail.sdu.edu.cn
Shandong University

Xiangxu Meng
mxx@sdu.edu.cn
Shandong University

Lei Meng*
lmeng@sdu.edu.cn
[1]Shandong University
[2]Shandong Research Institute of
Industrial Technology

ABSTRACT

Representation learning for images has been advanced by recent progress in more complex neural models such as the Vision Transformers and new learning theories such as the structural causal models. However, these models mainly rely on the classification loss to implicitly regularize the class-level data distributions, and they may face difficulties when handling classes with diverse visual patterns. We argue that the incorporation of the structural information between data samples may improve this situation. To achieve this goal, this paper presents a framework termed **C**lass-level **S**tructural **R**elation **M**odeling and **S**moothing for Visual Representation Learning (CSRMS), which includes the Class-level Relation Modelling, Class-aware Graph Sampling, and Relational Graph-Guided Representation Learning modules to model a relational graph of the entire dataset and perform class-aware smoothing and regularization operations to alleviate the issue of intra-class visual diversity and inter-class similarity. Specifically, the Class-level Relation Modelling module uses a clustering algorithm to learn the data distributions in the feature space and identify three types of class-level sample relations for the training set; Class-aware Graph Sampling module extends typical training batch construction process with three strategies to sample dataset-level sub-graphs; and Relational Graph-Guided Representation Learning module employs a graph convolution network with knowledge-guided smoothing operations to ease the projection from different visual patterns to the same class. Experiments demonstrate the effectiveness of structured knowledge modelling for enhanced representation learning and show that CSRMS can be incorporated with any state-of-the-art visual representation learning models for performance gains. The source codes and demos have been released at https://github.com/czt117/CSRMS.

*Corresponding author.

CCS CONCEPTS

• **Computing methodologies** → **Classification and regression trees**; **Learning latent representations**; **Neural networks**.

KEYWORDS

Image Classification, Representation Learning, Relational Modelling, Curriculum Construction

1 INTRODUCTION

In recent years, deep learning has witnessed remarkable advancements across various fields, including classification[11, 19, 22, 24, 31, 47, 49], recommendation[27–30, 32], image generation[20, 21, 41, 48] federal learning[25, 36] and video analysis[7, 39, 52–54]. Currently, deep learning-based image classification methods typically involve extracting visual representations using a visual backbone network and then mapping these representations to their corresponding classes. However, the high complexity of visual information poses a challenge for the model to extract discriminative representations for each class effectively. Previous works have primarily focused on increasing the complexity of the network architecture at the model level [8, 13, 35] or introducing instance-level image transformations to enhance the model's learning capacity [10, 40, 56, 59]. Although these approaches have brought incremental improvements, the intra-class diversity problem in the image limits the models' learning of representations in each class, and the inter-class similarity also affects the decision-making of models.

To mitigate the aforementioned issues, it would be beneficial to clarify the role of samples in model training as well as the interrelation among samples of each class. This can assist the model in achieving more effective optimization. However, the exploration of sample relationships is still in its early stages. Existing methods mainly focus on exploring instance-level relations between samples, such as identifying outlier representations through clustering algorithms and performing multiple rounds of iteration for correction [51], or using attention mechanisms or graph construction methods
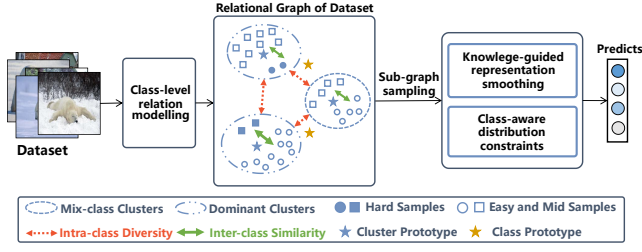
Figure 1: Illustration to CSRMS. CSRMS constructs a dataset-level relation graph using visual patterns, conducts sub-graph sampling based on relations, and performs representation learning with knowledge smoothing and class-level constraints to aggregate information.

to implicitly explore the relations between samples and enhance representation learning [14, 16, 37, 46]. Although these methods can utilize intra-class information to improve intra-class diversity, the attention to inter-class similarity is still inadequate. Contrastive learning [34, 45] is a method that explores the relationships between images by using images from the same class as positive samples and images from different classes as negative samples, to narrow the gap between images from the same class and increase the gap between images from different classes. However, how to delve deeper into the relationships and construct reasonable and efficient constraint methods is still a challenge. Existing methods often exhibit inadequate sample selection and do not pay sufficient attention to the difficulty of representation learning, which limits their ability to handle complex image data.

To address these limitations, we propose a novel approach for enhancing image representation, named CSRMS, to address the aforementioned challenges. The main concept of the approach is presented in Figure 1. The method constructs dataset-level hierarchical relation graphs by mining class-level knowledge, thereby achieving an explicit and effective sample relationship. In the Offline Learning Process (OLP) stage, CSRMS explores the distribution patterns of images and constructs the relation graph to guide the representation learning. In the Training Process (TP) stage, Class-aware Graph Sampling(CaGS) is conducted to acquire a batch-level sub-graph from the dataset-level hierarchical graph, based on sampling strategies. Moreover, curriculum construction is conducted based on easy-hard estimation to further enhance the representation learning. Finally, in Relational Graph-Guided Representation Learning (RGRL) the above relations are constrained and regularized visual representations are generated through graphical smoothing and class-level constraint. By constructing sampling strategies, "selective" sampling can be performed, reducing redundancy and enhancing the efficiency of representation learning. Through the construction of curriculum and relational graph-guided representation learning, representations can be regularized progressively, which can alleviate the problems of intra-class differences and inter-class similarities and achieve better representation learning effects as illustrated in Figure 2.

Experiments are conducted on the CIFAR10, CIFAR100, Vireo172, and NUS-WIDE datasets in terms of performance comparison. Ablation study of the key components of CSRMS and in-depth analyses of different positive and negative sampling strategies demonstrate that CaGS and RGRL can better model sample relations, construct
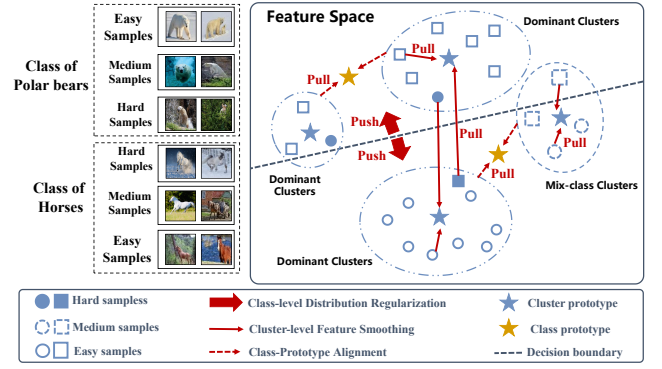

Figure 2: Illustration to strategies of CSRMS to mitigate intra-class diversity of visual patterns.

constraints of representations and then continuously improve classification accuracy. Case studies for visual distribution patterns and behaviours in different successful and failure cases of CSRMS demonstrate that CSRMS can effectively alleviate the problems of intra-class differences and inter-class similarities. To summarize, the main contributions are:

- A model-agnostic framework CSRMS is proposed, which models a relational graph and performs class-aware smoothing and regularization to address inter-class similarity and intra-class diversity, resulting in enhanced classification accuracy.
- A novel class-level strategy is proposed, which models the class-level relationships of visual representations, constructs effective sampling strategies and curriculum and utilizes explicit smoothing and constraints to enhance representation learning.
- Experiments demonstrate that the relation-based sampling and smoothing, the class-aware regularization can effectively alleviate intra-class diversity and inter-class similarity, and improve classification performance. This verifies the effectiveness of CSRMS in enhancing representation learning.

2 RELATED WORK

Sample Relationship in Image Classification. The relations between samples have been extensively studied using implicit or explicit approaches. The former [14–16] leverages cross-attention mechanisms, such as transformers or attention modules, to implicitly capture batch-level inter-sample relationships. Explicit schemes, such as Graph Neural Networks (GNN), explicitly model the relation graph among same-class samples to enhance visual representations [9, 37, 46]. However, these methods only address intra-class diversity, while neglecting inter-class similarity. Contrastive learning methods [12, 34, 45] explicitly explore sample relationships through the creation of a unified contrast constraint, pushing apart representations of different categories while drawing closer those of the same class. Despite the effectiveness of contrastive learning in mitigating both intra-class diversity and inter-class similarity, current approaches generally lack efficient sampling strategies for constructing comparison constraints, which may limit their efficiency and capability for representation learning.

Class-aware Information Guided Image Classification. The training of classification models in conventional methods heavily relies on classification loss, which can be easily disrupted by differences
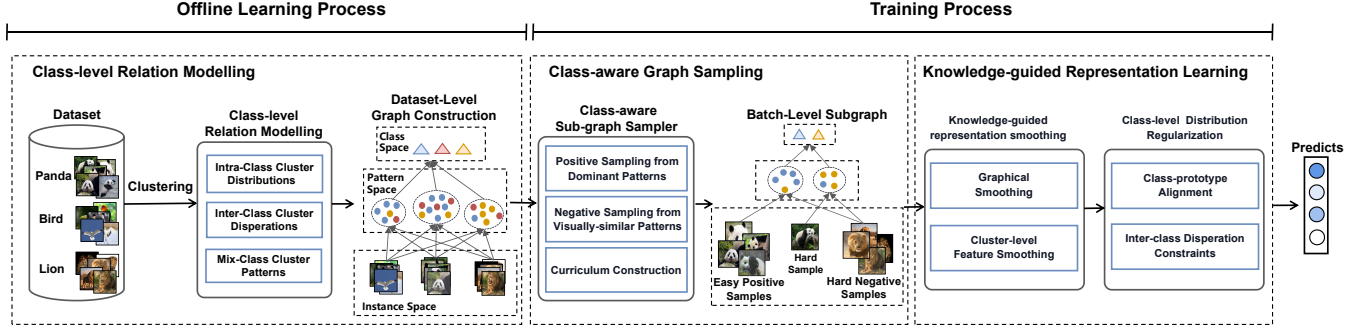
Figure 3: Illustration of CSRMS. CSRMS constructs a dataset-level relation graph by exploring the class-level relations. The Class-aware Graph Sampling module constructs a relational sub-graph based on class-aware curriculum sampling. The Relational Graph-Guided Representation Learning module acquires enhanced representations by smoothing and regularization.

within classes and similarities between classes. To address this problem, class-aware methods extract extra class-level information for enhancing image classification and approaches can be typically divided into two categories. The methods of the first type utilize the classification information generated by the model to integrate class-specific representations, such as prototypes, into representation learning [5, 55, 61], or as a constraint for improving contrastive learning optimization [2]. The methods of the second type utilize class labels as guidance, such as clustering to obtain the visual-spatial distribution of images, followed by introducing label information to constrain model learning for specific clusters [50]; other methods in this approach involve incorporating labels into the constraints of contrastive learning [57], augmenting images regarding the class [1], or using labels as more detailed guiding conditions in multi-label classification [4, 23].

3 PROBLEM FORMULATION

In conventional image classification algorithms, the dataset comprises a set of images $\mathcal{I} = \{I_i | i = 1, 2...N\}$ and corresponding labels $\mathcal{Y} = \{y_i | i = 1, 2...N\}$. Traditional deep learning-based methods usually learn a visual mapping $\mathcal{M}_v(\cdot)$ for visual inference $I_i \rightarrow \mathbf{F}_i$, where $\mathbf{F}_i$ represents the visual representation vector of image I_i. Subsequently, conventional algorithms generally train a classifier $C(\cdot)$ for predicting the class of I_i, i.e., $\mathbf{F}_i \rightarrow \mathbf{P}_i$, where $\mathbf{P}_i$ denotes the predicted scores associated with image I_i. During the training, a classification loss function, such as cross-entropy $\mathcal{L}_{ce}$, is employed to limit the discrepancy between the prediction and the label.

Different from conventional approaches, the proposed CSRMS firstly clusters visual representations $\mathbf{F}$ and incorporates class information $\mathcal{Y}$ to create a relation graph G_t of the training set offline. During the online training, CSRMS extends the typical batch construction process to sample sub-graphs G_b from G_t. By leveraging a graphical convolutional mapping $\mathcal{G}(\cdot, \cdot)$, CSRMS performs knowledge-guided graphical smoothing to form better representations $\mathbf{F}_G$, i.e., $\mathcal{G}(G_b, \mathbf{F}) \rightarrow \mathbf{F}_G$. To achieve smooth representations, CSRMS introduces cluster prototypes $w_{cluster}$ and class prototypes w_{class} for cluster-level alignment $\mathbf{F}_G \odot w_{cu} \rightarrow \mathbf{F}_u$ and class-level alignment $\mathbf{F}_u \odot p_{ca} \rightarrow \mathbf{F}_a$. Finally, CSRMS trains a classifier $C(\cdot)$ for class predictions $\mathbf{F}_a \rightarrow \mathbf{P}$, where $\mathbf{P}$ represents the predicted scores of images.

4 METHOD

The CSRMS proposes a targeted approach to alleviate the issue of intra-class visual diversity and inter-class similarity. The pipeline of CSRMS is illustrated in Figure 3. Specifically, during the Offline Learning Process, the Class-level Relation Modelling module uses a clustering algorithm to learn the data distributions in the feature space and identify three types of class-level sample relations for the training set. In the Training Process, the Class-aware Graph Sampling module extends the typical training batch construction process with three strategies to sample sub-graphs; and the Relational Graph-Guided Representation Learning module employs a graph convolution network with knowledge-guided smoothing operations to ease the projection from different visual patterns to the same class.

4.1 Class-level Relation Modeling (CRM)

To model the relationships among samples, CSRMS utilizes adaptive clustering in the CRM module to identify the aggregation of samples in the visual space, thereby capturing the distribution of patterns in the dataset. CRM then incorporates the class information of the samples to extract the "class-pattern-instance" relationships and organize them into a relational graph.

4.1.1 Definations of Class-Level Sample Relations. To clarify the relationship between samples and further strengthen their coherence, we first identify three types of bad relationships based on their feature distributions during clustering and their labels. For samples I_a and I_b, the relationship is defined as follows:

(a) **Intra-Class Diversity of Visual Patterns**: Samples within a class usually scatter across different dominant clusters, i.e., clusters that dominantly consist of samples in a certain class, in the visual space. These hard samples make it challenging to learn the class's representation and this relation is defined as:

$$R_{ID} = \{(I_a, I_b) | y_a = y_b, PC_a \neq PC_b\} \quad (1)$$

where PC_a and PC_b represent the dominant clusters that I_a and I_b belonging to, respectively.

(b) **Inter-Class Similarity of Visual Patterns**: In dominant clusters, visually similar samples belonging to different classes can make it difficult to differentiate. That can be defined as:

$$R_{IS} = \{(I_a, I_b) | y_a \neq y_b, PC_a = PC_b\} \quad (2)$$

(c) **Mix-class Cluster of Visual Patterns**: Since the variability in vision, visually-clustered clusters may not have a dominant class. Such mixed-class clusters are complex to model, as they involve the relationship between different sample classes. This relationship can be termed as:

$$R_{MC} = \{(I_a, I_b) | y_a \neq y_b, MC_a = MC_b\} \tag{3}$$

where MC_a and MC_b represent the mix-class clusters that I_a and I_b belonging to, respectively.

4.1.2 Relational Graph Construction of Dataset. To utilize and enhance the relationship outlined in Section 4.1.1, we depict the samples on a dataset-level relational graph G_t, which comprises class level, pattern level, and instance level. The connections between instance level and class level of G_t rely on the corresponding between images I and their labels Y. At the pattern level, the adaptive clustering algorithm ART [33] constructs the feature space and identifies clusters C to represent patterns based on the aggregation relationship of samples.

$$C = \text{ART}(\mathcal{M}_v(I)) \tag{4}$$

where $\mathcal{M}_v(\cdot)$ is the visual mapping, the clusters $C = \{c_1, ..., c_J\}$ and J is the number of clusters. Consequently, G_t can be represented by $\{Y, C, I\}$. For each instance I_i that is labeled y_i, we can get the pattern c_j by matching the clusters and find the connection $I_i \rightarrow c_j \rightarrow y_i$. Moreover, examples of various types of relationships can be located, by referring to the relationship specified in Section 4.1.1.

4.2 Class-aware Graph Sampling (CaGS)

In the CaGS module, CSRMS acquires a batch-level sub-graph from the dataset-level relation graph, based on sampling strategies. Moreover, curriculum construction is conducted based on easy-hard estimation to further enhance the representation learning.

4.2.1 Positive Sampling from Dominant Patterns. The relation R_{ID} indicates samples of the same class may be spread out in the feature space. To help the model learn the representations of an image I_i in a specific class y_i, CaGS uses dominant pattern sampler S_{dp} to select top-n positive samples that get the maximum distance from I_i in the largest cluster c_j dominated by class y_i:

$$\Omega_{posi} = S_{dp}(I_i, \varphi(I_i, I_i^j), n) \tag{5}$$

where Ω_{posi} is the set of positive samples, and $\varphi(\cdot)$ calculates the distance between I_i and the samples I_i^j of class i in cluster c_j. Within a cluster, the number of images of a certain class can reflect the *representativeness*, which is proportional to the number. So choosing positive samples from the biggest cluster enhances the chances of obtaining representative features of the class. Additionally, the distance condition improves the effectiveness of the selected positive samples in guiding.

4.2.2 Negative Sampling from Visually-similar Patterns. The inter-class relations R_{IS} and R_{MC} impact the classification mainly because of the visual similarity. Thus, CaGS uses a visually-similar pattern sampler S_{vp} to select top-m negative samples that get the minimum distance from I_i in the nearest cluster c_l dominated by class excluding y_i:

$$\Omega_{nega} = S_{vp}(I_i, \varphi(I_i, I_i^l), m) \tag{6}$$

where Ω_{nega} is the set of negative samples, and $\varphi(\cdot)$ calculates the distance between I_i and the samples I_i^l of class i in cluster c_l.

4.2.3 Curriculum Learning for Batch Construction. The previous section described how to collect positive and negative samples that can enhance representation learning through sample relations. However, introducing hard samples too early is not conducive to model learning. Therefore, in this section, we introduce batch construction based on curriculum learning. Unlike traditional reinforcement learning-based methods [1, 58], we quantify the *representativeness* of clusters to estimate the difficulty of the samples. To be specific, for cluster c_k, we denote the number of images of class y_j in the cluster by N_K^j and denote the number of all images in the cluster by N_K. Thus, the *representativeness* of cluster c_k for class y_j is denoted by $\frac{N_K^j}{N_K}$. And then, for an input image I_i of class y_j, its level of difficulty is mapped to:

$$f(I_i) = \begin{cases} \Omega_{easy} & \text{if } \frac{N_K^j}{N_K} > \rho_1 \\ \Omega_{medium}, & \text{if } \forall \frac{N_K^j}{N_K} < \rho_1 \\ \Omega_{hard}, & \text{if } \frac{N_K^h}{N_K} > \rho_1 \text{ and } j \neq h \end{cases} \tag{7}$$

where ρ_1 is a threshold that we set.

CSRMS takes a "decay method" to adjust the representation learning of samples. To be specific, we set different penalty coefficients for samples: λ_e for easy, λ_m for medium and λ_h for hard. α_i and α_f are utilized to regulate the coefficients, defined by:

$$\alpha_i * \lambda_e + (1 - \alpha_i) * (\alpha_f * \lambda_m + (1 - \alpha_f) * \lambda_h) = 1 \tag{8}$$

Empirically, α_i and α_f are initialized close to 1. When the loss converges to less than 0.01 and the average loss difference between two consecutive rounds of iteration is less than 0.0001, we start to decrease α_i. Similarly, when the loss converges again, we start to decrease α_f.

4.3 Relational Graph-Guided Representation Learning (RGRL)

Under the guidance of the relational graph constructed in Section 4.1 and the curriculum constructed in Section 4.2, RGRL explicitly aggregates the representations of images of the same class and constrains the distance between representations of images in different classes to better alleviate the negative impact of visual noise.

4.3.1 Cluster-aware Representation Smoothing. This module aims to perform batch-level and cluster-level representation smoothing by aggregating the information of representations, thereby alleviating the "intra-class diversity".

Graphical smoothing: In order to aggregate the information of images of the same class to complete intra-class representation smoothing, CSRMS utilizes a graphical smoothing $\mathcal{G}(\cdot)$, using batch-level subgraph as knowledge to guide the information aggregation between input image I_j and positive samples to generate enhanced representation. KNN algorithm is utilized to construct the symmetric adjacency $\hat{A}$ between the representation of I_j and the representation of positive samples.

$$F = \text{concat}(\mathcal{M}_v(I_j), \mathcal{M}_v(\Omega_{posi})) \tag{9}$$

$$F_g = \mathcal{G}\left(\hat{A}, I_j, \Omega_{posi}\right) = \text{softmax}\left(\hat{A} \, \text{ReLU}\left(\hat{A}XW^{(1)}\right)W^{(0)}\right) \quad (10)$$

where $\mathcal{M}_v(\cdot)$ denotes the visual encoder, $\hat{A}$ is symmetric adjacency, $W^{(0)} \in \mathbb{R}^{C \times H}$ denotes the input-to-hidden weight matrix for a hidden layer with H feature maps and $W^{(1)} \in \mathbb{R}^{H \times F}$ denotes the hidden-to-output weight matrix.

Cluster-level feature smoothing: In order to further aggregate representations of the same class at the cluster-level, CSRMS utilizes a cluster-level alignment to explicitly aggregate representations F_g and cluster prototype w_{cu} to generate aligned representations. Specifically, for I_i of class y_j, CSRMS aggregates the representations in the largest cluster c_j dominated by class y_j to get the cluster prototype w_{cu}^j. The alignment process is defined by:

$$F_u = \alpha_u \odot F_g + \beta_u \odot w_{cu} \quad (11)$$

where $\odot$ denotes the dot product and α_u and β_u are the coefficients.

4.3.2 Class-level Distribution Regularization. This module aims to complete the instance-level and class-level constraints by explicitly constructing dispersion loss, thereby alleviating the "inter-class similarity".

Class-level Representation Alignment: This module aims to complete representation smoothing by aggregating information at the class-level, thereby further alleviating the "intra-class diversity" of images. For I_i of class y_j, we aggregate the representations in all clusters dominated by class y_j to get the class-level prototype p_{ca}. To be specific, in Section 4.2.3, we quantify the *representativeness* of cluster c_k for class y_j: $\frac{N_K^j}{N_K}$. If $\frac{N_K^j}{N_K} > \rho_2$, we identify the cluster c_k is dominated by class y_j. The alignment process is defined by:

$$F_a = \alpha_a \odot F_u + \beta_a \odot p_{ca} \quad (12)$$

where $\odot$ denotes the dot product and α_a and β_a are the coefficients.

After that, CSRMS learns a classifier for class prediction and uses the cross-entropy loss as the supervised loss defined by:

$$\mathcal{L}_{ce} = \frac{1}{N}\sum_i \mathcal{L}_i = -\frac{1}{N}\sum_i\sum_{c=1}^{C} y_{ic}\log\left(P_{ic}\right) \quad (13)$$

where P denotes the prediction of CSRMS, y denotes the labels of images and C denotes the number of classes.

Negative sampling constraint: CSRMS constructs loss $\mathcal{L}_{nega}$ between the representation of the input image I_j and the representations of negative samples I_{nega}, explicitly constraining the distance between representations, defined by:

$$\mathcal{L}_{nega} = \sum_{i=1}^{N} -\log\left(\mu_i \frac{\theta}{\sum_{q=1}^{m}\left\|\mathcal{M}_v\left(I_i\right) - \mathcal{M}_v\left(\Omega_{nega}^q\right)\right\|_2 + \theta}\right) \quad (14)$$

where $\mathcal{M}_v(\cdot)$ denotes the visual encoder, θ denotes a fixed parameter, N and m denotes the number of images in the dataset and the number of negative samples.

Inter-class constraint: On this basis, CSRMS constructs dispersion loss between smoothed and aligned representations F_a of different categories, and explicitly constrains the distance of inter-class representations.

$$F_v^j = \sigma_v\left(\sum_1^m F_a^j\right) \quad (15)$$

Table 1: Statistics of the datasets used in the experiments.

Datasets	#Classes	#Image Size	#Training	#Testing
CIFAR10	10	32*32	50,000	10,000
CIFAR100	100	32*32	50,000	10,000
Vireo172	172	224*224	66,114	33,072
NUS-WIDE	81	224*224	121,962	81,636

$$\mathcal{L}_{inter} = \sum_{i,j}\{i = j\}\|\mathbf{0}\| + \{i \neq j\} - \log\left(\mu\frac{\theta}{\left\|F_v^i - F_v^j\right\|_2 + \theta}\right) \quad (16)$$

where σ_v denotes the aggregation process, F_a^j denotes the representation after the class-level alignment of class j and θ denotes a fixed parameter.

5 EXPERIMENTS

5.1 Experiment Settings

5.1.1 Datasets. In order to verify the effectiveness of CSRMS, we study our models based on publicly available datasets: CIFAR10, CIFAR100, Vireo172, and NUS-WIDE, and the statistics are shown in Table 1. In detail, CIFAR10 and CIFAR100 are colour image datasets that are closer to a universal object, both containing 60000 image samples. Vireo172 comprises a collection of food images that feature various Chinese dishes, containing 110,241 image samples, we refer to the original paper [3] and divide the training/testing set. NUS-WIDE is a multi-label classification dataset, originally containing 269,648 image samples, we refer to the original paper [6] and related works [43, 44] to divide the training/testing set and remove the sample without label or tag.

5.1.2 Evaluation Protocol. For the single-label datasets CIFAR10, CIFAR100 and Vireo172, we followed conventional measures of Top-1 and -5 accuracies to evaluate the classification performance. While for the NUS-WIDE multi-label dataset, we followed the original setups [6] to use Top-1 and -5 precision and recall.

5.1.3 Implementation Details. To verify the applicability of CSRMS, we investigate the performance of CSRMS on four visual backbones LeNet5 [18], ResNet18 [13], ResNet50 [13] and ViT [8] denoted as CSRMS(LeNet5), CSRMS(ResNet18), CSRMS(ResNet50) and CSRMS(ViT). The batch-size is fixed at 32. During training, we choose to use the SGD optimizer and the learning rate is selected from 5e-3 to 1e-1. The decay rate of the learning rate is selected from 0.1 and 0.5, and the decay interval is 20 epochs. The distance mentioned in 4.2 can be expressed using Euler distance or cosine similarity. The number of positive samples n and negative samples m are chosen from [5,10,20]. Regarding the threshold ρ_1 in Curriculum Construction, we conducted multiple experiments and choose it from 0.75 to 0.85 for the CIFAR100 and Vireo172 datasets, and choose it from 0.85 to 0.95 for the CIFAR10 and NUS-WIDE datasets. Moreover, for threshold ρ_2 mentioned in Class-level Representation Alignment, it is chosen from 0.5 to 0.55 for the CIFAR100 and Vireo172 datasets and chosen from 0.55 to 0.65 for the CIFAR10 and NUS-WIDE datasets based on experimental results.

Table 2: Performance comparison of algorithms. Metrics are Top-1/Top-5 Accuracy (Acc), Precision (P), and Recall (R).

Algorithm	CIFAR10		CIFAR100		Vireo172		NUS-WIDE			
	ACC@1	ACC@5	ACC@1	ACC@5	ACC@1	ACC@5	P@1	P@5	R@1	R@5
LeNet5	72.77	97.20	40.54	80.73	20.33	42.77	41.86	24.36	21.45	53.67
ResNet18	90.70	99.60	70.12	90.80	76.31	93.26	73.71	37.19	40.12	79.94
ResNet50	92.03	99.79	71.97	91.44	77.64	93.55	73.78	37.29	40.30	80.13
RE (AAAI'2020)	93.21	99.82	72.16	91.98	78.49	93.75	74.18	37.46	40.66	80.36
DLSA (ECCV'2022)	93.37	99.74	72.31	92.40	78.55	93.77	74.21	37.43	40.73	80.21
ACmix (CVPR'2022)	93.43	99.85	72.27	92.24	78.85	93.11	74.33	37.53	40.92	80.27
Resizer (ICCV'2021)	93.49	99.83	73.01	92.16	79.37	93.98	74.49	37.25	41.25	80.33
BatchFormer (CVPR'2022)	93.54	99.84	73.13	92.76	79.96	94.20	74.52	37.68	41.63	80.72
CUDA (ICLR'2023)	93.58	99.86	73.66	92.68	81.13	94.57	74.60	38.26	42.02	80.86
ViT (ICLR'2021)	98.68	99.99	81.70	96.02	85.92	96.47	79.75	39.86	44.64	86.10
CSRMS(LeNet5)	74.88	98.01	44.77	82.15	22.47	45.47	44.26	26.87	26.37	59.33
CSRMS(ResNet18)	94.62	99.99	74.57	94.47	82.60	95.92	75.12	39.42	42.25	81.17
CSRMS(ResNet50)	95.48	99.99	76.13	95.51	84.72	96.87	75.33	40.12	42.41	81.20
CSRMS(ViT)	**99.44**	**99.99**	**84.93**	**98.08**	**88.99**	**98.85**	**80.68**	**41.33**	**46.02**	**87.29**

5.2 Performance Comparison

This section reports the experimental performance of CSRMS and various baseline algorithms for image classification. The algorithms were evaluated using widely-used visual backbones, including LeNet [18], ResNet-18 [13], ResNet-50 [13] and ViT [8], as well as state-of-the-art algorithms such as RE [60], ACmix [35], Resizer [42], Batch-Former [16], DLSA [51] and CUDA [1] combined with ResNet18. The hyperparameters for CSRMS and the baselines were carefully tuned to achieve the best performance. From the performance as reported in Table 2, we can observe the followings:

- **The proposed CSRMS method consistently improves the generalization ability of visual backbones.** In different domains and datasets of different sizes, CSRMS can combine various types of backbone, including convolution-based models and Transform-based models, to improve classification performance.
- **CSRMS achieves more significant improvement in the case of high class complexity.** Most of the methods achieve good performance when facing the CIFAR-10 dataset with low classification complexity, while on CIFAR-100, VireoFood-172, and NUS-WIDE with the higher number of categories, CSRMS addresses the intra-class diversity and inter-class similarity problems, thus obtaining a 3%-11% improvement.
- **Visual backbones with more complex structures demonstrate significant advantages on larger datasets.** In the case of models with relatively basic architectures, such as LeNet5, there is a higher likelihood of overfitting when training on datasets with a higher number of categories such as CIFAR-100, or with a larger number of samples such as VireoFood-172 and NUS-WIDE. And the performance gap is more pronounced when comparing ResNet and ViT models.
- **Methods that introduce sample relationship or class-aware information can improve classification performance.** For example, compared with the advanced data augmentation method RE, the performances of CUDA that combine class-aware information have been improved for 2%-4% on larger datasets VireoFood-172 and NUS-WIDE; for relational modelling methods, BatchFormer and ACmix, which increases attention between samples, also make performance gains.

5.3 Ablation Study

In this section, we further studied the working mechanisms of different modules of CSRMS, as shown in Table 3. The following findings could be observed:

- **Positive sampling with graphical smoothing can alleviate intra-class diversity:** the accuracy obtained after adding DomPattern-based Intra-Class Sampling (CS(D)) and Graphical Smoothing (G) is always better than base, which verifies that graphical smoothing can alleviate effectiveness in terms of intra-class diversity.
- **Negative sampling with explicit constraints can mitigate inter-class similarity:** adding "+CS(S)" can push away representations of different categories through explicit constraints to enhance representation learning so that achieves better classification accuracy.
- **Curriculum construction can help the model better learn complex features and knowledge:** "+CS(D)+G+CS(S)+CS(C)" can significantly improve the accuracy of both models. This is because curriculum construction enhances representation learning by making the model converge faster and avoiding overfitting and local optimal solutions.
- **Cluster-level, class-level alignment and inter-class constraints can further alleviate intra-class differences and inter-class similarities:** the introduction of "A" and "IC" can further improve the accuracy of the two models. This is because "A" and "IC" can provide aggregation and constraints of more explicit schemes, thus enhancing representation learning.

5.4 In-depth Analysis

5.4.1 Analysis of the Impact of Different Sampling Strategies on Classification Accuracy. As illustrated in Table 4, we have evaluated the influence of distinct sample sampling strategies on the classification outcomes. **Positive sampling from images with dominant patterns (DC-posi) can better alleviate intra-class diversity:** "OC-posi" involves positive sampling from clusters other than the input image's own cluster, surpassing random sampling in mitigating intra-class diversity and improving classification. "DC-posi," on

Table 3: Ablation study of CSRMS with ResNet18 and ViT backbone. CS(D):Positive Sampling from Dominant Patterns; CS(S): Negative Sampling from Visually-simi; CS(C): Curriculum construction strategy; G: Graphical smoothing module; A: Cluster-level and class-level alignment; IC: Class-level Distribution Regularization.

Backbone	Models	CIFAR100		Vireo172	
		ACC@1	ACC@5	ACC@1	ACC@5
ResNet18	Base	70.12	91.80	77.31	93.26
	+CS(D)+G	71.65	93.76	78.69	93.82
	+CS(D)+G+CS(S)	72.95	94.01	80.62	94.03
	+CS(D)+G+CS(S)+CS(C)	73.30	94.28	81.25	94.96
	+CS(D)+G+CS(S)+CS(C)+A	73.62	94.40	81.79	95.38
	+CS(D)+G+CS(S)+CS(C)+A+IC	74.57	94.47	82.60	95.92
ViT	Base	81.70	96.02	85.92	96.47
	+CS(D)+G	82.55	96.51	86.60	97.32
	+CS(D)+G+CS(S)	83.68	97.04	87.61	98.52
	+CS(D)+G+CS(S)+CS(C)	84.05	97.43	88.10	98.69
	+CS(D)+G+CS(S)+CS(C)+A	84.42	97.85	88.45	98.77
	+CS(D)+G+CS(S)+CS(C)+A+IC	**84.93**	**98.08**	**88.99**	**98.85**

the other hand, goes a step further by selecting positive samples from the most representative cluster, unifying sample representation learning, reducing intra-class differences, and boosting classification performance. **Negative sampling from images with visually similar patterns (SC-posi) can better alleviate inter-class similarity:** "OC-nega" employs negative sampling from different clusters for the input image. This strategy leads to an "attempt to make sufficiently dissimilar images even more dissimilar", which results in a slight decrease in performance compared to random sampling. On the other hand, "SC-nega" selects negative samples from the same cluster as the input image. This targeted approach effectively addresses inter-class similarity issues, outperforming random sampling and leading to better classification results.

Table 4: Ablation of different sampling strategies.

Backbone	Strategies	CIFAR100		Vireo172	
		ACC@1	ACC@5	ACC@1	ACC@5
ResNet18	Base	70.12	91.80	77.31	93.26
	Random-posi	70.82	91.99	77.69	93.42
	OC-posi	71.06	92.03	77.94	93.58
	DC-posi	71.65	93.76	78.69	93.82
	Random-nega	71.82	93.85	79.03	93.67
	OC-nega	71.77	93.82	78.82	93.63
	SC-nega	72.95	94.01	80.62	94.03
ViT	Base	81.70	96.02	85.92	96.47
	Random-posi	81.99	96.20	86.17	96.86
	OC-posi	82.24	96.36	86.37	96.99
	DC-posi	82.55	96.51	86.60	97.32
	Random-nega	82.76	96.62	86.84	97.79
	OC-nega	82.73	96.90	86.80	97.77
	SC-nega	**83.68**	**97.04**	**87.61**	**98.52**

5.4.2 Analysis of the Impact of Different Smoothing Algorithms on Classification Accuracy. As illustrated in Table 5, we have evaluated the influence of distinct Smoothing algorithms on the classification outcomes. **Graphical smoothing can better aggregate the**

Table 5: Performance of CSRMS with smoothing algorithms.

Backbone	Algorithms	CIFAR100		Vireo172	
		ACC@1	ACC@5	ACC@1	ACC@5
ResNet18	Base	70.12	91.80	77.31	93.26
	JS-loss	70.88	92.35	78.24	93.46
	GNN	71.33	92.87	79.09	94.12
	Tail-GNN	71.38	92.85	79.13	94.01
	GCN	71.65	93.76	78.69	93.82
ViT	Base	81.70	93.38	85.92	96.47
	JS-loss	82.06	93.51	86.65	96.76
	GNN	82.25	93.89	86.88	97.05
	Tail-GNN	82.27	93.88	86.91	97.05
	GCN	**82.55**	**96.51**	**87.69**	**97.32**

representation of the same class and better improve the classification effect than explicit constraint: Utilizing graphical smoothing algorithms: GNN [38], Tail-GNN [26] and GCN [17] to aggregate representations can achieve better classification accuracy than using explicit constraints such as js-divergence. **Incorporating convolution when aggregating node information provides stronger expressive power and generalization ability:** GNN computes a weighted sum of adjacent nodes' features, while GCN uses convolutional operations for better representation fusion. This enables GCN to fuse relation information more effectively, resulting in better representation aggregation. Tail-GNN, though useful for imbalanced data, achieves similar performance to conventional GNN as it focuses on handling difficult samples rather than optimizing representation learning.

5.4.3 Analysis of the Impact of Different Clustering Parameters on Classification Accuracy. **Based on our evaluation (Table 6), different clustering results minimally impact the final classification accuracy**. Although varying clustering outcomes can affect image recognition accuracy, the impact is limited because our clustering relies on visual commonality, and erroneous clusters constitute only a small fraction. These inaccuracies diminish gradually with increased training epoch iterations, as the network learns more precise feature representations and corrects erroneous clusters. Thus, for instances of poor clustering quality, excessive concern is unnecessary; instead, we should focus on enhancing the accuracy and generalization capacity of feature expression to achieve superior recognition results.

Table 6: Image classification performance of CSRMS with different clustering parameters.

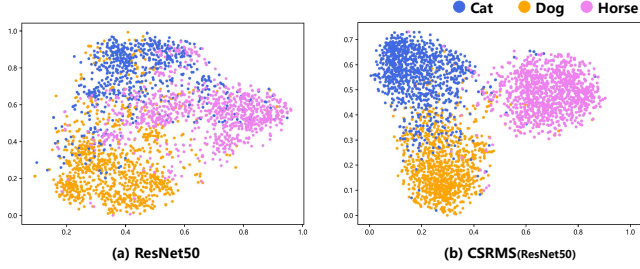| Backbone	Vigilance Parameter	Number of Clusters	Number of Dominant Clusters	Vireo172	
				ACC@1	ACC@5
ResNet18	Base	-	-	77.31	93.26
	0.5	503	178	82.56	95.86
	0.7	462	207	82.58	95.90
	0.85	424	220	82.60	95.92
	0.95	406	261	82.58	95.90
ViT	Base	-	-	85.92	96.47
	0.5	436	185	85.95	98.84
	0.7	398	196	88.95	98.83
	0.85	375	203	88.99	98.85
	0.95	369	212	88.95	98.83

Figure 4: Visualization of representations encoded by ResNet50 and CSRMS. To qualitatively verify the effect of introducing relational modelling on representation learning, we choose the images of three highly confusing categories: Cat, Dog and Horse, and compared the visualization results of representations encoded by ResNet50 and CSRMS.

5.5 Case Study

5.5.1 Effect analysis of Relationship Modelling. This section evaluates the impact of relational modelling on representation learning. The confusion matrix analysis reveals that the cifar10 dataset's cat, dog, and horse categories experience the highest level of misclassification. To investigate this further, we select and visualize these images' representations encoded by ResNet50 and CSRMS using 2D-tsne, as presented in Figure 4. In Figure 4 (a), ResNet50 demonstrates relatively good discrimination among the three categories. However, the representation distribution across different categories is mixed, and representations of the same class are scattered widely. In contrast, Figure 4 (b) illustrates that CSRMS effectively addresses this issue by bringing representations of the same class closer together while pushing representations of different categories apart. This provides compelling evidence that introducing relationship modelling significantly enhances representation learning.

5.5.2 Error Analysis of Prediction. In this section, we present a 2D t-SNE visualization and error analysis to demonstrate the effectiveness of CSRMS in addressing intra-class diversity and inter-class similarity. Fig 5 compares the visual representations generated by CSRMS and ResNet50 for four images, along with their predicted ingredients and confident scores. In image (a), where the composition is clear, both models provide accurate predictions. However, CSRMS shows a higher confident score in predicting the "bird" position. Moreover, in the representation space, CSRMS places the representation of image (a) closer to the center of the "bird" representation set. For image (b), where the components are unclear, ResNet50 tends to produce incorrect predictions. In contrast, CSRMS places the representation closer to the center of the "plane" representations set, leading to more accurate predictions. In image (c), although both models produce incorrect predictions, CSRMS generates results highly similar to ResNet50. However, the representation space derived from ResNet50 exhibits an abnormally chaotic pattern, possibly occurring by chance. For image (d), neither model accurately predicts the image due to its blurry components and small pixel range. Nevertheless, CSRMS places the representation closer to the "plane" representations set, resulting in a higher confident score. These findings strongly support the efficacy of CSRMS in improving representation learning by mitigating intra-class diversity and inter-class similarity.
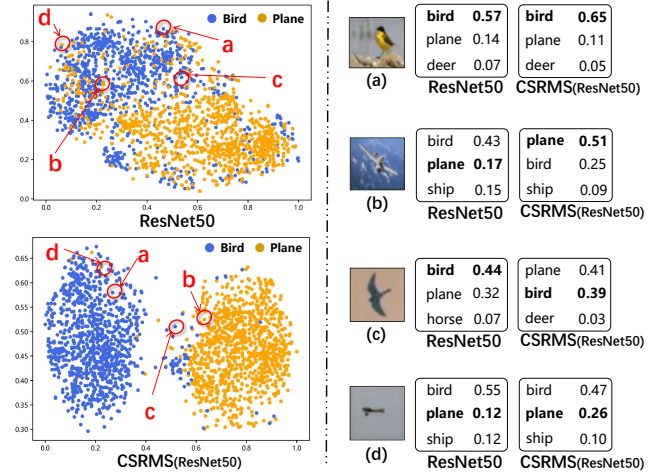

Figure 5: Case Study on CSRMS in success and failure cases. To qualitatively demonstrate the efficacy of CSRMS in alleviating intra-class diversity and inter-class similarity, we select the images of two categories that are highly confused in the CIFAR10 dataset: "plane" and "bird". (a) Both models achieve reasonable performance. (b) ResNet50 fails. (c) CSRMS performs worse. (d) both models perform badly.

6 CONCLUSION

This paper proposes a novel approach CSRMS to alleviate the issue of intra-class visual diversity and inter-class similarity in representation learning by modelling a relational graph of the entire dataset and performing class-aware smoothing and regularization operations. This approach learns the data distributions in the feature space and extends the typical training batch construction process. A graph convolution network with knowledge-guided smoothing operations is utilized to ease the projection from different visual patterns to the same class. Experiments conducted on CIFAR10, CIFAR100, Vireo172, and NUS-WIDE datasets demonstrate the effectiveness of CSRMS in improving classification accuracy and verifying the effectiveness of structured knowledge modelling for enhanced representation learning.

Future work of this study may focus on two directions. First, the model can integrate rich multimodal information, including semantic cues, to optimize the sampling strategy and foster a deeper understanding of visual-semantic relationships. Second, by incorporating self-supervised learning pre-training techniques, CSRMS can leverage unlabeled data to learn powerful representations, thereby improving data efficiency and generalization across tasks and domains. These avenues of exploration offer promising opportunities for the continued improvement of CSRMS.

7 ACKNOWLEDGMENTS

REFERENCES

[1] Sumyeong Ahn, Jongwoo Ko, et al. 2023. CUDA: Curriculum of Data Augmentation for Long-tailed Recognition. *arXiv preprint arXiv:2302.05499* (2023).

[2] Mathilde Caron et al. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS* (2020).

[3] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *ACM MM*.

[4] Zhao-Min Chen, Xiu-Shen Wei, et al. 2019. Multi-Label Image Recognition with Joint Class-Aware Map Disentangling and Label Correlation Embedding. In *IEEE*.

[5] Gong Cheng, Pujian Lai, et al. 2023. Class attention network for image recognition. *Science China Information Sciences* (2023).

[6] Tat-Seng Chua, Jinhui Tang, et al. 2009. Nus-wide: a real-world web image database from national university of singapore. In *ACM MM*.

[7] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual Encoding for Video Retrieval by Text. *TPAMI* (2021).

[8] Alexey Dosovitskiy, Lucas Beyer, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Victor Garcia and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043* (2017).

[10] Golnaz Ghiasi, Yin Cui, et al. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*.

[11] Qing-Ling Guan, Yuze Zheng, Lei Meng, Li-Quan Dong, and Qun Hao. 2023. Improving the Generalization of Visual Classification Models Across IoT Cameras via Cross-modal Inference and Fusion. *IEEE Internet of Things Journal* (2023).

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[14] Ruibing Hou, Hong Chang, et al. 2019. Cross attention network for few-shot classification. *NeurIPS* (2019).

[15] Zhi Hou, Xiaojiang Peng, et al. 2020. Visual compositional learning for human-object interaction detection. In *ECCV*.

[16] Zhi Hou, Baosheng Yu, and Dacheng Tao. 2022. Batchformer: Learning to explore sample relationships for robust representation learning. In *CVPR*.

[17] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[18] Yann LeCun, Léon Bottou, et al. 1998. Gradient-based learning applied to document recognition. *IEEE* (1998).

[19] Xiangxian Li, Haokai Ma, Lei Meng, and Xiangxu Meng. 2021. Comparative study of adversarial training methods for long-tailed classification. In *ADVM*.

[20] Xiang Li, Lei Wu, Xu Chen, Lei Meng, and Xiangxu Meng. 2022. Dse-net: Artistic font image synthesis via disentangled style encoding. In *ICME*.

[21] Xiang Li, Lei Wu, Changshuo Wang, Lei Meng, and Xiangxu Meng. 2023. Compositional Zero-Shot Artistic Font Synthesis. *IJCAI* (2023).

[22] Xiangxian Li, Yuze Zheng, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. 2023. Cross-modal Learning Using Privileged Information for Long-tailed Image Classification. *CVM* (2023).

[23] Chengliang Liu, Jie Wen, et al. 2023. Incomplete Multi-View Multi-Label Learning via Label-Guided Masked View-and-Category-Aware Transformers. *arXiv preprint arXiv:2303.07180* (2023).

[24] Jinxing Liu, Junjin Xiao, Haokai Ma, Xiangxian Li, Zhuang Qi, Xiangxu Meng, and Lei Meng. 2022. Prompt Learning with Cross-Modal Feature Alignment for Visual Domain Adaptation. In *CAAI*.

[25] Tianhan Liu, Zhuang Qi, Zitan Chen, Xiangxu Meng, and Lei Meng. 2023. Cross-Training with Prototypical Distillation for improving the generalization of Federated Learning. *ICME* (2023).

[26] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. 2021. Tail-gnn: Tail-node graph neural networks. In *KDD*.

[27] Haokai Ma, Xiangxian Li, Lei Meng, and Xiangxu Meng. 2021. Comparative study of adversarial training methods for cold-start recommendation. In *ADVM*.

[28] Haokai Ma, Zhuang Qi, Xinxin Dong, Xiangxian Li, Yuze Zheng, and Xiangxu Mengand Lei Meng. 2023. Cross-Modal Content Inference and Feature Enrichment for Cold-Start Recommendation. *IJCNN* (2023).

[29] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2023. Exploring False Hard Negative Sample in Cross-Domain Recommendation. In *Recsys*.

[30] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2023. Triple Sequence Learning for Cross-domain Recommendation. *arXiv preprint arXiv:2304.05027* (2023).

[31] Lei Meng, Long Chen, Xun Yang, Dacheng Tao, Hanwang Zhang, Chunyan Miao, and Tat-Seng Chua. 2019. Learning using privileged information for food recognition. In *ACM MM*.

[32] Lei Meng, Fuli Feng, Xiangnan He, Xiaoyan Gao, and Tat-Seng Chua. 2020. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *Proceedings of MM*.

[33] Lei Meng, Ah-Hwee Tan, and Donald C Wunsch. 2015. Adaptive scaling of cluster boundaries for large-scale social media data clustering. *TNNLS* (2015).

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[35] Xuran Pan, Chunjiang Ge, et al. 2022. On the integration of self-attention and convolution. In *CVPR*.

[36] Zhuang Qi, Yuqing Wang, Zitan Chen, Ran Wang, Xiangxu Meng, and Lei Meng. 2022. Clustering-based Curriculum Construction for Sample-Balanced Federated Learning. In *CAAI*.

[37] Pau Rodríguez, Issam Laradji, et al. 2020. Embedding propagation: Smoother manifold for few-shot classification. In *ECCV*.

[38] Franco Scarselli, Marco Gori, et al. 2008. The graph neural network model. *IEEE transactions on neural networks* (2008).

[39] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating Objects and Relations in User-Generated Videos. *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (2019).

[40] Shiv Shankar, Vihari Piratla, et al. 2018. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745* (2018).

[41] Weilin Sun, Xiangxian Li, Manyi Li, Yuqing Wang, Yuze Zheng, Xiangxu Meng, and Lei Meng. 2022. Sequential Fusion of Multi-view Video Frames for 3D Scene Generation. In *CAAI*.

[42] Hossein Talebi and Peyman Milanfar. 2021. Learning to resize images for computer vision tasks. In *ICCV*.

[43] Jinhui Tang, Xiangbo Shu, et al. 2016. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. In *ICCV*.

[44] Jinhui Tang, Xiangbo Shu, et al. 2016. Tri-clustered tensor completion for social-aware image tag refinement. *TPAMI* (2016).

[45] Petar Veličković, Guillem Cucurull, et al. [n. d.]. Graph Attention Networks. In *ICLR*.

[46] Chu Wang, Babak Samari, et al. [n. d.]. Affinity graph supervision for visual recognition. In *CVPR*.

[47] Yuqing Wang, Xiangxian Li, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. 2022. Causal Inference with Sample Balancing for Out-of-Distribution Detection in Visual Classification. In *CAAI*.

[48] Yuqing Wang, Xiangxian Li, Zhuang Qi, Jingyu Li, Xuelong Li, Xiangxu Meng, and Lei Meng. 2022. Meta-causal feature learning for out-of-distribution generalization. In *ECCV*.

[49] Yuqing Wang, Zhuang Qi, Xiangxian Li, Jinxing Liu, Xiangxu Meng, and Lei Meng. 2023. Multi-channel Attentive Weighting of Visual Frames for Multimodal Video Classification. *IJCNN* (2023).

[50] Yue Xu, Yong-Lu Li, Jiefeng Li, and Cewu Lu. 2022. Constructing Balance from Imbalance for Long-Tailed Image Recognition. In *ECCV*.

[51] Yue Xu, Yong-Lu Li, et al. 2022. Constructing balance from imbalance for long-tailed image recognition. In *ECCV*.

[52] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. *SIGIR* (2020).

[53] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. *SIGIR* (2021).

[54] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-Supervised Video Object Grounding by Exploring Spatio-Temporal Contexts. *ACM MM* (2020).

[55] Zhixiong Yang, Junwen Pan, et al. 2022. ProCo: Prototype-Aware Contrastive Learning for Long-Tailed Medical Image Classification. In *MICCAI*.

[56] Hongyi Zhang, Moustapha Cisse, et al. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[57] Zhen Zhao, Luping Zhou, et al. 2022. LaSSL: Label-Guided Self-Training for Semi-supervised Learning. In *AAAI*.

[58] Yu Zheng, Jiahui Zhan, et al. 2023. Curricular contrastive regularization for physics-aware single image dehazing. *arXiv preprint arXiv:2303.14218* (2023).

[59] Zhun Zhong, Liang Zheng, et al. 2020. Random erasing data augmentation. In *AAAI*.

[60] Zhun Zhong, Liang Zheng, et al. 2020. Random erasing data augmentation. In *AAAI*.

[61] Jianggang Zhu, Zheng Wang, et al. 2022. Balanced Contrastive Learning for Long-Tailed Visual Recognition. In *CVPR*.