

A dual cross-modal interactive guided common representation method for fine-grained cross-modal retrieval

Hongchun Lu^{1,2}, Min Han^{1,2}, Xue Li³, Le An⁴

¹ School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, Sichuan 610031, China

² Manufacturing Industry Chains Collaboration and Information Support Technology Key Laboratory of Sichuan Province, Chengdu, Sichuan 610031, China

³ School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

⁴ School of Foreign Languages, Xinjiang University, Urumqi 830046, China

hanmin@swjtu.edu.cn

Abstract. In fine-grained cross-modal retrieval tasks, the huge heterogeneity gap between different modalities is a key factor leading to low retrieval performance. Therefore, addressing the media divide (i.e., inconsistent representation of different media types) is an important way to improve retrieval performance. Although previous research has yielded some results, the standard model still has some shortcomings. First, the information interaction between different modalities is ignored when learning common representations of different media data. Second, discriminative fine-grained features are not fully exploited. To address this challenge, we propose a dual cross-modal interaction-guided common representation network (DCINet) to enhance the information interaction between different modalities while mining discriminative features in media data. Specifically, we construct a common representation network and use pre-interaction and post-interaction multimodal feature inputs into the network for training, respectively. The two training strategies guide the learning of the common representation network through a maximal-minimal game, effectively enhancing cross-media semantic consistency and improving retrieval accuracy. Finally, extensive experiments and ablation studies conducted on public datasets demonstrate the effectiveness of our proposed method.

Keywords: Fine-grained, Cross-media retrieval, Cross-media spatial interaction, Cross-media channel interaction.

1 Introduction

Fine-grained cross-modal retrieval has become an urgent technical problem in the field of cross-modal retrieval [1] [2] [3]. It has great research value and application demand in both academic and industrial fields. Different from coarse-grained cross-media retrieval, the purpose of a fine-grained cross-modal retrieval task is to retrieve results that

are completely relevant to the fine-grained subcategories of the submitted query according to the fine-grained retrieval requirements [2]. The difference between the two is shown in Fig.1. When a user inputs an image of a subcategory of a bird, the coarse-grained cross-media retrieval can only categorize it simply as a "bird", without going further to differentiate its subcategories. In contrast, fine-grained cross-media retrieval can retrieve specific subcategories of birds and return relevant results. This enables users to find the exact information they need. The difficulty lies in its (1) small inter-class differences, but large heterogeneity gaps between modalities, and (2) large intra-class differences, but small differences in data representation within the same modality. Therefore, this task is more challenging than general coarse-grained cross-modal retrieval.

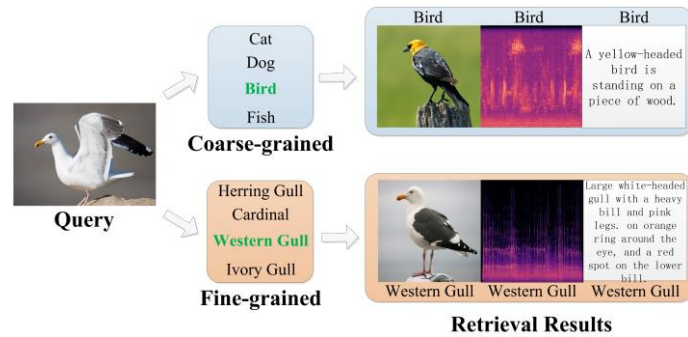


Fig. 1. The comparison of the coarse-grained cross-media retrieval and fine-grained.

In recent years, with the rapid development of the Internet, all kinds of media data have shown explosive growth. As a result, the traditional single-modal retrieval methods appear to be not flexible and practical enough. Researchers and scholars have begun to conduct a lot of research on cross-modal retrieval techniques and achieved remarkable results [4][5][6][7]. However, existing research predominantly focuses on coarse-grained cross-modal retrieval, which may not be able to meet the diverse and complex needs of human life [2]. To address these challenges, recent research has shifted its focus to fine-grained cross-modal retrieval methods. For example, in [2], the first fine-grained cross-media retrieval dataset was constructed, providing a foundation for fine-grained cross-modal research. Additionally, a unified deep-learning algorithm has been proposed. Shan et al. [3] proposed a generalized attention space learning method to learn the common attention space of different modalities. These methods simply input different modal features into the common network for training and do not handle the differences between different modalities well. To alleviate this problem, Shen et al. [1] proposed a channel blending algorithm, which effectively enhances the information interaction between different modalities. However, the algorithm only performed fusion in the channel dimension and did not consider the interaction in the spatial dimension. To better address the significant heterogeneity gap between different modalities, we propose a dual cross-modal interaction learning approach to effectively enhance cross-

media semantic consistency and improve retrieval accuracy. The network structure is depicted in Fig.2.

Specifically, we first construct a common representation network (e.g., ResNet50) and train it using two different training strategies to guide the learning of the common spatial network via a min-max game. In the first step, the output feature maps of ResNet50 (with the fully connected and pooled layers removed) are directly input into the second sub-module of the common representation network for training without further processing. In the second step, the output feature maps generated by ResNet50 undergo computation through the cross-modal spatial interaction (CMSI) and cross-modal channel interaction (CMCI) modules before being transmitted to the second sub-module of the common representation network for learning. We believe that this dual cross-modal interaction approach to learning has multiple benefits. Firstly, for cross-modal spatial interaction computation, we do not simply exchange feature maps of different modalities, but rather by mask coverage. For example, by replacing a portion of the feature map on the image modality with a portion of the feature map on the text modality, this approach ensures that the spatial features between different modalities are effectively interacted with, and discriminative features are mined simultaneously. As part of the features in the original modality are covered, the model will be biased to learn the missing features during training, prompting the network to learn discriminative features. Secondly, cross-channel interactive computation further reduces heterogeneous differences across modalities. Our extensive experiments on a widely used benchmark demonstrate that our approach provides superior feature representation and excellent performance in cross-modal fine-grained retrieval.

Overall our main contributions can be summarized as follows:

1. We propose a novel and highly reliable fine-grained cross-modal retrieval deep learning framework DCINet, aiming to effectively enhance cross-media semantic consistency and improve retrieval accuracy.
2. We construct a fine-grained cross-modal spatial interaction module to enhance spatial information interaction between different modalities while mining discriminative features in media data. It effectively solves the problem of small inter-class differences but significant heterogeneity gaps between different modalities in fine-grained cross-modal retrieval.
3. We developed a fine-grained cross-modal channel interaction module to enhance the interaction of channel information between different modalities, effectively improving the inter-class separability and intra-class compactness of multimodal data.

The rest of the paper is organized as follows: Section 2 reviews the pertinent works regarding coarse-grained cross-modal retrieval and fine-grained cross-modal retrieval. Section 3 elaborates the details of the proposed framework. We conduct the experiment and evaluation of our proposed methods in Section 4 and conclude in Section 5.

2 Related Work

2.1 Coarse-Grained Cross-Modal Retrieval

With the increasing size of multimedia databases, traditional single-modal retrieval methods are no longer sufficient to meet human needs. To address these challenges, researchers have proposed various cross-modal retrieval methods. For instance, Li et al. [8] introduced a cross-modal factor analysis method, utilizing statistical correlation analysis to learn cross-media correlations. Wang et al. [9] presented a graph regularization algorithm that integrates inter-modal and intra-modal similarities into joint graph regularization to efficiently learn cross-modal correlations. In recent years, with the advancement of deep learning, numerous studies have delved into cross-media retrieval research based on convolutional neural networks. Liu et al. [10] enhanced the accuracy of cross-modal image retrieval by integrating text and image generation features into the model. Yu et al. [11] employed graph neural networks (GNNs) and introduced a cross-modal feature matching network, which facilitates cross-modal retrieval by integrating various modal features into the model, thereby mitigating the degradation of retrieval performance caused by information misalignment. However, these approaches are primarily designed for two types of media, resulting in suboptimal performance when directly applied to a broader range of media. Consequently, researchers have embarked on exploring novel approaches aimed at developing generalized spatial learning strategies applicable to various media types. For instance, Huang et al. [12] introduced a Modal-Reverse Hybrid Transmission Network (MHTN), with the aim of facilitating knowledge transfer between a single-modal source domain and a cross-modal target domain, while learning common representations across modalities. Experimental validation is conducted on five different types of modal data, including images, text, audio, video, and 3D.

2.2 Fine-Grained Cross-Modal Retrieval

Most existing research is biased toward coarse-grained cross-modal retrieval, with less attention given to fine-grained cross-modal retrieval. Notable works in this area include He et al. [2], who co-learned the cross-media relevance of different media data using the ResNet network and employed classification loss, center loss, and ranking loss for improved feature learning in general-purpose networks. Wang et al. [13] proposed a dual-branch fine-grained cross-media network (DBFC-Net) and introduced a new distance metric, Cosine+, to bridge the gap between different media types. Bai et al. [14] utilized proprietary and general-purpose networks for finer feature extraction, where the proprietary network extracts a single feature for each media, and the generic network extracts common feature representations for different media. Shan et al. [3] proposed an attentional hybrid network to efficiently learn common representations for different media data. Shen et al. [1] proposed a channel hybrid approach to enhance the information interaction of fine-grained objects in different modalities. Chen et al. [15] employed LAGC-Attention to mine and fuse the feature information of different modalities and map them to a common space, thereby narrowing the semantic gap between

different features. Hong et al. [16] utilized a generative adversarial network to learn the common semantic space between different modalities. Different from the above approaches, we propose a novel approach of dual cross-modal interaction, which facilitates the comprehensive interaction and fusion of different media data in both spatial and channel dimensions, to efficiently learn the common representations of different media data.

3 Method

In this section, we first present an overview of the framework in Section 3.1. Subsequently, we will describe the CMSI and CMCI modules in Sections 3.2 and 3.3, respectively. Finally, the loss function is discussed in Section 3.4.

3.1 Framework Overview

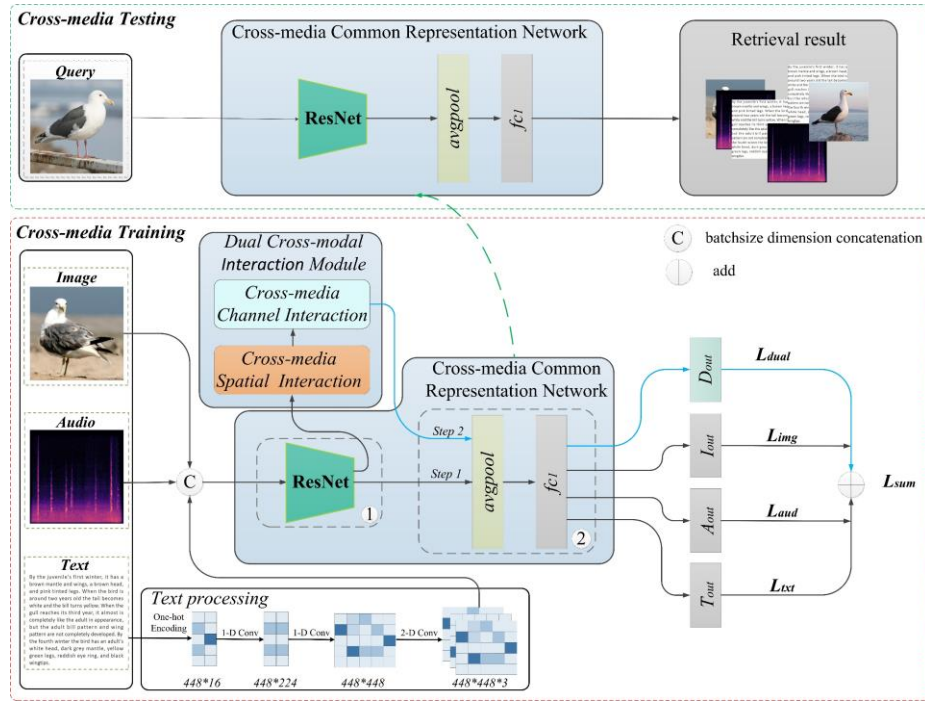


Fig. 2. DCINet architecture, including text processing, cross-media common representation network, cross-modal spatial interaction, and cross-modal channel interaction. The red dashed box represents the cross-modal training network, while the green dashed box indicates the cross-modal testing network.

The architecture of the proposed dual cross-modal interaction network (DCINet) is illustrated in Fig.2. DCINet comprises four components: text processing, cross-media

common representation network (CCRN), cross-modal spatial interaction (CMSI), and cross-modal channel interaction (CMCI). Text processing aims to standardize raw text modal data into the same size and dimension as images and audio, facilitating efficient computation of inter-modal features. CCRN extracts feature information from different media using ResNet50 and projects them into a common space. The CMSI component replaces the masked markers of one modal feature map with the visible markers of the other modal feature map to form a mixed-modal feature map, which is embedded into the second sub-module of CCRN for training. This effectively narrows the significant heterogeneity gap between different modalities and improves the intra-class separability of modalities. CMCI employs tags to retrieve different modal feature maps belonging to the same class in each mini-batch and channels their feature maps for channel interactions to further enhance the inter-class separability and intra-class compactness of the multimodal modalities.

Specifically, the red dashed box in Fig.2 shows the training process of the model. The data of different modalities are spliced by batch-size dimensions. Then, the model undergoes training in two different strategies. In the second strategy, the feature map output from the first sub-module in the CCRN is first processed by CMSI. Then, the resulting feature map from CMSI undergoes further processing by CMCI before being sent to the second sub-module in the CCRN for classification. The green dashed box in Fig.2 shows the testing process of the model. It is worth noting that in the testing phase, only the CCRN is used. As a result, the network has a smaller number of parameters and reduces the inference time.

3.2 Cross-media spatial interaction

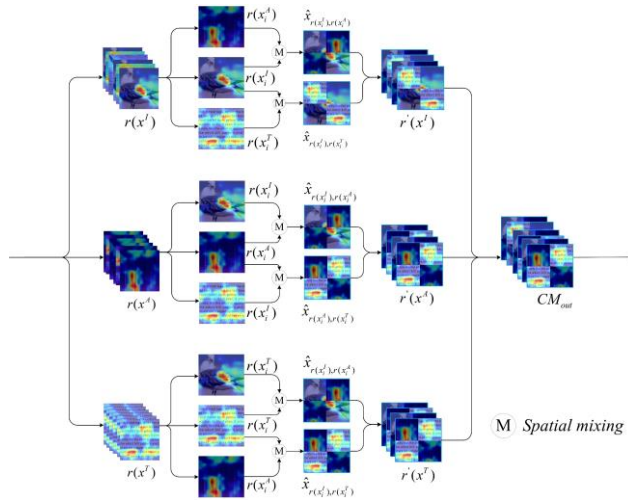


Fig. 3. The architecture of the CMSI module.

Inspired by the success of CutMix [17] in object detection, we propose the CMSI method to narrow the huge heterogeneity gap between different modalities and improve

the intra-class separability of modalities. The network structure is shown in Fig.3. Before the CMSI calculation, the raw text modal data is transformed to obtain x_k^T by the text processing module shown in Fig.2. Its expression can be calculated as:

$$x_k^T = \text{Conv}_{2D}(\text{Conv}_{1D}(\text{Conv}_{1D}(\delta(x_k^{T_o})))) \quad (1)$$

where $x_k^{T_o}$ denotes the kth image in the input text modal data; δ denotes One-hot Encoding; Conv_{1D} denotes 1-dimensional convolutional operation; and Conv_{2D} denotes 2-dimensional convolutional operation. Subsequently, the data from different modalities is concatenated along the batchsize dimension and input to the first sub-module of the Cross-media Common Representation Network (CCRN) for feature mapping learning. The output features are then used as inputs to the CMSI module for spatial interaction between different modal data. It is important to note that CMSI is applied in each mini-batch to operate on different modal features. Specifically, in each mini-batch, we take the features $r(x^I)$, $r(x^A)$, $r(x^T)$ output from ResNet50 as inputs to the CMSI module, where $x^I = (x_1^I, \dots, x_i^I, \dots, x_b^I)$, $x^A = (x_1^A, \dots, x_j^A, \dots, x_b^A)$, $x^T = (x_1^T, \dots, x_k^T, \dots, x_b^T)$, $x^I, x^A, x^T \in \mathbb{R}^{b \times c \times h \times w}$. Thus, the CMSI module is computed as follows: for example, between two types of modal data, namely, image and audio. Let $M=1$ represent the mask marker of feature map $r(x_i^I)$, then $1-M$ represents the mask marker of feature map $r(x_j^A)$. Therefore, the mixed modal feature map $CM_{out} \in \mathbb{R}^{b \times c \times h \times w}$ after spatial interaction is expressed as follows:

$$\hat{x}_{r(x_i^I), r(x_j^A)} = r(x_i^I) \odot M + r(x_j^A) \odot (1-M) \quad (2)$$

$$\hat{x}_{r(x_i^I), r(x_j^A)} = r(x_i^I) \odot M + r(x_j^A) \odot (1-M) \quad (3)$$

$$\hat{x}_{r(x_j^A), r(x_k^T)} = r(x_j^A) \odot M + r(x_k^T) \odot (1-M) \quad (4)$$

$$CM_{out} = (r'(x^I), r'(x^A), r'(x^T)) = \sum_{i,j,k=1}^{i,j,k=b} (\hat{x}_{r(x_i^I), r(x_j^A)}, \hat{x}_{r(x_i^I), r(x_k^T)}, \hat{x}_{r(x_j^A), r(x_k^T)}) \quad (5)$$

where \odot represents element-wise multiplication; r represents the ResNet50 operation; r' represents the feature map of each modality computed by CMSI; and i, j, k denote the subscript indexes of the media features.

Finally, the feature map CM_{out} after spatial interaction is sent to CMCI for further cross-modal channel interaction and feature refinement. Together with the original branch, CCRN network learning is guided by two different training strategies. It should

be noted that the labels of the feature maps after cross-modal space interaction are converted accordingly using the method described in CutMix [17]. We validate the CMSI module in the experimental section. From the results, it effectively enhances cross-media semantic consistency and improves retrieval accuracy significantly.

3.3 Cross-media channel interaction

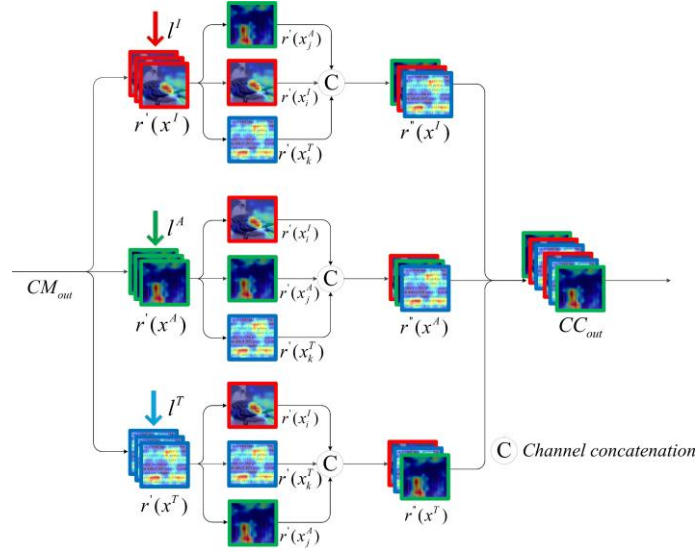


Fig. 4. The architecture of the CMCI module.

In cross-modal fine-grained retrieval tasks, the majority of existing algorithms directly map data from different modalities into a unified common space. Following the calculation of inter-modal spatial feature interaction by the CMSI module, and aiming to enhance the interaction of inter-modal feature information, we propose a straightforward and efficient method for cross-media channel interaction (CMCI), inspired by the paper [1]. The network structure is shown in Fig.4. The output feature map CM_{out} of the CMSI module is used as an input to the CMCI, so the label corresponding to CM_{out} is $L = (l^I, l^A, l^T)$, where $l^I = (l_1^I, \dots, l_i^I, \dots, l_b^I)$, $l^A = (l_1^A, \dots, l_j^A, \dots, l_b^A)$, $l^T = (l_1^T, \dots, l_k^T, \dots, l_b^T)$, $l^I, l^A, l^T \in R^{b \times 1}$. The same categories of different modalities are matched in each mini-batch by using the label L . If identical labels are present in both modalities, 1/3 of the channel features from one modality are swapped with the corresponding 1/3 features of the same category in the other modality. If identical labels are present in all three modalities, 2/3 of the channel features from the first modality are replaced with the 1/3 features of the second and third modalities, respectively. The expression for the feature map after cross-modal channel interaction, denoted as CC_{out} , is calculated as follows:

$$\begin{aligned}
CC_{out} &= (r^n(x^T), r^n(x^T), r^n(x^T)) \\
&= \sum_{i,j,k=1}^{i,j,k=b} (\hat{x}_{r(x_i^I), r(x_j^A)}, \hat{x}_{r(x_i^I), r(x_k^T)}, \hat{x}_{r(x_j^A), r(x_k^T)}) \\
&= \begin{cases} 1/3cf(r'(x_i^I) \Leftrightarrow r'(x_j^A)), & l_i^I = l_j^A \\ 1/3cf(r'(x_i^A) \Leftrightarrow r'(x_k^T)), & l_i^A = l_k^T \\ 1/3cf(r'(x_i^I) \Leftrightarrow r'(x_j^A), r'(x_i^I) \Leftrightarrow r'(x_k^T), r'(x_j^A) \Leftrightarrow r'(x_k^T)), & l_i^I = l_j^A = l_k^T \end{cases}
\end{aligned} \tag{6}$$

where cf denotes cross-modal channel feature computation; \Leftrightarrow represents channel feature interaction; r' and r^n represent the feature maps before and after each modality is computed via CMCI, respectively; and i, j, k denote the subscript indexes of the media features. Finally, the output feature map CC_{out} after cross-modal channel interaction is fed into the second sub-module of the CCRN for learning and network loss calculation.

3.4 Loss function

In Fig.2. We utilized five loss functions, L_{img} , L_{aud} , L_{txt} , L_{dual} , to drive effective training and feature learning in the network. Among these, L_{img} , L_{aud} , L_{txt} represent the loss calculation for the two different modalities: image, audio, and text, respectively. L_{dual} denotes the loss calculation for dual cross-modal interactive modules (CMSI and CMCI). To ensure simplicity and efficiency, all the losses are calculated using the cross-entropy loss with the following expressions.

$$\begin{aligned}
L_{img}, L_{aud}, L_{txt}, L_{dual} &= \frac{1}{m} \sum_{i=1}^m l_{ce}(y_i, f_i^s) \\
&= \frac{1}{m} \sum_{i=1}^m -y_i^T \log \frac{\exp(f_i^s)}{\sum_{j=1}^c \exp(f_{ij}^s)}
\end{aligned} \tag{7}$$

where y_i represents the true label; f_{ij}^s represents the j th element of f_i^s ; c represents the number of categories in the dataset; and m represents the number of samples in the dataset. Finally, our total loss L_{sum} is the sum of four losses, which drives the network to map features from different modalities into a unified space and enhance the cross-modal retrieval performance of the model. The expression is as follows:

$$L_{sum} = L_{img} + L_{aud} + L_{txt} + L_{dual} \tag{8}$$

4 Experiments

In this section we first present the dataset and evaluation metrics in section 4.1, and describe the implementation details and hyperparameter settings for the experiments in section 4.2. Finally, in section 4.3 the ablation experiments of the DCINet are shown, validating the effectiveness of each module, which is analyzed and discussed.

4.1 Datasets and Evaluation Metrics

The proposed DCINet method was evaluated on the publicly available cross-modal fine-grained dataset PKU FG-XMedia [2], which is currently the only dataset for fine-grained cross-modal retrieval [2]. The dataset comprises data in four modalities: image, audio, text, and video, each containing 200 fine-grained bird categories. The image modal data is sourced from CUB200-2011 [18], consisting of 5,994 training images and 5,794 test images. The audio modal data includes 6,000 training samples and 6,000 test samples, with each audio is processed by the short-time Fourier transform [19] to generate a spectrogram [20], ultimately converted into a 448x448x3 image for representation. For text modality, 4,000 training samples and 4,000 test samples are used. We followed the approach of FGCrossNet [2], where text features are converted into a matrix representation of size 448x448x3 through one-hot coding and convolution (i.e., text processing in Fig.2). The video modal data, obtained from YouTube Bird [21]. It is worth noting that we tried to download it and found that some of the videos no longer existed, so we removed the video modal data for the fairness of the experiment. The evaluation was performed using the mean average precision (MAP), calculated by determining the average precision (AP) of the results returned by query samples, followed by the average of AP values across all queries to compute the final MAP.

4.2 Implementation Details

Training Settings. In our experiments, we utilize ResNet50 pre-trained on ImageNet21K [22] as the Cross-media Common Representation Network, with the input size of different modal data resized to 448x448x3. Specifically, the image modal data undergoes preprocessing by scaling it to 510x510, followed by resizing to 448x448x3 using Centre Crop. During training, Momentum SGD is chosen as the optimizer with initial learning rate 0.001 and weight decay 0.0001. and the batch size is set to 8; a total of 150 epochs are trained.

The CMSI computation utilizes the Mixup function from the timm¹ library. Importantly, before applying the Mixup function, we randomly shuffle feature maps within each mini-batch along the batchsize dimension to ensure comprehensive interaction among features from different modalities. Through multiple training and debugging, the Mixup function hyper-parameters `mixup_alpha`, `cutmix_alpha`, `cutmix_minmax`, `prob`, `switch_prob` is set as 0.80, 1.00, [0.50, 0.60], 1.00, 1.00. Using the

¹ <https://www.cnpython.com/pypi/timm>

Pytorch framework as the main implementation substrate, and all experiments are performed on a single Nvidia GeForce RTX 3090.

It's important to emphasize that in the testing process, we only use the Cross-media Common Representation Network in Fig.2, and the CMSI module and CMCI module are not involved in the testing. As a result, the network has a low number of parameters and inference time, in addition, Cross-media Common Representation Network can be replaced using common backbone networks, such as vgg16, transformer, etc.

4.3 Ablation Study

To ascertain the effectiveness of the proposed method, we performed ablation experiments on the CMSI and CMCI modules in bimodal and multimodal fine-grained cross-modal retrieval tasks, respectively, employing ResNet50 as the baseline model. The results are presented in Tables 1 and 2.

Table 1. mAP scores for fine-grained cross-modality retrieval of different components in bimodality.

Models	i2t	i2a	t2i	t2a	a2i	a2t	Average
ResNet50	0.225	0.562	0.274	0.217	0.579	0.185	0.340
ResNet50+CMCI	0.223	0.566	0.275	0.221	0.584	0.186	0.342
ResNet50+CMSI	0.230	0.572	0.282	0.229	0.591	0.193	0.350
ResNet50+CMSI+CMCI	0.242	0.572	0.294	0.242	0.585	0.205	0.357

In the bimodal fine-grained retrieval task, we observed that integrating the CMCI module generally enhanced the bimodal retrieval results compared to the baseline, as CMCI effectively integrates the channel information between different modalities. Furthermore, the addition of the CMSI module significantly improved the per-bimodal retrieval results with a 1.0% increase in average mAP. This demonstrates the effectiveness of the CMSI module in narrowing the heterogeneity gap between different modalities by facilitating spatial feature interactions. Additionally, the simultaneous integration of both modules consistently improved network performance, resulting in an average mAP improvement of 1.7%. This improvement is attributed to the enhanced interaction of spatial and channel information between different modalities through the combination of the CMSI and CMCI modules, effectively enhancing the semantic coherence between them.

Table 2. mAP scores for fine-grained cross-modality retrieval of different components in multimodality.

Models	i2all	t2all	a2all	Average
ResNet50	0.530	0.217	0.443	0.397
ResNet50+CMCI	0.532	0.219	0.447	0.399
ResNet50+CMSI	0.535	0.225	0.457	0.406
ResNet50+CMSI+CMCI	0.535	0.238	0.458	0.411

In the multimodal fine-grained retrieval task, we interestingly find that each component consistently improves the performance of the network. Compared to the baseline, the addition of CMCI improves the average mAP by 0.2% and the addition of CMSI improves the average mAP by 0.9%. Combining the results in Table 1, we find that CMSI achieves significant improvement in both bimodal fine-grained retrieval and multimodal fine-grained retrieval tasks. This is because CMSI not only narrows the heterogeneity gap between modalities through inter-modal spatial feature interactions, but it also has another benefit, i.e., it effectively motivates the network to learn discriminative features within modalities by means of mask coverage. In other words, CMSI effectively solves the problem of small inter-class differences but large heterogeneity gaps across modalities in cross-modal fine-grained retrieval, which is consistent with our previous analysis.

We also observe that by further adding CMCI to CMSI, the mAP of multimodal fine-grained retrieval reaches 41.1%. This indicates that CMCI can effectively improve the inter-class separability and intra-class compactness of multimodal in both bimodal fine-grained retrieval and multimodal fine-grained retrieval tasks. Therefore, the proposed DCINet approach combines CMSI and CMCI to solve the problem of small inter-class differences but large heterogeneity gaps between different modalities, and also alleviate the problem of large intra-class differences but small differences in data representation within the same modality.

Table 3. mAP scores of CMSI and CMCI embedded after different layers of ResNet50.

Models	Layer	i2t	i2a	t2i	t2a	a2i	a2t	Average
DCINet	layer1	0.222	0.559	0.276	0.223	0.580	0.186	0.341
	layer2	0.220	0.562	0.275	0.222	0.580	0.187	0.341
	layer3	0.224	0.574	0.279	0.229	0.583	0.191	0.347
	layer4	0.242	0.572	0.294	0.242	0.585	0.205	0.357

To assess the optimal performance of the CMSI and CMCI modules, we tried to access them behind different layers of ResNet50 and the results are presented in Table 3. It can be observed that the effect of CMSI and CMCI is more significant when they are embedded in deeper layers. When they are embedded behind layer 4, the average mAP reaches 35.7%, which is the optimal performance.

Table 4. DCINet mAP scores on different backbone Cross-media Common Representation Networks.

Models	backbone	i2t	i2a	t2i	t2a	a2i	a2t	Average
DCINet	ResNet50	0.242	0.572	0.294	0.242	0.585	0.205	0.357
	VGG16	0.121	0.460	0.165	0.127	0.480	0.100	0.242
	VGG19	0.151	0.480	0.194	0.149	0.499	0.121	0.266
	Transformer	0.198	0.467	0.253	0.160	0.501	0.136	0.286

In Table 4, we show the mAP scores of DCINet on different backbone Cross-media Common Representation Networks. It is observed that when Transformer is used as the

Cross-media Common Representation Network, the performance is significantly better than the two backbone models, VGG16 and VGG19. It should be noted that due to the dimensionality issue, when Transformer is used as the backbone, we first use bilinear interpolation to dimensionally transform the output features of Transformer, and then use the Dual Cross-modal Interaction Module to perform the computation. As mentioned above, Cross-media Common Representation Network can be replaced using common backbone networks.

In Fig. 5, we plot the variation of each loss curve of DCINet during training. Through observation, we can find that with the increase of Epoch, the loss curve of DCINet gradually flattens out, indicating that the model gradually converges. In Fig. 6, we show the change of the loss curve of each model in the ablation experiment during the training process. We find that DCINet's loss float is slightly higher than the other models in the first 30 training rounds, but converges significantly after 30 rounds. This further validates the feasibility of the model.

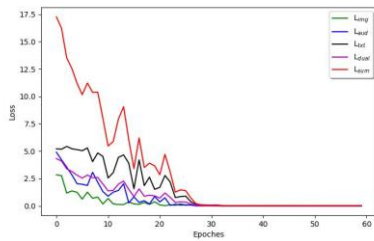


Fig. 5. Changes in each loss profile of the DCINet model.

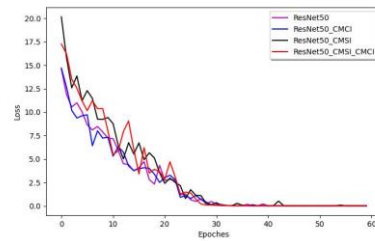


Fig. 6. Change in loss profile for each model in the ablation experiment.

5 Conclusion

In this paper, we propose a new dual cross-modal interaction-guided common representation network (DCINet), aiming to effectively bridge the large heterogeneity gap between different modalities and improve intra-class separability within modalities. The spatial and channel features between different modalities are targeted through CMSI and CMCI strategies to interact and refine feature information from multiple dimensions. By fully integrating the advantages of each module, we are able to effectively perform fine-grained cross-modal data retrieval. Comprehensive experimental results show that our approach has significant effectiveness in fine-grained cross-media retrieval tasks. In the future, we would like to further explore the retrieval efficiency and accuracy of fine-grained cross-modal retrieval tasks and conduct experiments on a wider range of datasets for better application in real-world environments.

Acknowledgments. This research is supported by the National Key Research and Development Program of China (2023YFB3308500). We would also like to thank our tutor for the careful guidance and all the participants for their insightful comments.

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

References

1. Shen, Y., Sun, X., Wei, X. S., Hu, H., & Chen, Z.: A channel mix method for fine-grained cross-modal retrieval. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 01-06. IEEE, (July 2022).
2. He, X., Peng, Y., & Xie, L.: A new benchmark and approach for fine-grained cross-media retrieval. In: Proceedings of the 27th ACM international conference on multimedia, pp. 1740-1748. (October 2019).
3. Shan, W., Huang, D., Wang, J., Zou, F., & Li, S.: Self-attention based fine-grained cross-media hybrid network. *Pattern Recognition*, 130, 108748 (2022).
4. Yao, T., Wang, R., Wang, J., Li, Y., Yue, J., Yan, L., & Tian, Q.: Efficient Supervised Graph Embedding Hashing for large-scale cross-media retrieval. *Pattern Recognition*, 145, 109934 (2024).
5. Li, L., Shu, Z., Yu, Z., & Wu, X. J.: Robust online hashing with label semantic enhancement for cross-modal retrieval. *Pattern Recognition*, 145, 109972 (2024).
6. Li, F., Wang, B., Zhu, L., Li, J., Zhang, Z., & Chang, X.: Cross-Domain Transfer Hashing for Efficient Cross-modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).
7. Zhang, D., Wu, X. J., & Chen, G.: ONION: Online Semantic Autoencoder Hashing for Cross-Modal Retrieval. *ACM Transactions on Intelligent Systems and Technology*, 14(2), 1-18 (2023).
8. Li, D., Dimitrova, N., Li, M., & Sethi, I. K.: Multimedia content processing through cross-modal association. In: Proceedings of the eleventh ACM international conference on Multimedia, pp. 604-611. (November 2003).
9. Wang, K., Wang, W., He, R., Wang, L., & Tan, T.: Multi-modal sub-space learning with joint graph regularization for cross-modal retrieval. In: 2013 2nd IAPR Asian Conference on Pattern Recognition, pp. 236-240. IEEE, (November 2013).
10. Liu, J., Yang, M., Li, C., & Xu, R.: Improving cross-modal image-text retrieval with teacher-student learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8), 3242-3253 (2020).
11. Yu, H., Yao, F., Lu, W., Liu, N., Li, P., You, H., & Sun, X.: Text-image matching for cross-modal remote sensing image retrieval via graph neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 812-824 (2022).
12. Huang, X., Peng, Y., & Yuan, M.: MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Transactions on Cybernetics*, 50(3), 1047-1059 (2018).
13. Wang, Q., Guo, Y., & Yao, Y.: DBFC-Net: a uniform framework for fine-grained cross-media retrieval. *Multimedia Systems*, 28(2), 423-432 (2022).
14. Bai, J., Yao, Y., Wang, Q., Zhou, Y., Yang, W., & Shen, F.: Multi-model network for fine-grained cross-media retrieval. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 187-199. Springer International Publishing, Cham (October 2020).

15. Chen, Q., Zhang, Y., Liu, J., Wang, Z., Deng, X., & Wang, J.: Multi-modal Fine-grained Retrieval with Local and Global Cross-Attention. In: 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), pp. 1-7. IEEE, (July 2023).
16. Hong, J., Luo, H., Yao, Y., & Tang, Z.: Generative adversarial and self-attention based fine-grained cross-media retrieval. In: Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing, pp. 1-8 (December 2020).
17. Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6023-6032 (2019).
18. Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S.: The caltech-ucsd birds-200-2011 dataset. (2011).
19. Gröchenig, K.: Foundations of time-frequency analysis. Springer Science & Business Media (2013).
20. Wu, Z., Jiang, Y. G., Wang, X., Ye, H., & Xue, X.: Multi-stream multi-class fusion of deep networks for video classification. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 791-800. ACM, New York, NY, USA (2016).
21. Zhu, C., Tan, X., Zhou, F., Liu, X., Yue, K., Ding, E., & Ma, Y.: Fine-grained video categorization with redundancy reduction attention. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 136-152. (2018).
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L.: Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 211-252 (2015)