Multi-modal Fine-grained Retrieval with Local and Global Cross-Attention

1st Qiaosong Chen
Key Laboratory of Data Engineering and Visual Computing
Chongqing University of Posts and Telecommunications
Chongqing 400065, P.R.China
chenqs@cqupt.edu.cn

2nd Ye Zhang
Key Laboratory of Data Engineering and Visual Computing
Chongqing University of Posts and Telecommunications
Chongqing 400065, P.R.China
S210231266@stu.cqupt.edu.cn

3rd Junzhuo Liu
Key Laboratory of Data Engineering and Visual Computing
Chongqing University of Posts and Telecommunications
Chongqing 400065, P.R.China
ljz15231510075@163.com

4th Zhixiang Wang
Department of Radiation Oncology (Maastro), GROW-School
Maastricht University Medical Centre+, Maastricht
The Netherlands.
zhixiang.wang@maastro.nl

5th Xin Deng
Key Laboratory of Data Engineering and Visual Computing
Chongqing University of Posts and Telecommunications
Chongqing 400065, P.R.China
dengxin@cqupt.edu.cn

6th Jin Wang
Key Laboratory of Data Engineering and Visual Computing
Chongqing University of Posts and Telecommunications
Chongqing 400065, P.R.China
wangjin@cqupt.edu.cn

***Abstract*—The goal of cross-modal retrieval is that the user gives any sample as a query sample, and the system retrieves and feeds back various modal samples related to the query sample. At present, the cross-modal retrieval method mainly focuses on coarse-grained, which is far from being satisfied in practical application. However, there are many difficulties in fine-grained retrieval, such as the heterogeneous gap and semantic gap between multi-modal data, the difficulty of similarity measurement, and the small difference in fine-grained sample features. To overcome these limitations, we propose a novel multi-modal fine-grained retrieval method with the LAGC-Attention module, which can fully extract and fuse feature information from different modalities and represent them in a common space. Specifically, we use local and global cross self-attention to extract the neighboring and global context information for each single modal data, which greatly enhances the feature representation capability of each modality (image, text, audio, video), and especially reduce the gap between different feature distributions. Finally, Extensive experiments and ablation studies demonstrate that our method achieves state-of-the-art on the public dataset PKU FG-XMedia.**

***Index Terms*—Cross-media fine-grained retrieval, Local and global cross-attention, Heterogeneity gap**

I. Introduction

with the development of society, the way people see and understand the world has undergone many changes. In the current era of big data, images, texts, audios and videos are the main medium for information exchange in people's daily life. With the exponential growth of these multimedia data, it is necessary to establish efficient and accurate cross-modal retrieval technology. As shown in Fig. 1, when the user inputs data of any modal type, the cross-modal fine-grained retrieval technology can return retrieval results of other modal types. However, current cross-modal retrieval tasks usually focus on coarse-grained, which is far from being satisfied in the needs of practical applications. In contrast, whether in industry or academia, fine-grained retrieval has greater application needs and research value. Therefore, fine-grained retrieval has become an important research direction!

The biggest difficulty in cross-modal fine-grained retrieval is the difference in data heterogeneity. The data representations from different modalities are inconsistent and belong to different feature spaces [1]–[8].Therefore, the mainstream ideas of existing cross-modal fine-grained retrieval methods are generally similar. They map different types of input data into the embedding space of common features, calculating retrieval results based on the similarity between the features of the input query and the candidate features in the database [7]–[12].

In the past cross-modal retrieval methods, different neural networks are generally used to extract features for different modalities, or only one backbone network is used to extract feature vectors of different modalities at the same time [7], [8], [13]. For examples, [11] established strongly supervised learning between sample images and text by preprocessing the input information, and performed multi-modal representation learning for fine-grained cross-modal retrieval. [9] adopted an attention space training method to learn common representa-

This work is supported by the National Key Research and Development Program of China (No.2022YFE0101000).

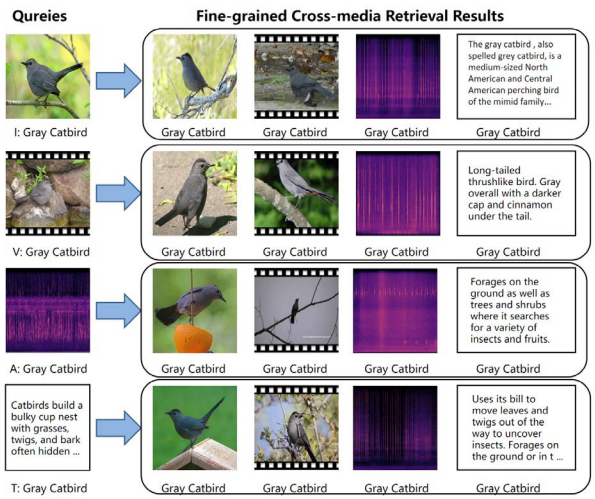

Fig. 1. Overview of cross-media fine-grained retrieval. From left to right, when the user inputs an image (or video, audio, text) of a certain category, he or she will receive the data of other media types.

tions between different models. Specifically, they introduced local self-attention layers and a similarity stitching method to understand the content relationship between features. [10], [12] algorithm model consists of two networks, a private network, and a public network. The proprietary network extracts unique features for each modality separately and obtains an accurate feature representation, whereas the public network aims to extract common features of four modalities. [9], [11] focus on the use of modal-specific information, but it is difficult to extract the connections and commonalities between different modalities. [10], [12] focus on extracting the connections and commonalities between modalities, but the commonalities and connections are only a small part of the entire information, which results in a large amount of effective modal-specific information loss.

In response to the above problems, we propose a novel multi-modal fine-grained retrieval method with the LAGC-Attention module, which can effectively extract the features of multi-modal samples and perform cross-modal fine-grained retrieval. Specifically, we use local and global cross self-attention to extract the neighboring and global context information for each single modal data, which greatly enhances the feature representation capability of each modality (image, text, audio, video), and especially reduce the gap between different feature distributions. By combining modality-specific features and modality-common features and projecting them into a common space, clustering and matching are performed to accomplish cross-modal fine-grained retrieval tasks.

Compared with previous works, the biggest advantage of our method is that it fully considers the modality-specific feature extraction process, and significantly utilizes local and global attention information for feature enhancement and integration. In this work, we adapt CNN-based Resnet50 as the backbone. In CNN, the reception field of its convolution kernel is usually small. Although the residual structure is used for stacking, it is not efficient in capturing global feature information, and the lack of fine-grained feature information often directly affects the retrieval results. Therefore, we adapt the multi-head attention mechanism, which utilizes (Key, Query, Value) triples to compute attention matrix with global context information. Due to the inherent position information in the CNN feature extraction process, we do not use additional position encoding. In the mode-specific feature extraction process, the LAGC-Attention module first obtains local feature information through convolution operations, then extracts global feature information, and finally uses channel attention to correct feature. During the extraction of modality-common feature, channel attention is also used to filter and enhance features. It is worth noting that the above processes are completed simultaneously. Finally, the model fuses the obtained single-modal specific feature and multi-modal common feature, and maps them into the common space.

To summarize, we make the following contributions.

- We propose a novel multi-modal fine-grained retrieval method with the LAGC-Attention module. It makes full use of local and global feature information and can be adapted to cross-modal fine-grained retrieval tasks in different application scenarios.
- Our newly proposed LAGC-Attention module is a feature-aggregation architecture that is suitable for any type of data for fine-grained feature extraction.
- Proved by ablation experiments and comparative experiments, our method achieves the best retrieval results on the public dataset.

II. Related works

A. Cross-modal representation learning

Cross-modal retrieval has received a lot of attention in the past many years [9]–[12], [14]–[18], but its core idea remains unchanged. Taking the method based on deep neural network(DNN) as an example, it makes full use of the powerful feature extraction ability of DNN to extract effective representations of samples from different modalities. It maps the extracted feature vectors into common space, and then establishes semantic associations of different modalities in the common space [1]. Among them, the excellent multi modal representation learning method can extract effective information from samples of different modalities, so that the feature vector contains rich semantic information in the original samples, which can greatly improve the accuracy of the subsequent retrieval work. For example, [15] adopts a knowledge-transfer approach to transfer single-modal source-domain feature representations to multi modal target domains to jointly learn common representations of multi modal data. [14] employs sparse and semi-supervised regularization to learn a joint representation of multi modal data, which not only reduces the dimension of the original features, but also incorporates cross modal correlations into the final representation. It accomplishes cross-modal retrieval tasks on datasets of up to 5 modality types with great success. [17] proposes a

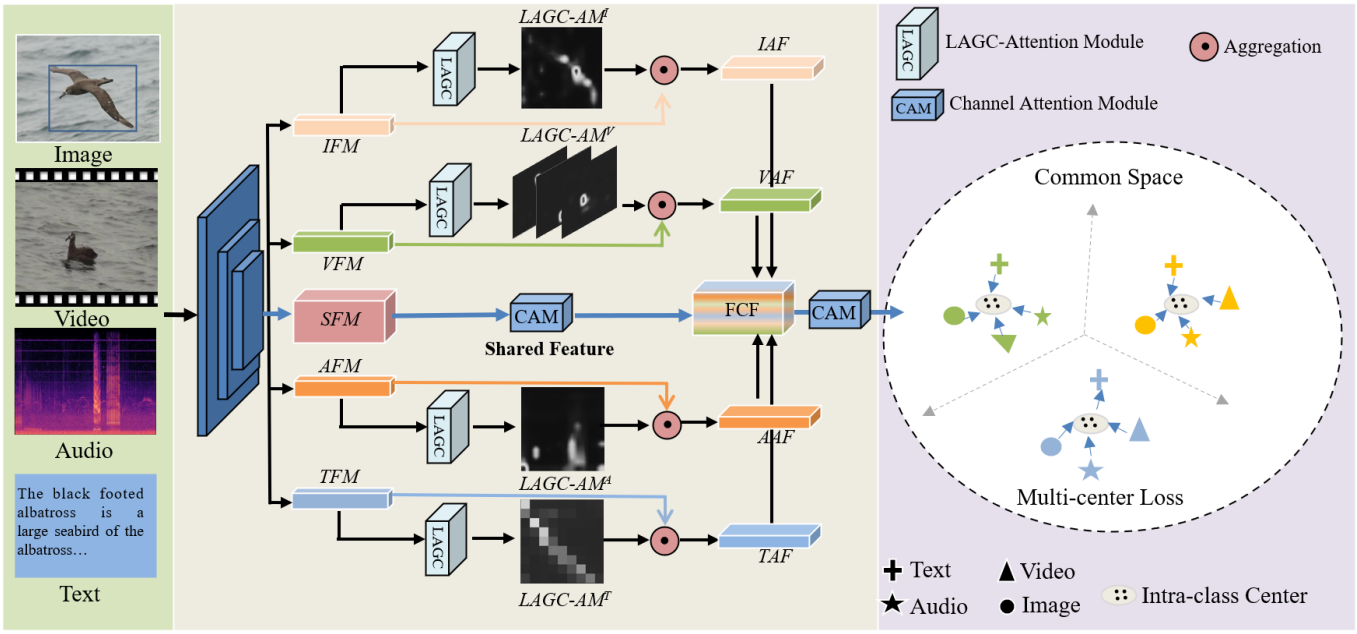

Fig. 2. A schematic illustration of our proposed method with LAGC-Attention module for cross-modal retrieval. The letter I stands for Image, and the letter V stands for Video, and the letter A stands for Audio, and the letter T stands for Text. So the acronym IFM stands for Image Feature Map, the acronym IAF stands for Image Aggregation Feature, and the acronym $LAGC-AM^I$ stands for Local And Global Cross-Attention Map. The rest modal abbreviations representation is in the same way. Besides, the acronym SFM stands for Shared Feature Map, and the acronym FCF stands for Fused Common Feature.

simple yet effective general hashing framework, which can be applied to all different scenarios while maintaining the semantic distance between data points. The method first learns optimal hash codes for both modalities simultaneously to preserve semantic similarity between data points, and then learns a hash function to map features to hash codes. Different from the above methods, our method adopts an end-to-end structure to efficiently extract modality-specific features with local and global cross-attention, and correct the final common feature representation with channel attention.

B. Multi-head Self-attention

With the success of the transformer structure in the field of Natural Language Processing (NLP) [19], [20], the self-attention mechanism has also received extensive attention in the field of Computer Vision (CV) [21]–[28]. First of all, [21] augment convolutional networks with self-attention by concatenating convolutional feature maps with a set of feature maps generated by a novel relative self-attention mechanism. After the Vision Transformer (ViT) was proposed [22], new network models based on the self-attention mechanism emerge in an endless stream. [23] further proposes a hierarchical Transformer structure, and uses a shifted window to calculate the attention representation. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. It successfully applies the ViT structure to a variety of visual downstream tasks and achieves excellent results. Our work fully considers the advantages and disadvantages of CNN and Transformer structures: CNN can effectively extract local feature information of images, while the self-attention mechanism in transformer can effectively extract global feature information of images. Therefore, we combine their advantages to design local and global cross-attention, and apply it to cross-modal fine-grained retrieval task.

III. Methodology

In this section, we first introduce the whole framework of our method in Fig. 2, then introduce the LAGC-Attention module proposed in detail, and finally introduce the loss function.

A. Framework Overview

In Fig. 2, our method can be divided into two parts, modal specific feature extraction part, and modal shared feature extraction part. It is noteworthy that they use a unified backbone network. In the modal specific feature extraction part, we input the data of four modalities into the backbone to calculate the feature map of each modal(IFM, VFM, AFM, and TFM). Then the feature map of every modal will be sent to the LAGC-Attention module respectively. Finally, through the local and global cross-attention mechanism, the aggregation feature map will be calculated and received(IAF, VAF, AAF, and TAF). In the modal shared feature extraction part, we input the data of four modalities into the backbone to calculate the shared feature map(SFM), through the Channel Attention Module(CAM) to calculate the shared feature. Finally, the shared feature of all modalities and the aggregation feature

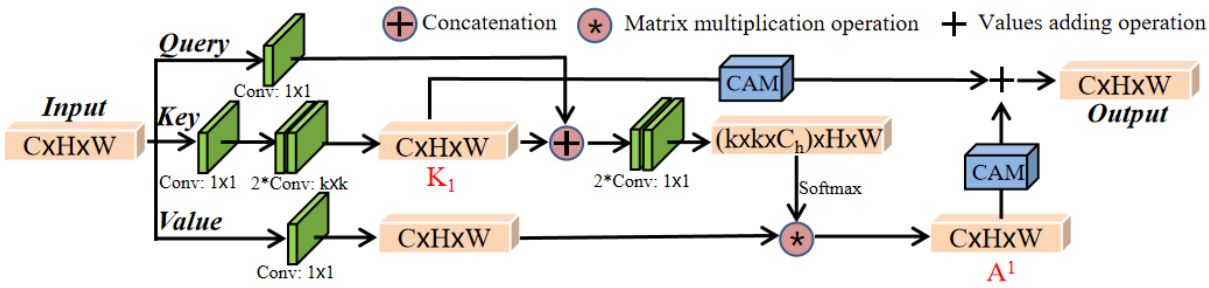

Fig. 3. The architecture of our proposed LAGC-Attention module.

of each modal will be concatenated together through CAM, and map them into a common space.

For example, we assume that there are a total of n pairs of quaternary samples consisting of image-video-audio-text data, set as $\phi = \left(x_i^I, x_i^V, x_i^A, x_i^T\right)_{i=1}^n$, where x_i^I is an image sample, x_i^V is a video sample, x_i^A is an audio sample, and x_i^T is a text sample. They all belong to the $i-th$ category of samples. For each quaternary sample pair, they share a label $y_i \in R^c$, where c represents the total number of categories. After the data samples of all modalities are fed into the backbone, two kinds of loss functions are adopted for supervision, which are the cross-entropy loss function and the multi-center loss function.

B. Visual Representation Learning With Attention

We use Resnet50 as the backbone and load its pretrained weights. Resnet50 is a very common backbone in computer vision, with excellent feature extraction capability. Fig. 3 shows the structure of our proposed LAGC-Attention module in detail. It is obvious that our proposed attention module is more focused on the construction of local and global attention, which is more conducive to extracting features of fine-grained data samples. Our attention module will not change the dimension of the input feature, so it can be expressed as:

$$y = A_{att}(x) \tag{1}$$

Suppose we have input feature map $x \in R^{c\times h\times w}$, and the output feature is $y \in R^{c\times h\times w}$, where h, w, c are the height, width and channel of the feature map. The A_{att} is feature mapping matrix of the LAGC-Attention module. Taking the image feature map $x \in R^{c\times h\times w}$ as an example, the keys, queries, values are encoded via 1×1 convolution operation:

$$Key = W_k(x), Query = W_q(x), Value = W_v(x) \tag{2}$$

The key map will employs twice $k \times k$ convolution calculation to extract local contextual feature information firstly. It is worth noting that its feature dimension has not changed, and we defined it as K_1. After that, K_1 is the concatenation of query map. By this way, the attention matrix A is calculated based on the query map and the key map at the same time. In other words, local contextual attention is added to the subsequent global self-attention learning process successfully. It is realized via twice 1×1 convolution operation:

$$A \in R^{(k\times k\times C_h)\times h\times w} = W_2W_1(x) \tag{3}$$

$A \in R^{(k\times k\times C_h)\times h\times w}$ means that A's each spatial position contains a $k \times k \times C_h$ vector, that consists of C_h local query-key relation maps for all heads, where C_h is the head number, and W_2, W_1 means the two mapping matrices. Next, the global attention matrix A^1 is calculated by normalizing the attention matrix A with Softmax operation along channel dimension for each head.

$$A^1 = V \circledast Softmax(A) \tag{4}$$

The $\circledast$ denotes the matrix multiplication operation. Finally, the K_1 and A^1 will be aggregated by the CAM module. After the values adding operation, the final output y of our LAGC-Attention module will be calculated. Note that y is calculated via local and global attention, which is working on spatial level. So the CAM module will correct the channel level feature information.

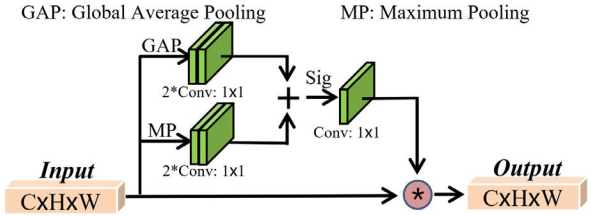

Fig. 4. The architecture of Channel Attention module.

Channel attention is a common attention mechanism, which can correct the local and global attention features of fine-grained data samples [29]. The corrected features can retain valuable features and eliminate unworthy ones. It is can be divided into three steps, Maximum Pooling(MP), Global Average Pooling(GAP), and Excitation operation. Taking the A^1 as an example, the equation is as follows:

$$y = Sigmoid(MP(A^1) + GAP(A^1)) \circledast A^1 \tag{5}$$

TABLE I
COMPARISON TO THE STATE-OF-ARTS FOR CROSS-MODAL RETRIEVAL(ONE-TO-ONE) ON PKU FG-XMEDIA DATASET.

Method	I→T	I→A	I→V	T→I	T→A	T→V	A→I	A→T	A→V	V→I	V→T	V→A	Average
CMDN [16]	0.099	0.009	0.377	0.123	0.007	0.078	0.017	0.008	0.010	0.446	0.081	0.009	0.105
GSPH [17]	0.140	0.098	0.413	0.179	0.024	0.109	0.129	0.024	0.073	0.512	0.126	0.086	0.159
JRL [14]	0.160	0.085	0.435	0.190	0.028	0.095	0.115	0.035	0.065	0.517	0.126	0.068	0.160
ACMR [18]	0.162	0.119	0.477	0.075	0.015	0.081	0.128	0.028	0.068	0.536	0.138	0.111	0.162
MHTN [15]	0.116	0.195	0.281	0.124	0.138	0.185	0.196	0.127	0.290	0.306	0.182	0.306	0.204
FGCrossNet [11]	0.210	0.526	0.606	0.255	0.181	0.208	0.553	0.159	0.443	0.629	0.195	0.437	0.366
DBFC-Net [12]	0.298	0.563	0.626	0.346	0.252	0.284	0.580	0.223	0.462	0.669	0.286	0.483	0.423
SAFGCM [9]	0.293	0.625	0.636	0.335	0.263	0.272	0.629	0.231	0.495	0.677	0.270	0.525	0.437
FGCMR [10]	0.355	0.629	0.660	0.409	0.324	0.335	0.643	0.287	0.515	0.706	0.335	0.544	0.478
Ours	**0.357**	**0.635**	**0.646**	**0.431**	**0.328**	**0.352**	**0.675**	**0.313**	**0.611**	**0.691**	**0.343**	**0.601**	**0.499**

TABLE II
COMPARISON TO THE STATE-OF-ARTS FOR CROSS-MODAL RETRIEVAL(ONE-TO-ALL) ON PKU FG-XMEDIA DATASET.

Method	I→ALL	T→ALL	A→ALL	V→ALL	Average
ACMR [18]	0.245	0.039	0.041	0.279	0.151
CMDN [16]	0.321	0.071	0.016	0.229	0.159
JRL [14]	0.344	0.080	0.069	0.275	0.192
GSPH [17]	0.387	0.103	0.075	0.312	0.219
MHTN [15]	0.208	0.142	0.237	0.341	0.232
FGCrossNet [11]	0.549	0.196	0.416	0.485	0.412
DBFC-Net [12]	0.602	0.284	0.545	0.461	0.474
SAFGCM [9]	0.618	0.270	0.546	0.497	0.482
FGCMR [10]	0.637	0.333	0.577	0.509	0.514
Ours	**0.651**	**0.342**	**0.585**	**0.544**	**0.531**

C. Training Objectives

The role of the loss function is to correctly guide the network structure parameters for effective learning. It is composed of a supervised learning cross-entropy loss and a multi-center loss.

(1) The supervised loss function classification constraints are defined as follows:

$$\begin{aligned} L_{cls} = & \frac{1}{N_I}\sum_{k=1}^{N_I} l\left(x_k^I, y_k^I\right) + \frac{1}{N_V}\sum_{k=1}^{N_V} l\left(x_k^V, y_k^V\right) \\ & + \frac{1}{N_A}\sum_{k=1}^{N_A} l\left(x_k^A, y_k^A\right) + \frac{1}{N_T}\sum_{k=1}^{N_T} l\left(x_k^T, y_k^T\right) \end{aligned} \tag{6}$$

In formula (6), $l\left(x_k, y_k\right)$ is the cross-entropy loss function, *I*, *T*, *V* and *A* represent media types for image, text, video, and audio, respectively. Taking text as an example, N_T is the number of samples of text data in the training set, y_k^T denotes the label of the $k-th$ text data, x_k^T denotes the feature of $k-th$ text data.

(2) The idea of the multi-center loss function can be summarized as follows: it first gathers the samples of the same modality and the same category around the intra-class center of the modality. Then by narrowing the distance between the intra-class centers of different modalities and the same category, the samples of different modalities and the same category are indirectly gathered together. The equation of multi-center loss is defined as follows:

$$\mathcal{L}_{\mathrm{m}} = \frac{1}{4}\sum_{m=1}^{\mathrm{I,V,A,T}}\sum_{i=1}^{N}\left(\left\|\boldsymbol{x}_i^m - \boldsymbol{c}_y^m\right\|_2^2 + \mu D\left(\boldsymbol{C}_y\right)\right) \tag{7}$$

In formula (7), x_i^m is the $i-th$ sample vector of m modal. c_y^m is the intra-class center of the sample vector for m modal. C_y stands for intra-class center of all modalities under y category. μ is the weight parameters. $D(\cdot)$ is the distance function of intra-class center, whose equation is as follows:

$$D\left(\boldsymbol{C}_y\right) = \frac{1}{4}\sum_{m_2=1}^{\mathrm{I,V,A,T}}\left\|\boldsymbol{c}_y^m - \boldsymbol{c}_y^{m_2}\right\|_2^2 \tag{8}$$

In formula (8), the c_y^m and $c_y^{m_2}$ represent the intraclass centers of different modalities under the same category. So the total loss function can be defined as:

$$L_{total} = \alpha L_{cls} + \beta L_m \tag{9}$$

IV. EXPERIMENTS

A. Dataset

To verify the effectiveness of the method, we conduct experiments on the public dataset PKU FG-XMedia. As far as we know, it is the only cross-modal fine-grained retrieval dataset [11]. To eliminate the heterogeneity of different modal data types and improve the performance of the algorithm. We preprocess the data for the four modalities separately. For image data, to avoid being interfered with by background noise, before being input to the network, it will be cropped according to the pixel coordinates of the target frame, so only a part containing the target will be intercepted. For audio data, it is converted to a spectrogram by using Short-time Fourier Transform. For video data, it uses the method of frame extraction, and each video extracts 10 frames of images to represent the video samples. For text data, it is transformed into a one-dimensional word vector through word embedding, then spliced into an image-like matrix form through convolution operation. In this way, the four types of data can be organized into a four-dimensional matrix and input to the backbone network. The experiments use the mean average precision (MAP) score to measure its performance. The MAP query score is calculated by the average of the average accuracy of each query sample.

TABLE III
ABLATION STUDY FOR OUR METHOD ON PKU FG-XMEDIA DATASET(ONE-TO-ONE).

Method	I→T	I→A	I→V	T→I	T→A	T→V	A→I	A→T	A→V	V→I	V→T	V→A	Average
W/O both	0.235	0.537	0.618	0.322	0.221	0.248	0.595	0.255	0.481	0.678	0.311	0.529	0.419
W/O CAM	0.341	0.613	0.622	0.407	0.302	0.374	0.654	0.307	0.584	0.669	0.343	0.588	0.483
Ours	**0.357**	**0.635**	**0.646**	**0.431**	**0.328**	**0.352**	**0.675**	**0.313**	**0.611**	**0.691**	**0.343**	**0.601**	**0.499**

TABLE IV
ABLATION STUDY FOR OUR METHOD ON PKU FG-XMEDIA DATASET(ONE-TO-ALL).

Method	I→ALL	T→ALL	A→ALL	V→ALL	Average
W/O both	0.581	0.309	0.517	0.499	0.476
W/O CAM	0.643	0.325	0.564	0.528	0.515
Ours	**0.651**	**0.342**	**0.585**	**0.544**	**0.531**

B. Retrieval Tasks

To verify the effectiveness of our method, we compare our method with state-of-the-art methods on the PKU FG-XMedia dataset. Retrieving other media data using any modality as input data can be divided into one-to-one retrieval and one-to-all retrieval. For example, in a one-to-one retrieval experiment, if a user submits a picture of Herring Gull, he will get a sample of the Herring Gull video, which is called Image Retrieval Video (I→V). Similarly, the video retrieval image can be represented as (V→I). Therefore, one-to-one fine-grained cross-media retrieval includes I→V, I→T, I→A, V→I, V→T, V→A, T→I, T→V, T→A, A→ I, A→V, A→T. In the one-to-all retrieval experiment, if the user submits a picture of Herring Gull, he will get samples of audio, video, and text about Herring Gull, which is called image retrieval and all remaining media types (I→all). Similarly, the other three modalities' one-to-all retrieval tasks can be expressed as V→All, T→All, and A→All.

C. Implement Details

We use resnet50 as the backbone network. After the data of the four modalities are preprocessed, the dimension of the data sample will be fixed at $448 \times 448 \times 3$. The entire program code is written by Pytorch, and the graphics card is RTX A5000. During training, we set the batchsize as 16, and the initial learning rate is set to 0.0001, and AdamW is selected by the optimizer, and the learning rate schedule is the cosine learning rate with a warm-up of 1000 steps.

D. Experimental Results and Ablation Analysis

Tables I and II show the experiment results of one-to-one retrieval and one-to-all retrieval, respectively. It can be seen that compared with other methods, our method has a obvious degree of performance improvement. To further demonstrate the effectiveness of our method, we conduct ablation experiments seriously. As shown in Tables III and IV, We have conducted detailed experiments on whether to use LAGC-Attention module, including using CAM module or not. (1) *W/O both* means that both modules are not included at the same time. (2)*W/O CAM* refers to adding LAGC-Attention module without adopting the CAM module. We can observe that the retrieval MAP scores from *W/O CAM* are basically improved and the final result also reflects the correction effect of the CAM module on fine-grained features.

V. CONCLUSION

In conclusion, we propose a novel multi-modal fine-grained retrieval method with LAGC-Attention module. It makes full use of local and global feature information and can be adapted to cross-modal fine-grained retrieval tasks in different application scenarios. We have done a lot of experiments on the public dataset PKU FG-XMedia, and our method achieved state-of-the-art. To further demonstrate the reliability of our method, we conduct ablation experiments seriously. However, in the course of the experiment, we also observe some unsolved problems. For example, MAP scores involving text modal are generally poor and we will try more datasets in the future to prove the effectiveness of our method comprehensively!

REFERENCES

[1] X. He and Y. Peng, "Fine-grained visual-textual representation learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 520–531, 2019.

[2] P. Huang, T. Li, G. Gao, Y. Yao, and G. Yang, "Collaborative representation based local discriminant projection for feature extraction," *Digital Signal Processing*, vol. 76, pp. 84–93, 2018.

[3] Y. Yao, J. Zhang, F. Shen, W. Yang, X.-S. Hua, and Z. Tang, "Extracting privileged information from untagged corpora for classifier learning." in *IJCAI*, 2018, pp. 1085–1091.

[4] Y. Yao, F. Shen, J. Zhang, L. Liu, Z. Tang, and L. Shao, "Extracting multiple visual senses for web learning," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 184–196, 2018.

[5] ——, "Extracting privileged information for enhancing classifier learning," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 436–450, 2018.

[6] Y. Yao, W. Yang, P. Huang, Q. Wang, Y. Cai, and Z. Tang, "Exploiting textual and visual features for image categorization," *Pattern Recognition Letters*, vol. 117, pp. 140–145, 2019.

[7] L. Ying, G. Yingying, F. Jie, F. Jiulun, H. Yu, and L. Jiming, "Survey of research on deep learning image-text cross-modal retrieval," *Journal of Frontiers of Computer Science & Technology*, vol. 16, no. 3, p. 489, 2022.

[8] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] C. Wang, Y. Yao, Q. Wang, and Z. Tang, "Local self-attention on fine-grained cross-media retrieval," in *ACM Multimedia Asia*, 2021, pp. 1–7.

[10] J. Bai, Y. Yao, Q. Wang, Y. Zhou, W. Yang, and F. Shen, "Multi-model network for fine-grained cross-media retrieval," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2020, pp. 187–199.

[11] X. He, Y. Peng, and L. Xie, "A new benchmark and approach for fine-grained cross-media retrieval," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1740–1748.

[12] Q. Wang, Y. Guo, and Y. Yao, "Dbfc-net: a uniform framework for fine-grained cross-media retrieval," *Multimedia Systems*, vol. 28, no. 2, pp. 423–432, 2022.

[13] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao *et al.*, "Vision-language pre-training: Basics, recent advances, and future trends," *Foundations and Trends® in Computer Graphics and Vision*, vol. 14, no. 3–4, pp. 163–352, 2022.

[14] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2013.

[15] X. Huang, Y. Peng, and M. Yuan, "Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE transactions on cybernetics*, vol. 50, no. 3, pp. 1047–1059, 2018.

[16] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks." in *IJCAI*, 2016, pp. 3846–3853.

[17] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for n-label cross-modal retrieval," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4076–4084.

[18] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 154–162.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[25] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*. PMLR, 2020, pp. 1691–1703.

[26] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[27] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[28] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.