Multi-Modal Knowledge Hypergraph for Diverse Image Retrieval

Yawen Zeng[1], **Qin Jin**[2*], **Tengfei Bao**[1], **Wenfeng Li**[1]

[1] ByteDance AI Lab
[2] School of Information, Renmin University of China
yawenzeng11@gmail.com, qjin@ruc.edu.cn, {baotengfei,liwenfeng.x}@bytedance.com

Abstract

The task of keyword-based diverse image retrieval has received considerable attention due to its wide demand in real-world scenarios. Existing methods either rely on a multi-stage re-ranking strategy based on human design to diversify results, or extend sub-semantics via an implicit generator, which either relies on manual labor or lacks explainability. To learn more diverse and explainable representations, we capture sub-semantics in an explicit manner by leveraging the multi-modal knowledge graph (MMKG) that contains richer entities and relations. However, the huge domain gap between the off-the-shelf MMKG and retrieval datasets, as well as the semantic gap between images and texts, make the fusion of MMKG difficult. In this paper, we pioneer a degree-free hypergraph solution that models many-to-many relations to address the challenge of heterogeneous sources and heterogeneous modalities. Specifically, a hyperlink-based solution, Multi-Modal Knowledge Hyper Graph (MKHG) is proposed, which bridges heterogeneous data via various hyperlinks to diversify sub-semantics. Among them, a hypergraph construction module first customizes various hyperedges to link the heterogeneous MMKG and retrieval databases. A multi-modal instance bagging module then explicitly selects instances to diversify the semantics. Meanwhile, a diverse concept aggregator flexibly adapts key sub-semantics. Finally, several losses are adopted to optimize the semantic space. Extensive experiments on two real-world datasets have well verified the effectiveness and explainability of our proposed method.

Introduction

Keyword-based image retrieval has been a classic task in the multimedia field, because in real-world applications (Ionescu et al. 2021), using keywords as queries is the most convenient and effective retrieval manner for users. Due to the relatively broad semantics of keywords, the diversity of retrieval results is particularly important for keyword-based image retrieval to meet the needs of users. In fact, keyword-based queries are usually short and semantically incoherent (e.g. 2.7 words on average in the Div150AdHoc dataset), which imposes higher diversity requirements than general retrieval (queries with more than 20 words) (Wang et al. 2017; Cao et al. 2020; Zeng 2022). Therefore, the task of keyword-based diverse image retrieval, which aims to diversify the retrieved image results for a given short query, has attracted much research attention.

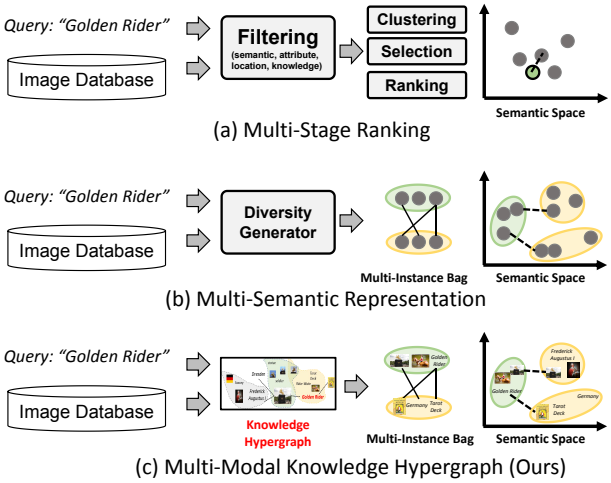

Figure 1: Examples of the multi-stage, multi-instance and our multi-modal knowledge hypergraph solution.

Previous works on keyword-based diverse image retrieval can be roughly divided into two branches: multi-stage ranking based approaches and multi-semantic representation based approaches. Multi-stage ranking based methods follow "first filter then rank", where manual strategies are designed in the ranking stage to adjust the order of retrieved images (Seddati et al. 2017; Peng et al. 2017). As shown in Fig. 1(a), such methods rely heavily on human experience and are prone to cascading errors. On the other hand, multi-semantic representation based methods directly learn diverse semantics via multiple instance learning (Song and Soleymani 2019; Zeng et al. 2022b), i.e. enrich a query with multiple sub-semantics (or called instances). However, as shown in Fig. 1(b), these type of approaches are mostly implicit, which cannot explain what an instance means (explainability), and how many instances are sufficient to cover a keyword-based query with broad semantics (quality). Along the multiple instance learning direction, in this work, we aim to learn more diverse and better semantic representations to improve both quality and explainability.

*Corresponding Author
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

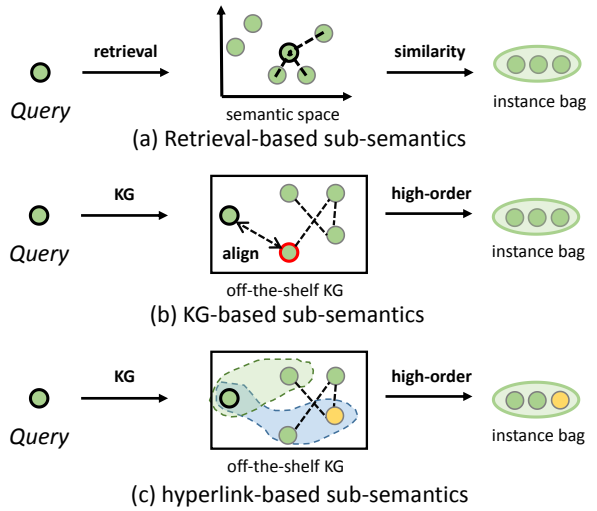

Figure 2: Strategies of explicitly modeling sub-semantics.

Intuitively, there are several ways to **explicitly** extend diverse sub-semantics end-to-end, e.g. adopting semantic similarity or external knowledge. The strategy of adopting semantic similarity, such as retrieval in Fig. 2(a), can capture similar semantic patterns, but it is difficult to cover long-tailed sub-semantics. Adopting external knowledge, especially knowledge graph (KG), can naturally expand sub-semantics via multi-hop graph propagation, as shown in Fig. 2(b). We argue that multi-modal knowledge graph (MMKG), which consists of rich multi-modal entities and relations, and is capable of connecting visual and textual (Zhu et al. 2022), is a better knowledge source for the diverse image retrieval task. Therefore, as illustrated in Fig. 1(c), our motivation stems from replacing sub-semantics in multiple instance bags with knowledge-aware multi-modal instances, which enables more diverse representations. However, it is not trivial to replace due to the heterogeneous aligning and linking of multi-modal knowledge. In fact, there is a huge domain gap between the off-the-shelf multi-modal knowledge graph and retrieval datasets, coupled with the semantic gap between images and texts. Therefore, how to link and align heterogeneous sources and heterogeneous modalities becomes a bottleneck.

To get around the quagmire of heterogeneous linking and aligning for our task, we propose a novel hyperlink-based solution to conveniently bridge heterogeneous sources and modalities, namely, **M**ulti-Modal **K**nowledge **H**yper**G**raph (**MKHG**). In fact, the degree-free hypergraph has the ability to encode more complex many-to-many relations, which is well suited for broad short queries and multi-semantic images in diverse retrieval scenarios. In this way, a query no longer needs to be aligned with an entity in the off-the-shelf MMKG as in Fig. 2(b), or construct a new local sub-graph according to relations in the MMKG. Instead, as shown in Fig. 2(c), a query is only integrated into the graph via hyperlinks of similar semantic patterns, and then diverse semantics are found with the help of high-order connectivity. This simple but effective approach is highly scalable, even capable of handling out-of-domain queries.

Specifically, our proposed MKHG consists of four key components as depicted in Fig. 3: knowledge hypergraph construction, multi-modal instance bagging, diverse concept aggregator, and semantic space optimizer. Among them, the hypergraph construction module aims to link the off-the-shelf MMKG and retrieval database via various types of hyperedges. Afterwards, the multi-modal instance bagging module explicitly selects multiple instances to diversify the semantics. Meanwhile, to guarantee the reliability of the instances, the diverse concept aggregator flexibly adapts key sub-semantics. Finally, we design several losses to help the semantic space optimization, namely graph matching loss, instance-level loss and knowledge-level loss. Experiments on two real-world benchmark datasets demonstrate that our proposed MKHG not only achieves better diverse image retrieval results, but also has better explainability.

The main contributions are summarized as follows:

- To the best of our knowledge, this is the first work that introduces a multi-modal knowledge graph to explicitly solve the task of diverse image retrieval.
- We propose a simple but effective solution to fuse the off-the-shelf knowledge graph and retrieval database via hyperlinks. Meanwhile, it has the ability to handle out-of-domain queries, which sheds some light on the representation learning for knowledge-enhanced applications.
- Extensive experiments conducted on two real-world datasets demonstrate the effectiveness and explainability of our solution.

Related Work

Diverse Image Retrieval

Existing works can be mainly divided into multi-stage ranking and multi-semantic representation methods. Most multi-stage models employ a filtering stage and a ranking stage. The filtering stage applies attribute information to obtain candidate images, and the ranking stage designs several strategies to diverse the image order. Seddati et al. (2017) provide a three-stage scheme in which irrelevant images are filtered firstly, then DBSCAN is leveraged to increase diversity in the second stage, and finally re-ranking is performed. However, such multi-stage methods are complex in design and rely more on human experience. Multi-semantic representation methods learn multiple features for short keywords, i.e. one-to-many sub-semantics (Zhao et al. 2017; Song and Soleymani 2019; Zeng et al. 2022a). The VMIG approach introduces a variational multiple instance graph, which packs multiple features in one instance bag to represent sub-semantics (Zeng et al. 2022b). Wu and Ngo (2020) design an inactive words loss to expand the semantic concepts, and Su et al. (2021) provide a dynamic intent graph to balance content and intent.

Multi-Modal Knowledge Graph

Multi-modal knowledge graph (MMKG) is an extension of plain text knowledge graph, which connects texts and images to build a more general knowledge system (Zhu et al. 2022; Zeng et al. 2021). Enriching entities and

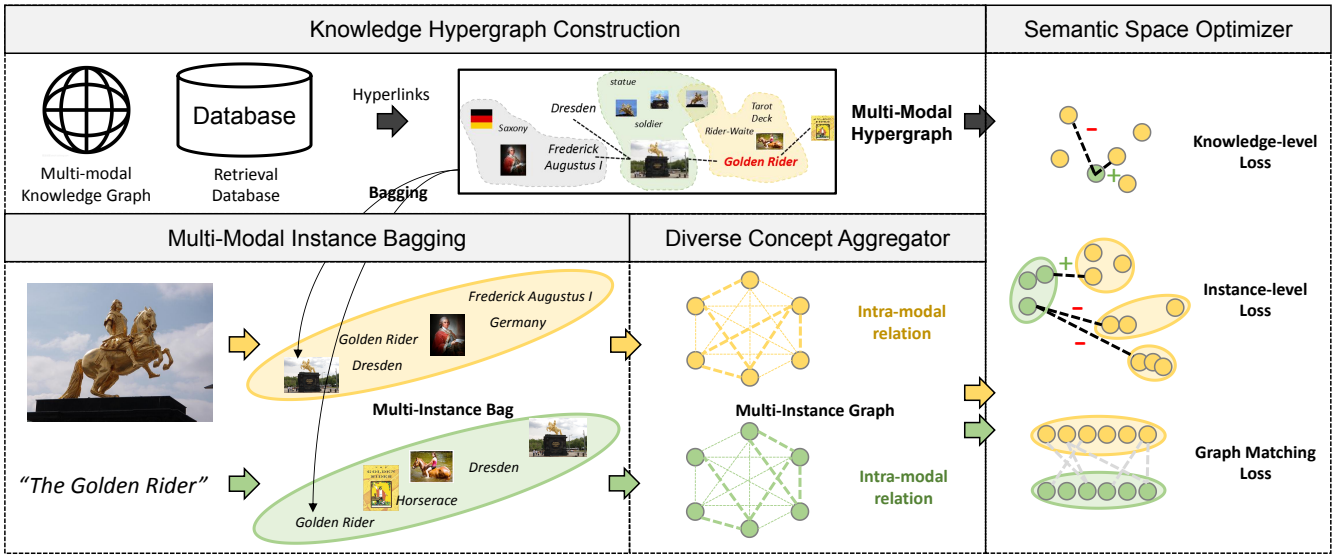

Figure 3: The overall architecture of MKHG for keyword-based diverse image retrieval. The input is a keyword-based query and images in a dataset, and the matching score is considered as the ranking score during the testing phase.

concepts in the knowledge graph can help solve the long tail problem, and multi-modal information is adopted for visual complementation and textual disambiguation. Furthermore, it can even provide commonsense knowledge to perform multi-modal reasoning, which has been successful in tasks such as VQA. Sun et al. (2020) apply the multi-modal knowledge graph in recommendation system to address cold start and data sparsity issues. Xu et al. (2021) construct an e-commerce multi-modal knowledge graph, which has been launched online in the Taobao app to serve customers. Zhao et al. (2021) introduce a multi-modal knowledge graph with external knowledge collected from the web for entity-aware image captioning task.

Proposed Method

Fig. 3 illustrates the overall framework of our method, which consists of four key components: 1) knowledge hypergraph construction; 2) multi-modal instance bagging; 3) diverse concept aggregator; and 4) semantic space optimizer. The hypergraph construction aims to link an off-the-shelf MMKG and diverse retrieval databases. Among them, we design various types of hyperedges to cover relevant sub-semantics comprehensively. The multi-modal instance bagging module then performs instance selection via higher-order relations of hyperedges to represent a keyword-based query. Notably, a mixed instance bag containing both images and texts makes the multi-semantic representation more diverse. Furthermore, the concept aggregator balances noises among the sub-semantics in instance bags. Finally, we further design multiple losses to constrain the semantic space optimization.

Knowledge Hypergraph Construction

The motivation of this work stems from modeling sub-semantics in an explicit manner with the help of MMKG, which enables more diverse representations. However, when a heterogeneous entity (e.g. a retrieval query) wants to extend sub-semantics from an off-the-shelf MMKG, the strategies of alignment and linking are necessary but labor-intensive. Furthermore, the semantic relations among visual and textual entities are often in the form of many-to-many, which is difficult to be statically captured by the one-to-one connected graph. Naturally, each entity should be connected to entities of various semantics and various modalities, and each other entity should also be associated with multiple entities. Therefore, these facts inspire us to explore the hypergraph structure. Unlike simple graphs where all edges must be of degree two, the hypergraph can utilize their degree-free hyperedges to encode higher-order semantics. In this subsection, a multi-modal knowledge hypergraph is constructed to unleash the power of MMKG.

Hypergraph Construction Firstly, we select the off-the-shelf MMKG ImageGraph (Wang, Qi, and Zheng 2020) as a base graph, which contains 15K entities with 55.8 images per entity. Formally, we define it as $\mathcal{G} = (\mathcal{X}, \mathcal{R}, \mathcal{E})$, where $\mathcal{X}, \mathcal{R}, \mathcal{E}$ refer to the set of mutli-modal entities, relations (including attributes), and edges respectively. The edges between entities are one-to-one, connecting a pair of entities by certain relation, $\mathcal{E}_i = (\mathcal{X}_{i_1}, \mathcal{R}_k, \mathcal{X}_{i_2})$.

To bridge the heterogeneous gap between the MMKG and retrieval datasets, a multi-modal knowledge hypergraph is constructed via several hyperlinks, i.e. semantic groups, to capture many-to-many relations. Formally, a hyperedge, $\mathcal{E}_i^h = (\mathcal{X}_{i_1}, ..., \mathcal{X}_{i_j})$ indicates a semantically similar group among multiple entities. In this work, the off-the-shelf MMKG will not be modified, and three types of hyperedges are added to easily and quickly obtain the groups of the same semantic pattern, i.e. visual hyperedges, textual hyperedges, and attribute hyperedges. Specifically, we flatten all images and texts in the knowledge graph $\mathcal{G}$ and the diverse retrieval dataset as visual pool and text pool. Then, the visual pool and the text pool are clustered separately to capture the same

semantic patterns, that is, the samples in the same cluster will share a hyperedge. Meanwhile, samples with the same attributes should also share an attribute hyperedge (e.g. city), which can further enrich relations among entities. Finally, an extended knowledge hypergraph $\mathcal{G}^h = (\mathcal{X}, \mathcal{R}, \mathcal{E}^h)$ is produced under this simple but efficient manner. In this way, an out-of-domain query can also be linked into a off-the-shelf graph easily, and then obtain multiple sub-semantics via higher-order propagation.

Hypergraph Representation Since higher-order correlations in the hypergraph $\mathcal{G}^h$ are complex to learn with ordinary graph structure methods, we employ a hypergraph convolutional layer to encode entities (Bai, Zhang, and Torr 2021). Specifically, the hyperedges will be concatenated to generate a hypergraph adjacency matrix $\mathbf{H}$, which is fed into a graph convolutional network for encoding along with the entity features $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_n\}$,

$$\mathbf{X}^{(l+1)} = \sigma(\mathbf{D}^{-1/2}\mathbf{H}\mathbf{O}\mathbf{B}^{-1}\mathbf{H}^T\mathbf{D}^{-1/2}\mathbf{X}^{(l)}\mathbf{P}), \quad (1)$$

where $\mathbf{O}, \mathbf{D}, \mathbf{B}, \mathbf{P}$ are the hyperedge weight matrix, vertex degree matrix, edge degree matrix and the weight matrix between the (l) layer and $(l+1)$ layer, respectively. Notably, to narrow the cross-modal semantic gap, we adopt a multi-modal extractor (Radford et al. 2021) to obtain n features including visual $\mathbf{x}^v$ and textual entities $\mathbf{x}^t$.

Multi-modal Instance Bagging

As the core of multi-semantic representation, explicitly selecting diverse instances to represent sub-semantics is critical for diverse retrieval. In fact, there are many higher-order relations in the multi-modal hypergraph that capture comprehensive sub-semantics. To simplify the process of constructing a multi-modal multi-instance bag, we directly treat nodes with nearest m higher-order connections as sub-semantics. As shown in Fig. 3, the input sample is represented as a mixed instance bag consisting of itself and other higher-order related entities. Notably, since keyword-based queries are short, applying a multi-semantic representation on textual modality is necessary, but optional on images. However, we consider that when both texts and images are modeled as multi-semantic representations, the many-to-many relations will be understood more thoroughly, which is verified in experiments. A bag of instances is denoted as $\mathcal{B}_i = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$, where m is the number of instances.

Diverse Concept Aggregator

The concept aggregator is designed to make instances more reasonable, i.e. diverse concepts are less redundant and evenly distributed. Since we roughly treat nearest m higher-order nodes as sub-semantics in an instance bag, it needs to be further refined. In particular, a small number of instances may result in insufficient diversity, while an excessive number may introduce redundancy and noise. Therefore, following (Zeng et al. 2022b), we employ a multi-instance graph to learn the relations and remove redundant instances among concepts. Formally, the local multi-instance graph is defined as $\mathcal{G}^x = (\mathcal{X}, \mathcal{E}^x)$, where instances are treated as nodes, and edges are determined by the affinity score $d(a,b) = \frac{a \cdot b}{||a||||b||}$ of the two instances. To alleviate the influence of noisy instances, two nodes are linked only if the affinity score exceeds the threshold τ. Furthermore, the message propagation (Hamilton, Ying, and Leskovec 2017) is executed to aggregate neighbors,

$$\mathbf{x}_{\mathcal{N}(\mathbf{x}_i)} = \mathrm{AGGREGATE}(\mathbf{x}_j, \forall j \in \mathcal{N}(\mathbf{x}_i))), \quad (2)$$

$$\widetilde{\mathbf{x}}_i = \sigma(\mathbf{W}_\mathbf{x} \cdot \mathrm{CONCAT}(\mathbf{x}_i, \mathbf{x}_{\mathcal{N}(\mathbf{x}_i)})), \quad (3)$$

where $\mathbf{W}_\mathbf{x}$ is a weight matrix, and the "AGGREGATE" operator is implemented through a max pooling to reduce redundancy and refine features.

Semantic Space Optimizer

To make the diverse multi-semantic features learn smoothly, we design several losses, namely graph matching loss $\mathcal{L}_{mat}$, instance-level loss $\mathcal{L}_{ins}$ and knowledge-level loss $\mathcal{L}_{kno}$.

Graph Matching Loss The graph matching measures the semantic scores of textual instance bags $\mathcal{B}_i^t$ and visual instance bags $\mathcal{B}_j^v$. Concretely, the discrepancy between each instance in the bags is measured by,

$$S(\mathcal{B}_i^t, \mathcal{B}_j^v) = \min_i \sum_j^m a_{ij}(\widetilde{\mathbf{x}}_i^t - \widetilde{\mathbf{x}}_j^v) = \min_i (\widetilde{\mathbf{x}}_i^t - \sum_j^m a_{ij}\widetilde{\mathbf{x}}_j^v), \quad (4)$$

$$a_{ij} = \frac{exp((\widetilde{\mathbf{x}}_i^t)^T \widetilde{\mathbf{x}}_j^v))}{\sum_j exp((\widetilde{\mathbf{x}}_i^t)^T \widetilde{\mathbf{x}}_j^v))}. \quad (5)$$

where $S(\mathcal{B}_i^t, \mathcal{B}_j^v)$ is the matching score, m is the number of instances. For a textual bag of a query $\mathcal{B}_i^t$, we calculate the attention score a_{ij} of each instance relative to the visual instance bag. Then graph matching loss $\mathcal{L}_{mat}$ is defined as the triplet form,

$$\mathcal{L}_{mat} = \sum_{\mathcal{C}} |\Delta + S(\mathcal{B}_i^t, \mathcal{B}_j^{v-}) - S(\mathcal{B}_i^t, \mathcal{B}_j^{v+}|_+, \quad (6)$$

where Δ is the margin. The positive pair $(\mathcal{B}_i^t, \mathcal{B}_j^{v+})$ is semantically related, while the unmatched $(\mathcal{B}_i^t, \mathcal{B}_j^{v-})$ is a negative pair in training batches. This loss encourages the positive pair to have a higher score than negative pairs.

Instance-level Loss The instance-level loss $\mathcal{L}_{ins}$ is adopted to constrain multiple instance learning, which makes instance bags more distinguishable. Furthermore, inspired by contrastive self-supervised learning (Gutmann and Hyvärinen 2010; Chen et al. 2020), a contrastive loss has the ability to further improve feature representation. Specifically, the loss is formulated as,

$$\mathcal{L}_{ins} = -log(\frac{e^{\widetilde{\mathbf{x}}_i^t \widetilde{\mathbf{x}}_j^v}}{e^{\widetilde{\mathbf{x}}_i^t \widetilde{\mathbf{x}}_j^v} + \sum_{(t,v) \in \mathcal{N}_{ins}} e^{\widetilde{\mathbf{x}}_i^t \widetilde{\mathbf{x}}_j^v}}), \quad (7)$$

where $\mathcal{N}_{ins}$ is the set of negative pairs randomly sampled from the training batches. This instance-level loss constrains the semantic space so that the semantically related entities are pulled closer and otherwise pushed farther apart.

Type	Method	Div150Adhoc								
		accuracy				diversity				speed
		P@10	P@20	D@10	D@20	A@10	A@20	C@10	C@20	S-qt
multi-stage	DTF	79.54%	77.05%	54.63%	52.79%	28.55%	24.41%	37.57%	51.58%	2.8 s
	DESA	76.25%	74.52%	52.24%	51.32%	31.41%	26.52%	35.50%	48.60%	1.0 s
	GRAPH4DIV	78.11%	74.82%	54.23%	52.31%	32.22%	27.22%	32.67%	46.56%	1.1 s
multi-semantic	DMIH	76.37%	73.31%	52.89%	50.24%	33.78%	28.49%	30.74%	42.46%	0.4 s
	PVSE	77.83%	74.44%	54.02%	52.13%	32.04%	27.30%	31.11%	44.32%	0.4 s
	FCA-Net	81.85%	79.57%	55.60%	54.80%	43.79%	37.82%	26.02%	31.55%	0.5 s
	VMIG	84.84%	82.12%	60.36%	58.03%	24.35%	20.42%	43.88%	57.64%	0.5 s
	Ours	**85.55%**	**84.06%**	**61.76%**	**59.71%**	**23.88%**	**19.91%**	**44.46%**	**58.42%**	**0.3 s**

Table 1: Performance comparison of various state-of-the-art baselines on Div150Adhoc dataset.

Type	Method	Div400								
		accuracy				diversity				speed
		P@10	P@20	D@10	D@20	A@10	A@20	C@10	C@20	S-qt
multi-stage	DTF	78.14%	76.20%	53.83%	51.92%	28.55%	24.12%	41.51%	55.27%	2.8 s
	DESA	76.50%	73.82%	51.65%	50.52%	30.91%	25.25%	39.47%	52.82%	1.0 s
	GRAPH4DIV	77.55%	74.50%	54.33%	52.70%	30.24%	25.91%	35.66%	45.77%	1.1 s
multi-semantic	DMIH	76.55%	73.66%	52.70%	50.31%	31.44%	26.39%	34.66%	44.50%	0.4 s
	PVSE	77.29%	74.45%	54.28%	52.56%	32.07%	26.05%	35.10%	45.09%	0.4 s
	FCA-Net	79.98%	78.42%	54.88%	53.38%	41.33%	34.61%	29.91%	34.90%	0.5 s
	VMIG	81.51%	78.27%	56.75%	55.27%	24.68%	21.34%	46.59%	59.01%	0.5 s
	Ours	**82.68%**	**80.66%**	**57.50%**	**56.81%**	**23.63%**	**20.60%**	**47.47%**	**59.89%**	**0.3 s**

Table 2: Performance comparison of various state-of-the-art baselines on Div400 dataset

Knowledge-level Loss Similar to instance-level loss, the knowledge-level loss $\mathcal{L}_{kno}$ constrains the learning of the hypergraph, which is formulated as,

$$\mathcal{L}_{kno} = -log(\frac{e^{\widetilde{\mathbf{x}}_i^t \widetilde{\mathbf{x}}_j^v}}{e^{\widetilde{\mathbf{x}}_i^t \widetilde{\mathbf{x}}_j^v} + \sum_{(t,v)\in\mathcal{N}_{kno}} e^{\widetilde{\mathbf{x}}_i^t \widetilde{\mathbf{x}}_j^v}}). \quad (8)$$

We pay more attention to which entities have links, thereby the negative set $\mathcal{N}_{kno}$ is defined as no link or hyperlink between the two entities. Finally, the overall loss is defined as follows,

$$\mathcal{L} = \mathcal{L}_{mat} + \lambda_1 \mathcal{L}_{ins} + \lambda_2 \mathcal{L}_{kno}, \quad (9)$$

where λ_1 and λ_2 are balance factors. Notably, to maintain the stability of training, we perform the knowledge-level loss $\mathcal{L}_{kno}$ first as a warm-up.

Experiments

Experimental Settings

Datasets Two public datasets, one on daily life and one on tourist locations, are adopted for evaluation.

Div150AdHoc[1] is a dataset for the competition of diverse social image retrieval (Ionescu et al. 2016), which contains a variety of keyword-based queries with an average length of only 2.7 words. In total, there are 134 daily life queries with $39,474$ images. Some extra meta information, such as photo title, tags, date etc, are also available.

Div400[2] is constructed by MediaEval Workshop (Ionescu et al. 2014), whose queries about tourist locations have an average length of 3.7 words. This dataset contains 396 queries with $43,418$ images, and also includes meta information such as GPS coordinates, Wikipedia pages and URLs in Flickr.

Evaluation Protocol We adopt five evaluation metrics from accuracy, diversity and speed perspectives. For accuracy, the precision and normalized discounted cumulative gain (NDCG) are utilized to measure top-k ranking performance. Meanwhile, the average diverse rank (ADR) and cluster recall (CR) are used for diversity measurement, while S-qt is the inference time per query (Bo and Gao 2019; Renders and Csurka 2017; Peng et al. 2017). In the rest of this paper, we denote Precision, NDCG, ADR and CR as $P@k, D@k, A@k, C@k$, with top-k results of 10, 20, respectively. Notably, the higher $P@K$, $D@K$ and $C@K$ indicate better accuracy, while the lower value of $A@K$ and S-qt represents richer image diversity and faster speed.

Implementation Details We implement our solution based on the Tensorflow framework[3]. In the feature representation, textual feature $\mathbf{x}^t$ and visual feature $\mathbf{x}^v$ are both $1,024$-dimensional vectors. For specific hyper-parameters in our method, the instance numbers m, the margin Δ, the threshold τ and the balance factors λ_1, λ_2 are set as $(8, 0.4, 0.4, 0.4, 0.3)$ and $(8, 0.4, 0.3, 0.3, 0.3)$ on Div150AdHoc and Div400 datasets, respectively.

Overall Performance Comparison

To verify the effectiveness of our proposed MKHG, we compare it with the following methods: 1) multi-stage retrieval algorithms: DTF (Bo and Gao 2019), DESA (Qin, Dou, and Wen 2020) and GRAPH4DIV (Su et al. 2021); 2)

[1]http://campus.pub.ro/lab7/bionescu/Div150Adhoc.html

[2]http://multimediaeval.org/mediaeval2014/diverseimages2014

[3]https://www.tensorflow.org/

Type	Method	Div150Adhoc							
		accuracy				diversity			
		P@10	P@20	D@10	D@20	A@10	A@20	C@10	C@20
Retrieval	CLIP	83.46%	81.45%	57.74%	56.25%	39.64%	33.76%	26.35%	31.97%
Knowledge	retrieval-based	75.34%	72.85%	51.45%	50.24%	36.41%	31.50%	28.04%	37.93%
	KG-based	79.36%	76.07%	55.42%	53.25%	31.28%	27.40%	36.45%	49.47%
	MMKG-based	83.24%	81.63%	58.27%	57.29%	25.75%	23.14%	41.21%	54.24%
Hyperedge	v-hyperedge	84.04%	82.22%	59.54%	58.06%	25.07%	22.03%	42.33%	55.40%
	t-hyperedge	84.22%	82.34%	59.64%	58.21%	24.76%	21.41%	42.63%	55.77%
	a-hyperedge	84.10%	82.24%	59.48%	58.11%	24.89%	21.75%	42.37%	55.55%
Channel	Ours w/o image-diverse	76.67%	74.34%	52.67%	51.33%	34.94%	29.38%	28.65%	39.12%
Optimization	Ours w/o aggregator	79.11%	75.89%	55.21%	53.14%	31.03%	27.29%	33.11%	45.30%
	Ours w/o $\mathcal{L}_{ins}$	81.88%	79.21%	56.99%	55.81%	28.58%	26.25%	36.61%	49.48%
	Ours w/o $\mathcal{L}_{kno}$	81.77%	79.32%	57.35%	55.85%	28.61%	26.09%	36.54%	49.32%
ALL	Ours	**85.55%**	**84.06%**	**61.76%**	**59.71%**	**23.88%**	**19.91%**	**44.46%**	**58.42%**

Table 3: Performance comparison of various component combinations on Div150Adhoc dataset.

Type	Method	Div400							
		accuracy				diversity			
		P@10	P@20	D@10	D@20	A@10	A@20	C@10	C@20
Retrieval	CLIP	79.76%	79.01%	56.11%	54.97%	39.79%	33.90%	26.05%	31.63%
Knowledge	retrieval-based	75.14%	72.52%	51.16%	49.79%	36.10%	31.33%	27.74%	37.36%
	KG-based	78.83%	76.97%	55.14%	53.01%	29.95%	26.20%	42.19%	52.29%
	MMKG-based	81.46%	78.86%	56.32%	55.91%	25.45%	23.22%	43.34%	57.35%
Hyperedge	v-hyperedge	81.95%	78.93%	57.01%	56.02%	24.98%	22.76%	44.71%	58.15%
	t-hyperedge	81.72%	78.99%	56.81%	56.09%	25.20%	22.96%	44.20%	58.00%
	a-hyperedge	82.06%	79.25%	57.03%	56.13%	24.34%	22.60%	44.88%	58.43%
Channel	Ours w/o image-diverse	75.40%	73.75%	51.59%	51.17%	36.61%	31.04%	28.63%	38.38%
Optimization	Ours w/o aggregator	78.78%	75.59%	55.01%	52.84%	31.25%	27.54%	32.51%	55.46%
	Ours w/o $\mathcal{L}_{ins}$	79.46%	77.11%	55.57%	54.39%	28.18%	25.81%	39.29%	51.13%
	Ours w/o $\mathcal{L}_{kno}$	79.07%	77.72%	55.59%	54.61%	28.21%	25.76%	39.11%	51.80%
ALL	Ours	**82.68%**	**80.66%**	**57.50%**	**56.81%**	**23.63%**	**20.60%**	**47.47%**	**59.89%**

Table 4: Performance comparison of various component combinations on Div400 dataset.

multi-semantic approaches: DMIH (Zhao et al. 2017), PVSE (Song and Soleymani 2019), FCA-Net (Han et al. 2021) and VMIG (Zeng et al. 2022b). Among them, the DTF, DESA and GRAPH4DIV combine auxiliary information (e.g. title or subtopic) to perform filtering and ranking. DMIH and PVSE are the multiple instance learning methods that capture multi-semantics, while FCA-Net and VMIG are graph-based retrieval networks to match sub-semantics.

Experimental results are reported in Table 1 and Table 2. 1) The multi-stage diverse retrieval methods DTF, DESA and GRAPH4DIV rely on auxiliary information to perform well in accuracy and diversity, but they are slow. Compared with DESA that uses a self-attention mechanism to capture implicit sub-semantics, both DTF with extending title and GRAPH4DIV for modeling intent achieve better performance. This proves that the introduction of external knowledge is a complement to keyword-based short queries. 2) The multi-semantic approaches take advantage of their diversity modeling and improve the metrics. Among them, graph matching-based methods FCN-Net and VMIG outperform general multiple instance learning solutions. These results show that the graph learning among sub-semantics removes redundant concepts, and makes diverse representations more reasonable. 3) Our approach achieves the best performance on both Div150AdHoc and Div400 datasets, which relates to our explicit multi-semantic modeling by introducing multi-modal knowledge hypergraph.

Ablation Study

To thoroughly investigate all components of our model, we carry out ablation studies from the following perspectives: 1) Advanced retrieval model. The CLIP (Radford et al. 2021), one of the popular frameworks in cross-modal retrieval, is directly applied for this diverse task. 2) Knowledge strategy. To explore how to expand the sub-semantics of queries, we design variants as in Fig. 2, namely, 'retrieval-based', 'KG-based', and 'MMKG-based'. The 'retrieval-based' variant retrieves items directly from a knowledge pool as instances, 'KG-based' variant introduces plain-text wikipedia, and 'MMKG-based' variant implements another public multi-modal knowledge graph: richpedia. 3) Hypergraph design. We separately implement visual hyperedge, text hyperedge and attribute hyperedge based on the 'MMKG-based' variant. 4) Multi-semantic channel. This variant 'Ours w/o image-diverse' only performs textual multi-semantic representations without diversifying images. 5) Optimization module. The three variants 'w/o aggregator', 'w/o $\mathcal{L}_{ins}$', and 'w/o $\mathcal{L}_{kno}$', ablate the instance graph and two contrastive losses, respectively.

The ablation results are shown in Table 3 and Table 4, and we have the following observations: 1) CLIP has high accuracy but low diversity, indicating the single semantic is insufficient in diverse scenarios. 2) After applying retrieval strategy or general knowledge graph for semantic expansion, the diversity of 'retrieval-based' and 'KG-based'

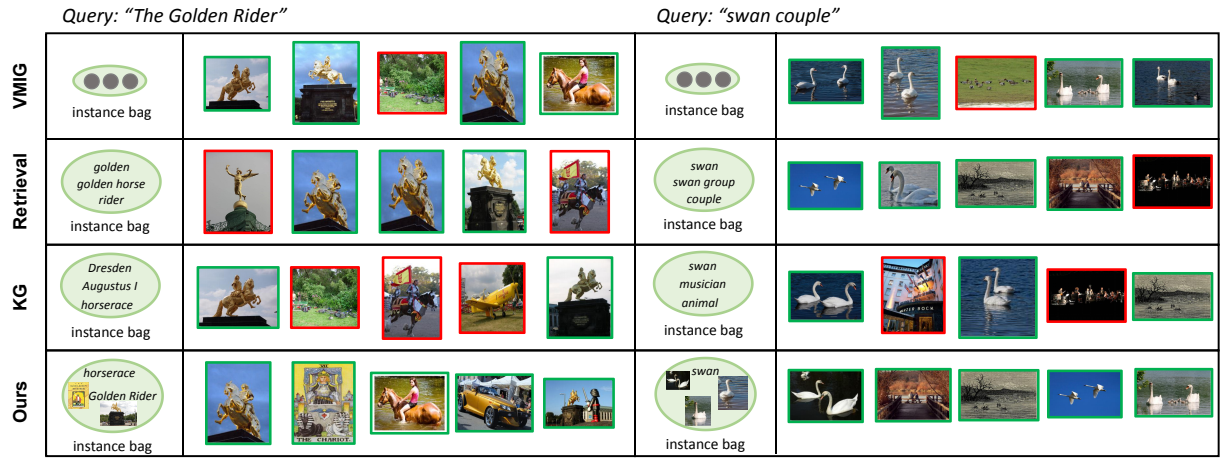

Figure 4: The visualization under the query "The Golden Rider" in Div400 and "swan couple" in Div150AdHoc.

is improved, which proves that explicit representation is effective. Furthermore, the performance of 'MMKG-based' is further enhanced, and other off-the-shelf MMKG can also be borrowed. This shows that graphs with richer information and more diverse relationships are capable of capturing more complex relations. 3) The score of three hyperedges increases, which proves that capturing the same semantic patterns is simple but effective. 4) The performance drops after ablating multi-semantic representations for images, which means that modeling both text and image can thoroughly understand many-to-many relations. 5) The ablation of the instance graph and the two loss functions degrades scores, which demonstrates that every module in our model is necessary. Among them, the instance graph reduces redundant sub-semantics, while the loss functions help multiple instance learning and hypergraph learning.

Retrieval Visualization

We visualize the details of instance bags and retrieval results for several methods on two datasets, i.e. the implicit modeled VMIG and the explicit extended retrieval-based, KG-based and ours. As shown in Fig. 4, the green boxes represent the correct results, and the red boxes are the mismatched images. Among them, VMIG has good performance on diversity, but its captured sub-semantics are not explainable. Furthermore, the retrieval-based method can only obtain semantically similar words (e.g. "rider", "animal"), while the KG-based method is prone to introduce noise (e.g. "musical"), resulting in errors. Our method has a better balance between accuracy and diversity, making full use of the rich information of the off-the-shelf MMKG and reasonably controlling the noise.

Robustness Discussion

To explore the robustness and scalability of our method, we conduct an experiment on out-of-domain queries. Specifically, the larger MSCOCO dataset[4] is adopted for diverse retrieval, with no training or tuning at all. Although the MSCOCO is not applied for this task, its supercategory labels are in the form of keywords (e.g."animal"), which are suitable for diverse retrieval scenarios. Therefore, we treat the supercategory label as the keyword-based query, and the subcategory coverage of the result images as the diverse score $C@k$ for robustness discussion.

Method	P@10	P@20	C@10	C@20
retireval-based	66.30%	65.99%	19.08%	27.55%
KG-based	65.65%	64.11%	22.07%	29.19%
MMKG-based	66.03%	65.43%	22.54%	29.45%
Ours	**75.34%**	**72.65%**	**27.03%**	**33.83%**

Table 5: Robustness Discussion on MSCOCO dataset.

As shown in Table 5, the retrieval-based strategies assign instances to out-of-domain queries via similarity measure, it has robust accuracy but poor diversity. KG- and MMKG-based strategies perform simple entity linking via cosine function and improve the scores. Moreover, our hypergraph-based method significantly outperforms other strategies on MSCOCO, which proves that the designed hyperedges are capable of generalizing to out-of-domain queries.

Conclusions

In this work, we contribute a novel multi-modal knowledge hypergraph for keyword-based diverse image retrieval to explicitly extend sub-semantics end-to-end. Specifically, a hypergraph construction module customizes various hyperedges to link the heterogeneous MMKG and retrieval databases. Then a multi-modal instance bagging and a diverse concept aggregator is designed to explicitly select reasonable instances. Finally, several losses optimize the semantic space. In this way, off-the-shelf MMKGs can be linked quickly and easily. Experimental results show that our method achieves superior diversity and explainability.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (No. 62072462) and National Key R&D Program of China (No. 2020AAA0108600).

[4]https://cocodataset.org/

References

Bai, S.; Zhang, F.; and Torr, P. H. S. 2021. Hypergraph convolution and hypergraph attention. *Pattern Recognit.*, 110: 107637.

Bo, Y.; and Gao, X. 2019. Diversified textual features based image retrieval. *Neurocomputing*, 357: 116–124.

Cao, D.; Zeng, Y.; Wei, X.; Nie, L.; Hong, R.; and Qin, Z. 2020. Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization. In *Proceedings of the ACM MM*, 898–906. ACM.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 1597–1607. ICML.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 297–304.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In *Proceedings of the NeurIPS*, 1024–1034.

Han, N.; Chen, J.; Xiao, G.; Hao, Z.; Zeng, Y.; and Chen, H. 2021. Fine-grained Cross-modal Alignment Network for Text-Video Retrieval. In *Proceedings of the ACM MM*, 3826–3834. ACM.

Ionescu, B.; Gînsca, A.; Zaharieva, M.; Boteanu, B.; Lupu, M.; and Müller, H. 2016. Retrieving Diverse Social Images at MediaEval 2016: Challenge, Dataset and Evaluation. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, volume 1739.

Ionescu, B.; Radu, A.; Menéndez, M.; Müller, H.; Popescu, A.; and Babak, L. 2014. Div400: a social image retrieval result diversification dataset. In *Multimedia Systems Conference 2014*, 29–34.

Ionescu, B.; Rohm, M.; Boteanu, B.; Gînsca, A.; Lupu, M.; and Müller, H. 2021. Benchmarking Image Retrieval Diversification Techniques for Social Media. *IEEE Trans. Multim.*, 23: 677–691.

Peng, L.; Bin, Y.; Fu, X.; Zhou, J.; Yang, Y.; and Shen, H. T. 2017. CFM@MediaEval 2017 Retrieving Diverse Social Images Task via Re-ranking and Hierarchical Clustering. In *Proceedings of the Working Notes Proceedings of the MediaEval 2017 Workshop*, volume 1984.

Qin, X.; Dou, Z.; and Wen, J. 2020. Diversifying Search Results using Self-Attention Network. In *Proceedings of the CIKM*, 1265–1274.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the ICML*, 8748–8763.

Renders, J.; and Csurka, G. 2017. NLE@MediaEval'17: Combining Cross-Media Similarity and Embeddings for Retrieving Diverse Social Images. In *Proceedings of the Working Notes Proceedings of the MediaEval 2017 Workshop*, volume 1984.

Seddati, O.; Ben-Lhachemi, N.; Dupont, S.; and Mahmoudi, S. 2017. UMONS @ MediaEval 2017: Diverse Social Images Retrieval. In *Proceedings of the Working Notes Proceedings of the MediaEval 2017 Workshop*, volume 1984.

Song, Y.; and Soleymani, M. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *Proceedings of the CVPR*, 1979–1988. IEEE.

Su, Z.; Dou, Z.; Zhu, Y.; Qin, X.; and Wen, J. 2021. Modeling Intent Graph for Search Result Diversification. In *Proceedings of the SIGIR*, 736–746. ACM.

Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; and Zheng, K. 2020. Multi-modal Knowledge Graphs for Recommender Systems. In *Proceedings of the CIKM*, 1405–1414.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *Proceedings of the ACM MM*, 154–162. ACM.

Wang, H., Mengand Wang; Qi, G.; and Zheng, Q. 2020. Richpedia: A Large-Scale, Comprehensive Multi-Modal Knowledge Graph. *Big Data Res.*, 22: 100159.

Wu, J.; and Ngo, C. 2020. Interpretable Embedding for Ad-Hoc Video Search. In *Proceedings of the ACM MM*, 3357–3366.

Xu, G.; Chen, H.; Li, F.; Sun, F.; Shi, Y.; Zeng, Z.; Zhou, W.; Zhao, Z.; and Zhang, J. 2021. AliMe MKG: A Multi-modal Knowledge Graph for Live-streaming E-commerce. In *Proceedings of the CIKM*, 4808–4812.

Zeng, Y. 2022. Point Prompt Tuning for Temporally Language Grounding. In *SIGIR*, 2003–2007.

Zeng, Y.; Cao, D.; Lu, S.; Zhang, H.; Xu, J.; and Zheng, Q. 2022a. Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning. *ACM Trans. Multim. Comput. Commun. Appl.*, 18: 56:1–56:21.

Zeng, Y.; Cao, D.; Wei, X.; Liu, M.; Zhao, Z.; and Qin, Z. 2021. Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval. In *Proceedings of the CVPR*, 2215–2224. IEEE.

Zeng, Y.; Wang, Y.; Liao, D.; Li, G.; Huang, W.; Xu, J.; Cao, D.; and Man, H. 2022b. Keyword-Based Diverse Image Retrieval with Variational Multiple Instance Graph. *IEEE Trans. Neural Networks Learn. Syst.*

Zhao, W.; Guan, Z.; Luo, H.; Peng, J.; and Fan, J. 2017. Deep Multiple Instance Hashing for Object-based Image Retrieval. In *Proceedings of the IJCAI*, 3504–3510.

Zhao, W.; Hu, Y.; Wang, H.; Wu, X.; and Luo, J. 2021. Boosting Entity-aware Image Captioning with Multi-modal Knowledge Graph. *arXiv preprint arXiv:2107.11970*, 1–10.

Zhu, X.; Li, Z.; Wang, X.; Jiang, X.; Sun, P.; Wang, X.; Xiao, Y.; and Yuan, N. J. 2022. Multi-Modal Knowledge Graph Construction and Application: A Survey. *arXiv preprint arXiv:2202.05786*, 1–10.