Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060395

ARTICLE

MMCSD: Multi-Modal Knowledge Graph Completion Based on Super-Resolution and Detailed Description Generation

Huansha Wang*, Ruiyang Huang*, Qinrang Liu, Shaomei Li and Jianpeng Zhang

National Digital Switching System Engineering & Technological R&D Center, Information Engineering University, Zhengzhou, 450001, China

*Corresponding Authors: Huansha Wang. Email: whs123@mail.ustc.edu.cn; Ruiyang Huang. Email: gisexpert@163.com

Received: 31 October 2024; Accepted: 20 January 2025; Published: 26 March 2025

ABSTRACT: Multi-modal knowledge graph completion (MMKGC) aims to complete missing entities or relations in multi-modal knowledge graphs, thereby discovering more previously unknown triples. Due to the continuous growth of data and knowledge and the limitations of data sources, the visual knowledge within the knowledge graphs is generally of low quality, and some entities suffer from the issue of missing visual modality. Nevertheless, previous studies of MMKGC have primarily focused on how to facilitate modality interaction and fusion while neglecting the problems of low modality quality and modality missing. In this case, mainstream MMKGC models only use pre-trained visual encoders to extract features and transfer the semantic information to the joint embeddings through modal fusion, which inevitably suffers from problems such as error propagation and increased uncertainty. To address these problems, we propose a Multi-modal knowledge graph Completion model based on Super-resolution and Detailed Description Generation (MMCSD). Specifically, we leverage a pre-trained residual network to enhance the resolution and improve the quality of the visual modality. Moreover, we design multi-level visual semantic extraction and entity description generation, thereby further extracting entity semantics from structural triples and visual images. Meanwhile, we train a variational multi-modal auto-encoder and utilize a pre-trained multi-modal language model to complement the missing visual features. We conducted experiments on FB15K-237 and DB13K, and the results showed that MMCSD can effectively perform MMKGC and achieve state-of-the-art performance.

KEYWORDS: Multi-modal knowledge graph; knowledge graph completion; multi-modal fusion

1 Introduction

Multi-modal knowledge graph (MMKG) refers to the large-scale graph-structured knowledge base that utilizes semantic knowledge networks to describe multi-modal entities in the form of triples like $(\boldsymbol{head\ entity}, \boldsymbol{relation}, \boldsymbol{tail entity})$. Numerous studies have shown that multi-modal data can introduce more supervised information and external knowledge into deep neural networks. Thus, MMKGs have become one of the hottest research fields of knowledge graphs, with important applications in multi-modal knowledge question answering system, retrieval-augmented generation, and so on. Due to the continuous growth of data and knowledge, as well as the quality limitations of online encyclopedias (which are the main knowledge source of knowledge graphs), the existing mainstream knowledge graphs inevitably have problems such as insufficient knowledge and incomplete content. A specific manifestation is the lack of entities or relations in the triples like $(\boldsymbol{head\ entity}, \boldsymbol{relation}, ?)$ or $(\boldsymbol{head\ entity}, ?, \boldsymbol{tail entity})$. The absence of entities or relations in the triples will lead to a break in the knowledge chain, affecting the accuracy of reasoning and searching for knowledge. In order to correctly fill in the missing content in the

triples and thus improve the richness of the knowledge, more and more researchers are focusing on Multi-modal Knowledge Graph Completion (MMKGC). MMKGC task aims to discover the unknown triples in the MMKG and complete the entities and relations that should be contained in incomplete triples, which is one of the keys to the construction and extension of the multi-modal knowledge graph.

Mainstream MMKGC models are accomplished based on representation learning, which encodes attributes and relations of each modality into low-dimensional embedding. Subsequently, various modal fusion methods are employed to obtain multi-modal embeddings for both entities and relations, and joint embeddings are formed for all candidate triples, and then decoded to obtain the correct results. Xie et al. [1] used the pre-trained AlexNet to extract visual features from entity description images and sought consistency of the vector space by translation between multi-modal embeddings. Mousselly Sergieh et al. [2] adopted Imagine and DeViSE to fuse multi-modal information. Wang et al. [3] proposed a specific multi-modal auto-encoder module to extract multi-modal joint embeddings directly. Some researchers [4–6] also focus on mitigating inter-modal interference and improving the quality of multi-modal embedding by modifying the sampling strategy, refining the modal embedding, or introducing the large language model.

However, existing models generally ignore the poor quality and lack of multi-modal information present in mainstream datasets as well as in reality MMKGs. As shown in Fig. 1, not every multi-modal entity has visual attributes that highly correspond to its semantics, and this is mainly reflected in the following aspects: (1) Not all entities have multi-modal information, and for entities without visual information, mainstream models typically use random, mean, or zero tensors as visual embeddings, which has less positive effect on the modal interaction and modeling. (2) The semantic relationship between multi-modal information and the entity itself is poor. Unlike the image-text matching task, the text modality information of entities in the MMKGC task may only be the entity name (some datasets have a few sentences of entity description), making it difficult to align the text modality and visual modality well. (3) Multi-modal information itself has quality issues. Due to the fact that mainstream MMKGC datasets are built on semi-structured data from open-source encyclopedias, many entity description images themselves are not clear enough, and fail to effectively describe entities. In this case, previous MMKGC models only use pre-trained visual models to extract visual features, and transfer the information to the joint embeddings through modal fusion, which inevitably have problems such as error propagation and increased uncertainty. Zhang et al. [7] have recognized the negative impact of missing modalities and have proposed the use of generative adversarial networks (GAN) to supplement the missing features. However, training generative adversarial networks consumes a lot of additional computational resources and is disconnected from the original triples. Moreover, it does not solve the problem of poor visual features due to the low quality of the described images.

To address the above issues, we propose MMCSD, a Multi-modal knowledge graph Completion model based on Super-resolution and Detailed Description Generation. The core idea is to optimize the visual information in the joint embedding by improving the quality of the visual embedding as well as completing the missing features, thus facilitating modal interaction. Specifically, we utilize the pre-trained residual network [8] to enhance the resolution and improve the quality of the visual modality. Moreover, we design multi-level visual semantic extraction and entity description generation, thereby further extracting entity semantics from structural triples and visual images. Meanwhile, we train a variational multi-modal auto-encoder and utilize the pre-trained multi-modal language model to complement the missing visual features. We conduct experiments to verify the effectiveness of the model on FB15K-237 [9] and DB13K. The results show that MMCSD can effectively perform MMKGC and achieve state-of-the-art performance.

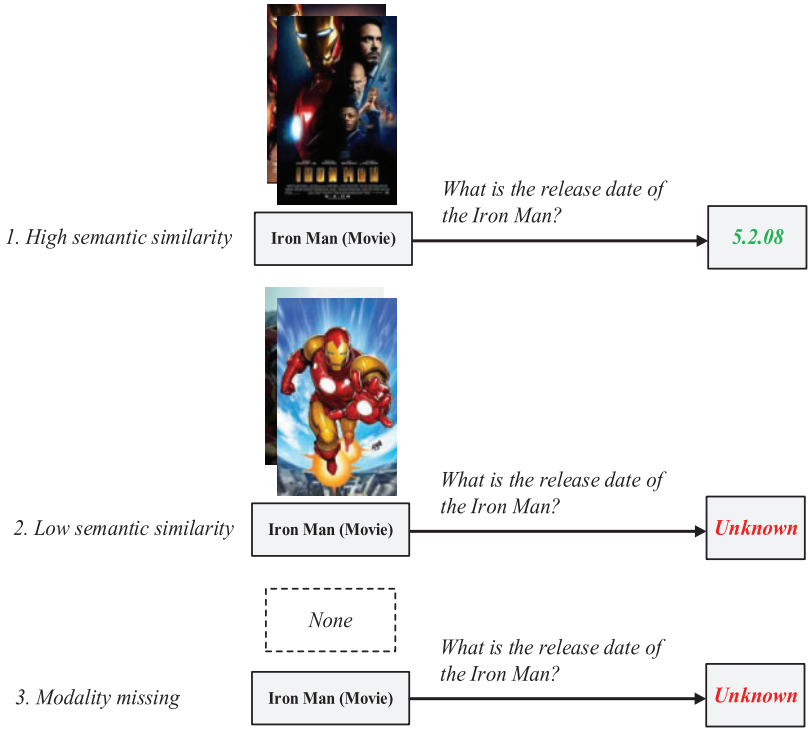

Figure 1: Three different quality levels of multi-modal entities and their impact on downstream tasks

The contributions of this paper are as follows:

1. To address the problem of poor visual embedding due to the low quality of descriptive images and primitive way of visual feature extraction, we propose to utilize a pre-trained residual network to super-resolve the original images and employ the multi-level visual semantic extraction mechanism to extract the deep semantics hidden in the visual modality.
2. To address the problem of modality missing, we introduce the modality imagination mechanism, i.e., by training the variational multi-modal auto-encoder to simulate the missing visual embedding. Meanwhile, we propose to adopt the multi-modal large model to generate features semantically similar to the triples as a complement.
3. We structure the knowledge of DB13K in the multi-modal knowledge graph MMKB[1] (Multi-modal Knowledge Bases) [10] to construct a multi-modal knowledge completion dataset for model evaluation. The dataset will be available at Github.

2 Related Work

2.1 Knowledge Graph Completion

With the proposal of knowledge graph, researchers have long recognized the importance of knowledge completeness for knowledge graph applications, and thus knowledge graph completion has been one of the most critical related tasks. Traditional knowledge completion tasks mainly rely on manual effort. Although this method has high quality, with the continuous increase in the scale of knowledge graphs, manual completion is no longer applicable to existing knowledge graphs. Representation learning technology excels at extracting strong features and learning optimal representations of objects, thereby simplifying downstream

[1] https://github.com/mniepert/mmkb (accessed on 11 May 2018).

task steps and improving their effectiveness. Due to the ability to extract strong features from complex data forms, representation learning has important applications in the knowledge graph, mainly including translational-distance based models and neural network based models.

Knowledge completion based on translational-distance involves two steps. First, it projects entities onto a low dimensional vector space and treats the relations between entities as translations of entity vectors. Then, it models entities and relations separately through iterative training. The earliest translational-distance based model is TransE [11]. It demonstrates its excellent modeling ability in sparse knowledge graph work and inspires many similar works based on translational-distance in the future. However, the assumptions are too simple, which makes it difficult to model complex relations of one-to-many and many-to-many better. TransH [12] makes the same entity or relation have different vector representations in different triples by introducing hyperplanes instead of relationship vectors. RotatE [13], on the other hand, treats relations as rotation angles between entity vectors, thus modeling entities and relations in more complex space. Translational-distance based models have the advantages of simplicity and a low number of parameters, but accordingly, their effectiveness generally fails to match that of neural network based models. Currently, translational-distance based models are usually used as the basis for other extended models, the entity and relation embeddings pre-trained by TransE are employed as the original features for training.

Moreover, with deep learning technology demonstrating its powerful feature extraction capabilities in fields such as computer vision and natural language processing, knowledge graph completion based on neural networks has become mainstream. ConvE [14] uses convolutional neural networks to aggregate entities and relations, and then calculate corresponding similarity scores. ConvKB [15] associates a unique feature embedding for different entities and relations to improve the transfer and representation ability of the model. With the continuous development of graph convolutional networks (GCN), researchers have attempted to utilize the advantages of GCNs in learning node and edge representations in graph-structured data to better learn entity-relation features of knowledge graphs. R-GCN [16] introduces the relation matrix in GCNs as a mapping transform when entities aggregate neighborhood features. KBAT [17] utilizes graph attention network as the encoder to integrate the information of multi-hop neighbors of entities and then uses ConvKB to decode entity representations. In recent years, a large number of knowledge completion models have emerged that adopt Transformer [18] or pre-trained models to model triples, and have achieved good results. Neural network based models can effectively extract hidden potential features from the knowledge graph with high accuracy, high inference scalability and efficiency. However, neural networks rely on a large amount of training data, which is a data-driven endeavor that usually performs poorly when applied to knowledge graphs with sparse data. In addition, these models suffer from common shortcomings of neural networks such as low interpretation and too many parameters.

2.2 Multi-Modal Knowledge Graph Completion

With the continuous development of multi-modal technology, researchers have gradually realized the positive effects of multi-modal data such as images and videos on improving the completeness and universality of knowledge graphs. Many multi-modal knowledge graphs based on large-scale image training sets, single-modal knowledge graphs, Wikipedia, and other data, such as ImgPedia [19], VisualSem [20], GAKG [21], have emerged in large numbers. At the same time, multi-modal knowledge completion tasks have also become one of the important subtasks of knowledge completion.

IKRL [1] for the first time introduces visual information to the knowledge graph completion task by using a pre-trained AlexNet to extract visual features from entity description images and seeks consistency in the vector space by transforming between multi-modal embeddings. MKBE [22] uses different encoders to process numerical and textual data to generate corresponding embeddings in addition to description images,

respectively, thus introducing more supervisory information. TransAE [3] processes pre-generated visual and text features through multi-modal auto-encoders to obtain multi-modal joint embeddings as entity representations. These methods are ineffective in extracting and representing multi-modal information, thus affecting the final embedding quality.

Since the modal interactions between the MMKGC tasks are still unclear, researchers have tried several mechanisms to facilitate them. VBKGC [23] adopts a transformer-based multi-modal pretraining model to simultaneously process multi-modal image-text pairs to encode entities. MANS [4] proposes modality aware negative sampling strategy to generate modality-level negative samples where the descriptive images are not correlated with the entity. MoSE [5] focuses on the difference of modality importance. It learns modality-split relation embeddings for each modality, which alleviates the modality interference. MACO [7] leverages the generative adversarial framework and trains a pair of generator-discriminators to generate missing modality features, preventing degradation of embedding quality due to missing multi-modal information. KoPA [6] integrates pre-trained structural embeddings with large language models (LLMs), and achieves structural-aware reasoning in the LLMs.

However, previous research has not focused on further exploration and mining for visual modality processing, and it is still common to use only pre-trained encoders to obtain visual embeddings. This simple visual feature extraction approach does not take into account the difference in relevance between images and entities, as well as the quality of the images themselves. Given the generally poor quality of multi-modal information in datasets and real-world MMKGs, the use of this approach will have a negative impact on the modeling of visual embeddings and joint embeddings. Moreover, most existing approaches do not consider the problem of missing modalities, and for entities with missing visual modality, mainstream models typically use random, mean, or zero tensors as their alternative visual embeddings, which do not contribute positively to modal interaction and fusion. Although there has been work that recognizes detrimental effects of modality missing and proposes to generate auxiliary embeddings using generative adversarial networks, this approach still has limitations. Training generative adversarial networks requires significant additional computational resources, and the auxiliary embeddings are divorced from the knowledge of the entities themselves and thus lack interpretation.

3 Preliminary

A multi-modal knowledge graph (MMKG) can be denoted as:

$$G = (E, R, V, T, A), \tag{1}$$

where E is the set of entities, R is the set of relations, V is the set of visual knowledge, T is the set of triples, and A is the set of entity attributes (note that in some datasets, entity attributes are also provided in triple form). A triple can be represented as (h, r, t), where $h, t \in E$ are head entity and tail entity, and $r \in R$ is relation.

Multi-modal knowledge graph completion (MMKGC) models represent entities and relations as embeddings, and then use a score function $f(h, r, t)$ to assess the likelihood of triples. For the evaluation query $q = (h, r, ?)$ or $(?, r, t)$, the model will sort all candidate entities and output a ranked list of preferences.

4 Method

To solve the problem of modality missing and low quality visual embedding in MMKGC, we propose MMCSD, a multi-modal knowledge graph completion model based on super-resolution and detailed description generation. It follows the established framework of previous MMKGC models, which involves

first extracting embeddings for different modalities separately, then generating multi-modal embeddings through modal fusion, and finally decoding the concatenated features of all candidate triples.

On this basis, we use a pre-trained residual network to improve the resolution of the visual modalities and prevent the degradation of the quality of the visual embedding due to the low clarity of the original image. Compared to the traditional approach of extracting features using only visual encoders, we employ a multi-level visual semantic extraction mechanism to comprehensively capture the semantic information within the visual modality by combining the outputs of multiple mature models. We also feed the triples into the large language model, aiming to obtain knowledge from a natural language perspective beyond structural features. For the missing modality problem, we train a variational multi-modal auto-encoder to simulate the missing visual embeddings and adopt a multi-modal large language model to supplement the original visual features.

In this chapter, we elaborate on the specific architecture and composition of MMCSD. Sections 4.1–4.3 detail the generation of embeddings for three different modalities respectively, with a particular emphasis on the visual embeddings. Following that, Sections 4.4 and 4.5 describe the modality fusion and decoding mechanisms. The missing modality completion mechanism we propose is explained in Section 4.6.

The entire multi-modal embedding generation process of MMCSD is shown in Fig. 2.

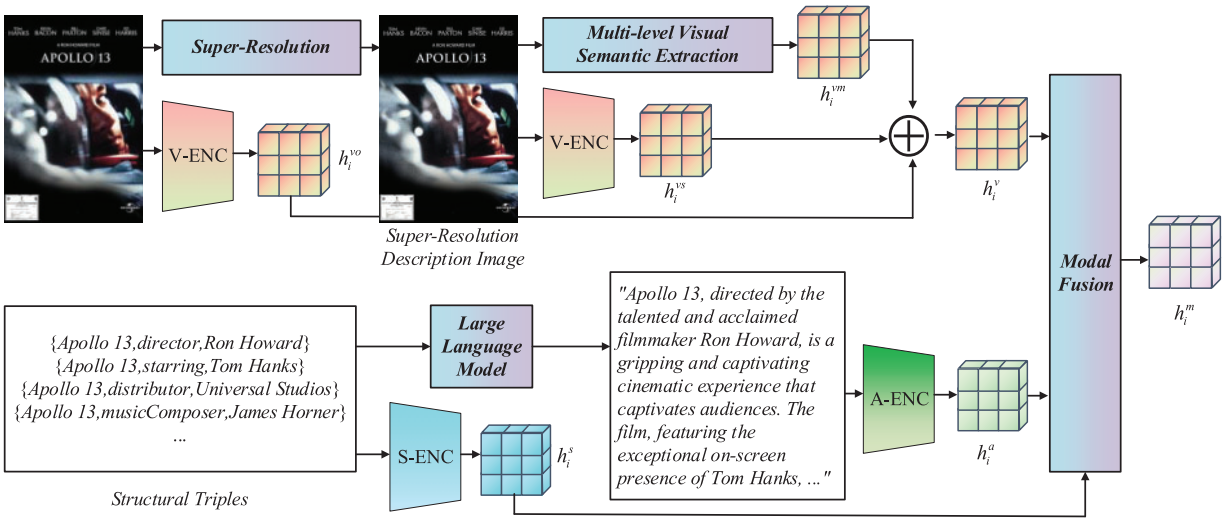

Figure 2: The multi-modal embedding generation process of MMCSD. As detailed in Chapter 4, we adopt VGG (Visual Geometry Group) and CLIP (Contrastive Language-Image Pre-Training) as visual encoder, SpGAT as structural encoder, and BLIP2 (Bootstrapping Language-Image Pre-training with frozen unimodal models) as attribute encoder

4.1 Neighborhood Structural Embedding

Most mainstream knowledge completion models use graph convolutional networks (GCN) [24] or graph attention networks (GAT) [25] to extract entity structural information from knowledge graph triples.

GAT exhibits excellent feature extraction capabilities for graph-structured data, but it also entails relatively high computational complexity. In MMKGC, the relevant edges of most entities are sparse, meaning that the encoder only needs to compute the weights between a relatively small number of nodes for each entity. Therefore, in order to effectively obtain structural embeddings while conserving computational resources, we utilize Sparse GAT (SpGAT) [17] to calculate the attention weights of each relevant triple and iteratively obtain multi-hop knowledge.

Specifically, for entity e_i, in order to obtain its structural embedding $\vec{h}_i^s$, SpGAT firstly initializes its related triple embeddings $\vec{t}_{i,j,k}$ through embedding concatenation and linear operations:

$$\vec{t}_{i,j,k} = W_1[\vec{h}_i \| \vec{h}_j \| \vec{g}_k], \tag{2}$$

vectors $\vec{h}_i$, $\vec{h}_j$, and $\vec{g}_k$ denote original embeddings of entities e_i, e_j, and relation r_k of triple (e_i, r_k, e_j), respectively. W_1 denotes the linear transformation matrix.

Similar to vanilla GAT, SpGAT then quantifies and calculates the importance $\beta_{i,j,k}$ and attention weight $\alpha_{i,j,k}$ of each triple for entity structural embedding:

$$\beta_{i,j,k} = LeakyReLU(W_2\vec{t}_{i,j,k}), \tag{3}$$

$$\alpha_{i,j,k} = softmax(\beta_{i,j,k}), \tag{4}$$

where W_2 refer to a weight matrix, $softmax$ and $LeakyReLU$ represent the corresponding functions.

Subsequently, the model sums up the triple embeddings based on the calculated attention weights $\alpha_{i,j,k}$ to obtain final structural entity embeddings $\vec{h}_i^s$:

$$\vec{h}_i^s = \sigma\left(\sum_{j\in N_i}\sum_{k\in R_{ij}} \alpha_{i,j,k}\vec{t}_{i,j,k}\right). \tag{5}$$

N_i denotes the neighborhood of entity e_i and R_{ij} denotes the set of relations connecting entities e_i and e_j.

We use multi-head attention to get more comprehensive and complete knowledge about the neighborhood. SpGAT employs averaging calculation results of M attention mechanisms to get final embedding vectors for entities:

$$\vec{h}_i^s = \sigma\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{j\in N_i}\sum_{k\in R_{ij}} {\alpha_{i,j,k}}^m {\vec{t}_{i,j,k}}^m\right). \tag{6}$$

4.2 Visual Embedding

Visual modality, as the most unique modality in MMKGs, contains a large amount of semantics related to entities or triples. The mainstream models, which only use pre-trained visual models (such as VGG or ResNet) to encode visual information, have significant limitations in the situation of low visual modality quality in the MMKG. To address this issue, we incorporate an additional step of adopting a residual network to perform super-resolution on the visual modality prior to using a pre-trained visual encoder to extract original visual embeddings. Following this, we leverage a variety of well-established techniques from the fields of computer vision and multi-modal learning to extract visual semantics at multiple levels. This approach enhances the semantic similarity between structural embeddings and visual embeddings, thereby facilitating subsequent modal interaction and fusion.

4.2.1 Original Visual Embedding

We use a pre-trained visual model (PVM), e.g., VGG-16 (Visual Geometry Group 16-layer network), to encode the described images of the entities, and then adopt the final layer output of it as the original visual features.

Afterward, the original visual features are input into a feed-forward neural network to obtain the original visual embeddings with the same dimension as the entity structural features:

$$\vec{h}_i^{vo} = W^v PVM(Img_i) + b^v, \tag{7}$$

where $\vec{h}_i^{vo}$ is the original visual embeddings of e_i, W^v and b^v are learnable weight and the bias matrix of corresponding feed-forward neural network, PVM means pre-trained visual model, Img_i denotes the visual image of e_i. When a single entity has multiple descriptive images, we apply mean pooling to aggregate visual features.

4.2.2 Super-Resolution

The resolution of an image greatly affects the accuracy of its semantic analysis. Most mainstream MMKGs are constructed based on open-source crowd-sourcing semi-structured encyclopedias, where the sources of visual modality are extensive and complex. It leads to certain visual modalities within the knowledge graph having low clarity and resolution. Some MMKGs are even unable to provide original images and only present pre-trained visual features. Taking the dataset FB15K-237 as an example, MMKGC models that do not focus on original feature extraction typically only utilize its visual features pre-extracted by VGG. This over-reliance on general visual encoders is based on an unrealistic assumption that all visual modalities and entity ontologies have high semantic similarity, and it ignores the issue of low modality quality, greatly affecting the effectiveness of modal fusion.

We first consider addressing this issue from the perspective of improving the quality of visual modalities. Most of the original images in the datasets are only 30–50 KB in size. Therefore, we adopt super-resolution technology. This helps enhance the resolution of the images in the MMKG and optimize the subsequent semantic extraction and modeling effects. Super-resolution technology aims to preserve the semantics of the original image while increasing its resolution to enhance its clarity. EDSR [8] is a renowned work in the field of super-resolution, which enhances performance by removing redundant structures in traditional residual network based super-resolution methods. Although it was an early work, it was a state-of-the-art work at that time, and did not consume too much computing resources. Therefore, considering all factors, we use EDSR as the super-resolution tool in the model architecture.

Considering that the entity description images in the MMKB, which is the data source for FB15K-237 and DB13K, usually come from three different sources (Google, Bing, and Yahoo) and are often repeated. In order to save computational resources while maintaining the comparability of the final experimental results, we do not simply perform super-resolution on all images (FB15K-237 has approximately 270,000 entities, each with 0–15 descriptive images). Instead, we pre-use VGG-16 to extract semantic features from each descriptive image and compare them with the visual features of the entities provided in the dataset. Due to the small size of the images within the graph, extracting semantic features does not require extensive computational resources or time. Then, we select the image with the highest similarity for processing and super-resolution.

Specifically, we adopt EDSR to perform super-resolution on the target image Img_o, which increases the length and width by four times each, in order to obtain a clear image Img_s that preserves semantic information. Afterwards, we use the multi-modal pre-trained model CLIP to extract the features of Img_s, serving as a supplement to $\vec{h}_i^{vo}$. The primary reason for choosing CLIP is to maintain vector space consistency with the embeddings generated in the subsequent missing modality generation mechanism, and CLIP also has better visual semantic extraction capability than VGG-16:

$$\vec{h}_i^{vs} = CLIP(Img_s). \tag{8}$$

4.2.3 Multi-Level Visual Semantic Extraction

Most previous models simply used pre-trained visual encoders to extract visual features, which only allowed for coarse-grained semantic extraction of semantically rich visual modality, resulting in insufficient semantic similarity between visual embeddings and other embeddings. To address this issue, inspired by Monkey [26], we propose a multi-level visual semantic extraction mechanism. It aims to create rich and high-quality image descriptions by effectively mixing outputs from various generators.

Similar to Monkey, we use several advanced technologies or systems to combine: BLIP2 [27], which understands image semantics and generates descriptive text, PPOCR [28] extracts possible text information from images based on its powerful optical character recognition capability, GRiT [29] identifies targets and their positions in the image and performs detailed image text matching, FastSAM [30] segments images based on semantics, and Qwen1.5 [31], utilizing its powerful contextual understanding and generation capabilities.

Fig. 3 shows the pipeline for multi-level visual semantic extraction: (a) We first extract the name of the entity from the datasets as the core of the entire mechanism, hoping that the content generated will not deviate too much. (b) Subsequently we use BLIP2 to generate the global descriptions of entity visual information so as to analyze and describe the specific content of the image from an overall perspective. (c) Next, we adopt GRiT to identify specific areas and object coordinates in the image and generate detailed descriptions, PPOCR to extract possible textual information, FastSAM to segment objects. Then BLIP2 is adopted to generate descriptions of them. (d) A filter based on BLIP2 for image matching is used to delete low-confidence objects and areas recognized by GRiT and FastSAM. As for optical characters, we directly use the confidence output by PPOCR to screen them. (e) Finally, detailed information including global description, text extraction, and objects with spatial coordinates are fed into Qwen1.5 for fine-tuning, enabling it to generate semantically rich descriptive text. For the final generated summary description text, we use BLIP2 to extract its semantic features $\vec{h}_i^{vm}$ as the other auxiliary to the original visual features.

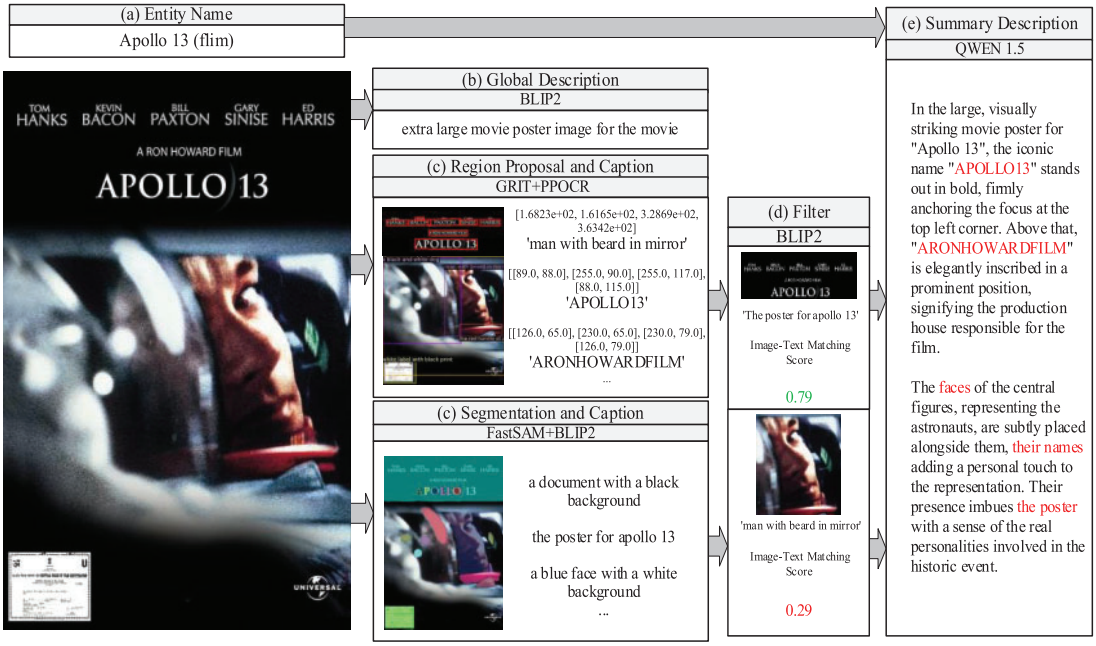

Figure 3: The pipeline for the multi-level visual semantic extraction

When constructing the prompt for inputting Qwen1.5, we imitated GRiT[2] by inputting captions into ChatGPT to generate image scene descriptions, and additionally informing it of the entity name and the task of generating entity descriptions, therefore achieving good generation results.

By aggregating the unique features of these systems and models, the multi-level visual semantic extraction mechanism captures the details in the visual modality and generates concise, accurate, and contextual descriptive text. Thus supplementing the entity images that were originally semantically sparse, and preventing poor multi-modal fusion caused by low modality quality.

4.3 Triple Semantic Embedding

As the main component of the MMKGC datasets, triple data contains the most important attributes and relations of entities. Previous models only extract adjacency relationships between entity nodes using graph neural networks. Based on these relationships, they generate structural embeddings for entity nodes and relationship edges. However, triples still embody a large amount of semantic information and the graph neural network cannot fully understand them.

Some studies have similarly recognized this issue and proposed using language models to extract the semantics of entity names as a supplement. Yet, the semantic information contained solely in entity names is not sufficiently rich and can easily lead to semantic confusion with entities that have similar names. All triples related to an entity contain all the knowledge associated with that entity in MMKGs. Therefore, we attempt to leverage large language models to extract all the entity-related semantics contained within them.

In detail, we input the name and all triples of each entity into the Qwen1.5 and prompt it to generate a small text description. The triple is highly correlated with the entity, and there must be a head or tail entity in the triple that is consistent with the target entity. Therefore, using the LLM to analyze triples for generating entity description yields better results. Small-scale testing revealed that when the number of triples related to an entity is three or more, the text description generated by the LLM retains the semantics of the triples effectively without introducing excessive noise.

A specific example of attribute embedding generation is provided in Fig. 2. We also use BLIP2 to extract features of text descriptions generated based on triples, treating them as alternative attribute embeddings $\vec{h}_i^a$ and serving as an external supplement to structural and visual embeddings:

$$\vec{h}_i^a = BLIP2(Qwen1.5(triples_i)). \tag{9}$$

4.4 Modal Fusion

Multi-modal fusion aims to integrate information from different modalities into a unified representation. Following the previous works, we project the structural, visual, and attribute embeddings of entities onto the same vector space based on the translational-distance function used in TransE. Afterwards, each uni-modal embedding is further input into a single-layer neural network for weighted fusion to get original multi-modal embedding $\vec{h}^m$, promoting them to also meet the translational-distance strategy.

For a specific triple $t_{i,j,k} = (e_i, r_k, e_j)$, the uni-modal translational-distance function can be denoted as:

$$f(t_{i,j,k}) = -||\vec{h}_i + \vec{g}_k - \vec{h}_j||_1. \tag{10}$$

[2] https://github.com/JialianW/GRiT (accessed on 01 December 2022).

Due to the general difficulty in finding appropriate visual images to describe the abstract relations within triples, simple feature concatenation or weighted summation is generally not used alone in modal fusion in MMKGC that focus on entity-entity relationships. Therefore, after obtaining the initial multi-modal embeddings through weighted summation, we train the network to ensure that the structural, visual, attribute, and joint multi-modal embeddings all meet the translational-distance strategy, thereby promoting spatial consistency among the embeddings:

$$f_{intramodal}(t_{i,j,k}) = -\|\vec{h}_i^{n_1} + \vec{g}_k - \vec{h}_j^{n_2}\|_1, n_1 \& n_2 \in \{s, v, a\}, n_1 = n_2, \tag{11}$$

$$f_{intermodal}(t_{i,j,k}) = -\|\vec{h}_i^{n_1} + \vec{g}_k - \vec{h}_j^{n_2}\|_1, n_1 \& n_2 \in \{s, v, a\}, n_1 \neq n_2, \tag{12}$$

$$f_{multimodal}(t_{i,j,k}) = -\|\vec{h}_i^{m} + \vec{g}_k - \vec{h}_j^{m}\|_1. \tag{13}$$

We additionally use the Margin Ranking Loss function, hoping that the model's ranking score for positive samples is higher than that for the negative samples generated by randomly modifying the head or tail entities:

$$L = \sum_{t_{i,j,k} \in S} \sum_{t_{i,j,k}' \in S'} max\{f(t_{i,j,k}) - f(t_{i,j,k'}) + \gamma, 0\}, \tag{14}$$

where $\gamma > 0$ is a margin hyper-parameter, S is the set of positive triples, S' is the set of negative triples, and $\vec{h}^m$ refer to the multi-modal joint embedding obtained through weighted sum.

The final loss function can be expressed as:

$$L_{all} = L_{intermodal} + L_{intramodal} + L_{multimodal}. \tag{15}$$

4.5 Decoder

We use ConvE to decode the triple embeddings obtained by concatenating multi-modal joint embeddings of entities and relations. For the triple $t_{i,j,k} = (e_i, r_k, e_j)$, the decoding process can be denoted as:

$$C^m(t_{i,j,k}) = \sigma(f(W_3\left(Vec(f\left(\left[\vec{h}_i^m, \vec{g}_k\right] * \omega\right))\right))\vec{h}_j^m), \tag{16}$$

where ω represents the convolutional filter, $*$ is the convolution operator, σ and f are the activation function and W_3 represents a linear transformation matrix used to compute the final score of the triple. Inspired by KBAT, we adopt the approach that fuses the final trained entity embeddings with the original embeddings to prevent knowledge and information loss during the training process. We additionally decode the final structural embeddings and the weighted sum of the two:

$$C(t_{i,j,k}) = C^m(t_{i,j,k}) + C^s(t_{i,j,k}). \tag{17}$$

The ConvE decoder is trained with binary cross-entropy loss:

$$L(C, t) = -\frac{1}{N} \sum \left(t_i \cdot log\left(C_i\right) + \left(1 - t_i\right) \cdot log\left(1 - C_i\right)\right), \tag{18}$$

where t is the label vector, the elements of t are 1 for relations existing and 0 otherwise.

4.6 *Additional Missing Modality Completion*

To address the issue of modality missing, we propose the ***additional missing modality completion*** mechanism, which primarily consists of two parts: ***missing modality generation*** and ***missing modality imagination***. Firstly, we consider leveraging the powerful natural language understanding capabilities of large language models to extract semantic knowledge from textual information of entities and generate simulated multi-modal joint embeddings to replace the missing modalities. Additionally, inspired by VAE [32] and UMAEA [33], we train a variational multi-modal auto-encoder to generate potential probability distributions of the missing modalities, thereby assisting in the training process. Fig. 4 shows the architecture of the additional missing modality completion.

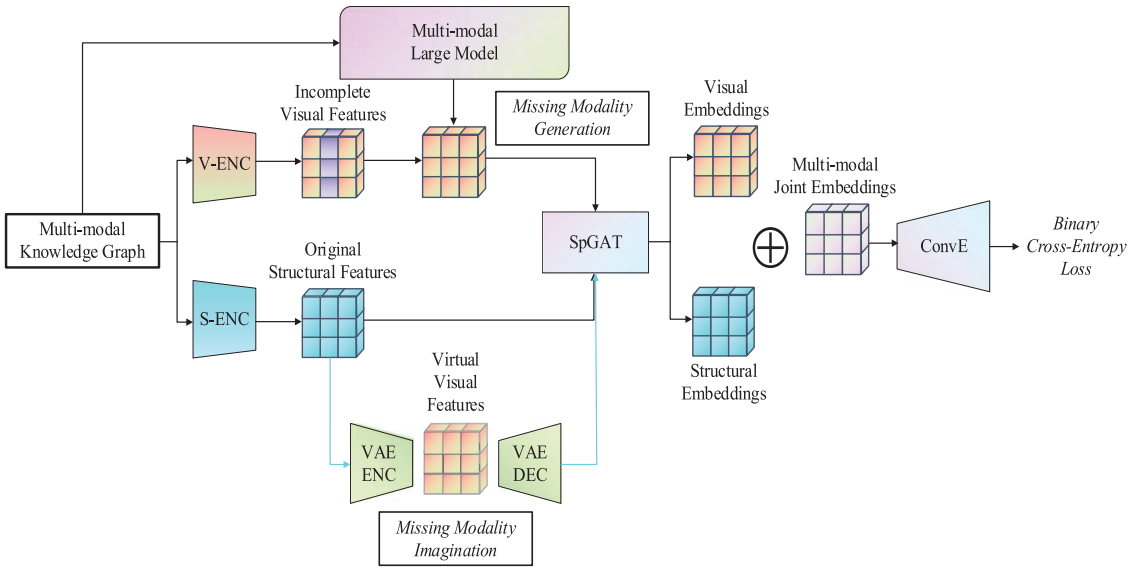

Figure 4: The overall architecture of the missing modality completion

4.6.1 *Missing Modality Generation*

Given the powerful natural language understanding capabilities of large language models, we leverage the latent multi-modal knowledge embedded in multi-modal LLMs to address the missing modality issue in MMKGC. To be specific, we employ CLIP [34] as the missing modality generator. CLIP is a multi-modal large model that utilizes 400 million image-text pairs for contrastive training. Through large-scale training, multi-modal semantic information is implicitly stored in its model parameters.

For a modality missing entity e_i, we extract its entity names and text descriptions, concatenate them, and then truncate the result for use as input $d_i = (w_1, w_2, \ldots, w_n)$ to ensure it does not exceed the maximum input length of CLIP. CLIP first converts it to corresponding tokens:

$$\vec{d}_i = ([CLS], w_1, w_2, \ldots, w_n, [SEP]), \tag{19}$$

where $[CLS]$ and $[SEP]$ are special tokens representing the beginning and end of the text, respectively. The embedding of $[CLS]$ is usually used to represent the entire text semantic information. We use the text encoder in CLIP to directly obtain the input text features, and further obtain multi-modal joint embedding as virtual visual embedding $\vec{h}_i^{vir1}$ through normalization:

$$\vec{h}_i^{vir1} = LayerNorm(CLIP(\vec{d}_i)). \tag{20}$$

The obtained virtual visual embeddings $\vec{h}_i^{vir1}$ will replace the original visual embeddings for entities that lack visual modalities.

4.6.2 Missing Modality Imagination

Inspired by VAE and UMAEA, based on a variational multi-modal auto-encoder framework, we train a set of Multi-Layer Perceptron (MLP) as encoder-decoder, input the original structural embedding $\vec{h}_i^s$, outputs the generated structural embedding $\vec{h}_i^{s,}$. We adopt the hidden layer between them as another virtual visual feature $\vec{h}_i^{vir2}$:

$$\left[\mu_i \oplus log\left(\sigma_i\right)^2\right] = MLP_{Enc}\left(\vec{h}_i^s\right), \tag{21}$$

$$\vec{h}_i^{vir2} = z \odot \sigma_i + \mu_i, z \sim \mathcal{N}(0, I), \tag{22}$$

$$\vec{h}_i^{s,} = MLP_{Dec}\left(\vec{h}_i^{vir2}\right). \tag{23}$$

μ_i and σ_i represent the mean and variance of the simulated Gaussian distribution z, respectively. And $\vec{h}_i^{s,}$ is the generated structural embedding output by decoder.

In order to improve the quality of generated virtual visual features, we set three different loss functions to train the MLPs. Firstly, the KL divergence L_{KL} is minimized to train the variational autoencoder with a potential space that is close to a Gaussian distribution. Secondly, the differences between the generated virtual visual and structural features and the real features are minimized:

$$L_{KL} = \mathbb{E}\left(\left(\mu_i\right)^2 + \left(\sigma_i\right)^2 - log\left(\sigma_i\right)^2 - 1\right)/2, \tag{24}$$

$$L_{re}^{vis} = |\vec{h}_i^{vir2} - \vec{h}_i^{v}|, \tag{25}$$

$$L_{re}^{str} = |\vec{h}_i' - \vec{h}_i|, \tag{26}$$

$$L_2 = L_{KL} + L_{re}^{vis} + L_{re}^{str}. \tag{27}$$

5 Experiment

In this chapter, we will report the experiment details including datasets, evaluation index, parameter settings, baselines and the results. We conduct experiments to answer the following three questions about MMCSD:

1. Question 1 (Q1): Has MMCSD shown improvement compared to the previous baselines?
2. Question 2 (Q2): Compared to other models that focus on the issue of missing modalities, does our proposed missing modality completion mechanism exhibit superiority?
3. Question 3 (Q3): Is the design of each part of MMCSD valid?

5.1 Datasets

We employed two multi-modal knowledge completion datasets for training and evaluation: the public benchmark FB15K-237, which is widely used in MMKGC, and DB13K, which we constructed based on knowledge from MMKB [3]. The specific statistics of the datasets are shown in Table 1.

Table 1: The statistics of datasets

Datasets	Entities	Relations	Train	Valid	Test	Multi-modal prop.
FB15K-237	14541	237	272115	17535	20466	91.05%
DB13K	12842	279	59434	19800	19796	99.96%

MMKB is a comprehensive multi-modal knowledge graph that encapsulates triples and visual data sourced from esteemed knowledge graphs like FB15K, DBpedia15K, and Yago15K. Inspired by the data format and structure of the mainstream MMKGC dataset FB15K-237, we have reconstructed the triples and visual modalities of DBpedia15K. We reassigned unique identifiers to the entities and relations in DBpedia15K and structured the triples accordingly. Additionally, we extracted pre-processed visual features associated with these entities and organized them into a visual feature matrix with the ordering determined by the newly assigned entity identifiers. This approach ensures consistency between visual features and triples in the newly created dataset, thereby enhancing its usability and versatility.

5.2 Evaluation Index

Following mainstream MMKG research, We use *Hit@n*, and Mean Reciprocal Rank (*MRR*) to objectively evaluate the effectiveness of the model. The larger *Hit@n* and *MRR* indicate the better performance of the model:

$$Hits@n = \frac{1}{|S|}\sum_{i}^{|S|} ||(rank_i \leq n), \tag{28}$$

$$MRR = \frac{1}{|S|}\sum_{i}^{|S|} \frac{1}{rank_i}. \tag{29}$$

Hit@n represents the probability that the top n items of the candidate triple prediction possibility rank have correct results, and *MRR* represents the average of the reciprocal of correct ranking in the candidates.

5.3 Implementation Details

MMCSD is trained on single NVIDIA RTX 3090 GPU. The proposed approach was implemented using Python 3.8.16, PyTorch 1.12.0, CUDA 12.1. We follow the step-by-step training method of KBAT, first training SpGAT to obtain embeddings of entities, relations, and visual information, then training the ConvE decoder for specific knowledge completion tasks.

The number of SpGAT training epochs is 3000, while for ConvE it is 300. We use Adam to optimize all the parameters with initial learning rate set at 0.001. Both the entity and relation embeddings of the final SpGAT layer are set to 200. The activation functions used for training ConvE are *Sigmoid* and *ReLU*.

For visual embedding, we use pre-processed visual features by trained VGG16 as original visual features [3]. We use the CLIP with ViT-B/32 as the multi-modal large model in MMCSD.

We follow the training approach of the UMAEA for missing modality imagination. In the first 1500 epochs of training SpGAT, only L_{all} was used for training, without involving missing modality imagination mechanisms. In the latter 1500 epochs, the modality imagination mechanism is added and SpGAT be trained using $L_{all} + L_2$. Because the missing modality imagination mechanism is used after multi-modal large model generation, we assume that all entities already contain their multi-modal information (including virtual multi-modal information), and during the decoding process, we weight the original embedding

onto the final embedding. Therefore, we do not freeze the main model during training missing modality imagination mechanisms.

5.4 Baselines

We compared our MMCSD against several knowledge graph completion models, encompassing both uni-modal and multi-modal approaches to demonstrate the superiority of our MMCSD. Firstly, we choose the conventional text-based uni-modal models for comparison to demonstrate the improvement brought by the visual information. For uni-modal models, we included TransE [11] and RotatE [13], which consider the relation in triples as translation vector from head entities to tail entities in different vector spaces. Additionally, we considered DistMult [35] and ComplEx [36], which attach great importance to mining the potential semantics of entities and relations. We also considered neural network-based models such as KBAT [17].

Secondly, we compared MMCSD with multi-modal models, including TransAE [3] and IKRL [1], which respectively extend TransE based on visual representations and multi-modal AutoEncoder. VBKGC [23] employs VisualBERT as a multi-modal encoder to capture the deeply fused multi-modal features of entities. MKBE [22] uses uni-modal embeddings as the context for attribute-specific decoders to generate the missing values of triples. MKGformer [37] leverages a hybrid transformer architecture with unified input-output and performs multi-modal fusion in multi-level. MoSE [5] learns modality-split embeddings for each modality to alleviate the modality interference. IMF [38] employs a two-stage multi-modal fusion framework to preserve modality-specific knowledge as well as to take advantage of the complementarity between different modalities. LAFA [39] designs a modality interaction attention mechanism that dynamically measures the contribution of images to entity embedding based on link information, thereby mitigating the impact of extraneous information in the visual modality on the complementation effect. CMR [40] proposes a unified cross-modal contrast learning to simultaneously capture multi-modal correlations of query-entity pairs in a unified representation space to improve the similarity of representations of useful semantic neighbors and then support the semantic neighbor retrieval. HKA [41] proposes macro- and micro-knowledge alignment module to capture the global semantic relevance between modalities and more effectively reveal the local consistency information through multi-modal supervisory effects.

In addition, we also compared MMCSD with MACO [7], which also considers the problem of modality missing and proposes to generate auxiliary embeddings by generative adversarial networks, to analyze the effectiveness of the additional missing modality completion method.

5.5 Main Results and Analysis (Q1 & Q2)

Table 2 shows the main results of the MMKGC experiments on two datasets (*: There is a label leakage error in KBAT, so the corrected result is poor compared with the result in their paper [17,23]), some baselines were not tested on the DB13K due to their code not being open-source or requiring additional external knowledge.

The comparison of results between MMCSD and uni-modal models in Table 2 demonstrates the importance of visual modality in MMKGC. Compared to uni-modal models, the majority of models that incorporate visual modalities show some degree of performance improvement. However, it is noticeable that the use of multi-modal information does not necessarily enhance the effectiveness of knowledge completion. For instance, some indices for IKRL and TransAE are even inferior to those of TransE. We believe that the poor performance of some models that add visual modalities may be attributed to their use of rough visual information and primitive modal fusion methods.

Table 2: Comparative results of MMCSD against other baseline methods on two datasets

Models	FB15K-237			DB13K		
	Hits@1	**Hits@3**	**MRR**	**Hits@1**	**Hits@3**	**MRR**
TransE	0.198	0.376	0.279	0.155	0.390	0.292
KBAT*	0.183	0.317	0.287	0.177	0.283	0.249
DistMult	0.199	0.301	–	0.185	0.299	0.256
ComplEx	0.194	0.297	–	–	–	–
RotatE	0.241	0.375	0.338	0.228	0.358	0.317
TransAE	0.199	0.317	–	0.161	0.376	0.289
IKRL	0.194	0.284	–	0.159	0.370	0.282
VBKGC	0.213	0.332	0.301	–	–	–
MKBE	0.258	–	0.347	–	–	–
MKGformer	0.256	0.367	–	0.241	0.369	0.323
MoSE	0.281	0.411	–	–	–	–
IMF	0.287	–	0.389	–	–	–
LAFA	0.269	0.398	–	–	–	–
CMR	0.263	0.395	–	–	–	–
HKA	0.291	0.424	–	–	–	–
MMCSD (ours)	**0.294**[1]	**0.425**	**0.395**	**0.291**	**0.405**	**0.384**

Note: [1]Bold font in the table indicates the best results. *There is a label leakage error in KBAT, so the corrected result is poor compared with the result in their paper [7,32].

Compared with multi-modal models, MMCSD achieved significant improvement in all indices. Specifically, compared to the best baseline IMF, *Hits@1* has increased by 0.7%, and *MRR* has improved by 0.6%. This demonstrates the positive effect of deep semantic exploration in various modalities and missing modality completion on MMKGC.

On our self-constructed DB13K, we similarly obtained comparable results to those from the FB15K-237, proving the generalizability of DB13K. Although the training set of DB13K is relatively small, it is difficult to obtain more supervised information compared to FB15K-237. However, due to the superiority of its data format and structure and the high richness of multi-modal information, the experiments conducted on it still achieved the same results as we expected.

To verify the effectiveness of missing modality completion on MMKGC, we set the proportion of multi-modal information in FB15K-237 to 20%, 40%, 60%, and 80% for experiments, while comparing MMCSD which only used additional missing modality completion (MMC-only) with MACO which employed IKRL as score function (IKRL-based). The experimental results for MACO, as reported in Fig. 3 of their paper [7], due to the lack of specific numerical values, only display the approximate maximum value. The contrast results are shown in Fig. 5.

It can be seen that compared to MACO, significant improvements have been made. Due to the different encoder-decoder and multi-modal fusion methods used, it cannot definitively be concluded that the missing modality generation in MMCSD is superior to the mechanism used in MACO. However, using pre-trained multi-modal large models and training VAE can save computational resources more, which is one of the advantages of our proposed method. It is worth noting that the completeness of the original visual features

does not necessarily correlate with the effectiveness of MMKGC. The specific promotion mechanism of visual modality for MMKGC is still not fully understood. Therefore, the methods and timing for integrating multi-modal information, as well as its specific impact on knowledge completion, should be key areas of focus in our future research.

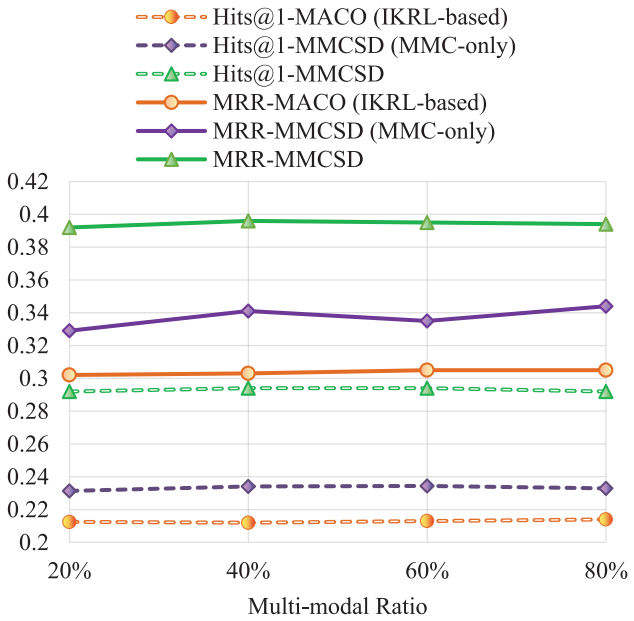

Figure 5: Comparison results of MMCSD which only used additional missing modality completion (MMC-only) with MACO which employed IKRL as score function (IKRL-based) in FB15K-237 with different multi-modal ratios

To evaluate the robustness of MMCSD, we conducted experiments on the FB15K-237 where entity description images were subjected to adversarial attacks. Considering that the original visual features were extracted by VGG16, we performed the Projected Gradient Descent (PGD) attack on VGG16 trained on the CIFAR10 targeting the image classification task and subsequently used the attacked model to generate adversarial samples. The PGD is implemented based on the open source Python library Adversarial Robustness Toolbox (ART).

The experimental results are presented in Table 3. It can be seen that MMCSD has better robustness compared to the model that only employs multi-modal fusion. By analyzing some of the cases, we find that although the original visual features of the described image are greatly changed after the attack, the accuracy of detail extraction by mature models such as object recognition and optical character recognition is not significantly reduced, so that the quality of Multi-level visual Semantic Extraction and Triple Semantic Embedding is still guaranteed.

Table 3: Comparative results of MMCSD on FB15K-237 after adversarial attacks

Datasets	**MMCSD (only modal fusion)**		**MMCSD**	
	Hits@1	**MRR**	**Hits@1**	**MRR**
FB15K-237	0.215	0.307	0.294	0.395
FB15K-237 (B.A.)	0.198	0.288	0.283	0.386
Decline	*0.017*	*0.019*	*0.011*	*0.009*

5.6 Ablation Study (Q3)

The ablation study is further conducted to prove the effects of different improvement methods. We designed experiments to compare the models without Super-Resolution (w/o SR), without Missing Modality Completion (w/o MMC), without Multi-level visual Semantic Extraction (w/o MSE), and without Triple Semantic Embedding (w/o TSE) respectively.

As shown in Table 4, the improvement effect of adopting MMC on the model is relatively weak, possibly because at least 91% of the entities in FB15K-237 contain descriptive image information, and other mechanisms also have a certain complementary effect on semantics. Meanwhile, the comparison of the performance with models that only applied multi-modal fusion also proves the effectiveness of enhancement methods proposed in this paper for MMKGC.

Table 4: The results of the ablation experiment

Models	FB15K-237		
	Hits@1	**Hits@3**	**MRR**
MMCSD (only modal fusion)	0.215	0.329	0.307
MMCSD (w/o MSE)	0.272	0.405	0.359
MMCSD (w/o SR)	0.279	0.405	0.364
MMCSD (w/o TSE)	0.278	0.410	0.362
MMCSD (w/o MMC)	0.287	0.416	0.388
MMCSD	**0.294**[1]	**0.425**	**0.395**

Note: [1]Bold font in the table indicates the best results.

Except for Missing Modality Completion, the other three mechanisms have effectively improved various indicators of MMKGC. Among them, Multi-level Visual Semantic Extraction, which aims at exploring semantics in visual modalities, has the best performance, consistent with our expectations. Visual modality, as the most unique and semantically rich information in MMKGs, has a significant promoting effect on the representation of multi-modal entities and triples. Therefore, further analysis and processing of visual modalities should be the key to improving the effectiveness of MMKGC in our future work.

Fig. 6 shows the decrease in indices after removing various optimization methods. The greater the decrease, the more important the optimization method is for improving the model performance. MSE, SR, and TSE add more semantic information of entities from different levels, which obviously improves the effectiveness of knowledge completion. MSE and SR respectively optimized the detail analysis mechanism and global semantic features of entity images, which have better improvement effects compared to the TSE method using triple text semantics. The MMC method, as a supplement and enhancement for modality missing entities, has a certain effect, but due to the limited number of modality missing entities in FB15K-237, the optimization is not as high as the other three methods.

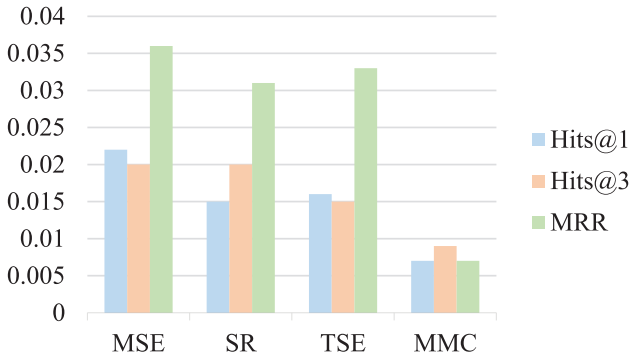

Figure 6: The impact of removing various optimization methods on MMCSD

5.7 Complexity Analysis

The whole model implementation can be divided into two phases: feature pre-extraction and model training. The feature pre-extraction phase involves the acquisition of the original visual features (usually the original visual features extracted by VGG16 are already available in the dataset), the super-resolution of the original image, the multi-level visual semantic extraction, the generation of triple semantic embeddings, and the implementation of the missing modality generation mechanism. The model training phase includes the training of the graph attention network encoder, the multi-modal fusion layer, the decoder, and the missing modality imagination mechanism.

Table 5 shows the performance and efficiency of the model training phase of MMCSD. Compared to the vanilla uni-modal knowledge completion model, MMCSD only adds the multi-modal fusion layer and the variational multi-modal auto-encoder, and therefore does not require significantly more computation and time in the model training phase.

Table 5: Performance evaluation metrics for MMCSD

Datasets	Avg. training time per epoch	FLOPs	Params.
FB15k-237	3.458 s	16.1964 G	1.1108 M
DB13K	1.326 s	14.3258 G	

As we use a number of well-established techniques to extract semantic information from the original multi-modal data, some additional time is required beyond the model training phase. Table 6 shows the average time to process a single piece of data for the established methods covered in this paper when using a single 3090 GPU, and the application of large language model to summarize the semantics of entities extracted by other methods and generate descriptive text takes more time and computational resources. As the dataset expands, the time required for the feature pre-extraction phase will increase accordingly. The total time depends only on the number of entities in the dataset and the number of descriptive images, but not on the number of triples.

Table 6: Average processing time per data item (text or image) for the each mature model

Models	Average time (s)	GPU memory usage (MiB)
ESDR	2.075	3690
BLIP2	2.406	15360
PPOCR	3.686	1460
GRiT	0.21	6340
FastSAM	0.052	3080
Qwen1.5	15.478	16450
CLIP	1.972	1808

6 Conclusion

Aiming at the problem of low modality quality and modality missing, we propose MMCSD, a multi-modal knowledge graph completion model based on super-resolution and detailed description generation. The core idea is to use super-resolution and multi-level image semantic extraction to further enrich semantic information of multi-modal embeddings. Moreover, we additionally use the missing modality generation and imagination mechanisms to complete the visual feature of modality missing entities. The experiment shows that our proposed methods have a significant positive effect on MMKGC.

In future work, we will attempt to use more novel graph neural networks, transformer architecture, and even large language models to extract deep information in graph-structured data to build better-quality structural embeddings. Secondly, it can be seen from the experiment that the promotion effect of multi-modal information and its fusion timing and method on MMKGC is not clear enough. We will strive to explore its principles and further optimize the multi-modal knowledge completion effect.

Acknowledgement: The authors are grateful to all the editors and reviewers for their detailed review and insightful advice.

Funding Statement: This research was funded by Research Project, grant number BHQ090003000X03.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Huansha Wang, Ruiyang Huang; data collection: Huansha Wang; analysis and interpretation of results: Huansha Wang, Qinrang Liu; draft manuscript preparation: Huansha Wang; visualization: Shaomei Li, Jianpeng Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Huansha Wang, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

Abbreviations

MMKG	Multi-modal knowledge graph
MMKGC	Multi-modal knowledge graph completion
MMCSD	Multi-modal knowledge graph Completion model based on Super-resolution and Detailed Description Generation
LLM	Large language model

GCN	Graph convolutional network
GAT	Graph attention network
SpGAT	Sparse graph attention network
PVM	Pre-trained visual model
MLP	Multi-Layer Perceptron

References

1. Xie R, Liu Z, Luan H, Sun M. Image-embodied knowledge representation learning. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017 Aug 19–26; Melbourne, VIC, Australia: International Joint Conferences on Artificial Intelligence Organization. Vol. 2017, p. 3140–6. doi:10.24963/ijcai.2017/438.
2. Mousselly Sergieh H, Botschen T, Gurevych I, Roth S. A multimodal translation-based approach for knowledge Graph Representation learning. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics; 2018; New Orleans, Louisiana, Stroudsburg, PA, USAACL. p. 225–34. doi:10.18653/v1/s18-2027.
3. Wang Z, Li L, Li Q, Zeng D. Multimodal data enhanced representation learning for knowledge graphs. In: 2019 International Joint Conference on Neural Networks (IJCNN); 2019 Jul 14–19; Budapest, Hungary: IEEE. Vol. 2019, p. 1–8. doi:10.1109/ijcnn.2019.8852079.
4. Zhang Y, Chen M, Zhang W. Modality-aware negative sampling for multi-modal knowledge graph embedding. In: 2023 International Joint Conference on Neural Networks (IJCNN); 2023 Jun 18–23; Gold Coast, QSL, Australia: IEEE. Vol. 2023, p. 1–8. doi:10.1109/IJCNN54540.2023.10191314.
5. Zhao Y, Cai X, Wu Y, Zhang H, Zhang Y, Zhao G, et al. MoSE: modality split and ensemble for multimodal knowledge graph completion. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022; Abu Dhabi, United Arab Emirates. p. 10527–36. doi:10.18653/v1/2022.emnlp-main.
6. Zhang Y, Chen Z, Guo L, Xu Y, Zhang W, Chen H. Making large language models perform better in knowledge graph completion. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024; Melbourne VIC Australia: ACM. p. 233–42. doi:10.1145/3664647.3681327.
7. Zhang Y, Chen Z, Zhang W. MACO: a modality adversarial and contrastive framework for modality-missing multi-modal knowledge graph completion. In: Natural language processing and Chinese computing. Cham: Springer Nature Switzerland; 2023. p. 123–34. doi: 10.1007/978-3-031-44693-1_10.
8. Lim B, Son S, Kim H, Nah S, Lee KM. Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2017 Jul 21–26; Honolulu, HI, USA: IEEE. Vol. 2017, p. 1132–40. doi:10.1109/CVPRW.2017.151.
9. Toutanova K, Chen D, Pantel P, Poon H, Choudhury P, Gamon M. Representing text for joint embedding of text and knowledge bases. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon, Portugal, Stroudsburg, PA, USAACL; p. 1499–509. doi:10.18653/v1/d15-1174.
10. Liu Y, Li H, Garcia-Duran A, Niepert M, Onoro-Rubio D, Rosenblum DS. MMKG: multi-modal knowledge graphs. In: The semantic web. Cham: Springer International Publishing; 2019. p. 459–74. doi: 10.1007/978-3-030-21348-0_30.
11. Bordes A, Usunier N, Duran AG, Weston J, Yakhnenko O. Translating embeddings for modeling multi relational data. In: NIPS'13: Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS); 2013 Dec 5–8; Lake Tahoe, Nevada, USA. p. 2787–95.
12. Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. Proc AAAI Conf Artif Intell. 2014:1112–9. doi:10.1609/aaai.v28i1.8870.
13. Sun Z, Deng Z, Nie J, Tang J. RotatE: knowledge graph embedding by relational rotation in complex space. arXiv:1902.10197. 2019.
14. Dettmers T, Minervini P, Stenetorp P, Riedel S. Convolutional 2D knowledge graph embeddings. Proc AAAI Conf Artif Intell. 2018;1811–8. doi:10.1609/aaai.v32i1.11573.

15. Nguyen DQ, Nguyen TD, Nguyen DQ, Phung D. A novel embedding model for knowledge base completion based on Convolutional neural network. In: Proceedings of the 2018 Conference of the North American Chapter Of the Association for Computational Linguistics: Human Language Technologies; 2018; New Orleans, LA, USA; p. 327–33. doi:10.18653/v1/n18-2053.
16. Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling relational data with graph convolutional networks. In: The semantic web. Cham: Springer International Publishing; 2018. p. 593–607. doi: 10.1007/978-3-319-93417-4_38.
17. Nathani D, Chauhan J, Sharma C, Kaul M. Learning attention-based embeddings for relation prediction in knowledge graphs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence, Italy, Stroudsburg, PA, USAACL; p. 4710–23. doi:10.18653/v1/p19-1466.
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 5998–6008.
19. Ferrada S, Bustos B, Hogan A. IMGpedia: a linked dataset with content-based analysis of wikimedia images. In: The semantic web-ISWC 2017. Cham: Springer International Publishing; 2017. p. 84–93. doi: 10.1007/978-3-319-68204-4_8.
20. Alberts H, Huang N, Deshpande Y, Liu Y, Cho K, Vania C, et al. VisualSem: a high-quality knowledge graph for vision and language. In: Proceedings of the 1st Workshop on Multilingual Representation Learning; 2021; Punta Cana, Dominican Republic. Stroudsburg, PA, USAACL. p. 138–52. doi:10.18653/v1/2021.mrl-1.13.
21. Deng C, Jia Y, Xu H, Zhang C, Tang J, Fu L, et al. GAKG: a multimodal geoscience academic knowledge graph. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management; 2021 Nov 1–5; Queensland, Australia. p. 4445–54. doi:10.1145/3459637.3482003.
22. Pezeshkpour P, Chen L, Singh S. Embedding multimodal relational data for knowledge base completion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018; Brussels, Belgium, Stroudsburg, PA, USAACL. p. 3208–18. doi:10.18653/v1/d18-1359.
23. Zhang Y, Zhang W. Knowledge graph completion with pre-trained multimodal transformer and twins negative sampling. arXiv:2209.07084. 2022.
24. Kipf TN, Welling M. Semi supervised classification with graph convolutional networks. arXiv:1609.02907. 2017.
25. Velickovic P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv:1710.10903. 2018.
26. Li Z, Yang B, Liu Q, Ma Z, Zhang S, Yang J, et al. Monkey: image resolution and text label are important things for large multi-modal models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE. Vol. 2024, p. 26753–63. doi:10.1109/CVPR52733.2024.02527.
27. Li J, Li D, Savarese S, Hoi SCH. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 19730–42.
28. Du Y, Li C, Guo R, Yin X, Liu W, Zhou J, et al. PP-OCR: a practical ultra lightweight OCR system. arXiv:2009.09941. 2020.
29. Wu J, Wang J, Yang Z, Gan Z, Liu Z, Yuan J, et al. GRiT: a generative region-to-text transformer for object understanding. In: Computer Vision-ECCV 2024-18th European Conference; 2024 Sep 29–Oct 4; Milan, Italy. p. 207–24. doi:10.1007/978-3-031-72989-8_12.
30. Zhao X, Ding W, An Y, Du Y, Yu T, Li M, et al. Fast segment anything. arXiv:2306.12156. 2023.
31. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, et al. Qwen technical report. arXiv.2309.16609. 2024.
32. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:1312.6114. 2022.
33. Chen Z, Guo L, Fang Y, Zhang Y, Chen J, Pan J, et al. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In: 22nd International Semantic Web Conference; 2023 Nov 6–10; Athens, Greece; p. 121–39. doi:10.1007/978-3-031-47240-4.

34. Radford A, Kim J, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (PMLR 139); 2021 Jul 18–24; p. 8748–63.
35. Yang B, Yih W, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. In: 3rd International Conference on Learning Representations, ICLR 2015; 2015 May 7–9; San Diego, CA, USA.
36. Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. In: Proceedings of the 33rd International Conference on Machine Learning; 2016 Jun 19–24; New York City, NY, USA. p. 2071–80.
37. Chen X, Zhang N, Li L, Deng S, Tan C, Xu C, et al. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2022; Madrid Spain: ACM. p. 904–15. doi:10.1145/3477495.3531992.
38. Li X, Zhao X, Xu J, Zhang Y, Xing C. IMF: interactive multimodal fusion model for link prediction. In: Proceedings of the ACM Web Conference 2023; 2023; Austin TX USA: ACM. p. 2572–80. doi:10.1145/3543507.3583554.
39. Shang B, Zhao Y, Liu J, Wang D. LAFA: multimodal knowledge graph completion with link aware fusion and aggregation. Proc AAAI Conf Artif Intell. 2024;38(8):8957–65. doi:10.1609/aaai.v38i8.28744.
40. Zhao Y, Zhang Y, Zhou B, Qian X, Song K, Cai X. Contrast then memorize: semantic neighbor retrieval-enhanced inductive multimodal knowledge graph completion. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024; Washington, DC, USA: ACM. p. 102–11. doi:10.1145/3626772.3657838.
41. Xu Y, Li Y, Xu M, Zhu Z, Zhao Y. HKA: a hierarknowledge alignment framework for multimodal knowledge graph completion. ACM Trans Multimed Comput Commun Appl. 2024;20(8):1–19. doi:10.1145/3664288.