Multi-modal Siamese Network for Few-shot Knowledge Graph Completion

Yuyang Wei Wei Chen* Xiaofang Zhang Pengpeng Zhao Jianfeng Qu Lei Zhao*
School of Computer Science and Technology, Soochow University
yyweisuda@stu.suda.edu.cn
{robertchen,xfzhang,ppzhao,jfqu,zhaol}@suda.edu.cn

2024 IEEE 40th International Conference on Data Engineering (ICDE) | 979-8-3503-1715-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDE60146.2024.00061

***Abstract*—Multi-modal data have recently been utilized to improve the performance of knowledge graph completion (KGC), attracting widespread research interest. However, they have been ignored in few-shot knowledge graph completion (FKGC), which aims to discover potential facts involving unseen relations that only appear in few-shot triples. The most relevant FKGC study simply concatenates various modal features, but the performance is still limited due to the following problems: (1) lack of exploiting significant multi-modal features in neighborhoods, and (2) ineffectively modeling inter-modal interactions in a few-shot setting. To tackle these problems, we propose a novel relational learning model entitled MMSN (Multi-Modal Siamese Network) for few-shot knowledge graph completion, which is composed of the following two primary modules: the Siamese multi-modal neighbor encoder (SMNE) and the meta-learning multi-modal knowledge representation decoder (MKRD). The module SMNE is developed to encode diverse modalities of neighbors by a Siamese attention network, fuse multi-modal information through a gating fusion network, and learn effective relational embeddings using an aggregator. The module MKRD is introduced to handle inter-modal interactions between multiple modalities and train the proposed model in a few-shot scenario. Extensive experiments demonstrate that our proposed model MMSN outperforms the state-of-the-art FKGC models, including uni-modal and multi-modal models, on two real-world few-shot multi-modal datasets.**

***Index Terms*—Multi-modal Knowledge Graph, Few-shot Knowledge Graph Completion, Unseen Relations, Meta-learning, Multi-modal Fusion**

I. Introduction

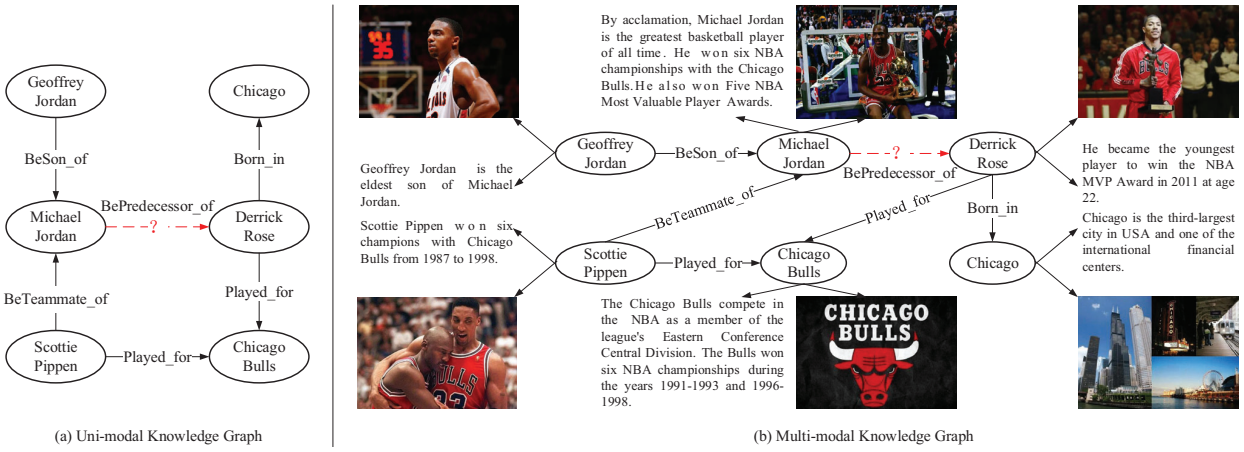
Knowledge graphs (KGs) like FreeBase [1], Wikidata [2], and YAGO [3] are widely used to improve the performance of downstream tasks, e.g., semantic search [4] [5] [6], question answering [7] [8], and entity recognition [9] [10]. Despite the successful applications of KGs, most of them still suffer from incompleteness. To automatically complete KGs, knowledge graph completion (KGC) [11] [12] [13] [14] is proposed to discover potential facts and has achieved great contributions. With the development of multi-modal knowledge graphs (MMKGs), some multi-modal KGC methods [15] [16] [17] [18] have been proposed to utilize multi-modal data, such as image features [19] [20] and text descriptions [21], to further improve KGC performance. However, the above methods only focus on completing static knowledge graphs and ignore the emergence of new relations in the updated knowledge graphs.

* These authors are corresponding authors.

The emergence of new relations, which are unseen in background KGs, introduces new potential facts that need to be completed, further exacerbating the incompleteness of KGs. Especially, a large portion of unseen relations only appear in a few facts, formulated as few-shot relations [22], making the KGC task more challenging. To tackle this challenge, several meta-learning approaches [23] [24] [25] have been proposed to conduct few-shot knowledge graph completion (FKGC). These existing methods primarily focus on leveraging the local knowledge graph structure to improve entity representations in the support set and aggregate these enhanced representations to learn the core information for FKGC, i.e., few-shot relational vectors, ultimately achieving effective FKGC. For example, given a uni-modal knowledge graph presented in Fig. 1(a), the neighbor "Scottie Pippen", a teammate of "Michael Jordan", played for "Chicago Bulls", thereby the structured knowledge of "Scottie Pippen" provides information for inferring that "Michael Jordan" played for "Chicago Bulls". Encoding this neighborhood structured knowledge, the existing methods enhance entity representations and facilitate the completion of the new relation "BePredecessor_of" from "Michael Jordan" to "Derrick Rose". Despite the significant contributions of the above studies, they neglect the rich auxiliary multi-modal information in FKGC scenarios. To address this problem, MULTIFORM [25] concatenates the pre-trained representations of multiple modalities to further enhance entity representations and improve FKGC performance.

Although MULTIFORM has achieved a promising performance, it is just a preliminary attempt that exploits multi-modal features to enhance entity representations by simple concatenation operation. The multi-modal FKGC performance is still limited by the following two problems. (1) **Lack of exploiting significant multi-modal features in neighborhoods**. For example, given a multi-modal knowledge graph presented in Fig. 1(b), we can learn from the multi-modal knowledge of "Scottie Pippen" that he played for the "Chicago Bulls" from 1987 to 1998. As the teammate of "Scottie Pippen", "Michael Jordan" played for "Chicago Bulls" in the 1990s, while "Derrick Rose" won the Most Valuable Player Award in 2011 with the Bulls. We can infer that "Michael Jordan" is a predecessor of "Derrick Rose". Based on this inference, the multi-modal information of neighbors can provide valuable information to enhance entity representation and learn effective few-shot relation representations in few-shot

2375-026X/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDE60146.2024.00061



(a) Uni-modal Knowledge Graph

(b) Multi-modal Knowledge Graph

Fig. 1. A motivating example of neighborhood multi-modal features. Both the uni-modal knowledge graph on the left and the multi-modal knowledge graph on the right have the same entities. The multi-modal knowledge graph has auxiliary data including images and texts. The "BePredecessor_of" is an unseen relation to be completed.

knowledge graph completion. Therefore, how to effectively encode neighborhood multi-modal knowledge to enhance entity representations is significant in promoting the performance of FKGC in multi-modal scenarios. (2) **Ineffectively modeling inter-modal interactions in a few-shot setting**. In this study, representations of unseen relations are learned by an encoder that exploits neighborhood multi-modal features. However, the commonly used uni-modal decoders are not reasonable for translating relations, which contain multi-modal information, on structured entities. Thus, it is essential to introduce a decoder with the capability of exploiting multiple modalities to complete facts of unseen relations. Additionally, different modalities of an entity are dependent, leading to inter-modal interactions between multiple modalities [16] [26] [27]. Although IKRL [28] and TransAE [29] with high complexity can handle inter-modal interactions, training these models requires a large number of instances, which are infeasible in a few-shot scenario. Therefore, these multi-modal models are challenging to be directly applied to model inter-modal between multiple modalities in few-shot settings.

To address the above-mentioned problems, we propose a novel relational learning model entitled MMSN (**M**ulti-**M**odal **S**iamese **N**etwork) for few-shot knowledge graph completion, which contains the following two modules: Siamese multi-modal neighboring encoder (SMNE) and meta-learning multi-modal knowledge representation decoder (MKRD). These two modules are designed to complement each other in tackling the FKGC problem with multiple modalities, where SMNE is a novel module to integrate the structured and multi-modal information in neighborhoods to learn few-shot relational representations, and MKRD is a meta-learning multi-modal decoder to train FKGC model in few-shot scenarios. Specifically, SMNE first utilizes a Siamese attention network to encode structured information and multi-modal data within the neighborhood, where modalities share the same attention weights. The Siamese attention network technique [30] has been proven to be an effective approach for collectively capturing valuable information from multiple modalities and is extensively employed in other knowledge graph tasks. In this study, the neighboring structured knowledge and neighboring multi-modal knowledge are fed into the Siamese attention network in pairs. In general, various modal data of key neighbors typically contain valuable information. For example, the multi-modal data and structured knowledge of "Scottie Pippen" complement each other in improving the inference that "Michael Jordan" is the predecessor of "Derrick Rose". Therefore, we use the Siamese attention network to assign the same weight to structured and multi-modal data of a neighbor, successfully avoiding assigning abnormal weights to specific modalities of key neighbors, and ensuring robust and accurate results. Next, SMNE fuses different modal features to enhance entity representations with a gating fusion network that filters the noise of modalities. Furthermore, SMNE employs an LSTM-based inner-attention network to aggregate enhanced support entity pair representations and learn few-shot relation representations. By incorporating multi-modal information into the few-shot relational representations learned by SMNE, the module MKRD introduces meta-learning into a multi-modal KGE method IKRL to model multi-modal features within relations and entities. MKRD enables modeling inter-modal interactions between structured and multi-modal features by combining four score functions in the few-shot scenario. Moreover, extensive experiments are conducted on two real-world few-shot datasets, and the results demonstrate that our proposed model outperforms the state-of-the-art baselines. The contributions of this paper can be summarized as follows.

- To the best of our knowledge, we are the first to explore neighborhood multi-modal features and inter-modal interactions in few-shot knowledge graph completion.
- We propose a multi-modal FKGC model MMSN, which includes the following two main modules. Firstly, the

module SMNE is designed to encode neighborhood multi-modal information by a Siamese network, and fuse different modalities with a gating fusion network, learn unseen relation representations using an aggregator. Secondly, the module MKRD considers the multi-modal knowledge within relational representations and models inter-modal interactions between structured features and multi-modal features to perform FKGC.
- Extensive experiments on two few-shot multi-modal datasets have been conducted, and the results demonstrate the superiority of MMSN compared with the state-of-the-art models.

The remainder of this paper is structured as follows. Firstly, we briefly review the related work in Section II. Then, we introduce the preliminary and formulate the problem in Section III and present MMSN to complete facts of unseen relations in Section IV. Extensive experiments are conducted on two real-world multi-modal datasets in Section V. Lastly, we conclude the paper with future scope in Section VI.

II. Related Work

A. Traditional KGC Models

Traditional KGC models [11] [13] [14] aim at completing potential facts for seen relations. A promising approach is knowledge graph embedding [31] [32] [33] which embeds entities and relation into a low continuous vector space. Traditional KGE models are categorized into three groups: translation-based models [31] [32] [34], bilinear-based models [35] [36] [37] and neural network models [38] [39] [40].

Translation-based models measure the plausibility of facts by calculating the distance between entities through relations. The well-known model of this work TransE [31] interprets each relation as a translation from a head to a tail. Following the key idea of TransE, the later translation-based models [32] [34] formulate each relation as multiple vectors to deal with complex relations, such as 1-N, N-1, and N-N. For example, TransH [32] initializes each relation with a translation vector and a hyperplane, projecting entities into corresponding relation-specific hyperplanes to achieve translation. TransR [34] projects the entity into corresponding relational space, and translates relations from heads to tails in various vector spaces. The bilinear model is started by RESCAL [35], which employs relational matrices to capture interactions between entities. To solve the overfitting of RESCAL, DisMult [36] defines the relations as diagonal matrices and significantly reduces parameters. ComplEx [37] argues that the real space is inadequate to model interactions between entities and extends bilinear model to the complex space. The last typical models are neural network approaches, which obtain more advanced performance using more complex operations. ConvE [39] utilizes a convolutions neural network to jointly capture the features of entities and relations. R-GCN [38] argues the heterogeneity of knowledge graphs and considers different impacts of relation in the neighbor encoder.

All of the above models are uni-modal traditional models, which focus on the structured information of KGs. To further promote KGC performance, IKRL [28] exploits text and image that intuitively describes the appearances and behaviors of entities. IKRL interprets each relation as a translation from head to tail with four energy functions that model interaction between structured modality and other modalities. Compared with IKRL, which separately learns image information and structured information, TransAE [29] introduces an auto-encoder to integrate multi-modal information, and combines it and TransE to learn a joint objective that models both multi-modal knowledge and structured knowledge simultaneously. MKGFormer [41] is a hybrid Transformer network for multimodal knowledge graph completion, which presents a multi-modal encoder to model image-text incorporated entity representations with multi-level fusion at the last several layers of ViT and BERT. These traditional models have made great contributions to the KGC task. However, they require adequate instances to train models, ignoring knowledge graphs that emerge new relations, i.e., unseen relations, with insufficient instances.

B. Few-shot KGC Models

Unlike traditional models that primarily handle seen relations, few-shot KGC models aim to complete potential facts involving unseen relations [22] [24] [25]. Most existing few-shot KGC models employ a meta-learning framework [42] [43] [44] and seek to enhance entity representations. One line of work is started by GMatching [22], which uses a meta-learning framework to complete facts in one-shot scenarios. GMatching introduces a local neighbor encoder to enhance entity embeddings, but assumes all neighboring relations contribute equally to the entity embedding in the neighbor encoder. FSRL [45] argues that the knowledge graphs are heterogeneous, and neighboring relations may have different impacts on entity embeddings. To address this problem, FSRL proposes a heterogeneous neighbor encoder that assigns different weights to neighboring relations and designs an LSTM-based aggregator to model the interaction among few-shot instances. FAAN [46] argues that entities have dynamic natures and the weights should not remain constant across all few-shot relations. Therefore, FAAN introduces a novel neighbor encoder to adaptively encode entity pairs. In contrast to the above models, MetaR [23] transfers shared knowledge from support sets to the queries based on a novel optimization strategy. Combining entity representation enhancement and knowledge transfer, GAAN [24] proposes a novel gating neighbor encoder to capture value information in neighborhoods and employs a meta-learning TransH (MTransH) to model complex few-shot relations, such as N-1, 1-N, and N-N. HiRe [47] is a hierarchical relational learning framework for FKGC, which captures three levels of relational information, including entity-level, triplet-level, and context-level, to enrich entity and relation representations. However, these models focus on the structured features of knowledge graphs, and ignore auxiliary multi-modal information, such as images and text, in few-shot scenarios.

MULTIFORM [25] is the most relevant study specifically designed to handle few-shot MMKGC. This model enhances entity representations by simply concatenating a variety of modal representations, such as structure, text, and image. However, despite its advanced capabilities, the performance of MULTIFORM is still hindered by the following limitations. For instance, multi-modal features in neighborhoods containing abundant auxiliary information have not yet been considered in the process of enhancing entity representations. Furthermore, the large amount of noise inherent in these multi-modal features is not effectively filtered by mere concatenation, leading to suboptimal performance. Having observed the drawbacks of existing work, we propose a novel multi-modal few-shot knowledge graph completion model MMSN in this work. This model incorporates neighborhood multi-modal information in the neighbor encoder and considers inter-modal interactions in the meta-learning knowledge representation decoder. Unlike the former model, which depends on pre-trained models for various modalities, our proposed model focuses on the application and fusion of multi-modal information.

III. Preliminary and Definition

In this section, we first present definitions used throughout the paper and formulate the multi-modal few-shot knowledge graph completion, then detail the corresponding completion setting in a meta-learning framework.

A. Problem Definition

A background KG is expressed as a collection of triples $\mathcal{G} = \{(h, r, t)|h \in \mathcal{E}, r \in \mathcal{R}, t \in \mathcal{E}\}$, where $\mathcal{E}$ is the entity set and $\mathcal{R}$ is the relation set. In this work, the FKGC task aims to predict the tail given the head entity and the few-shot relation, i.e., $(h, r, ?)$, where r is unseen in the background KG, i.e., $r \notin \mathcal{G}$, and is associated with few-shot triples. To improve the completion performance, we explore the FKGC in a multi-modal knowledge graph scenario, where each entity has two auxiliary modalities, including text and image. To better illustrate our model in this study, some definitions are presented as follows.

Definition 1: **Multi-modal Background Knowledge Graph (MMBKG).** A background multi-modal knowledge graph $\tilde{\mathcal{G}} = \{(h, r, t)|h \in \tilde{\mathcal{E}}, r \in \mathcal{R}, t \in \tilde{\mathcal{E}}\}$ is an extension of the background knowledge graph $\mathcal{G}$ by adding multi-modal auxiliary data, where each entity $e \in \tilde{\mathcal{E}}$ has structured data and multi-modal data.

Definition 2: **Few-shot Relation.** If the number of triples connected by an unseen relation r is equal to K, the relation r is called a K-shot relation. If K is less than or equal to a given low-frequency threshold, the K-shot relation r is referred to as a few-shot relation in the few-shot knowledge graph completion.

Problem Definition 1: **Multi-modal Few-shot Knowledge Graph Completion (MFKGC).** Similar to the FKGC [22] [45], given a few-shot relation r and its support triple set $\mathcal{S}_r = \{(h_i, r, t_i)\}_{i=1}^{K}$, the MFKGC task is to rank all possible tail entity in the candidate set for a new triple $(h_j, r, ?)$ using structured data and multi-modal data, simultaneously.

During the test period, we build the candidate entity set C based on the entity type constraint [48] [22] to predict facts of new relations. The candidate entity set is a closed set of entities that precludes unseen entities in the background multi-modal knowledge graph.

B. Meta-learning Completion Setting

To complete facts for unseen relations in a few-shot scenario, the meta-learning framework is utilized to conduct MFKGC. Following the standard FKGC setting, we first build a meta-training task set $\mathcal{T}_{mtr}$, where each task is a set of triples connected by a few-shot relation $r \in \mathcal{R}_{mtr}$ and has its own support/query set, i.e., $\mathcal{T}_{mtr} = \{\mathcal{S}_r, \mathcal{Q}_r\}$. To simulate the few-shot scenario, each $\mathcal{S}_r = \{(h_0, r, t_0), ..., (h_k, r, t_k), ...(h_K, r, t_K)\}$ contains K support triples, and $\mathcal{Q}_r = \{(h_i, r, t_i, C_{h_i,r})\}$ is the query triples of r, with the tail entities comprising the ground-true tail and the corresponding candidate entities $t_j \in C_{h_i,r}$. The proposed model could be tested on the query set $\mathcal{Q}_r$ by ranking all candidate entities given a query $(h_i, r, ?)$.

To evaluate the model performance, we construct a meta-testing task set $\mathcal{T}_{mte}$, where each task consists of triples connected by a few-shot relation $r \in \mathcal{R}_{mte}$. Note that, each relation $r \in \mathcal{R}_{mte}$ is unseen from meta-training set, i.e., $\mathcal{R}_{mte} \cap \mathcal{R}_{mtr} = \emptyset$. Similar to meta-training relations, each meta-testing relation $r \in \mathcal{R}_{mte}$ also has its own support triple set $\mathcal{S}_r$ and query triple set $\mathcal{Q}_r$. In addition, we construct a meta-valid task set $\mathcal{T}_{mv}$ with a relation set $\mathcal{R}_{mv}$ to tune hyper-parameters during the training period, where $\mathcal{R}_{mv} \cap \mathcal{R}_{mte} = \emptyset$ and $\mathcal{R}_{mv} \cap \mathcal{R}_{mtr} = \emptyset$. Moreover, all relations in $\mathcal{T}_{mtr}$, $\mathcal{T}_{mte}$ and $\mathcal{T}_{mv}$ are unseen in the background multi-modal KG $\tilde{\mathcal{G}}$.

IV. The Proposed model

In this section, we first present an overview of our proposed model MMSN, consisting of two primary modules SMNE and MKRD. Then, we introduce the two modules and the model training objective in detail.

A. Overview

The model overview of MMSN, which consists of two modules SMNE and MKRD, is intuitively presented in Fig. 2. Specifically, to predict a target link for an unseen relation, the proposed model first separately pre-trains structured data and multi-modal data with corresponding methods to initialize various modalities embeddings. Then, the module SMNE encodes structured and multi-modal features in neighborhoods with a Siamese attention network, fuses the encoded features to enhance entity representations with a gating fusion network, and employs an LSTM-based inner-attention aggregator [49] to learn few-shot relation representations. Additionally, the module MKRD is designed to train the proposed model in the few-shot scenario, considering multi-modal information contained in few-shot relational representations and modeling inter-modal interactions between multiple modalities. SMNE

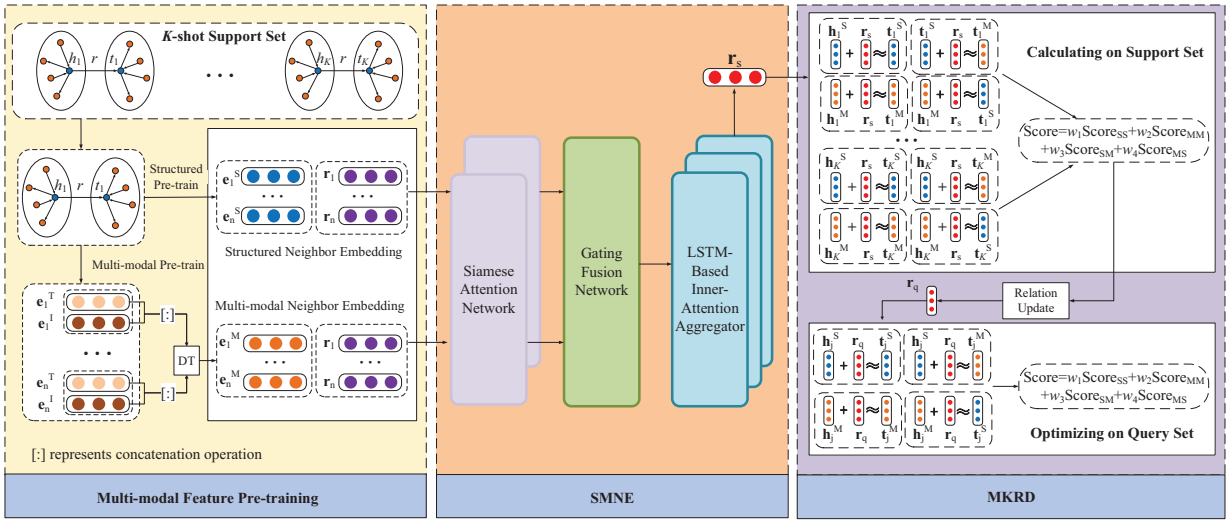


Fig. 2. Overview of MMSN.

and MKRD complementarily capture effective information from multi-modal data to complete triples with unseen relations.

B. Multi-modal Feature Pre-training

To exploit various modal features, these modalities have been pre-trained to be vectors by corresponding algorithms. Specifically, given a few-shot relation r, and its corresponding K-shot support set of triples $\mathcal{T} = \{(h_K, r, t_K)|h_K, t_K \in \mathcal{E}_{sup}\}$, where $\mathcal{E}_{sup}$ represents support entity set and each entity $e \in \mathcal{E}_{sup}$ has three features, including structure e^S, text e^T and image e^I. The structured features e^S and the relation r are initialized as d^S-dimensional vectors $\mathbf{e}^S$ and $\mathbf{r}$ by a popular pre-trained KGE model, such as TransE, which captures the global structure of MMBKG. The auxiliary features, i.e., text features e^I and image features e^T, are pre-trained as d^I-dimensional vector $\mathbf{e}^I$ and d^T-dimensional vector $\mathbf{e}^T$, respectively. Following [28], the image representation $\mathbf{e}^I$ is extracted from the last fully-connected layer before the softmax of VGG model [50], and $\mathbf{e}^T$ is initialized by the word2vec framework [51]. The final generated embeddings of the text and image are L_2-normalized.

C. Module SMNE

In the FKGC task, the existing models enhance the representations of support entity pairs and aggregate them to learn few-shot relational representations. These existing models focus on neighborhood structures, or exploit multi-modal features of the support entities. However, the multi-modal features in the support entity neighborhoods contain rich value information, which is ignored in enhancing the entity representations. In this module, a Siamese attention network, a popular technique to process multiple modalities [30] [52], is first designed to encode the structured and multi-modal features in the neighborhoods. Considering the varying contributions of neighbors

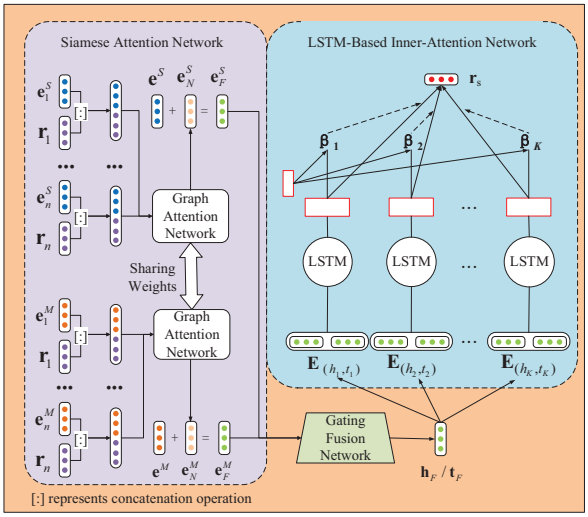


Fig. 3. The detailed schema of SMNE.

in enhancing the support entity representations, a graph attention network (GAT) [53] is employed to assign weights to neighbors, determining their importance. Both structured features and multi-modal features share the same attention weights of neighbors in the Siamese attention network, fully encoding neighborhood information. Then, a gating fusion network is utilized to fuse various modal neighborhood information, and improve the entity representations. Finally, an LSTM-based inner-attention aggregator is introduced to aggregate the enhanced entity pairs and learn representations of few-shot relations. The schematic diagram of the module SMNE is presented in Fig. 3.

Siamese attention network. To encode the neighborhood, the structured and multi-modal data are fed into the Siamese attention network in pairs. The structured and multi-modal data of a neighbor are assigned the same weight, avoiding

one of them being assigned an abnormal weight. Specifically, to obtain neighborhood information, the set of relational neighbors of the given head entity $h \in \mathcal{E}_{sup}$ is denoted as $\mathcal{N}_h = \{(r_i, t_i)|(h, r_i, t_i) \in \tilde{\mathcal{G}}\}$, and r_i and t_i represent the i-th neighboring relation and tail of h, respectively. Similarly, a given tail entity $t \in \mathcal{E}_{sup}$ also has a neighborhood $\mathcal{N}_t$. For clear description, the head and tail are generalized as entity $e \in \mathcal{E}_{sup}$, and its neighborhood is represented as $\mathcal{N}_e$. To enhance the representation of entity e, we design a Siamese attention network to separately encode the structured and multi-modal features of its neighborhood. Specifically, the neighborhood structured embeddings of e are first encoded by a GAT, which considers the different impacts of neighboring relations. The neighborhood structured embedding $\mathbf{e}_N^S$ is formulated as follows.

$$\begin{aligned} \mathbf{c}_i^S &= LeakyReLU(\mathbf{W}_1[\mathbf{r}_i : \mathbf{e}_i^S] + b_1) \\ \alpha_i^S &= softmax(\mathbf{U}_1^\mathsf{T}\mathbf{c}_i^S) \\ \mathbf{e}_N^S &= \sum_{(r_i,e_i)\in\mathcal{N}_e} \alpha_i^S \mathbf{c}_i^S \end{aligned} \tag{1}$$

where $\mathbf{r}_i$ and $\mathbf{e}_i^S$ are the structured embeddings of a neighboring relation r_i and its linked neighboring entity e_i, $[\mathbf{r}_i : \mathbf{e}_i^S]$ denotes a concatenation operation, $\mathbf{W}_1 \in \mathbb{R}^{2d^S \times d^S}$ represents a linear transformation matrix, $b_1 \in \mathbb{R}^{d^S}$, $LeakyReLU$ is an activation function, $\mathbf{U}_1 \in \mathbb{R}^{d^S \times 1}$ is a weight vector. To achieve the final enhanced structured embedding of e, the structured representation of e^S is combined with the neighborhood embedding as follows.

$$\mathbf{e}_F^S = \mathbf{e}_N^S + \mathbf{e}^S \tag{2}$$

In this work, the neighborhood multi-modal features and structured features of e are fed to the Simases network in pairs. Based on this network architecture, the neighborhood multi-modal features of e can be encoded by a GAT that shares weights with the GAT that encodes structured features.

Specifically, the text feature $\mathbf{e}_i^T$ and the image feature $\mathbf{e}_i^I$ of the neighbor entity $e_i \in \mathcal{N}_e$ are concatenated as $[\mathbf{e}_i^T : \mathbf{e}_i^I]$. This concatenation embedding is followed by a dimension transformation (DT) process to obtain multi-modal embedding $\mathbf{e}_i^M$ of neighbor e_i. The transferred representation $\mathbf{e}_i^M$ shares the same dimension as $\mathbf{e}_i^S$. To obtain multi-modal embedding $\mathbf{e}_N^M$ of the whole neighborhood, We concatenate relation representation $\mathbf{r}_i$ and the transferred representation $\mathbf{e}_i^M$ as $[\mathbf{r}_i : \mathbf{e}_i^M]$ and then encode it with the graph attention network as follows.

$$\begin{aligned} \mathbf{e}_i^M &= \mathbf{W}_{dt}[\mathbf{e}_i^T : \mathbf{e}_i^I] + b_{dt} \\ \mathbf{c}_i^M &= LeakyReLU(\mathbf{W}_1[\mathbf{r}_i : \mathbf{e}_i^M] + b_1) \\ \alpha_i^M &= softmax(\mathbf{U}_1^\mathsf{T}\mathbf{c}_i^M) \\ \mathbf{e}_N^M &= \sum_{(r_i,e_i)\in\mathcal{N}_e} \alpha_i^M \mathbf{c}_i^M \end{aligned} \tag{3}$$

where $\mathbf{W}_{dt} \in \mathbb{R}^{(d^I+d^T)\times d^S}$ is a linear transformation matrix and $b_{dt} \in \mathbb{R}^{d^S}$ is a bias to achieve dimension transformation. Besides, $\mathbf{W}_1$, b_1, and $\mathbf{U}_1$ are the same learnable parameters as those used in encoding neighborhood structures. With the multi-modal information encoded, the final enhanced multi-modal embeddings also equal neighborhood embedding plus its entity embedding as follows.

$$\mathbf{e}_F^M = \mathbf{e}_N^M + \mathbf{e}^M \tag{4}$$

Gating fusion network. After extracting structured and multi-modal information in the neighborhood, the final representation of entity e is obtained by a modality fusion operation. A simple concatenation or a plus process achieves a coarse fusion, but the multi-modal representations learned by a deep network contain a large amount of noise [28], which can degrade the fusion effectiveness. To capture the valuable information from different modalities and filter out noise, we employ a gating network to achieve multi-modal fusion $\mathbf{e}_F$ as follows.

$$\begin{aligned} g &= sigmoid(\mathbf{U}_g^\mathsf{T}\mathbf{e}_F^M + b_g) \\ \mathbf{e}_F &= (1-g)\mathbf{e}_F^S + g\mathbf{e}_F^M \end{aligned} \tag{5}$$

where $\mathbf{U}_g \in \mathbb{R}^{d^S \times 1}$ is represents a linear layer and $b_g \in \mathbb{R}^{d^S}$ is a scalar bias parameter. Based on the gating mechanism, the enhanced entity representations effectively capture valuable information from a variety of modalities.

LSTM-based inner-attention aggregator. By applying the Siamese network and gating fusion network to each head and tail of support triples, each entity pair $(h_k, t_k) \in \mathcal{S}_r$ are in the form $\mathbf{E}_{(h_k,t_k)}$, which represents the concatenation of $[\mathbf{h}_k^F : \mathbf{t}_k^F]$. To learn the few-shot relational representations, the entity pair $\mathbf{E}_{(h_k,t_k)}$ are fed to an LSTM-based inner-attention aggregator. The LSTM network [54] has been widely used for aggregating few-shot support entity pairs [45] [24], which can model the interactions and capture the similarities among few-shot support entity pairs. For a few-shot task, support entity pairs are connected by the same relation. Therefore, these support entity pairs share similarities, which are represented by a relation in a knowledge graph. To model similarities, we first employ an LSTM network to pass key information between support entity pairs, with adequate random sampling to ensure sufficient information passing among different entity pairs. Then, an attention network is utilized to learn the weights of different entity pairs and obtain the final relation representations. The LSTM-based inner-attention aggregator is formulated as follows.

$$\begin{aligned} \mathbf{m}_k &= \mathbf{LSTM}(\mathbf{E}_{(h_k,t_k)}, \mathbf{m}_{k-1}) \\ \mathbf{m}'_k &= \mathbf{W}_2\mathbf{m}_k \end{aligned} \tag{6}$$

where $\mathbf{m}_k$ and $\mathbf{m}_{k-1}$ denote the k-th and $(k-1)$-th hidden states of the LSTM, respectively. The final k-th hidden state $\mathbf{m}'_k$ is obtained by a linear transformation matrix $\mathbf{W}_2$.

Then, the weight of each final hidden state is calculated by the attention network, and the few-shot relational representations are obtained by accumulating these final states as follows.

$$\beta_k = softmax(\mathbf{U}_2^\mathsf{T}\mathbf{m}'_k) \tag{7}$$

where $\mathbf{U}_2$ is a weight vector, and β_k denotes the weight of k-th support entity pair embedding obtained from the

corresponding neighborhood. The final relation representation is learned by weights as follows.

$$\mathbf{r}_s = \mathbf{W}_s \sum_{K} \beta_k \mathbf{m}'_k \tag{8}$$

where $\mathbf{W}_s$ is a linear transformation, and $\mathbf{r}_s$ denotes the general embedding of the few-shot relation r, which integrates all the multi-modal neighborhood information from K-shot support triples.

Through the Siamese attention network, the modality fusion, and the aggregator, we can learn few-shot relational representations from a variety of modal features in the neighborhoods. This approach doesn't require prior knowledge of the number of relations to be completed, making it ideal for dealing with unseen relations.

D. Module MKRD

Considering the presence of multi-modal information in relational representations, popular uni-modal decoders, such as TransE and ComplEx, are not suitable for scoring triples in MFKGC scenarios. To conduct MFKGC, we propose the module MKRD, which introduces meta-learning into the multi-modal KGE model IKRL and models the inter-modal interaction between modalities. Specifically, given a few-shot relation r and its support set $\mathcal{S}_r$ and query set $\mathcal{Q}_r$, MKRD first calculates the loss on $\mathcal{S}_r$ with the learned relation representations $\mathbf{r}_s$, then obtains gradient meta to make a rapid relation update. MKRD exploits updated relation representations to calculate the loss on the query set and optimize the whole model. Inspired by the multi-modal KGE model IKRL that could model multi-modal features, the overall score function concerning a support triple $(h_i, r, t_i) \in \mathcal{S}_r$ in MKRD is designed as follows.

$$E_O(h_i, r, t_i) = w_1 E_{SS} + w_2 E_{SM} + w_3 E_{MS} + w_4 E_{MM} \tag{9}$$

where w_1, w_2, w_3, and w_4 are learnable hyperparameters for joint score function.

The overall score function is determined by modeling the inter-modal interaction between structured and multi-modal entity representations. Specifically, E_{SS} is the same score function as the uni-modal KGE model TransE, which only models structured representations. E_{MM} is a similar score function to TransE, but it only models multi-modal representations. Furthermore, E_{SM} and E_{MS} are the score functions to model inter-modal interactions between different modalities of the head and the tail. These score functions are given as follows.

$$\begin{aligned} E_{SS}(h_i, r, t_i) &= ||\mathbf{h}_i^S + \mathbf{r}_s - \mathbf{t}_i^S||_{L1/L2} \\ E_{SM}(h_i, r, t_i) &= ||\mathbf{h}_i^S + \mathbf{r}_s - \mathbf{t}_i^M||_{L1/L2} \\ E_{MS}(h_i, r, t_i) &= ||\mathbf{h}_i^M + \mathbf{r}_s - \mathbf{t}_i^S||_{L1/L2} \\ E_{MM}(h_i, r, t_i) &= ||\mathbf{h}_i^M + \mathbf{r}_s - \mathbf{t}_i^M||_{L1/L2} \end{aligned} \tag{10}$$

where $\mathbf{h}_i^S$ and $\mathbf{t}_i^S$ are pre-training structured embeddings of the head h_i and the tail t_i, $\mathbf{h}_i^M$ and $\mathbf{t}_i^M$ are multi-modal embeddings of the head h_i and the tail t_i learned from the pre-trained image and text representations, and $\mathbf{r}_s$ is the few-shot relational embedding derived from the module SMNE. In this paper, we utilize $L2$-norm to normalize all score functions.

Based on the overall score in Eq. (9), we employ a negative sampling loss to optimize the proposed model in MFKGC. Specifically, given a positive triple $(h_i, r, t_i) \in \mathcal{S}_r$, we corrupt this triple by replacing the tail t_i with a randomly sampling entity t'_i and ensure that the negative triple (h_i, r, t'_i) is incorrect. The negative sampling loss function is defined as follows.

$$\mathcal{L}(\mathcal{S}_r) = \sum_{(h_i, r, t_i) \in \mathcal{S}_r} [\gamma + E_O(h_i, r, t_i) - E_O(h_i, r, t'_i)]_+ \tag{11}$$

where $[x]_+ = max[0, x]$, i.e., it denotes the positive part of x. Here, $\gamma > 0$ is a fixed margin, and $(h_i, r, t'_i) \in \mathcal{S}'_r$ is a corresponding negative triple to the positive triple (h_i, r, t_i). To update the relation representations, we calculate the gradient of relation representation $\mathbf{r}_s$ as follows.

$$Grad(r_s) = \frac{dL(\mathcal{S}_r)}{d\mathbf{r}_s} \tag{12}$$

Then, the relation representation $\mathbf{r}_s$ can be updated for the query set following the stochastic gradient descent (SGD) as follows.

$$\mathbf{r}_q = \mathbf{r}_s - l_r Grad(r_s) \tag{13}$$

where l_r denotes the learning rate. With the updated relation representation $\mathbf{r}_q$, we transfer it to triples in the query set $\mathcal{Q}_r = \{(h_j, r, t_j)\}$ and calculate their score as follows.

$$\begin{aligned} E_O(h_j, r, t_j) &= w_1 E_{SS} + w_2 E_{SI} + w_3 E_{IS} + w_4 E_{II} \\ E_{SS}(h, r, t) &= ||\mathbf{h}_j^S + \mathbf{r}_q - \mathbf{t}_j^S||_{L1/L2} \\ E_{SM}(h, r, t) &= ||\mathbf{h}_j^S + \mathbf{r}_q - \mathbf{t}_j^M||_{L1/L2} \\ E_{MS}(h, r, t) &= ||\mathbf{h}_j^M + \mathbf{r}_q - \mathbf{t}_j^S||_{L1/L2} \\ E_{MM}(h, r, t) &= ||\mathbf{h}_j^M + \mathbf{r}_q - \mathbf{t}_j^M||_{L1/L2} \end{aligned} \tag{14}$$

Following the same way in the support set, the loss function on the query set is defined as:

$$\mathcal{L}(\mathcal{Q}_r) = \sum_{(h_j, r, t_j) \in \mathcal{Q}_r} [\gamma + E_O(h_j, r, t_j) - E_O(h_j, r, t'_j)]_+ \tag{15}$$

$\mathcal{L}(\mathcal{Q}_r)$ is the optimization objective for training the whole model.

In order to be applied to the score function, the multi-modal representation should be converted to the same dimensions as the structured representation. A simple linear transformation network loses detailed information. Inspired by [28], we employ an attention network to extract more information related to the knowledge graph structure from multi-modal features for the score function. The attention network considers the inter-modal interactions between multi-modal and structured data to perform dimension transformation. The formulation of the attention network is as follows.

$$\begin{aligned} \mathbf{e}_a^M &= \mathbf{W}_3^\top [\mathbf{e}^S : \mathbf{e}^M] + b_3 \\ \eta &= softmax(\mathbf{U}_3^\top \mathbf{e}_a^M) \\ \mathbf{e}_a^M &= \eta \mathbf{e}^M \end{aligned} \tag{16}$$

TABLE I
STATISTICS OF DATASETS.

Dataset	#Ents	#Rels	#Triples	#Train-Tasks	#Valid-Tasks	#Test-Tasks
FB-Img-Few	11728	959	327046	84	11	33
DB-Img-Few	10310	122	46811	61	4	10

where $[\mathbf{e}^S : \mathbf{e}^M]$ represent a concatenate operator of $\mathbf{e}^S$ and $\mathbf{e}^M$, which is learned by the weight $\mathbf{W}_{dt}$ and b_{dt} in Eq. (3), and $\mathbf{U}_3$ is a weight vector. Applying $\mathbf{e}_a^M$ in the score functions Eq. (9) and Eq. (14), h^M and t^M are learned as h_a^M and t_a^M.

V. EXPERIMENTS

In this section, we begin by introducing the datasets, baselines, and experimental settings, which include hyperparameter selections and evaluation metrics. Subsequently, extensive link prediction experiments are conducted to evaluate MFKGC performance of MMSN compared to state-of-the-art baselines. We further demonstrate the effectiveness of each primary module of our proposed model with ablation experiments. We also study the impact of each modality by ablation experiments. Moreover, the impact analysis of the few-shot size and maximum neighbor number are provided. The source code is available online [1].

A. Datasets

We evaluate our model and the baselines on two public multi-modal datasets which consist of structured features and two auxiliary modalities including text and image. The first dataset is extracted from FB-Img, named FB-Img-Few, and the second one is constructed based on DB-Img, named DB-Img-Few. Following the constructed rule for few-shot relational benchmarks [22] [25] [46], the relations associated with less than 500 but more than 50 are selected as few-shot relations. FB-Img-Few and DB-Img-Few include 128 and 75 task relations, respectively. The remaining relations and their triples are used to construct the background multi-modal knowledge graph. To adequately test the reliability of our proposed model, we carefully set an appropriate number of validation/testing tasks that align with the existing work, and categorize the remaining tasks as training tasks. Specifically, 11/33 and 4/10 few-shot relations for FB-Img-Few and DB-Img-Few are selected as validation/testing datasets, respectively, aligning with the quantity of uni-modal benchmarks like Wiki-One and NELL-One [22]. The remaining 84 and 61 tasks for FB-Img-Few and DB-Img-Few are designated as training tasks, respectively. In both datasets, each entity has one image pre-trained embedding and one text pre-trained embedding. The statistics of datasets are listed in detail in TABLE I. #Ents, #Rels, and #Triples represent the quantities of entities, relations, and triples in the background multi-modal knowledge graph, respectively.

In addition, the image embeddings of FB-Img-Few and DB-Img-Few are provided from [17] and [55], respectively. These embeddings have been pre-trained using the visual model VGG [50] and are widely used in KGC tasks. The image embeddings for both datasets are 4096-dimensional. For FB-Img-Few, the text embeddings can be found in [17]. These embeddings are created using a pre-trained model for Freebase entities provided by the Word2Vec framework [51], and finally generated by L_2 normalization. The text embeddings for FB-Img-Few are 1000-dimensional. As for DB-Img-Few, we set each entity's text as a collection of its inherent attributes, which contain the essential characteristics of the entity. To generate the text embeddings, we utilize the Doc2Vec model [56], an improved version of the Word2Vec framework, to pre-train the text data. The resulting embeddings are also subjected to L_2 normalization. The text embeddings for DB-Img-Few are also 1000-dimensional.

B. Baselines

We compare our model with four categories of baselines: (1) **Traditional uni-modal KGC models (TUKGC)**. This type of model completes triples for uni-modal knowledge graphs with seen entities and relations. We employ three widely used uni-modal models including TransE [31], DistMult [36], and ComplEx [37]. To implement this type of model on FKGC datasets, apart from all the triple of background KG and training set, the few-shot support triples of validate and test set are utilized to train the models.

(2) **Traditional multi-modal KGC models (TMKGC)**. This type of model achieves KGC with multi-modal features. We employ three widely used models, including IKRL [28], TransAE [29] and MKGFormer [41], as baselines. Similar to the above models, the training set of these two models consists of all the triple of background KG and training set, along with the few-shot support triples from validate and test sets. Each entity has structure, image, and text features.

(3) **Uni-modal FKGC models (UFKGC)**. This type of model aims to complete facts for unseen relations in few-shot scenarios. The existing FKGC models including GMatching [22], MetaR [23], FSRL [45], FAAN [46], GAAN [24], and HiRe [47] are set to be baselines. These models encode neighborhoods to enhance entity representations and learn embeddings of few-shot relations. Both GMatching and MetaR have two versions: one with random initialization and another with pre-trained embeddings. For a fair comparison, the results of GMatching and MetaR are achieved with pre-trained embedding instead of random initialization.

(4) **Multi-modal FKGC models (MFKGC)**. MULTIFORM [25] is the state-of-the-art MFKGC model, which focuses on pre-training a variety of modalities. MULTIFORM simply concatenates the pre-trained embeddings of different modalities. To make a fair comparison in evaluating FKGC and highlight the effective application of multiple modalities

[1] https://github.com/YuyangWei/MMSN

TABLE II
EVALUATION RESULTS OF MMSN ON FB-IMG-FEW AND DB-IMG-FEW.

FB-Img-Few		Hits@1		Hits@5		Hits@10		MRR	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
TUKGC	TransE	0.004	0.004	0.128	0.176	0.165	0.233	0.065	0.088
	DisMult	0.065	0.079	0.156	0.206	0.232	0.306	0.121	0.156
	ComplEx	0.025	0.176	0.079	0.355	0.112	0.420	0.055	0.261
TMKGC	MKGFormer	0.009	0.007	0.030	0.026	0.047	0.037	0.024	0.020
	IKRL	0.021	0.018	0.059	0.050	0.081	0.068	0.043	0.038
	TransAE	0.075	0.112	0.158	0.327	0.199	0.228	0.119	0.179
UFKGC	GMatching	0.186	0.228	0.386	0.433	0.471	0.554	0.287	0.329
	MetaR	0.174	0.244	0.293	0.382	0.362	0.464	0.239	0.316
	FSRL	0.174	0.246	0.444	0.517	0.587	0.649	0.304	0.377
	FAAN	0.172	0.279	0.362	0.554	0.475	0.641	0.278	0.403
	GANA	0.263	0.299	0.434	0.477	0.517	0.559	0.352	0.388
	HiRe	0.157	0.247	0.372	0.435	0.502	0.587	0.266	0.351
MFKGC	MULTIFORM	0.240	0.324	0.472	0.588	0.593	0.686	0.350	0.444
	MMSN	**0.677**	**0.729**	**0.692**	**0.817**	**0.697**	**0.846**	**0.684**	**0.769**

DB-Img-Few		Hits@1		Hits@5		Hits@10		MRR	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
TUKGC	TransE	0.103	0.121	0.264	0.299	0.290	0.341	0.177	0.206
	DisMult	0.091	0.073	0.183	0.174	0.213	0.208	0.134	0.123
	ComplEx	0.134	0.134	0.196	0.212	0.206	0.229	0.163	0.170
TMKGC	MKGFormer	0.028	0.028	0.064	0.077	0.086	0.102	0.049	0.054
	IKRL	0.180	0.199	0.260	0.327	0.288	0.372	0.218	0.259
	TransAE	0.119	0.154	0.247	0.296	0.286	0.332	0.177	0.221
UFKGC	GMatching	0.119	0.104	0.259	0.259	0.330	0.340	0.188	0.180
	MetaR	0.111	0.139	0.177	0.266	0.218	0.336	0.151	0.206
	FSRL	0.130	0.125	0.310	0.331	0.414	0.410	0.22	0.218
	FAAN	0.144	0.254	0.313	0.437	0.385	0.486	0.225	0.337
	GANA	0.063	0.154	0.215	0.245	0.274	0.295	0.130	0.207
	HiRe	0.062	0.079	0.164	0.200	0.260	0.272	0.123	0.145
MFKGC	MULTIFORM	0.148	0.194	0.315	0.351	0.358	0.428	0.224	0.273
	MMSN	**0.565**	**0.548**	**0.578**	**0.627**	**0.579**	**0.639**	**0.571**	**0.582**

in our proposed model, MULTIFORM utilizes the same pre-trained embeddings of each modality with our proposed model.

C. Experimental Setup

Hyperparameters. The structured information of entity and relation in the background KGs are pre-trained by TransE. The dimension of the pre-trained embeddings is set to 100 for both FB-Img-Few and DB-Img-Few. The image features of both datasets are pre-trained by VGG and their dimension is set to 4096. The text features are pre-trained by the Word2evc framework for the two datasets and the embedding dimension is set to 1000. The parameters of the proposed model are fine-tuned on the validation dataset. We utilize the SGD optimizer to update the model parameters. The maximum size of the neighborhood in the multi-modal neighbor encoder is set to 40 for both datasets. The learning rate l_r is set to 0.001 for FB-Img-Few and 0.01 for DB-Img-Few. The margin γ is set to 1.0. We employ mean reciprocal rank (MRR) on the validation dataset as the early stopping strategy and set it to 10 epochs.

Evaluation metrics. We employ link prediction to evaluate the performance of the proposed model and the baselines for the FKGC task. To be more specific, the objective is to rank the true tail entity higher than other candidate entities. We utilize Hits@n and MRR to evaluate the performance of all models. Hits@n represents the proportion of correct tail entities ranked within the top n positions. The value of n is set to 1, 5, and 10, in line with previous studies [22] [24] [47]. MRR stands for mean reciprocal rank and measures the average rank of the correct tail entities.

D. Main Results

To evaluate the performance of the proposed model, we conduct the link prediction in two FKGC scenarios including 1-shot and 5-shot. The overall results on FB-Img-Few and DB-Img-Few are presented in TABLE II, from which several observations can be obtained as follows.

1) MMSN consistently outperforms all baselines on multi-modal few-shot datasets. Compared to the best baseline MULTIFORM on FB-Img-Few, MMSN achieves 10.4% and 16% improvements in terms of Hits@10 on 1-shot and 5-shot, respectively. While compared to FAAN on DB-Img-Few, MMSN achieves 19.4% and 15.7% improvements in Hits@10 on 1-shot and 5-shot, respectively. MMSN also achieves significant improvements in the remaining metrics. This phenomenon benefits from the carefully designed neighbor encoder and knowledge graph representation decoder,

which complementarily capture valuable information from the multi-modal data.

2) The meta-learning FKGC models perform better than the traditional models on both datasets overall. This is due to the fact that traditional methods require a sufficient number of instances for training. Especially, compared to other traditional models, the Transformer-based multi-modal method MKGFormer, with more complex parameters, needs larger instances to train the model than other multi-modal methods. Therefore, the Transformer-based multi-modal method MKGFormer performs poorly. The meta-learning models excel at few-shot knowledge graph completion by learning a few support triples. Moreover, the meta-learning FKGC models exploit neighborhood information to enhance the representations of support entities, and further improve the ability to complete triples for unseen relations.

3) From the results, it can be observed that previous models exhibit a significant difference between Hits@1 and Hits@10. However, MMSN exhibits a smaller difference between these two metrics and achieves a high value in Hits@1, indicating its capability to make accurate link predictions in multi-modal few-shot scenarios. This is because MMSN extracts effective information from multi-modal data, resulting in entity representations with improved discrimination and enabling precise completion of missing elements.

4) Although MULTIFORM explores multi-modal data to enhance entity representations and achieves the best performance among the baselines on FB-Img-Few, its improvements are limited to a simple concatenation operation. As a result, the performance of MULTIFORM can be affected by noise in multi-modal features, and it heavily relies on modality-specific pre-training methods, resulting in poor generalization. As observed from the table, MMSN outperforms MULTIFORM on both datasets. This indicates that MMSN effectively filters out noise in multi-modal information through its gating fusion network and successfully extracts valuable information from various modalities to improve FKGC performance.

5) Since taking advantage of multi-modal information in neighborhoods to learn few-shot relational representations, MMSN gains substantial improvements on the datasets with rich neighborhood information, such as DB-Img-Few. On datasets such as FB-Img-Few, which have more complex relations in background KGs, MMSN also significantly outperforms all baseline models. This is because the multi-modal information in neighborhoods enhances the performance and robustness of our proposed model.

E. Effects of Multi-modal Features

To complete a multi-modal knowledge graph in a few-shot scenario, various modalities provide different aspects of information to enhance entity representations. To investigate the effects of each auxiliary modality, we compare MMSN with three versions where features of a type of modality are removed: (1) MMSN-S, a version where only structural features are considered; (2) MMSN-TS, a version where image features are not calculated by the neighbor encoder and KGR decoder; (3) MMSN-MS, a version where textual features are not input into the neighbor encoder and KGR decoder. For our proposed model and all variants, we conduct 5-shot FKGC on both datasets FB-Img-Few and DB-Img-Few.

Observed from TABLE III, MMSN, MMSN-TS, and MMSN-MS significantly outperform MMSN-S, confirming the importance of auxiliary multi-modal information for FKGC. From the results of inputting two modalities, the performance of MMSN-TS is better than that of MMSN-MS. This is because text information is better able to provide the nature of entities and has less noise compared to image information. Furthermore, MMSN outperforms the other three variants, validating the stable ability of the proposed model to extract valuable information from multiple modalities.

F. Ablation Studies

MMSN is a few-shot multi-modal knowledge graph completion model, which includes the following carefully designed components, i.e., Siamese attention network, gating attention fusion, and meta-learning multi-modal knowledge representation decoder. To investigate the contributions of different components, we compare MMSN with the following three types of variants. (1) MMSN-TAN is a version that replaces the Siamese attention network with two attention networks without sharing weights. (2) The gating fusion network is replaced with concatenation operations to achieve neighborhood multi-modal fusion, represented as MMSN-Cat. (3) MMSN-U is constructed by replacing the multi-modal knowledge graph representation decoder with a uni-modal model TransE. For our proposed model and all variants, we conduct 5-shot FKGC on both datasets FB-Img-Few and DB-Img-Few. The ablation results are presented in TABLE IV. Several observations are made from the results.

(1) We first analyze the impact of the Siamese attention network. According to the results in TABLE IV, the performance of MMSN is significantly better than that of MMSN-TAN. The reason for this improvement is that both the structured features and multi-modal features of key neighbors play crucial roles in enhancing entity representations. Therefore, it is reasonable to jointly extract structured features and multi-modal features by sharing the same weights. When utilizing different attention networks, MMSN-TAN may assign lower weights to certain modal information of key neighbors, diminishing the quality of encoding.

(2) In terms of multi-modal fusion, MMSN significantly outperforms MMSN-Cat. This is because MMSN exploits a gating mechanism to filter out the noise contained in multi-modal representations which are learned by multi-modal neighbor encoders. By fusing these filtered features with structured features, MMSN effectively enhances entity representations.

(3) A meta-learning multi-modal decoder is introduced in MMSN to model the multi-modal information in few-shot relation representations and the inter-modal interaction between different modalities of query entities. Compared to MMSN, MMSN-U only focuses on modeling the structured

TABLE III
EFFECTS OF FEATURES ON FB-IMG-FEW AND DB-IMG-FEW.

	FB-Img-Few				DB-Img-Few			
	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
MMSN-S	0.306	0.485	0.571	0.395	0.057	0.180	0.224	0.133
MMSN-TS	0.697	0.742	0.764	0.719	0.470	0.577	0.579	0.517
MMSN-MS	0.620	0.696	0.728	0.657	0.395	0.443	0.467	0.415
MMSN	**0.729**	**0.817**	**0.846**	**0.769**	**0.548**	**0.627**	**0.639**	**0.582**

TABLE IV
ABLATION STUDY RESULTS ON FB-IMG-FEW AND DB-IMG-FEW.

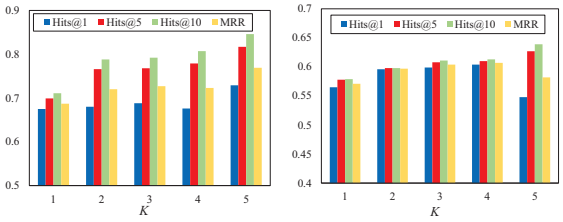
	FB-Img-Few				DB-Img-Few			
	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR
MMSN-TAN	0.577	0.732	0.776	0.651	0.423	0.458	0.472	0.441
MMSN-Cat	0.646	0.776	0.804	0.708	0.466	0.479	0.479	0.473
MMSN-U	0.307	0.490	0.570	0.397	0.083	0.185	0.24	0.147
MMSN	**0.729**	**0.817**	**0.846**	**0.769**	**0.548**	**0.627**	**0.639**	**0.582**

features in the decoder. As observed from TABLE IV, MMSN significantly outperforms MMSN-U. This is because multi-modal features provide crucial information that enhances the performance of FKGC. Additionally, the few-shot relation representations in MMSN are learned from SMNE, which considers both multi-modal and structured features. MMSN-U utilizes only structured features to decode the triples, but its performance is influenced by the absence of multi-modal information in few-shot relation representations.

G. Parameters Interpretability

To assess the sensitivity of the proposed model, we investigate the influence of several important parameters, including few-shot size and maximum number of neighbors.

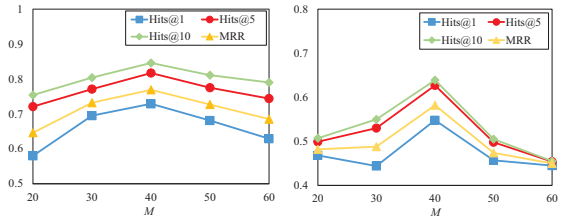
Impacts of few-shot size K. The few-shot size K controls the number of triples that the model can access in meta-learning training settings. To explore the impacts of K, we conducted experiments varying it from 1 to 5. The experimental results on both multi-modal datasets are presented in Fig. 4, which show that the proposed model consistently improves overall as the few-shot K increases. This is because, with the growth of few-shot instances, the model can access more training instances, allowing for the absorption of additional structured and multi-modal information from more neighbors.



(a) FB-Img-Few (b) DB-Img-Few

Fig. 4. Effects of few-shot size K.

Impacts of maximum number of neighbors M. This work designs a novel multi-modal neighbor encoder to enhance entity representations, capturing valuable multi-modal information from neighborhoods. M controls the amount of neighborhood information for the encoder. To detect the impact of the maximum number of neighbors, we conduct 5-shot FKGC experiments Varying M from 20 to 60. The experimental results are presented in Fig. 5, from which we can find that with the increase of the maximum number of neighbors, the model performance first increases and then decreases. This phenomenon occurs because, with a relatively small M, increasing its value allows the model to access more neighborhood information, leading to performance improvement. However, as M continues to increase, excessive noise from the neighborhood is introduced into the neighbor encoder. At a certain point, the negative impact caused by the noise outweighs the positive impact of the neighborhood information. As observed from Fig. 5, the model achieves the best performance on both datasets when M is set to 40.



(a) FB-Img-Few (b) DB-Img-Few

Fig. 5. Effects of the maximum number of neighbors M.

H. Impacts of Different Multi-modal Initialization Methods

To investigate the impact of multi-modal initialization methods, we propose two variants including MMSN-Bert&ViT and MMSN-Bert&ResNet. Specifically, MMSN-Bert&ViT initializes text features by Bert [57] and image features by ViT [58], respectively. MMSN-Bert&ResNet employs Bert to initialize text features and ResNet [59] to initialize image features. Notably, the image initialization methods are provided by [18].

TABLE V
RESULTS OF MULTI-MODAL INITIALIZATION MODELS.

	FB-Img-Few			
	Hits@1	Hits@5	Hits@10	MRR
MMSN-Bert&ViT	0.676	0.729	0.744	0.702
MMSN-Bert&ResNet	0.556	0.593	0.606	0.574
MMSN	**0.729**	**0.817**	**0.846**	**0.769**

	DB-Img-Few			
	Hits@1	Hits@5	Hits@10	MRR
MMSN-Bert&ViT	0.470	0.525	0.546	0.498
MMSN-Bert&ResNet	**0.615**	0.621	0.622	**0.618**
MMSN	**0.548**	**0.627**	**0.639**	**0.582**

In both variants, the dimensions of image embedding and text embedding are 1000 and 768, respectively.

As presented in TABLE V, MMSN consistently performs better than MMSN-Bert&ViT. This may be because the architectures of Bert and ViT are more complex compared to VGG and Word2Vec, implying that MMSN may require more training data to transfer knowledge from the support set to the query set with Bert and ViT. Additionally, the pre-trained weights of Bert and ViT are less relevant to FKGC tasks, while VGG, with its simple convolutional networks, exhibits better generalization. Therefore, MMSN outperforms MMSN-Bert&ViT in few-shot experimental settings. MMSN-Bert&ResNet performs better than MMSN in terms of Hits@1 and MRR on dataset DB-Img-Few. This may be because ResNet and Bert capture significant information from multi-modal data of DB-Img-Few to distinguish entities and achieve accurate link prediction. However, MMSN with the embeddings pre-trained by VGG and Word2Vec gains higher performance in most cases due to their superior generalization. Therefore, we utilize VGG and Word2Vec to initialize multi-modal features in this paper.

I. Performance on Different Relations

When testing different models, we observed significant variations in the results across different relations. Due to this observation, we conducted 5-shot FKGC experiments on the FB-Img-Few dataset test data to further investigate the model performance for each relation. In TABLE VI, we present the results of the proposed model and the best metric model MULTIFORM on the FB-Img-Few dataset. The superior result for each relation is highlighted in bold.

From TABLE VI, we can observe that the results of the baseline on different few-shot relations exhibit high variance. The reason may be that varying degrees of candidate entities associated with the few-shot relations have access to different levels of accessible neighborhood information. However, MMSN is able to achieve better and more stable results compared to the baseline in most cases. This is because MMSN captures valuable information from neighborhood multi-modal features and models inter-modal interactions between modalities to enhance the representation of entities, enabling the model to accurately predict the correct entities from candidates.

TABLE VI
RESULTS OF EACH RELATION (RID) IN FB-IMG-FEW TEST DATA.

Model	RId	Hits@10	RId	Hits@10	RId	Hits@10
MULTIFORM	1	1.000	12	0.460	23	0.066
MMSN		**1.000**		**0.860**		**0.829**
MULTIFORM	2	0.081	13	0.392	24	**0.978**
MMSN		**0.964**		**0.985**		0.225
MULTIFORM	3	0.461	14	0.233	25	**0.843**
MMSN		**0.712**		**0.439**		0.569
MULTIFORM	4	**1.000**	15	0.897	26	**0.510**
MMSN		0.786		**0.912**		0.117
MULTIFORM	5	0.921	16	0.472	27	0.470
MMSN		**0.979**		**0.980**		**0.946**
MULTIFORM	6	0.421	17	0.500	28	1.000
MMSN		**0.931**		**0.572**		**1.000**
MULTIFORM	7	0.850	18	**0.904**	29	0.965
MMSN		**1.000**		0.837		**1.000**
MULTIFORM	8	1.000	19	0.476	30	**0.986**
MMSN		**1.000**		**1.000**		0.984
MULTIFORM	9	**0.983**	20	0.225	31	0.873
MMSN		0.974		**0.902**		**1.000**
MULTIFORM	10	0.233	21	1.000	32	**1.000**
MMSN		**0.800**		**1.000**		0.971
MULTIFORM	11	0.819	22	0.673	33	0.883
MMSN		**0.958**		**0.832**		**0.958**

VI. CONCLUSION AND FUTURE WORK

In this paper, we explore the utilization of multi-modal neighborhood information and model inter-modal interaction between multiple modalities to improve MFKGC performance. To address this problem, a novel relational learning model MMSN is designed to accomplish the MFKGC task. In MMSN, the module SMNE first encodes neighborhood structured and multi-modal features with a Siamese attention network. Next, a gating fusion network is utilized to effectively fuse the encoded modalities, enhancing entity representations. Then, an LSTM-based inner-attention aggregator is employed to learn the few-shot relational representations. In addition, the module MKRD is designed to model inter-modal interactions between various modalities with four score functions and train the model in a meta-learning setting. Extensive experiments demonstrate that MMSN significantly outperforms the state-of-the-art baselines, including uni-modal and multi-modal models. In future work, we will explore how to model complex relations, such as 1-N, N-1, and N-N, in multi-modal few-shot knowledge graph completion scenarios.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China No. 62272332, the Major Program of the Natural Science Foundation of Jiangsu Higher Education Institutions of China No. 22KJA520006 and No. 22KJA520008.

REFERENCES

[1] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *SIGMOD*, 2008, pp. 1247–1250.

[2] D. Vrandecic and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[3] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW*, 2007, pp. 697–706.

[4] Z. Li, Y. Gu, Y. Shen, W. Hu, and G. Cheng, "TRAVERS: A diversity-based dynamic approach to iterative relevance search over knowledge graphs," in *WWW*, 2023, pp. 2560–2571.

[5] R. Zhu, Y. Zhao, W. Qu, Z. Liu, and C. Li, "Cross-domain product search with knowledge graph," in *CIKM*, 2022, pp. 3746–3755.

[6] Z. Li, X. Jin, S. Guan, W. Li, J. Guo, Y. Wang, and X. Cheng, "Search from history and reason for future: Two-stage reasoning on temporal knowledge graphs," in *ACL-IJCNLP*, 2021, pp. 4732–4743.

[7] J. Jiang, K. Zhou, X. Zhao, and J. Wen, "Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph," in *ICLR*, 2023.

[8] J. Dong, Q. Zhang, X. Huang, K. Duan, Q. Tan, and Z. Jiang, "Hierarchy-aware multi-hop question answering over knowledge graphs," in *WWW*, 2023, pp. 2519–2527.

[9] H. V. Nguyen, F. Gelli, and S. Poria, "DOZEN: cross-domain zero shot named entity recognition with knowledge graph," in *SIGIR*, 2021, pp. 1642–1646.

[10] Q. He, L. Wu, Y. Yin, and H. Cai, "Knowledge-graph augmented word representations for named entity recognition," in *AAAI*, 2020, pp. 7919–7926.

[11] A. Pavlovic and E. Sallinger, "Expressive: A spatio-functional embedding for knowledge graph completion," in *ICLR*, 2023.

[12] Z. Sun, Z. Deng, J. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *ICLR*, 2019, pp. 1–18.

[13] H. Chang, J. Cai, and J. Li, "Knowledge graph completion with counterfactual augmentation," in *WWW*, 2023, pp. 2611–2620.

[14] H. Ren, H. Dai, B. Dai, X. Chen, D. Zhou, J. Leskovec, and D. Schuurmans, "SMORE: knowledge graph completion and multi-hop reasoning in massive knowledge graphs," in *KDD*, 2022, pp. 1472–1482.

[15] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, and N. J. Yuan, "Multi-modal knowledge graph construction and application: A survey," *IEEE Trans. Knowl. Data Eng.*, pp. 1–20, 2022.

[16] Y. Ma, Z. Wang, M. Li, Y. Cao, M. Chen, X. Li, W. Sun, K. Deng, K. Wang, A. Sun, and J. Shao, "MMEKG: multi-modal event knowledge graph towards universal representation across modalities," in *ACL*, 2022, pp. 231–239.

[17] H. M. Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in **SEM@NAACL-HLT*, 2018, pp. 225–234.

[18] M. Wang, S. Wang, H. Yang, Z. Zhang, X. Chen, and G. Qi, "Is visual context really helpful for knowledge graph? A representation learning perspective," in *MM*, 2021, pp. 2735–2743.

[19] S. Ferrada, B. Bustos, and A. Hogan, "Imgpedia: A linked dataset with content-based analysis of wikimedia images," in *ISWC*, vol. 10588, 2017, pp. 84–93.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9726–9735.

[21] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," in *AAAI*, 2016, pp. 2659–2665.

[22] W. Xiong, M. Yu, S. Chang, X. Guo, and W. Y. Wang, "One-shot relational learning for knowledge graphs," in *EMNLP*, 2018, pp. 1980–1990.

[23] M. Chen, W. Zhang, W. Zhang, Q. Chen, and H. Chen, "Meta relational learning for few-shot link prediction in knowledge graphs," in *EMNLP-IJCNLP*, 2019, pp. 4216–4225.

[24] G. Niu, Y. Li, C. Tang, R. Geng, J. Dai, Q. Liu, H. Wang, J. Sun, F. Huang, and L. Si, "Relational learning with gated and attentive neighbor aggregator for few-shot knowledge graph completion," in *SIGIR*, 2021, pp. 213–222.

[25] X. Zhang, X. Liang, X. Zheng, B. Wu, and Y. Guo, "MULTIFORM: few-shot knowledge graph completion via multi-modal contexts," in *ECML-PKDD*, vol. 13714, 2022, pp. 172–187.

[26] Z. Lin, Z. Zhang, M. Wang, Y. Shi, X. Wu, and Y. Zheng, "Multi-modal contrastive representation learning for entity alignment," in *COLING*, 2022, pp. 2572–2584.

[27] Q. Fang, X. Zhang, J. Hu, X. Wu, and C. Xu, "Contrastive multi-modal knowledge graph representation learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8983–8996, 2023.

[28] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," in *IJCAI*, 2017, pp. 3140–3146.

[29] Z. Wang, L. Li, Q. Li, and D. Zeng, "Multimodal data enhanced representation learning for knowledge graphs," in *IJCNN*, 2019, pp. 1–8.

[30] L. Chen, Z. Li, T. Xu, H. Wu, Z. Wang, N. J. Yuan, and E. Chen, "Multi-modal siamese network for entity alignment," in *KDD*, 2022, pp. 118–126.

[31] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *NeurIPS*, 2013, pp. 2787–2795.

[32] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *AAAI*, 2014, pp. 1112–1119.

[33] X. You, B. Sheng, D. Ding, M. Zhang, X. Pan, M. Yang, and F. Feng, "Mass: Model-agnostic, semantic and stealthy data poisoning attack on knowledge graph embedding," in *WWW*, 2023, pp. 2000–2010.

[34] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *AAAI*, 2015, pp. 2181–2187.

[35] M. Nickel, V. Tresp, and H. Kriegel, "A three-way model for collective learning on multi-relational data," in *ICML*, 2011, pp. 809–816.

[36] B. Yang, W. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *ICLR*, 2015.

[37] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *ICML*, vol. 48, 2016, pp. 2071–2080.

[38] M. S. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *ESWC*, vol. 10843, 2018, pp. 593–607.

[39] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2d knowledge graph embeddings," in *AAAI*, 2018, pp. 1811–1818.

[40] R. Wang, B. Li, S. Hu, W. Du, and M. Zhang, "Knowledge graph embedding via graph attenuated attention networks," *IEEE Access*, vol. 8, pp. 5212–5224, 2020.

[41] X. Chen, N. Zhang, L. Li, S. Deng, C. Tan, C. Xu, F. Huang, L. Si, and H. Chen, "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion," in *SIGIR*, 2022, pp. 904–915.

[42] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017, pp. 1–11.

[43] T. Munkhdalai and H. Yu, "Meta networks," in *ICML*, 2017, pp. 2554–2563.

[44] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, vol. 70, 2017, pp. 1126–1135.

[45] C. Zhang, H. Yao, C. Huang, M. Jiang, Z. Li, and N. V. Chawla, "Few-shot knowledge graph completion," in *AAAI*, 2020, pp. 3041–3048.

[46] J. Sheng, S. Guo, Z. Chen, J. Yue, L. Wang, T. Liu, and H. Xu, "Adaptive attentional network for few-shot knowledge graph completion," in *EMNLP*, 2020, pp. 1681–1691.

[47] H. Wu, J. Yin, B. Rajaratnam, and J. Guo, "Hierarchical relational learning for few-shot knowledge graph completion," in *ICLR*, 2023, pp. 1–15.

[48] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon, "Representing text for joint embedding of text and knowledge bases," in *EMNLP*, 2015, pp. 1499–1509.

[49] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *EMNLP*, 2017, pp. 670–680.

[50] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[51] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *NeurIPS*, 2013, pp. 2121–2129.

[52] A. V. Chithra and D. Mishra, "Multi-modality fusion for siamese network based RGB-T tracking (mfsiamtrack)," in *CVIP*, vol. 1776, 2022, pp. 406–420.

[53] D. Nathani, J. Chauhan, C. Sharma, and M. Kaul, "Learning attention-based embeddings for relation prediction in knowledge graphs," in *ACL*, 2019, pp. 4710–4723.
[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[55] Y. Liu, H. Li, A. García-Durán, M. Niepert, D. Oñoro-Rubio, and D. S. Rosenblum, "MMKG: multi-modal knowledge graphs," in *ESWC*, vol. 11503, 2019, pp. 459–474.
[56] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, vol. 32, 2014, pp. 1188–1196.
[57] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.
[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021, pp. 1–21.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.