ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240612

LoginMEA: Local-to-Global Interaction Network for Multi-Modal Entity Alignment

Taoyu Su[a,b], Xinghua Zhang[a,b], Jiawei Sheng[a,*], Zhenyu Zhang[c] and Tingwen Liu[a,b]

[a]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[b]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[c]Baidu Inc.
{sutaoyu, zhangxinghua, shengjiawei, liutingwen}@iie.ac.cn, zhangzhenyu07@baidu.com

Abstract. Multi-modal entity alignment (MMEA) aims to identify equivalent entities between two multi-modal knowledge graphs (MMKGs), whose entities can be associated with relational triples and related images. Most previous studies treat the graph structure as a special modality, and fuse different modality information with separate uni-modal encoders, neglecting valuable relational associations in modalities. Other studies refine each uni-modal information with graph structures, but may introduce unnecessary relations in specific modalities. To this end, we propose a novel local-to-global interaction network for MMEA, termed as LoginMEA. Particularly, we first fuse local multi-modal interactions to generate holistic entity semantics and then refine them with global relational interactions of entity neighbors. In this design, the uni-modal information is fused adaptively, and can be refined with relations accordingly. To enrich local interactions of multi-modal entity information, we devise modality weights and low-rank interactive fusion, allowing diverse impacts and element-level feature interactions among modalities. To capture global interactions of graph structures, we adopt relation reflection graph attention networks, which fully capture relational associations between entities. Extensive experiments demonstrate superior results of our method over 5 cross-KG or bilingual benchmark datasets, indicating the effectiveness of capturing local and global interactions.

1 Introduction

Knowledge graphs (KGs) have emerged as a prominent data structure for representing factual knowledge in the form of triples, i.e., <entity, relation, entity>, where two entities are connected through relations. *Multi-modal knowledge graphs* (MMKGs) further extend traditional KGs by introducing representative multi-modal information, such as visual images, attributes, and relational data [21]. As a pivotal task for MMKG integration, *multi-modal entity alignment* (MMEA) aims to identify equivalent entities between two MMKGs. As shown in Figure 1(a), the model requires to identify entity `Oriental_Pearl_Tower` in *MMKG-1* is equivalent to `Oriental_Pearl` in *MMKG-2*, utilizing their multi-modal information. Such a task can benefit many downstream applications [17], such as recommendation systems [30] and question answering [44].

For MMEA, a crucial problem is how to exploit the consistency of equivalent entities between different MMKGs with their multi-modal information in relational graph structures [21]. To this end, most methods [4, 10, 8, 41] firstly encode uni-modal features, and then fuse them to obtain joint entity embedding for alignment. They can be roughly manifested into two groups (shown in Figure 2):

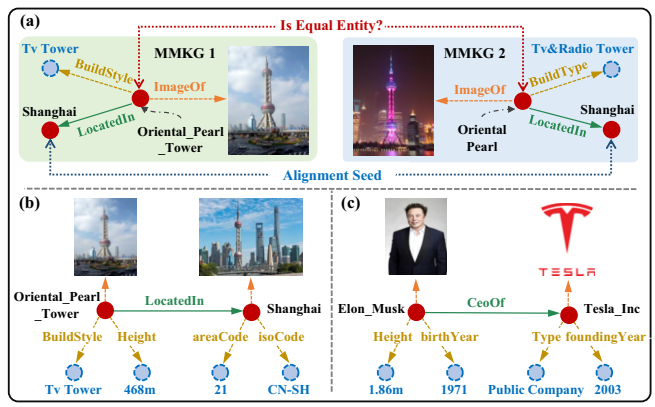

Figure 1. An example of the MMEA task between two MMKGs, where (b) and (c) showcase the relational associations of entity images in an MMKG.

- The first group of methods [20, 5, 19, 8], namely *graph-as-modality*, treat the graph structure as a special modality of entities, and separately encode entity uni-modal information (i.e., structures, images, attributes, relations). However, these methods cannot capture valuable relational associations between entity images without graph structures. As shown in Figure 1(b), in the MMKG, entity `Oriental_Pearl_Tower` is located in `Shanghai`, and their images of entities reflect this `LocatedIn` relation. Separately encoding the uni-modal features without graph structures would neglect these beneficial relational associations, which generates suboptimal uni-modal embeddings, and impedes the final joint multi-modal entity embeddings.
- Another group of methods [5, 41], namely *graph-upon-modality*, firstly encode uni-modal information (including images, attributes and relations), and then refine these uni-modal information with graph structures. However, there also exist relations that cannot be reflected in the entity images. As shown in Figure 1(c), the images of `Elon_Musk` and `Tesla_Inc` cannot reflect the `CeoOf` relation. Directly building the relations between images would introduce unnecessary relational inductive bias [1], leading to unstable performance in MMEA.

* Corresponding Author.

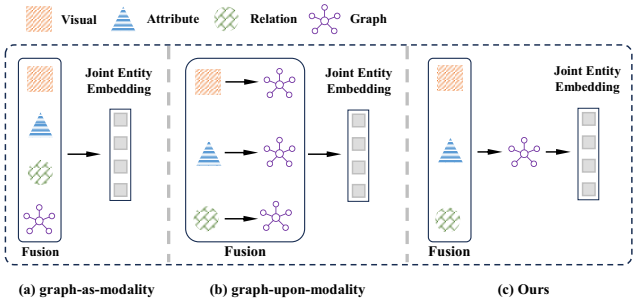

Figure 2. The schematic diagram of different modeling paradigms.

Therefore, a natural idea is to fuse multi-modal information firstly to obtain holistic entity semantics, and then refine them with relational graph structures (as Figure 2(c)). In this sense, the uni-modal information (e.g., images) is adaptively fused and accordingly refined by relations. However, there are still two vital challenges requiring further designs: (1) *How to fuse local different modality information of an entity considering its multi-modal interactions?* Existing multi-modal fusion mostly employs vector concatenation [21, 10, 4] or weighted attention mechanism [20, 19, 9, 26, 38]. Nevertheless, these methods only considers the importance of modality features, lacking multifarious feature interactions in entity multi-modal information. For instance, `Oriental_Pearl_Tower` has interactions between visual architectural cues and attribute details (e.g., height, function), ensuring its identification as a television tower. (2) *How to build global relational interactions between entities, while enhancing relational consistency between MMKGs?* Conventional MMEA studies [39, 8] mostly employ vanilla graph neural networks (GNNs) [13, 37], which can hardly capture relations between entity embeddings. Relational GNNs [27, 18] can be promising but may struggle in building relational consistency between different graph structures in the embedding space [25]. Since structural information is reported [4, 8] as pivotal information for MMEA results, it requires a further design for learning relational graph structures.

To this end, we propose a novel Local-to-global interaction network for Multi-modal Entity Alignment, termed as LoginMEA[1]. Particularly, it first fuses local multi-modal interactions to generate holistic entity semantics, then refines them with global relational interactions of entity neighbors. To fully fuse different modality information, we propose a *local multi-modal interactive fusion* module, which designs entity-specific adaptive weights and low-rank interactive fusion. Compared to existing methods [20, 19, 38, 8], this module discerns diverse weight impacts of different entity modality information and captures multifarious element-level modality feature interactions, allowing capability for building relational associations of uni-modal information between entities. Besides, we propose a *global multi-modal interactive aggregation* module, which adopts relational reflection graph attention networks to refine entity embeddings with entity neighbors. Compared to vanilla GNN-based methods [39, 8], this module fully utilizes relational interactions between entities in the MMKG, and can retain relational consistency between different MMKGs. Finally, we adopt a *contrastive alignment loss* to train the overall model, which ensures the consistency of equivalent entities from different MMKGs to achieve the MMEA task. The contributions of this paper are summarized as follows:

- We investigate the relational associations between entities in their multi-modal information, and we propose a novel MMEA framework, LoginMEA. To our knowledge, we are the first to build relational graph structures upon holistic entities to leverage relational associations of multi-modal information in MMEA studies.
- We develop the LoginMEA framework with local-to-global interaction networks, which builds multi-modal interactions of entity information with local multi-modal interactive fusion, and builds global relational interactions between joint multi-modal entity embeddings with global multi-modal interactive aggregation.
- Experimental results and extensive analyses confirm our significant improvements in comparison with previous state-of-the-art methods on 5 benchmark datasets.

2 Related work

Entity Alignment (EA). Existing EA methods aim to embed entities from different KGs into a unified vector space, and identify equivalent entities by measuring the distance between their embeddings. Early methods [6, 33, 45, 32] employ TransE [2] or its variants to learn entity embeddings and relations. Recognizing that entities with similar neighborhood structures are likely to be aligned, recent approaches [39, 14, 40, 3, 34] leverage graph neural networks (GNNs) [13, 37] to capture entity structure information to enhance entity embeddings. Besides, to compensate for limited graph structure signals in alignment learning, another line of recent studies [31, 35, 43, 7, 23] retrieve auxiliary supervision from side information such as attribute information and entity descriptions. Although the aforementioned methods attempt to improve entity representation by utilizing structural and side information, they can hardly directly utilize visual images of entities to enhance EA in MMKGs.

Multi-Modal Entity Alignment. Current MMEA methods can be roughly classified into two groups according to the utilization of graph structures: (1) In *graph-as-modality* approaches, the graph structure is considered as a distinct modality. Early researches [21, 10, 4] learn modality-specific embeddings by using separate encoders, then adopt direct or fixed-weight operations to combine multi-modal information. However, these approaches lack adaptability in learning the relative importance of different modalities. To address this, EVA [20] combines multi-modal information for MMEA with learnable weighted attention to model the importance of each modality. Building upon EVA's foundation, some subsequent works enhance entity alignment based on contrastive learning [19, 9, 15] or generated pseudo labels [26, 38]. Yet, these approaches fail to compute modality weights at the entity level. To address this limitation, recent works [8, 16] leverage transformer-based approaches for multi-modal fusion. However, the *graph-as-modality* methods face limitations by treating the graph structure as a single modality, hindering the capture of relational information between entity images. (2) In terms of *graph-upon-modality* methods, they firstly encode uni-modal information (including images, attributes, and relations), and then refine this uni-modal information with graph structures. MSNEA [5] utilizes TransE [2] for uni-modal information to guide relational feature learning. XGEA [41] employs the message-passing mechanism of GNNs to aggregate uni-modal information, thereby guiding structural embedding learning. However, learning relations directly from unimodal information that cannot reflect relations introduces unnecessary relational inductive bias. In this work, we introduce a local-to-global interaction network, fusing local multi-modal interactions to generate holistic entity semantics and then refine them with global relational interactions of entity neighbors.

[1] Our code is available at https://github.com/sutaoyu/LoginMEA, and the **supplementary material** are also available there.

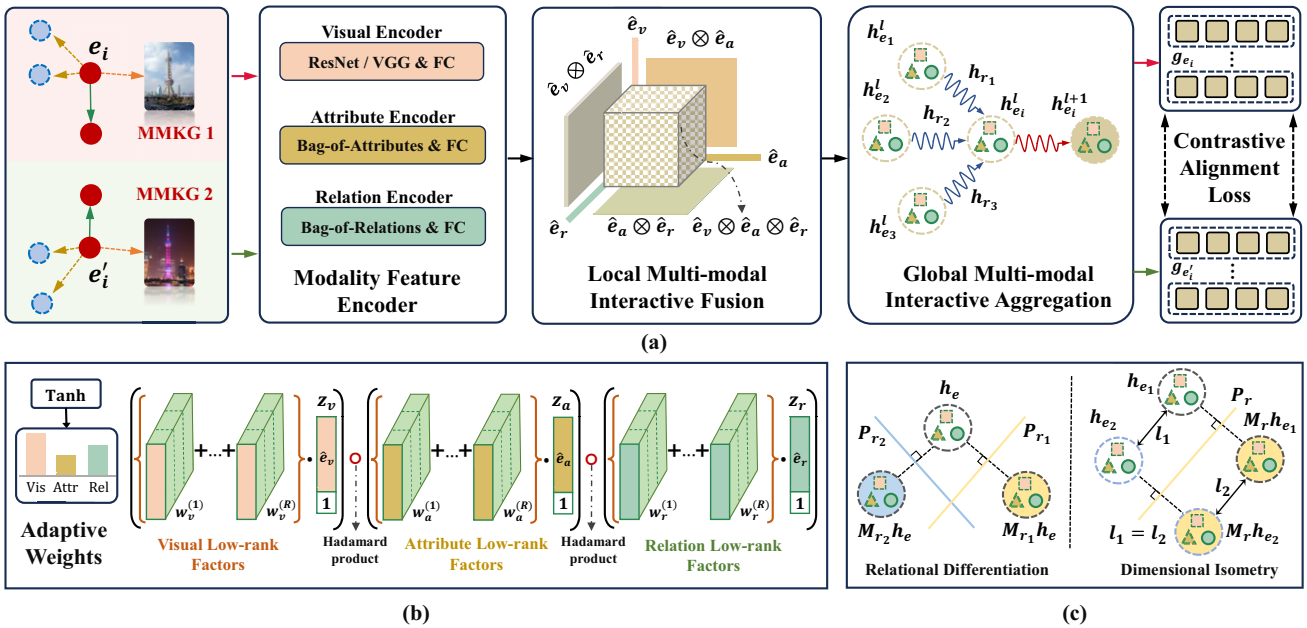

Figure 3. The LoginMEA framework for the multi-modal entity alignment, where (a) denotes the overall backbone, (b) the principle of the local multi-modal interactive fusion, (c) the principle of the global multi-modal interactive aggregation.

3 Problem Formulation

In general, a ***multi-modal knowledge graph (MMKG)*** is composed of relational triples with entities, attributes and images, which can be defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V})$, where $\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}$ are the sets of entities, relations, attributes and visual images, respectively. Therefore, the triples are defined as $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where each entity $e \in \mathcal{E}$ can have attributes and images. Following previous works [20, 19, 8], we focus on four kinds of entity information, including graph structure g, visual image v, neighboring relation r, and attribute a.

Based upon, ***multi-modal entity alignment (MMEA)*** [21, 4] aims to identify equivalent entities from two different MMKGs. Formally, given two MMKGs $\mathcal{G}_1$ and $\mathcal{G}_2$ with their relational triples and multi-modal attributes, the goal of MMEA task is to identify equivalent entity pairs $\{\langle e_1, e_2\rangle | e_1 \in \mathcal{G}_1, e_2 \in \mathcal{G}_2, e_1 \equiv e_2\}$. For model training, a set of pre-aligned entity pairs $\mathcal{S}$ (a.k.a, *alignment seed*) is provided. In the testing phase, given an entity $e \in \mathcal{G}_1$, the model requires to identify the equivalent $e_2 \in \mathcal{G}_2$ from all candidate entities.

4 Methodology

In this section, we present our LoginMEA framework, illustrated in Figure 3. Our model consists of four modules: 1) the *modality feature encoder* to generate uni-modal embeddings for each entity, 2) the *local multi-modal interactive fusion* to generate joint entity embeddings with multi-modal interactions, 3) the *global multi-modal interactive aggregation* to achieve relational interaction between joint multi-modal entity embeddings, and 4) the *contrastive alignment loss* to achieve entity alignment for equivalent entities.

4.1 Modality Feature Encoder

To capture entity features of all modalities, each modality takes a uni-modal encoder to generate the uni-modal embedding.

Visual Encoder. To capture visual features for entity images, we employ a pre-trained visual models (PVM) as the feature extractor, which generates embedded features of the visual modality through learnable convolutional layers. Specifically, for each image x_v, we input it into the PVM, and refine the output embedding with a feed-forward layer, which is formulated as follows:

$$\boldsymbol{e}_v = \boldsymbol{W}_v \mathrm{PVM}(x_v) + b_v, \tag{1}$$

where $\boldsymbol{W}_v \in \mathbb{R}^{d \times d_v}$ and $b_v \in \mathbb{R}^d$ are learnable parameters. Following previous works [20, 19, 8], we adopt pre-trained VGG-16 [28] for cross-KG datasets (FB15K-DB15K, FB15K-YAGO15K), and ResNet-152 [11] for bilingual datasets (DBP15K), respectively.

Attribute and Relation Encoder. In MMKGs, entities encompass diverse relation and attribute information. Consequently, we follow prior approaches [20, 19, 8] and employ a method akin to a bag-of-words model to capture entity attributes and relations. Specifically, we construct N-hot vectors for attributes and relations, where the corresponding position is set to 1 if the entity has the specific attribute or relation, otherwise 0, respectively. Note that, following previous works [8], we also consider the most frequently occurring top-K attributes and relations of all entities, leading to K-dimensional vectors. Afterward, we obtain the embeddings of entity attributes and relations as follows:

$$\boldsymbol{e}_l = \boldsymbol{W}_l \boldsymbol{x}_l + b_l, l \in \{a, r\}, \tag{2}$$

where $l \in \{a, r\}$ denotes the attribute or relation. $\boldsymbol{W}_l \in \mathbb{R}^{d \times d_l}$ and $b_l \in \mathbb{R}^d$ are learnable parameters, and $\boldsymbol{x}_l \in \mathbb{R}^{d_l}$ denotes the bag-of-attribute or bag-of-relation features, respectively.

4.2 Local Multi-modal Interactive Fusion

To capture multi-modal feature interactions, we propose local multi-modal interactive fusion, allowing each entity to incorporate complementary information for holistic entity semantics.

Entity-specific Adaptive Modality Weights. Conventional methods [21, 10] simply concatenate multi-modal embeddings or adopt global modality weights for all entities, which can hardly capture diverse importance of different modalities for each entity. Therefore, we design an entity-specific adaptive weighted mechanism for different modalities. For entity e and its modality embedding $\boldsymbol{e}_m, m \in \{v, a, r\}$, the modality weights α_m are derived by:

$$\alpha_m = \frac{\exp(\boldsymbol{w}_m^\top \operatorname{Tanh}(\boldsymbol{e}_m))}{\sum_{n\in\{v,a,r\}} \exp(\boldsymbol{w}_n^\top \operatorname{Tanh}(\boldsymbol{e}_n))}, \tag{3}$$

where $\boldsymbol{w}_m \in \mathbb{R}^d$ is a learnable vector of modality m. According to α_m of the entity, we can control the impacts of uni-modal features in the fusion. The weighted uni-modal embeddings is as follows:

$$\hat{\boldsymbol{e}}_m = \alpha_m \boldsymbol{e}_m, m \in \{v, a, r\}, \tag{4}$$

where $\hat{\boldsymbol{e}}_m$ is the weighted embedding of modality m of entity e.

Low-rank Interactive Fusion. To capture the multi-modal interactions of entity information, ***tensor fusion*** [42] is a successful approach for multi-modal fusion, which can enrich multifarious multi-modal interactions at the embedding vector element-level. Specifically, the input embeddings are firstly transformed into high-dimensional tensors, and then mapped into a low-dimensional embedding space. The joint entity embedding $\boldsymbol{h}_e$ is derived as follows:

$$\boldsymbol{h}_e = \boldsymbol{\mathcal{W}} \cdot \boldsymbol{\mathcal{Z}} + b, \tag{5}$$

where $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{(d_1\times d_2\times\cdots\times d_M)\times d_h}$ is an (M+1)-order weight tensor, $b \in \mathbb{R}^{d_h}$ is the bias, and $\boldsymbol{\mathcal{Z}} \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_M}$ is a high-dimensional M-order tensor for the interacted multi-modal feature. Note that the operation $\cdot$ denotes the tensor multiplication, leading to output embedding $\boldsymbol{h}_e \in \mathbb{R}^{d_h}$. Here, $\boldsymbol{\mathcal{Z}}$ is calculated by the mathematical outer product between the augmented vector of visual, attribute, and relation feature as $\boldsymbol{z}_v = [\hat{\mathbf{e}}_m\ 1]^\top$, $\boldsymbol{z}_a = [\hat{\mathbf{e}}_a\ 1]^\top$, and $\boldsymbol{z}_r = [\hat{\mathbf{e}}_r\ 1]^\top$, respectively. The extra constant dimension with value 1 retains the uni-modal features during the interaction, and thus $\boldsymbol{\mathcal{Z}}$ is defined as:

$$\begin{aligned}\boldsymbol{\mathcal{Z}} &:= \bigotimes_{m=1}^{M} \boldsymbol{z}_m := \begin{bmatrix} \hat{\boldsymbol{e}}_v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \hat{\boldsymbol{e}}_a \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \hat{\boldsymbol{e}}_r \\ 1 \end{bmatrix} \\ &= [\underbrace{\hat{\boldsymbol{e}}_v, \hat{\boldsymbol{e}}_a, \hat{\boldsymbol{e}}_r}_{\text{Uni-modal}}, \underbrace{\hat{\boldsymbol{e}}_v \otimes \hat{\boldsymbol{e}}_a, \hat{\boldsymbol{e}}_v \otimes \hat{\boldsymbol{e}}_r, \hat{\boldsymbol{e}}_a \otimes \hat{\boldsymbol{e}}_r}_{\text{Bi-modal}}, \underbrace{\hat{\boldsymbol{e}}_v \otimes \hat{\boldsymbol{e}}_a \otimes \hat{\boldsymbol{e}}_r}_{\text{Tri-modal}}]^\top,\end{aligned} \tag{6}$$

where $\otimes$ denotes the outer product between vectors. Here, we can observe that the uni-modal terms retain original information within each modality, the bi-modal terms retain interactions between two modalities, and the tri-modal terms retain interactions among three modalities. It is worth noting that, for 3 modalities, the total number of these terms is $C_3^1 + C_3^2 + C_3^3 = 7$. When there are n modalities, it will yield $C_n^1 + C_n^2 + \cdots + C_n^n = 2^n - 1$ terms. In this way, Eq. (6) enriches multifarious feature interactions at the vector element-level, and allows scalability for any number of modalities.

However, the dimensionality of the tensor $\boldsymbol{\mathcal{Z}}$ grows exponentially with the number of modalities as $\prod_{m=1}^{M} d_m$. Additionally, the number of parameters that need to be learned in $\boldsymbol{\mathcal{W}}$ will also increase accordingly. This situation not only results in severe computational overhead, but may also lead to overfitting in training. Therefore, to alleviate this, we adopt a ***low-rank multi-modal fusion*** [22] approach, which remains the enriched capability of feature interactions but develops an efficient manner. Specifically, we leverage a low-rank weight decomposition to approximate original weight tensor $\boldsymbol{\mathcal{W}}$:

$$\widetilde{\boldsymbol{\mathcal{W}}} := \sum_{i=1}^{R} \bigotimes_{m=1}^{M} \boldsymbol{w}_m^{(i)}, \tag{7}$$

where R is called *rank* of the tensor $\widetilde{\boldsymbol{\mathcal{W}}}$, which is the minimum value that makes the decomposition valid. There are r decomposition factors with $\boldsymbol{w}_m^{(i)} \in \mathbb{R}^{d_m\times d_h}$, and the outer product $\bigotimes_{m=1}^{M} \boldsymbol{w}_m^{(i)} \in \mathbb{R}^{(d_1\times d_2\times\cdots\times d_M)\times d_h}$. We can easily infer that each resulted tensor of the outer product has the rank of 1, since it is linearly dependent on vectors in the tensor. Based on the low-rank weight tensor, the tensor fusion in Eq. (5) can be derived as (b is omitted here):

$$\begin{aligned}\boldsymbol{h}_e &= (\sum_{i=1}^{R} \bigotimes_{m=1}^{M} \boldsymbol{w}_m^{(i)}) \cdot \boldsymbol{\mathcal{Z}} \\ &= \sum_{i=1}^{R} (\bigotimes_{m=1}^{M} \boldsymbol{w}_m^{(i)} \cdot \boldsymbol{\mathcal{Z}}) \\ &= \sum_{i=1}^{R} (\bigotimes_{m=1}^{M} \boldsymbol{w}_m^{(i)} \cdot \bigotimes_{m=1}^{M} \boldsymbol{z}_m) \\ &= \Lambda_{m=1}^{M} [\sum_{i=1}^{R} \boldsymbol{w}_m^{(i)} \cdot \boldsymbol{z}_m],\end{aligned} \tag{8}$$

where $\Lambda_{m=1}^{M}$ is defined as $\Lambda_{m=1}^{3} = \boldsymbol{z}_v \circ \boldsymbol{z}_a \circ \boldsymbol{z}_r$, and $\circ$ is Hadamard product. Therefore, in this paper, the low-rank multi-modal fused embedding $\boldsymbol{h}_e$ of 3 modalities can be expressed as follows:

$$\boldsymbol{h}_e = (\sum_{i=1}^{R} \boldsymbol{w}_v^{(i)} \cdot \boldsymbol{z}_v) \circ (\sum_{i=1}^{R} \boldsymbol{w}_a^{(i)} \cdot \boldsymbol{z}_a) \circ (\sum_{i=1}^{R} \boldsymbol{w}_r^{(i)} \cdot \boldsymbol{z}_r), \tag{9}$$

which enables to derive $\boldsymbol{h}_e$ directly based on the uni-modal embeddings and modal-specific decomposition factors, avoiding the heavy computation of large input tensor $\boldsymbol{\mathcal{Z}}$ and weight tensor $\boldsymbol{\mathcal{W}}$ in Eq. (5), while still allowing element-level multi-modal interactions.

Remarks. Actually, Eq. (8) reduces the computational complexity of tensorization and fusion from $O(d_h \times \prod_{m=1}^{M} d_m)$ to $O(d_h \times r \times \sum_{m=1}^{M} d_m)$, and adopts less parameters to avoid overfitting. Besides, Eq. (8) comprises fully differentiable operations, allowing the parameters $\boldsymbol{w}_m^{(i)}$ to be learned via back-propagation. Moreover, as $\boldsymbol{z}_m$ involves entity-specific weighted modality embedding $\hat{\boldsymbol{e}}_m$, the final fused multi-modal embedding $\boldsymbol{h}_e$ not only encompasses multifarious inter-modality interaction details as in Eq. (6), but also captures entity-specific importance of different modalities of each entity.

4.3 Global Multi-modal Interactive Aggregation

To perceive structural information of entities, we aggregate entity neighbors based on the holistic joint multi-modal entity embeddings, retaining relational graph structures for better entity embeddings.

Relational Reflection Graph Attention Network. To capture different importance of entity neighbors, it is intuitive to employ graph attention networks (GATs) [37]. However, vanilla GAT can hardly capture diverse relations between entities. To this end, we adopt a relational reflection graph attention network [25] to aggregate entity neighbors retaining relational structural information. Specifically, the l-th layer's embedding for entity e_i can be obtained as follow:

$$\boldsymbol{h}_{e_i}^{l+1} = \operatorname{Tanh}\left(\sum_{r_j, e_j\in\mathcal{N}(e_i)} \phi(r_j) \boldsymbol{M}_{r_j} \boldsymbol{h}_{e_j}^{l}\right), \tag{10}$$

where $\mathcal{N}(e_i)$ denotes the set of neighboring relations and entities. e_j and r_j denotes the neighboring entity and relation, respectively. $\phi(r_j)$ is a relation-specific scalar, controlling the importance of relation r_j in aggregating the corresponding neighboring entities. Here, $\boldsymbol{M}_{r_j}$ is a relational transformation matrix reflecting relation r_j, which naturally ensures the same entity is transformed by different relations distinguishable in different positions, namely *relational differentiation* property [25] (as shown in Figure 3(c)).

However, it is reported in Mao et al. [25] that a transformation matrix without constraints can hardly remain the *dimensional isometry* property [25], i.e., *when two entity embeddings are transformed by the same relation, their norms and relative distance should be retained* (as shown in Figure 3(c)). Therefore, if two entities from different MMKGs are aligned, their neighbors with the same relation can be easily aligned in the embedding space. To remain this relational consistency, a simple yet effective way is to constrain $\boldsymbol{M}_{r_j}$ as an orthogonal matrix. We refer the readers to the literature [25]. For implementation, $\boldsymbol{M}_{r_j}$ can be achieved by:

$$\boldsymbol{M}_{r_j} := \boldsymbol{I} - 2\boldsymbol{h}_{r_j}\boldsymbol{h}_{r_j}^T, \tag{11}$$

where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{h}_{r_j} \in \mathbb{R}^d$ denotes the learnable relation embedding of r_j. Here $\boldsymbol{h}_{r_j}$ is randomly initialized, and keeps normalized in learning to ensure $\left\|\boldsymbol{h}_{r_j}\right\|_2 = 1$. The proof for orthogonality is shown in Eq. (14). Using the relation embeddings, we can easily define the importance of neighbors with different relations. Similar to GAT [37], we achieve $\phi(r_j)$ by:

$$\phi(r_j) = \frac{\exp(\boldsymbol{q}^T\boldsymbol{h}_{r_j})}{\sum_{r_k,e_k\in\mathcal{N}(e_i)}\exp(\boldsymbol{q}^T\boldsymbol{h}_{r_k}))}, \tag{12}$$

where $\boldsymbol{q} \in \mathbb{R}^d$ denotes a learnable vector to measure the importance of the relation. To perceive global multi-hop structures, following previous studies [37, 34], we collect multi-hop neighboring information by stacking entity embeddings from different layers:

$$\boldsymbol{g}_{e_i} = \left[\boldsymbol{h}_{e_i}^0 \| \boldsymbol{h}_{e_i}^1 \| \ldots \| \boldsymbol{h}_{e_i}^l\right], \tag{13}$$

where $\|$ denotes concatenation. Note that, at the first layer, we initialize the input embeddings $\boldsymbol{h}_{e_i}^0$ with locally fused multi-modal embeddings, and then perceive global information with stacked layers.

Remarks. It is easy to prove that $\boldsymbol{M}_{r_j}$ is an orthogonal matrix with constraint $\left\|\boldsymbol{h}_{r_j}\right\|_2 = 1$, which can be derived by:

$$\begin{aligned}\boldsymbol{M}_{r_j}^T\boldsymbol{M}_{r_j} &= (\boldsymbol{I} - 2\boldsymbol{h}_{r_j}\boldsymbol{h}_{r_j}^T)^T(\boldsymbol{I} - 2\boldsymbol{h}_{r_j}\boldsymbol{h}_{r_j}^T)\\ &= \boldsymbol{I} - 4\boldsymbol{h}_{r_j}\boldsymbol{h}_{r_j}^T + 4\boldsymbol{h}_{r_j}\boldsymbol{h}_{r_j}^T\boldsymbol{h}_{r_j}\boldsymbol{h}_{r_j}^T = \boldsymbol{I}.\end{aligned} \tag{14}$$

The number of parameters of $\boldsymbol{M}_{r_j}$ is $|\mathcal{R}| \times d$ much less than $|\mathcal{R}| \times d^2$ in vanilla transformation matrix [27]. In this way, we not only build relational structures between multi-modal entity information, but also retain the relational consistency between diffent MMKGs.

4.4 Contrastive Alignment Loss

To ensure the consistency of equivalent entities for MMEA, inspired by contrastive learning works [12, 36], we define the training loss as:

$$\mathcal{L} = \sum_{(e_i,e_j)\in\mathcal{S}} -\log\frac{\exp(\text{sim}(\boldsymbol{g}_{e_i},\boldsymbol{g}_{e_j})/\tau)}{\sum_{(e_i,e_k)\notin\mathcal{S}}\exp(\text{sim}(\boldsymbol{g}_{e_i},\boldsymbol{g}_{e_k}))/\tau)}, \tag{15}$$

where $\mathcal{S}$ denotes the set of pre-aligned entity pairs, served as positive samples. For each positive entity pair, we create K negative entity pairs by replacing $e_j \in \mathcal{G}_2$ with false entity $e_k \in \mathcal{G}_2$. τ is a temperature factor, where a smaller τ emphasizes more on hard negatives, and we achieve $\text{sim}(\cdot)$ with cosine similarity for simplicity.

5 Experiments

5.1 Experimental Settings

Datasets. Following prior studies [19, 8], we employ two types of multi-modal entity alignment (EA) datasets. (1) Cross-KG datasets: we select FB15K-DB15K and FB15K-YAGO15K public datasets, which are deemed as the most typical datasets for MMEA task [4]. (2) Bilingual datasets: DBP15K [31, 20] is a commonly used benchmark for bilingual entity alignment, which contains three datasets built from the multilingual versions of DBpedia, including $\text{DBP15K}_{\text{ZH-EN}}$, $\text{DBP15K}_{\text{JA-EN}}$ and $\text{DBP15K}_{\text{FR-EN}}$. Each of the bilingual datasets contains about 400K triples and 15K pre-aligned entity pairs. We show the dataset details in **supplementary material** [29]. Notably, there are fewer relations, attributes and images in YAGO15K, which may lead to a sparser graph and the greater alignment difficulty. Following previous works [19, 8], we utilize 20%, 50%, 80% of true entity pairs as alignment seeds for training on cross-KG datasets, whereas we use 30% of entity pairs as alignment seeds for training on bilingual datasets. For the entities without corresponding images, we assign random vectors for the visual modality, as the setting of previous methods [19, 8].

Evaluation Metrics. In adherence to prior works [20, 19, 8], evaluation metrics utilized include Hits@1 (H@1), Hits@10 (H@10), and Mean Reciprocal Rank (MRR). Hits@N denotes the proportion of correct entities ranked in the top-N ranks, while MRR is the average reciprocal rank of correct entities. Higher values of Hits@N and MRR indicate better performance.

Baselines. We compare the proposed LoginMEA with the following competitive entity alignment baselines, including two groups: **Traditional EA Methods:** (1) **TransE** [2] assumes that the entity embedding ought to closely align with the sum of the attribute embedding and their relation. (2) **IPTransE** [45] introduces an iterative entity alignment mechanism, employing joint knowledge embeddings to encode entities and relations across multiple KGs into a unified semantic space. (3) **GCN-align** [39] utilizes Graph Convolutional Networks [13] to generate entity embeddings and combines them with attribute embeddings to align entities. (4) **KECG** [14] proposes a semi-supervised entity alignment method through joint knowledge embedding and cross-graph model learning. **Multi-modal EA Methods:** (1) **POE** [21] defines overall probability distribution as the product of all uni-modal experts. (2) **Chen et al.** [4] designs a multi-modal fusion module to integrate knowledge representations from multiple modalities. (3) **HMEA** [10] combines the structure and visual representations in the hyperbolic space. (4) **EVA** [20] integrates multi-modal information into a joint embedding, empowering the alignment model to auto-optimize modality weights. (5) **MSNEA** [5] develops a vision-guided relation learning mechanism for inter-modal knowledge enhancement. (6) **ACK-MMEA** [15] designs a multi-modal attribute uniformization method to generate an attribute-consistent MMKG. (7) **XGEA** [41] proposes a structural-visual attention network to guide the learning of embeddings. (8) **UMAEA** [9] introduces multi-scale modality hybrid for modality noise. (9) **PSNEA** [26] advocates an incremental alignment pool strategy to alleviate alignment seed scarcity issues. (10) **MCLEA** [19] performs contrastive learning to jointly model intra-modal and inter-modal interactions in MMKGs. (11)

Table 1. Experimental results on the 2 cross-KG datasets, including **FB15K-DB15K** (**FB-DB15K** for short) and **FB15K-YAGO15K** (**FB-YG15K** for short). We evaluate our model in different resource settings, with 20%, 50% and 80% seed alignments as in previous studies [8, 19]. The best result is **bold-faced** and the runner-up is underlined. $*$ indicates that the results are reproduced by the official source code.

Methods	FB-DB15K (20%)			FB-DB15K (50%)			FB-DB15K (80%)			FB-YG15K (20%)			FB-YG15K ((50%)			FB-YG15K (80%)		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
TransE [2]	.078	.240	.134	.230	.446	.306	.426	.659	.507	.064	.203	.112	.197	.382	.262	.392	.595	.463
IPTransE [45]	.065	.215	.094	.210	.421	.283	.403	.627	.469	.047	.169	.084	.201	.369	.248	.401	.602	.458
GCN-align [39]	.053	.174	.087	.226	.435	.293	.414	.635	.472	.081	.235	.153	.235	.424	.294	.406	.643	.477
KECG* [14]	.128	.340	.200	.167	.416	.251	.235	.532	.336	.094	.274	.154	.167	.381	.241	.241	.501	.329
POE [21]	.126	.151	.170	.464	.658	.533	.666	.820	.721	.113	.229	.154	.347	.536	.414	.573	.746	.635
Chen et al. [4]	.265	.541	.357	.417	.703	.512	.590	.869	.685	.234	.480	.317	.403	.645	.486	.598	.839	.682
HMEA [10]	.127	.369	-	.262	.581	-	.417	.786	-	.105	.313	-	.265	.581	-	.433	.801	-
EVA [20]	.134	.338	.201	.223	.471	.307	.370	.585	.444	.098	.276	.158	.240	.477	.321	.394	.613	.471
MSNEA [5]	.114	.296	.175	.288	.590	.388	.518	.779	.613	.103	.249	.153	.320	.589	.413	.531	.778	.620
ACK-MMEA [15]	.304	.549	.387	.560	.736	.624	.682	.874	.752	.289	.496	.360	.535	.699	.593	.676	.864	.744
XGEA* [41]	.475	.739	.565	.681	.857	.746	.791	.919	.840	.431	.691	.521	.585	.801	.666	.705	.873	.768
UMAEA* [9]	.533	.813	.633	.664	.868	.740	.817	.915	.853	.422	.695	.520	.599	.783	.668	.728	.862	.778
MCLEA [19]	.445	.705	.534	.573	.800	.652	.730	.883	.784	.388	.641	.474	.543	.759	.616	.653	.835	.715
MEAformer [8]	.578	.812	.661	.690	.871	.755	.784	.921	.834	.444	.692	.529	.612	.808	.682	.724	.880	.783
LoginMEA (Ours)	**.667**	**.854**	**.735**	**.758**	**.898**	**.810**	**.843**	**.942**	**.880**	**.622**	**.818**	**.691**	**.706**	**.865**	**.763**	**.780**	**.933**	**.833**

Table 2. Experimental results on 3 bilingual datasets, including $\textbf{DBP15K}_{\text{ZH-EN}}$, $\textbf{DBP15K}_{\text{JA-EN}}$ and $\textbf{DBP15K}_{\text{FR-EN}}$. The best result is **bold-faced** and the runner-up is underlined. $*$ indicates the results are reproduced by the official source code.

Methods	$\textbf{DBP15K}_{\text{ZH-EN}}$			$\textbf{DBP15K}_{\text{JA-EN}}$			$\textbf{DBP15K}_{\text{FR-EN}}$		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
GCN-align [39]	.434	.762	.550	.427	.762	.540	.411	.772	.530
KECG [14]	.478	.835	.598	.490	.844	.610	.486	.851	.610
BootEA [32]	.629	.847	.703	.622	.854	.701	.653	.874	.731
NAEA [46]	.650	.867	.720	.641	.873	.718	.673	.894	.752
EVA [20]	.761	.907	.814	.762	.913	.817	.793	.942	.847
MSNEA [5]	.643	.865	.719	.572	.832	.660	.584	.841	.671
XGEA* [41]	.803	.939	.854	.794	.942	.849	.821	.954	.871
UMAEA* [9]	.811	.969	.871	.812	.973	.873	.822	.981	.884
PSNEA [26]	.816	.957	.869	.819	.963	.868	.844	.982	.891
MCLEA [19]	.816	.948	.865	.812	.952	.865	.834	.975	.885
MEAformer [8]	.847	.970	.892	.842	.974	.892	.845	.976	.894
LoginMEA (Ours)	**.873**	**.978**	**.913**	**.866**	**.981**	**.911**	**.881**	**.988**	**.924**

MEAformer [8] utilizes a transformer-based fusion method to predict relatively mutual weights among modalities for each entity.

Among the methods, MCLEA and MEAformer are typical competitive methods, where MCLEA enhances single-modal representation relevance via contrastive learning, while MEAformer develops transformer-based attention multi-modal fusion method.

Implementation Details. In our experiments, the graph encoder is configured with a hidden layer size of $d_g = 300$ across 3 layers. The visual feature dimension d_v is allocated 4096, while the attribute and relation feature sizes d_a and d_r are configured at 1000. The graph embedding output is fixed at a size of 300, whereas the embedding sizes for other modalities are determined to be 100. The training was conducted over 600 epochs with a batch size of 3,500. AdamW optimizer [24] was employed with a learning rate of 5e-3 and a weight decay of 1e-2. Following previous works [20, 19, 8], we adopt an iterative training strategy to overcome the lack of training data in the same way, and also do not consider entity names for fair comparison. All the experiments are conducted on a 64-bit machine with two NVIDIA A100 GPUs, and 256 GB RAM memory. Our best hyperparameters are reported in the **supplementary material** [29], which are tuned by grid search according to MRR metric.

5.2 Overall Results

To verify the effectiveness of LoginMEA, we report overall average results on cross-KG and bilingual datasets as shown in Table 1 and Table 2, respectively. From the tables, we have several observations: (1) *Our proposed method outperforms all compared baseline models on 9 benchmarks in terms of three key metrics (H@1, H@10, and MRR).* Specifically, our model consistently outperforms state-of-the-art (SOTA) baselines, achieving significant improvements in Hits@1 scores across ZH-EN/JA-EN/FR-EN datasets with DBP15K, and elevates the existing high-performing Hits@1 scores from .847/.842/.845 to .873/.866/.881. Moreover, our model brings about an average increase of 3.9% and 8.1% in H@1 on cross-KG datasets at 80% and 50% seed settings, respectively. (2) *Our model achieves better results in relatively low-resource data scenario.* Compared to the runner-up method results, our model achieves an even more pronounced average gain of 13.3% in H@1 and 11.8% in MRR on cross-KG datasets with a 20% alignment seed setting. Our local multi-modal interacted fusion module enhances expressive entity embeddings in such scenarios by facilitating multi-modal information deep interaction. (3) *Compared to traditional EA models, the MMEA models all show significant enhancements.* Remarkably, our model exhibits substantial enhancements in H@1 scores, with an average increase of 47.6% (ranging from 37.4% to 53.9%) on Cross-KG datasets and an average improvement of 21.8% (ranging from 20.8% to 22.5%) on cross-lingual datasets. This demonstrates a significant improvement in entity alignment through the incorporation of multi-modal information. All the results demonstrate the effectiveness of our proposed LoginMEA model.

5.3 Ablation Study

To investigate the impact of each module in LoginMEA, we design two groups of variants in the ablation study: (1) LoginMEA with various components, such as removing or replacing specific modules. (2) LoginMEA without one specific modality, including visual, relation, and attribute. We conduct variant experiments on two Cross-KG datasets with 20% alignments seeds, showcasing in Table 3.

From the first group of variants, we remove the low-rank module and adaptive weights from the local interaction fusion module, causing a decline in performance. Notably, the absence of the low-rank module has a greater impact, highlighting the importance

Table 3. Variant experiments on **FB15K-DB15K** and **FB15K-YG15K** (20%). "*w/o*" means removing corresponding module from the complete model. "repl." means replacing corresponding module with the other module.

	Model	FB15K-DB15K			FB15K-YG15K		
		H@1	H@10	MRR	H@1	H@10	MRR
	LoginMEA	**.667**	**.854**	**.735**	**.622**	**.818**	**.691**
Component	*w/o* Low-rank	.607	.816	.683	.563	.763	.633
Component	*w/o* Adaptive weights	.639	.848	.714	.606	.817	.681
Component	*repl.* Concate Fusion	.517	.757	.603	.513	.735	.591
Component	repl. GAT	.474	.671	.542	.349	.546	.416
Component	repl. ICL	.605	.835	.693	.542	.785	.628
Modality	*w/o* Visual	.629	.847	.712	.595	.804	.670
Modality	*w/o* Attribute	.634	.832	.706	.579	.788	.653
Modality	*w/o* Relation	.612	.830	.692	.580	.799	.657

of effective inter-modality interaction over modality weight information. Furthermore, replacing the local interaction fusion module with a simple concatenation fusion module led to a more significant drop, confirming its effectiveness. It dramatically degrades the performance when replacing the global interaction aggregation module with GAT, which emphasizes its crucial role in learning relational global multi-modal information interaction. Lastly, substituting our alignment contrastive loss with Intra-modal Contrastive Loss (ICL) used in previous studies [19, 8], resulted in a decrease in overall performance, validating the effectiveness of our original loss function.

From the second group of variants, we observe varying degrees of performance decline upon removing different modalities. The removal of any modality information affects our model's local multi-modal interaction of the fusion module. Notably, we notice that the removal of relations has a relatively greater impact on overall performance compared to visual and attribute modalities. This can be attributed to the high frequency and importance of relations in Cross-KG knowledge graphs, influencing the global multi-modal information interaction within our aggregation module.

5.4 Performance under Different Modeling Paradigms

To further validate the effectiveness of our proposed modeling paradigm for MMEA task, we implement two variants according to the graph-as-modality and graph-upon-modality paradigm mentioned in Figure 2 based on the modules of LoginMEA: (1) **LoginMEA-GAM** is implemented according to the *graph-as-modality* paradigm, where the modality feature encoders, the structural modeling of graph, and the fusion module are all consistent with our LoginMEA method. (2) **LoginMEA-GUM** follows the *graph-upon-modality* paradigm, where all basic modules also maintain the same with LoginMEA to accurately explore the impact of paradigm that refines entities with relational structures on all modalities.

For LoginMEA, LoginMEA-GAM and LoginMEA-GUM, we conduct experiments under various alignment seed settings on FB15K-DB15K dataset, with results depicted in Figure 4. We can observe that our LoginMEA consistently achieves the best performance across different alignment seeds and metrics. This confirms the efficacy of our proposed paradigm, which first involves a local interactive fusion for more precise and holistic entity representations, and then follows by a global interactive aggregation upon the graph structure. Furthermore, compared with LoginMEA-GAM, LoginMEA-GUM shows better performance due to its full aggregation of all multi-modal information upon relational structures, which facilitates the learning of relational associations in modalities. However, the

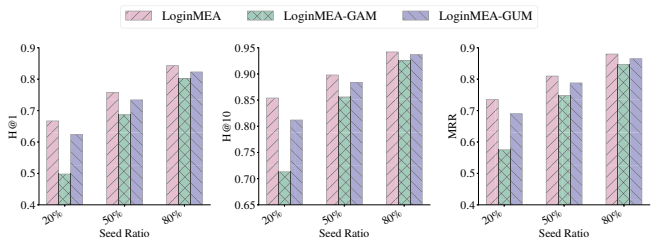

Figure 4. Results of different modeling paradigms for MMEA task on FB15K-DB15K with different seed ratios.

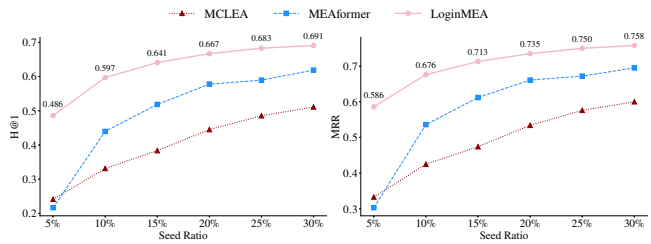

Figure 5. Results in the low-resource data scenario with proportions of seed alignments on FB15K-DB15K dataset.

absence of local interactions among entity modalities leads to sub-optimal results of LoginMEA-GUM compared to LoginMEA. Additionally, LoginMEA significantly outperforms LoginMEA-GAM and LoginMEA-GUM under the 20% alignment seed setting, confirming the obvious advantage of the modeling paradigm in LoginMEA that enhances the distinctiveness of entity embeddings.

5.5 Performance on Low-Resource Training Data

To further explore the performance with low-resource training data, we vary the alignment seed ratio from 5% to 30%. The latest baselines MCLEA [19] and MEAformer [8] are compared in Figure 5. We can observe that as the proportion of alignment seeds decreases, the performance of all methods tend to decrease in terms of metrics. However, it is obvious that our LoginMEA continuously outperforms MCLEA and MEAformer, which indicates the effectiveness of our proposed method especially under the low-resource scenarios. Moreover, it is worth noting that the gap between them is much more significant when the seed alignments are extremely few (5%), which guarantees the reliability and robustness of LoginMEA under extremely low-resource scenarios with local-to-global interactions.

6 Conclusion

In this paper, we propose a novel local-to-global interaction network for MMEA, termed as LoginMEA, by facilitating the interactions of multi-modal information and relational graph structures. Particularly, we develop a local multi-modal interactive fusion module to capture diverse impacts and element-level feature interactions among modalities. Besides, we devise a global multi-modal interactive aggregation module to fully capture relational associations between entities with their multi-modal information. Empirical results show that LoginMEA consistently outperforms competitors across all datasets and metrics. Further experiments demonstrate the effectiveness of the multi-modal fusion paradigm and the robustness of LoginMEA in low-resource scenarios.

Acknowledgements

We would like to thank the anonymous reviewers for their comments. This work was supported by the National Key Research and Development Program of China (Grant No.2021YFB3100600), the Youth Innovation Promotion Association of CAS (No.2021153), and the Postdoctoral Fellowship Program of CPSF (No.GZC20232968).

References

[1] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks, 2018.

[2] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In Proceedings of NeurIPS, pages 2787–2795, 2013.

[3] Y. Cao, Z. Liu, C. Li, Z. Liu, J. Li, and T. Chua. Multi-channel graph neural network for entity alignment. In Proceedings of ACL, pages 1452–1461, 2019.

[4] L. Chen, Z. Li, Y. Wang, T. Xu, Z. Wang, and E. Chen. Mmea: entity alignment for multi-modal knowledge graph. In Proceedings of KSEM, pages 134–147. Springer, 2020.

[5] L. Chen, Z. Li, T. Xu, H. Wu, Z. Wang, N. J. Yuan, and E. Chen. Multi-modal siamese network for entity alignment. In Proceedings of KDD, pages 118–126, 2022.

[6] M. Chen, Y. Tian, M. Yang, and C. Zaniolo. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In Proceedings of IJCAI, pages 1511–1517, 2017.

[7] M. Chen, Y. Tian, K. Chang, S. Skiena, and C. Zaniolo. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In Proceedings of IJCAI, pages 3998–4004, 2018.

[8] Z. Chen, J. Chen, W. Zhang, L. Guo, Y. Fang, Y. Huang, Y. Zhang, Y. Geng, J. Z. Pan, W. Song, et al. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In Proceedings of ACM MM, pages 3317–3327, 2023.

[9] Z. Chen, L. Guo, Y. Fang, Y. Zhang, J. Chen, J. Z. Pan, Y. Li, H. Chen, and W. Zhang. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In Proceedings of ISWC, pages 121–139, 2023.

[10] H. Guo, J. Tang, W. Zeng, X. Zhao, and L. Liu. Multi-modal entity alignment in hyperbolic space. Neurocomputing, 461:598–607, 2021.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of CVPR, pages 770–778, 2016.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of CVPR, pages 9726–9735, 2020.

[13] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In Proceedings of ICLR, 2017.

[14] C. Li, Y. Cao, L. Hou, J. Shi, J. Li, and T. Chua. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In Proceedings of EMNLP, pages 2723–2732, 2019.

[15] Q. Li, S. Guo, Y. Luo, C. Ji, L. Wang, J. Sheng, and J. Li. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In Proceedings of WWW, pages 2499–2508, 2023.

[16] Q. Li, C. Ji, S. Guo, Z. Liang, L. Wang, and J. Li. Multi-modal knowledge graph transformer framework for multi-modal entity alignment. In Findings of EMNLP, pages 987–999, 2023.

[17] K. Liang, L. Meng, M. Liu, Y. Liu, W. Tu, S. Wang, S. Zhou, X. Liu, F. Sun, and K. He. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[18] K. Liang, L. Meng, S. Zhou, W. Tu, S. Wang, Y. Liu, M. Liu, L. Zhao, X. Dong, and X. Liu. Mines: Message intercommunication for inductive relation reasoning over neighbor-enhanced subgraphs. In Proceedings of AAAI, volume 38, pages 10645–10653, 2024.

[19] Z. Lin, Z. Zhang, M. Wang, Y. Shi, X. Wu, and Y. Zheng. Multi-modal contrastive representation learning for entity alignment. In Proceedings of COLING, pages 2572–2584, 2022.

[20] F. Liu, M. Chen, D. Roth, and N. Collier. Visual pivoting for (unsupervised) entity alignment. In Proceedings of AAAI, pages 4257–4266, 2021.

[21] Y. Liu, H. Li, A. García-Durán, M. Niepert, D. Oñoro-Rubio, and D. S. Rosenblum. MMKG: multi-modal knowledge graphs. In Proceedings of ESWC, volume 11503, pages 459–474, 2019.

[22] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. Morency. Efficient low-rank multimodal fusion with modality-specific factors. In Proceedings of ACL, pages 2247–2256, 2018.

[23] Z. Liu, Y. Cao, L. Pan, J. Li, and T. Chua. Exploring and evaluating attributes, values, and structures for entity alignment. In Proceedings of EMNLP, pages 6355–6364, 2020.

[24] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In Proceedings of ICLR, 2019.

[25] X. Mao, W. Wang, H. Xu, Y. Wu, and M. Lan. Relational reflection entity alignment. In Proceedings of CIKM, pages 1095–1104, 2020.

[26] W. Ni, Q. Xu, Y. Jiang, Z. Cao, X. Cao, and Q. Huang. Psnea: Pseudo-siamese network for entity alignment between multi-modal knowledge graphs. In Proceedings of ACM MM, pages 3489–3497, 2023.

[27] 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 Proceedings of ESWC, volume 10843 of Lecture Notes in Computer Science, pages 593–607, 2018.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun, editors, Proceedings of ICLR, 2015.

[29] T. Su, X. Zhang, J. Sheng, Z. Zhang, and T. Liu. Loginmea: Local-to-global interaction network for multi-modal entity alignment, 2024. URL https://arxiv.org/abs/2407.19625. Full version of this paper.

[30] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng. Multi-modal knowledge graphs for recommender systems. In Proceedings of CIKM, pages 1405–1414, 2020.

[31] Z. Sun, W. Hu, and C. Li. Cross-lingual entity alignment via joint attribute-preserving embedding. In Proceedings of ISWC, pages 628–644, 2017.

[32] Z. Sun, W. Hu, Q. Zhang, and Y. Qu. Bootstrapping entity alignment with knowledge graph embedding. In Proceedings of IJCAI, pages 4396–4402, 2018.

[33] Z. Sun, J. Huang, W. Hu, M. Chen, L. Guo, and Y. Qu. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In Proceedings of ISWC, pages 612–629, 2019.

[34] Z. Sun, C. Wang, W. Hu, M. Chen, J. Dai, W. Zhang, and Y. Qu. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In Proceedings of AAAI, pages 222–229, 2020.

[35] B. D. Trisedya, J. Qi, and R. Zhang. Entity alignment between knowledge graphs using attribute embeddings. In Proceedings of AAAI, volume 33, pages 297–304, 2019.

[36] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019.

[37] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In Proceedings of ICLR, 2018.

[38] L. Wang, P. Qi, X. Bao, C. Zhou, and B. Qin. Pseudo-label calibration semi-supervised multi-modal entity alignment. In Proceedings of AAAI, pages 9116–9124, 2024.

[39] Z. Wang, Q. Lv, X. Lan, and Y. Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. In Proceedings of EMNLP, pages 349–357, 2018.

[40] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, and D. Zhao. Relation-aware entity alignment for heterogeneous knowledge graphs. In S. Kraus, editor, Proceedings of IJCAI, pages 5278–5284, 2019.

[41] B. Xu, C. Xu, and B. Su. Cross-modal graph attention network for entity alignment. In Proceedings of ACM MM, pages 3715–3723, 2023.

[42] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In Proceedings of EMNLP, 2017.

[43] Q. Zhang, Z. Sun, W. Hu, M. Chen, L. Guo, and Y. Qu. Multi-view knowledge graph embedding for entity alignment. In Proceedings of IJCAI, pages 5429–5435, 2019.

[44] Y. Zhang, S. Qian, Q. Fang, and C. Xu. Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering. In Proceedings of ACM MM, pages 1089–1097, 2019.

[45] H. Zhu, R. Xie, Z. Liu, and M. Sun. Iterative entity alignment via joint knowledge embeddings. In Proceedings of IJCAI, pages 4258–4264, 2017.

[46] Q. Zhu, X. Zhou, J. Wu, J. Tan, and L. Guo. Neighborhood-aware attentional representation for multilingual knowledge graphs. In Proceedings of IJCAI, pages 1943–1949, 2019.