SGMEA: Structure-Guided Multimodal Entity Alignment

Jingwei Cheng[†] [*], **Mingxiao Guo**[†], **Fu Zhang**
School of Computer Science and Engineering, Northeastern University, China
Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education,
Northeastern University, China
`{chengjingwei,zhangfu}@mail.neu.edu.cn`,
`guomingxiao818@163.com`

Abstract

Multimodal Entity Alignment (MMEA) aims to identify equivalent entities across different multimodal knowledge graphs (MMKGs) by integrating structural information, entity attributes, and visual data, thereby promoting knowledge sharing and deep multimodal data integration. However, existing methods often overlook the deeper connections between multimodal data. They primarily focus on the interactions between neighboring entities in the structural modality while neglecting the interactions between entities in the visual and attribute modalities. To address this, we propose a structure-guided multimodal entity alignment method (SGMEA), which prioritizes structural information from knowledge graphs to enhance the visual and attribute modalities. By fusing multimodal representations, SGMEA improves the accuracy of entity alignment. Experimental results demonstrate that SGMEA achieves state-of-the-art performance across multiple datasets, validating its effectiveness and superiority in practical applications.[1]

1 Introduction

Knowledge Graphs (KGs) organize and represent real-world knowledge through a graph structure, and they have become powerful tools in fields such as question answering (Chen et al., 2021, 2022b; Lan et al., 2021) entity linking (Radhakrishnan et al., 2018), text generation (Koncel-Kedziorski et al., 2019) and information retrieval (Han et al., 2018). In recent years, as application scenarios have become increasingly complex, Multimodal Knowledge Graphs (MMKGs) have emerged (Chen et al., 2020a). MMKGs integrate multimodal data, such as visual information, into traditional KGs (Lehmann et al., 2015; Vrandečić and Krötzsch, 2014; Liu et al., 2019;

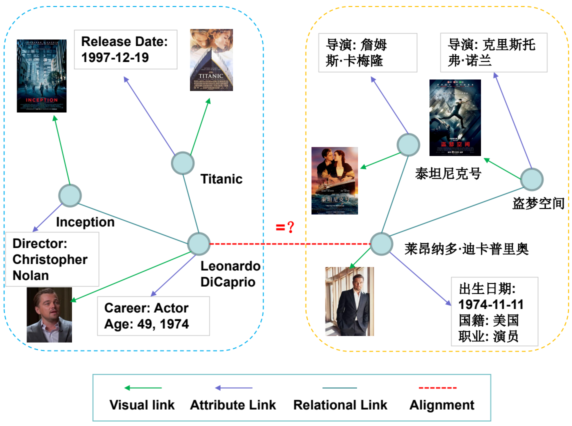

Figure 1: An example of multimodal entity alignment.

Chen et al., 2020a; Wang et al., 2021), thereby providing richer knowledge representations. In the process of MMKG integration, Multimodal Entity Alignment (MMEA) is a core task. As illustrated in Figure 1, MMEA aims to identify equivalent entities across different MMKGs by comprehensively considering the structural information of the graphs, entity attributes, and visual information. This process not only facilitates knowledge sharing between different MMKGs but also lays a solid foundation for the deep integration of multimodal data. However, existing methods typically leverage multimodal knowledge by simply combining unimodal features heuristically. These approaches overlook the deeper connections between multimodal data, resulting in the underutilization of potential cues within cross-modal information (Chen et al., 2022a). They focus only on the interaction between adjacent entities in the structural modality, while neglecting the interactions of entities in other modalities, such as visual and attribute modalities. MSNEA (Chen et al., 2022a) attempts to enhance this interaction through image-guided methods, yielding promising results. However, structural modality occupies a pivotal role among all modalities (Liu et al., 2021; Lin et al.,

[1]Code: `https://github.com/gmx1625/SGMEA`
[†]Equal contribution. [*]Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7851–7861
January 19–24, 2025.

2022). Consequently, structural modality should receive more attention and utilization during the multimodal alignment process. We speculate that a deeper exploration of intra-modality neighbor interactions will further enhance the accuracy and effectiveness of multimodal entity alignment.

To this end, we propose a structure-guided multimodal entity alignment method. This method prioritizes leveraging structural information from knowledge graphs to enhance both visual and attribute modalities, and by integrating multimodal representations, it more effectively identifies equivalent entities across different knowledge graphs. By emphasizing the core role of the structural modality, our method not only significantly improves alignment accuracy but also deeply explores potential connections between multimodal data, achieving more precise and comprehensive entity alignment. The experimental results clearly demonstrate the effectiveness of our approach.

In this paper, our main contributions are summarized as follows:

- We innovatively propose a method called SG-MEA, which prioritizes the use of structural information to enhance the visual and attribute modalities in knowledge graphs, by integrating multimodal representations to achieve more precise entity alignment.
- We particularly emphasize the importance of the structural modality in multimodal alignment and explore intra-modality interactions, thereby enhancing the accuracy and effectiveness of multimodal entity alignment.
- Our method achieves SOTA performance on three most widely used datasets, FB15K-DB15K, FB15K-YAGO15K and DBP15K datasets, validating its effectiveness and superiority in practical applications.

2 Related Work

2.1 Entity Alignment

Entity Alignment (EA) aims to identify equivalent entities across different Knowledge Graphs (KGs) to facilitate knowledge integration. Early work employed symbolic or schematic methods to address the EA problem (Wijaya et al., 2013; Suchanek et al., 2011). In recent years, embedding-based methods have gained increasing attention. These methods mainly fall into two categories: one category is translation-based methods (Bordes et al., 2013; Chen et al., 2017; Zhu et al., 2017; Sun et al., 2018; Trisedya et al., 2019; Zhang et al., 2019; Sun et al., 2019; Xin et al., 2022; Cai et al., 2022), which capture the structural information between entities through the translational properties of relations. They optimize the objective function to ensure that the distance between known aligned entity pairs in the embedding space is as small as possible, while the distance between non-aligned entity pairs is as large as possible. The other category is Graph Neural Networks (GNNs)-based methods (Wang et al., 2018; Li et al., 2019; Mao et al., 2020; Cao et al., 2019; Sun et al., 2020a; Mao et al., 2021; Sun et al., 2020b; Liu et al., 2020; Wu et al., 2020; Gao et al., 2022), which learn richer entity representations by aggregating the features of neighboring entities, effectively handling the structural information of knowledge graphs and enhancing alignment performance. Although embedding-based entity alignment methods have made significant progress in capturing the structural information of knowledge graphs and improving alignment performance, these methods mainly focus on single-modal (e.g., structural or textual) information. With the widespread application of multimodal data (e.g., images, audio, video, etc.), how to utilize multimodal information in knowledge graphs to further improve entity alignment performance has become a new research hotspot.

2.2 Multimodal Entity Alignment

Multimodal Entity Alignment (MMEA) effectively improves entity alignment performance by introducing multiple modalities of information. In recent years, researchers have proposed various methods to fully utilize these different modalities of information. PoE (Liu et al., 2019) integrates the outputs of single-modal experts by assigning probabilities to triples; MMEA (Chen et al., 2020a) generates multimodal entity representations and performs transfer learning;EVA (Liu et al., 2021)leverages visual knowledge and other auxiliary information to facilitate both supervised and unsupervised learning for entity alignment; MSNEA(Chen et al., 2022a) uses an image-guided multimodal Siamese network; MCLEA (Lin et al., 2022) explores intra- and inter-modal interactions through contrastive learning to bridge the gap between modalities; MEAformer (Chen et al., 2023a) is based on a multimodal Transformer architecture for alignment; and ACK-MMEA (Li et al., 2023) enhances knowledge graph entity alignment performance by considering multimodal attribute consistency. These methods significantly improve the accuracy and robustness of entity alignment by integrating multimodal information, providing a wealth of directions and ideas for multimodal entity alignment research.We propose a structure-guided multimodal entity alignment method that leverages knowledge graph structures to enhance visual and attribute modalities, improving the performance of knowledge graph entity alignment.

3 Method

3.1 Problem Definition

A multimodal knowledge graph can be represented as $G = (E, R, I, A, V, T_R, T_A)$, where E, R, I, A, and V are finite sets of entities, relations, images, attributes, and values, respectively. A knowledge graph consists of two types of triples: the set of relational triples T_R contains triples of the form (h, r, t), representing that entity h is related to entity t through relation r; the set of attribute triples T_A contains triples of the form (e, a, v), representing that entity e has an attribute a with value v. The goal of the multimodal entity alignment task is to identify equivalent entity pairs between two multimodal knowledge graphs. Given two multimodal knowledge graphs G^s and G^t, represented as $G^s = (E, R, I, A, V, T_R, T_A)$ and $G^t = (E', R', I', A', V', T'_R, T'_A)$, respectively, the cross-graph alignment seed set is defined as $H = \{(e, e') \mid e \in E, e' \in E', e \equiv e'\}$ where $\equiv$ denotes the equivalence between two entities. The objective of multimodal entity alignment is to find corresponding entity pairs that describe the same real-world concept in different multimodal knowledge graphs.

3.2 Framework Description

The overall framework is shown in Figure 2 and consists of three main components: the initial embedding acquisition module, the structure-guided module, and the modality fusion module.

3.3 Initial Embedding Acquisition

3.3.1 Structural Embedding

To model the structural relationships between modalities effectively, we employ a Graph Attention Network (GAT) for structural embedding (Velickovic et al., 2018). GAT adaptively assigns different attention weights to each node's neighbors, thereby capturing complex interaction information within the graph structure. For a given node in the graph, its initial feature representation is $h_i \in \mathbb{R}^d$. GAT generates a new representation h_i^g by aggregating weighted features of the node and its neighbors as follows:

$$h_i^g = \text{GAT}(W_g, M_g; x_i^g), \tag{1}$$

where M_g denotes the adjacency matrix of the graph, and W_g is a learnable diagonal matrix (Yang et al., 2015).

3.3.2 Relation, Attribute, and Visual Embedding

In the process of obtaining initial features, we employ a simple feedforward network to map relations, attributes, and visual features into a low-dimensional space. For relation features, we represent them using a bag-of-words model, where the core idea is to convert the relation name into a term frequency vector x^r. For attribute features, we utilize a pre-trained language model to process the textual information of attributes and attribute values, generating attribute features x^a through the BERT model. For visual features, we extract image features x^v using a pre-trained visual model such as ResNet-152(He et al., 2016). The mapping for each feature type can be expressed as:

$$h_i^m = \mathbf{W}_m \cdot x^m + b_m, \quad m \in \{r, a, v\}, \tag{2}$$

where $\mathbf{W}_m$ is the weight matrix for the linear transformation of relational, attribute, or visual feature, and b_m is the bias term.

3.4 Structure-guided

3.4.1 Structure-Guided Visual Embedding

To ensure that the image embedding not only captures visual information but also incorporates structural information from adjacent entities, we process the initial image embedding h_i^v with a one-layer Graph Attention Network (GAT) to generate a structure-guided image embedding h_i^{v+g}.

Specifically, we input the image embedding h_i^v and the adjacency matrix M_g into the GAT, and the updated image embedding representation is given by:

$$h_i^{v+g} = \text{GAT}(\mathbf{W}^v, M_g; h_i^v), \tag{3}$$

where $\mathbf{W}^v$ is the weight matrix for the linear transformation.Through multi-layer processing by the GAT, the image embedding not only integrates visual features but also incorporates graph structural

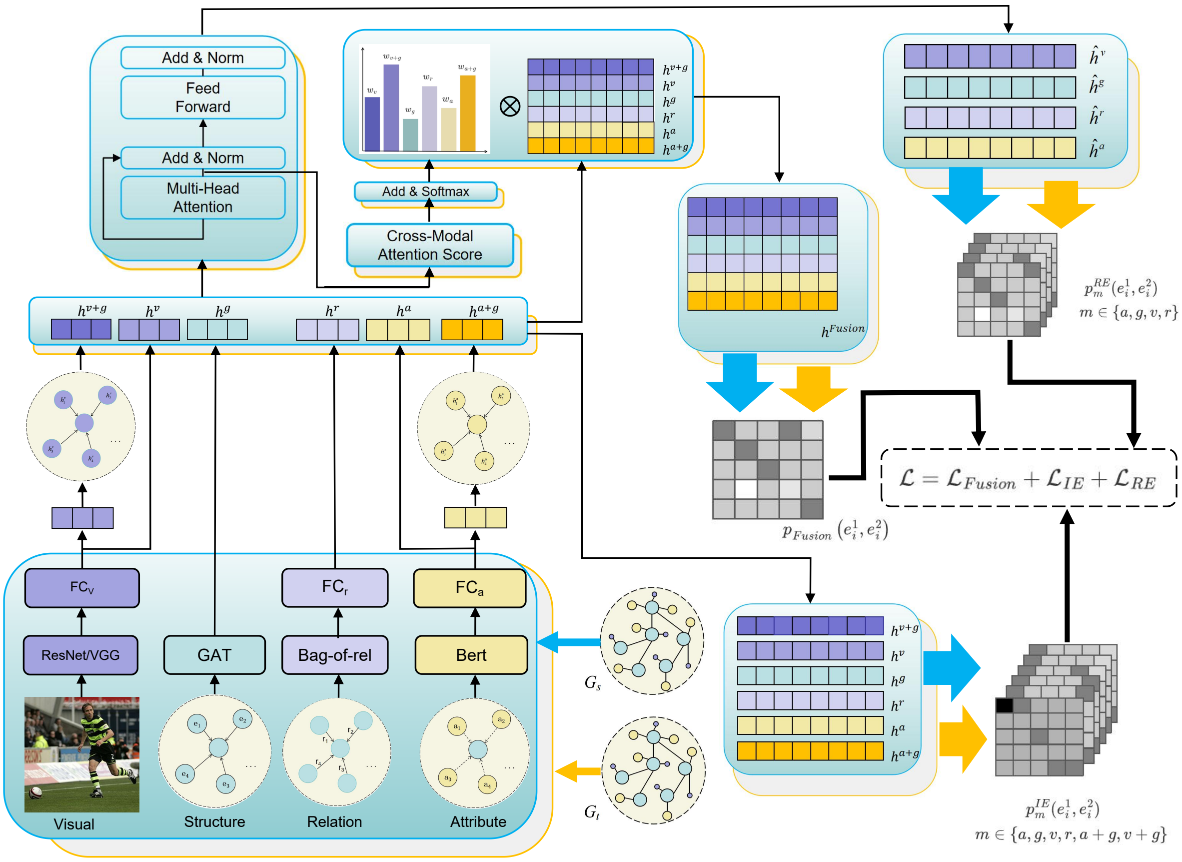

Figure 2: The overall framework of SGMEA

information, resulting in a more enriched embedding representation.

Ultimately, we obtain two levels of visual embeddings: the initial image embedding $h_i^v$and the structure-guided image embedding h_i^{v+g} obtained through further processing by the GAT. These provide a richer representation of the entity by integrating structural information at different levels.

3.4.2 Structure-Guided Attribute Embedding

Similar to the structure-guided image embedding, we also apply a Graph Attention Network (GAT) to guide the attribute embedding so that it can better integrate the structural information from neighboring entities. This results in a structure-guided attribute embedding, denoted as h_i^{a+g}.

3.4.3 Rationale for Not Applying Structure Guidance to Relations

We choose not to apply structural guidance to relations because relations inherently exist between two neighboring entities and are already explicitly modeled through their interactions. In the graph structure, relations naturally capture semantic information between entities, making additional GAT guidance unnecessary. Compared to attributes or image embeddings, the representation of relations is sufficiently robust, and further guidance may introduce redundancy or negatively impact the model's performance.

3.5 Modality Fusion

In this module, we follow Chen et al. (2023a) to adapt the vanilla Transformer (Zhou et al., 2021)

3.5.1 Modal representation generation and interaction

We first perform a linear transformation on the input representation of each modality h_m, mapping them into query vectors $Q_m^{(i)}$, key vectors $K_m^{(i)}m$, and value vectors $V_m^{(i)}$. The specific calculation formulas are as follows:

$$Q_m^{(i)}, K_m^{(i)}, V_m^{(i)} = h^m \mathbf{W}_q^{(i)}, h^m \mathbf{W}_k^{(i)}, h^m \mathbf{W}_v^{(i)}, \quad (4)$$

where $\mathbf{W}_k^{(i)}$, $\mathbf{W}_k^{(i)}$, and $\mathbf{W}_k^{(i)}$ are linear transformation matrices, and $m \in \{a, g, r, v, a+g, v+g\}$.

The interaction between modality m and modality j is computed using the scaled dot-product attention mechanism, as defined by the following formula:

$$\beta_{mj} = \text{Softmax}\left(\frac{Q_m^\top K_j}{\sqrt{d_h}}\right), \quad (5)$$

$$\text{Attention}(Q_m, K_j, V_j) = \sum\nolimits_{j\in\mathcal{M}} \beta_{mj} V_j\,, \quad (6)$$

where d_h is the hidden layer dimension, used to scale the dot product to keep it within a reasonable range.

3.5.2 Multi-head cross-attention and processing

To further enhance the model's ability to capture cross-modal interactions, we employ a Multi-Head Cross Attention (MHCA) mechanism. Multiple attention heads are calculated in parallel, and the formula for each head i is:

$$\text{head}_i = \text{Attention}(Q_m^{(i)}, K_j^{(i)}, V_j^{(i)}), \quad (7)$$

The outputs of all attention heads are then concatenated and mapped to the final output using a linear transformation matrix $\mathbf{W}_o$, as shown below:

$$\text{MHCA}(h_m) = \mathbf{W}_o\left(\bigoplus_{i=1}^{H} \text{head}_i^m\right), \quad (8)$$

H denotes the number of attention heads, and $\oplus$ represents the concatenation operation.

To further refine the modality representations, the output of MHCA is combined with the original input h_m through a residual connection and then processed with Layer Normalization, which is given by:

$$\hat{h}_m = \text{LayerNorm}(\text{MHCA}(h_m) + h_m), \quad (9)$$

After the multi-head cross-attention mechanism, a Feed-Forward Neural Network (FFN) further processes the modality representations. The FFN consists of two linear layers with ReLU activation to introduce non-linearity, defined as:

$$\text{FFN}(\hat{h}_m) = \text{ReLU}(\hat{h}_m \mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2, \quad (10)$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are linear transformation matrices, and b_1 and b_2 are bias terms. The output of the FFN is then combined with the input through a residual connection and processed with Layer Normalization:

$$\hat{h}_m = \text{LayerNorm}(\text{FFN}(\hat{h}_m) + \hat{h}_m), \quad (11)$$

3.5.3 Fusion representation generation

To generate the fused modality representation h_{Fusion}, we assign dynamic fusion weights w_m for each modality. The weights are dynamically calculated based on the interaction strength between modalities, as defined by:

$$w_m = \frac{\exp\left(\sum_{j\in M}\sum_{i=0}^{N_h}\beta_{mj}^{(i)} \Big/ \sqrt{|M|\times N_h}\right)}{\sum_{k\in M}\exp\left(\sum_{j\in M}\sum_{i=0}^{N_h}\beta_{kj}^{(i)} \Big/ \sqrt{|M|\times N_h}\right)}, \quad (12)$$

where M is the set of modalities, N_h is the number of attention heads, and $\beta_{mj}^{(i)}$ represents the interaction weight between modality m and j in the i-th attention head.

Finally, the fused representation is obtained by performing a weighted concatenation of each modality's unimodal representation h_m with its corresponding weight w_m, as shown below:

$$h_{\text{Fusion}} = \bigoplus_{m\in\{a,g,r,v,a+g,v+g\}} w_m \cdot h_m, \quad (13)$$

3.6 Optimization Objective

We employ contrastive learning to construct a loss function that ensures the representations of the same entity under different modalities are as close as possible in the vector space while enlarging the distance between different entities.We calculate the matching probability of entity pairs and design the loss function based on this probability.

Given an entity pair (e_i^1, e_i^2), where e_i^1 and e_i^2 represent the entity e_i under two different KG, we compute the matching probability of entity pair (e_i^1, e_i^2) as follows:

$$p_m(e_i^1, e_i^2) = \frac{\gamma_m(e_i^1, e_i^2)}{\gamma_m(e_i^1, e_i^2) + \sum_{e_j\in N_i^{neg}}\gamma_m(e_i^1, e_j^2)}, \quad (14)$$

where $\gamma_m(e_i^1, e_i^2) = \exp\left(h_i^{m^T} h_j^m/\tau\right)$ denotes the similarity measure between entities e_i^1 and e_i^2. N_i^{neg} represents the union of two negative sample sets (Sun et al., 2018; Chen et al., 2020b): N_i^{neg1}, which is the negative sample set from the source knowledge graph, containing all entities e_j^1 except for entity e_i^1; Similarly, N_i^{neg2} is the negative sample set from the target knowledge graph. This formulation allows us to measure the relative importance of entity pairs between positive and negative samples, thereby adaptively adjusting the model's focus on positive and negative samples.

To ensure matching consistency, i.e., the symmetry between $p_m(e_i^1, e_i^2)$ and $p_m(e_i^2, e_i^1)$, we take the average of the matching probabilities in both directions and using a logarithmic loss function. The specific loss function is defined as:

$$\mathcal{L}_m = -\log\left(\frac{p_m(e_i^1, e_i^2) + p_m(e_i^2, e_i^1)}{2}\right), \quad (15)$$

The goal of this loss function is to maximize the matching probability of positive sample pairs.

We need to consider not only the alignment loss of single modalities before cross-modal fusion but also the alignment loss of single modalities after cross-modal fusion, as well as the overall joint alignment loss. To this end, we compute the alignment loss of single-modality features before cross-modal fusion, $\mathcal{L}_{IE}$ (using the pre-fusion single-modality features h_m) (Lin et al., 2022), the alignment loss of multimodal features after cross-modal fusion, $\mathcal{L}_{RE}$ (using the post-fusion multimodal features $\hat{h_m}$), and the overall joint loss $\mathcal{L}_{Fusion}$ (for aligning multimodal features h_{Fusion}). For $\mathcal{L}_{RE}$, we do not compute the alignment loss for the modality guided by structure. We speculate that the Graph Attention Network (GAT) has already enhanced the structural representation of the features during the guiding stage, and further enforcing alignment may weaken the consistency by the Transformer layer (Zhou et al., 2021).

$$\mathcal{L}_{IE} = \sum_{m\in\{a,g,r,v,a+g,v+g\}} \mathcal{L}_m, \quad (16)$$

$$\mathcal{L}_{RE} = \sum_{m\in\{a,g,r,v\}} \hat{\mathcal{L}}_m, \quad (17)$$

where $\hat{\mathcal{L}}_m$ is a variant of $\mathcal{L}_m$, calculated using $\hat{\gamma}_m(e_i, e_j) = \exp\left(\hat{h}_i^{m^T}\hat{h}_j^m/\tau\right)$. Finally, our training objective is:

$$\mathcal{L} = \mathcal{L}_{Fusion} + \mathcal{L}_{IE} + \mathcal{L}_{RE} \quad (18)$$

4 Experiment

4.1 Experiment Setup

4.1.1 Datasets

We evaluate the performance of the model using three popular datasets, including the bilingual dataset DBP15K (ZH-EN, JA-EN, FR-EN) (Sun et al., 2017)and the monolingual datasets FB15K-DB15K and FB15K-YAGO15K (Liu et al., 2019). DBP15K contains around 400K triples and 15K aligned entity pairs, with 30% used as seed alignments. The monolingual datasets FB15K-DB15K and FB15K-YAGO15K cover different alignment ratios (20%, 50%, 80%). Additionally, we address the issue of missing images in our experiments by assigning a random vector sampled from a normal distribution to entities without images, where the distribution is parameterized by mean and standard deviation (Liu et al., 2021).

4.1.2 Iterative Training

We adopted a preparatory iterative training technique (Lin et al., 2022). Specifically, during each epoch ($K_e = 5$), we consider cross-KG entity pairs as mutual nearest neighbors in the vector space and add these pairs to the candidate list N^{cd}. Furthermore, if entity pairs remain as mutual nearest neighbors for consecutive K_s rounds ($K_s = 10$), they are included in the training set.

4.1.3 Baseline Methods

We use Hits@$N$and Mean Reciprocal Rank (MRR) to evaluate the performance of our model and the baseline methods. Hits@N (expressed as a percentage) represents the proportion of correctly aligned entities among the top N ranked candidates. MRR is the average of the reciprocal ranks of correctly aligned entities, where the reciprocal rank reports the rank of the correct entity alignment. Higher values for Hits@N and MRR indicate greater entity alignment accuracy.We selected the following baseline methods for comparison: MUGNN (Cao et al., 2019), AliNet (Sun et al., 2020b), BootEA (Sun et al., 2018), NAEA (Zhu et al., 2019), MMEA (Chen et al., 2020a), MSNEA (Chen et al., 2022a), MCLEA (Lin et al., 2022), MEAformer (Chen et al., 2023a), UMAEA (Chen et al., 2023b) and ACK-MMEA (Li et al., 2023).

4.1.4 Implementation Details

To ensure fairness and consistency in our experiments, all networks utilize a 300-dimensional hidden layer and are trained for 500 epochs (Chen et al., 2023a). We implement a cosine learning rate warm-up strategy with 15% warm-up, along with early stopping and gradient accumulation techniques. The optimizer used is AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the batch size is set to 3500. For the visual encoder (Chen et al., 2020a, 2023a), we follow the ResNet-152 architecture on DBP15K, with a visual dimension of $d_v = 2048$, and the VGG-16 (Simonyan and Zisserman, 2015) architecture on FBDB15K/FBYG15K, with

	DBP15K$_{ZH-EN}$			DBP15K$_{JA-EN}$			DBP15K$_{FR-EN}$		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
MUGNN (Cao et al., 2019)	.494	.844	.611	.501	.857	.621	.495	.870	.621
AliNet (Sun et al., 2020b)	.539	.826	.628	.549	.831	.645	.552	.852	.657
EVA (Liu et al., 2021)	.680	.910	.762	.673	.908	.757	.683	.923	.767
MSNEA (Chen et al., 2022a)	.601	.830	.684	.535	.775	.617	.543	.801	.630
MCLEA (Lin et al., 2022)	.715	.923	.788	.715	.909	.785	.711	.909	.782
MEAformer (Chen et al., 2023a)	.771	.951	.835	.764	.959	.834	.770	.961	.841
UMAEA (Chen et al., 2023b)	.800	.962	.860	.801	.967	.862	.818	.973	.877
SGMEA	**.852**	**.975**	**.899**	**.866**	**.979**	**.908**	**.882**	**.983**	**.920**

Table 1: Results without iteration on three bilingual datasets. The best results are marked with **bold**, and the second-best results are marked with underline.

	DBP15K$_{ZH-EN}$			DBP15K$_{JA-EN}$			DBP15K$_{FR-EN}$		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
BootEA (Sun et al., 2018)	.629	.847	.703	.622	.845	.701	.653	.874	.731
NAEA (Zhu et al., 2019)	.650	.867	.720	.641	.873	.718	.673	.894	.752
EVA (Liu et al., 2021)	.746	.910	.807	.741	.918	.805	.767	.939	.831
MSNEA (Chen et al., 2022a)	.643	.865	.719	.572	.832	.660	.583	.841	.671
MCLEA (Lin et al., 2022)	.811	.954	.865	.806	.953	.861	.811	.954	.865
MEAformer (Chen et al., 2023a)	.847	.970	.892	.842	.974	.892	.845	.976	.894
SGMEA	**.899**	**.984**	**.931**	**.901**	**.985**	**.933**	**.917**	**.990**	**.945**

Table 2: Results with iteration on three bilingual datasets.

$d_v = 4096$. A bag-of-words (BoW) model (Yang et al., 2019) is employed to encode relations into 1000-dimensional vectors, while pre-trained BERT is used to initialize attribute embeddings, with a dimension of 768. All experiments are conducted on an RTX 3090 GPU.

	Models	FB15K-DB15K			FB15K-YAGO15K		
		Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
20%	MMEA	.265	.541	.357	.234	.480	.317
	EVA	.199	.448	.283	.153	.361	.224
	MSNEA	.114	.296	.175	.103	.249	.153
	MCLEA	.295	.582	.393	.254	.484	.332
	ACK-MMEA	.304	.549	.387	.289	.496	.360
	MEAformer	.417	.715	.518	.327	.595	.417
	SGMEA	**.543**	**.777**	**.625**	**.587**	**.826**	**.670**
50%	MMEA	.417	.703	.512	.403	.645	.486
	EVA	.334	.589	.422	.311	.534	.388
	MSNEA	.288	.590	.388	.320	.589	.413
	MCLEA	.555	.784	.637	.501	.705	.574
	ACK-MMEA	.560	.736	.624	.535	.699	.593
	MEAformer	.619	.843	.698	.560	.778	.639
	SGMEA	**.716**	**.882**	**.775**	**.780**	**.924**	**.832**
80%	MMEA	.590	.869	.685	.598	.839	.682
	EVA	.484	.696	.563	.491	.692	.565
	MSNEA	.518	.779	.613	.531	.778	.620
	MCLEA	.735	.890	.790	.667	.824	.722
	ACK-MMEA	.682	.874	.752	.744	.676	.86
	MEAformer	.765	.916	.820	.703	.873	.766
	SGMEA	**.815**	**.931**	**.828**	**.857**	**.951**	**.894**

Table 3: The results on two monolingual datasets without iteration

4.2 Main Results

To ensure fair comparisons, we followed the approach of Chen et al. by excluding surface-level information interference(Chen et al., 2023b).

	Models	FB15K-DB15K			FB15K-YAGO15K		
		Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
20%	EVA	.231	.448	.318	.188	.403	.260
	MSNEA	.149	.392	.232	.138	.346	.210
	MCLEA	.395	.656	.487	.322	.546	.400
	MEAformer	.578	.812	.661	.444	.692	.529
	SGMEA	**.661**	**.847**	**.729**	**.750**	**.901**	**.805**
50%	EVA	.364	.606	.449	.325	.560	.404
	MSNEA	.358	.656	.459	.376	.646	.472
	MCLEA	.620	.832	.696	.563	.751	.631
	MEAformer	.690	.871	.755	.612	.808	.682
	SGMEA	**.752**	**.894**	**.802**	**.827**	**.938**	**.868**
80%	EVA	.491	.711	.573	.493	.695	.572
	MSNEA	.565	.810	.651	.593	.806	.668
	MCLEA	741	.900	.802	.681	.837	.737
	MEAformer	.784	**.921**	.834	.724	.880	.783
	SGMEA	**.828**	**.921**	**.861**	**.882**	**.967**	**.915**

Table 4: The results on two monolingual datasets with iteration

4.2.1 Non-Iterative Results

Under non-iterative training conditions, the results on the cross-lingual DBP15K dataset highlight the superior performance of our model. For example,in Table 1, on the DBP15K FR-EN dataset, our model achieved 88.2% Hits@1, outperforming the best baseline model UMAEA by 6.4%. Hits@10 and MRR were 98.3% and 0.920, respectively, leading across all metrics. These results fully demonstrate the significant advantages of our model in non-iterative training.

In monolingual tasks, such as in Table 3 the FB15K-DB15K and FB15K-YAGO15K datasets, our model also exhibited outstanding performance. Notably, Hits@1 with 20% of the training data surpassed the previous best baseline model by 26%. Furthermore, in many cases, our model using 20%

	DBP15K$_{ZH-EN}$			DBP15K$_{JA-EN}$			DBP15K$_{FR-EN}$		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
SGMEA	**.852**	.975	**.899**	**.866**	.979	**.908**	**.882**	.983	**.920**
w/o Guiding img and att	.834	.973	.886	.847	.975	.895	.862	.981	.907
w/o Guiding att	.841	.970	.889	.854	.975	.899	.875	.983	.916
w/o Guiding img	.850	**.977**	.898	.861	**.980**	.906	.877	**.985**	.917

Table 5: The ablation results on the DBP15K.

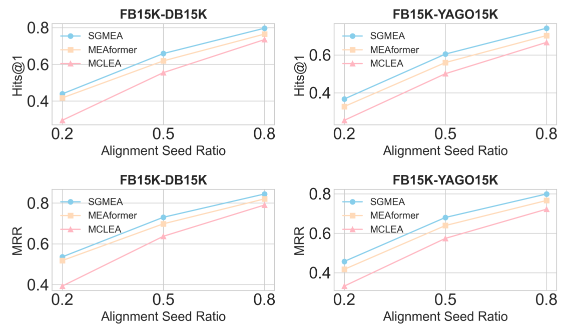

Figure 3: Attribute embedding uses the bag-of-words model on datasets FB15K-DB15K and FB15K-YAGO15K.

of the training data already outperformed the baseline models trained with 50% of the data.

4.2.2 Iterative Results

In the iterative training experiments, our model consistently demonstrated clear performance advantages across multiple datasets. As shown in Table 2, on the cross-lingual DBP15K datasets, our model excelled across all three language pairs (ZH-EN, JA-EN, FR-EN), significantly surpassing the best baseline model, MEAformer. In Table 4 On the monolingual FB15K-DB15K and FB15K-YAGO15K datasets, our model also performed exceptionally well across different training data ratios (20%, 50%, 80%). Particularly on the FB15K-YAGO15K dataset, with 20% of the training data, the model achieved 75% Hits@1, surpassing the MEAformer model by 30.6%.

4.3 Ablation Study

We conducted ablation experiments on the DBP15K dataset across three language pairs (ZH-EN, JA-EN, FR-EN). We removed the image guiding module, the attribute guiding module, and both modules together to analyze the impact of these components on the model's performance.The experimental results presented in Table 5.

First, we observe that whether image guidance or attribute guidance is added individually, the model's performance improves significantly compared to when no guidance is provided, with an average increase of 2%. This further demonstrates the importance of attribute values in multimodal entity alignment. While dual guidance from both images and attention enhances overall matching accuracy, in broader alignment metrics (such as Hits@10), attribute guidance may be more effective in some cases.

We speculate that although images provide high precision, they primarily rely on visual features. On the other hand, attribute information typically describes entities from multiple dimensions, covering a broader range of semantic features, thus helping the model match entities more effectively over a larger search space. The ablation study clearly shows that our proposed structural guidance modules play a crucial role in improving the performance of multimodal alignment tasks. These modules enable the model to better capture structural information across different languages, thereby enhancing matching accuracy.

4.4 Model Variants

To ensure a fair comparison and validate the superiority of our model, we developed a version of the model that uses a Bag-of-Words (BoW) representation for attribute embeddings, consistent with previous studies. We conducted experiments on the FB15K-DB15K and FB15K-YAGO15K datasets. The experimental results are shown in Figure 3, where we performed statistical analysis on Hits@1 and MRR under 20%, 50%, and 80% seed settings. The results demonstrate that our model architecture still achieves significant improvements across all metrics. Compared to the best-performing baseline model, MeaFormer, our model achieves average improvements of 4% and 0.04 in Hits@1 and MRR, respectively. Overall, the experimental results strongly demonstrate the effectiveness and robustness of our model, especially as it consistently outperforms across different seed ratios, further validating the effectiveness of the guiding theory.

5 Conclusion

This paper proposes a new method called SGMEA, which aims to address the issue of insufficient interaction between attribute and visual unimodal neighbors in multimodal entity alignment. SGMEA prioritizes the utilization of structural information from knowledge graphs to enhance the performance of the visual and attribute modalities. We conducted extensive experiments on several public datasets, and the results fully validate the effectiveness and soundness of SGMEA.

Limitation

In this study, while the proposed Structure-Guided Multimodal Entity Alignment (SGMEA) method achieves promising results in integrating structural information to enhance the performance of visual and attribute modalities, its over-reliance on structural information also reveals potential limitations. The structural information in knowledge graphs may suffer from insufficient heterogeneity, meaning that the structures of different knowledge graphs may not be completely consistent or may have partial omissions. This issue could lead to insufficiencies or inaccuracies during the alignment process. To address this problem, future research can draw inspiration from the concept of graph structure completion to further expand and refine the structural information in heterogeneous knowledge graphs, thereby improving the accuracy and robustness of entity alignment.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057), and Sponsored by CAAI MindSpore Open Fund, developed on OpenI Community.

References

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Weishan Cai, Wenjun Ma, Jieyu Zhan, and Yuncheng Jiang. 2022. Entity alignment with reliable path reasoning and relation-aware heterogeneous graph transformer. In *IJCAI*, pages 1930–1937. ijcai.org.

Yixin Cao, Zhiyuan Liu, Chengjiang Li, Juanzi Li, and Tat-Seng Chua. 2019. Multi-channel graph neural network for entity alignment. In *ACL (1)*, pages 1452–1461. Association for Computational Linguistics.

Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020a. MMEA: entity alignment for multi-modal knowledge graph. In *KSEM (1)*, volume 12274 of *Lecture Notes in Computer Science*, pages 134–147. Springer.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022a. Multi-modal siamese network for entity alignment. In *KDD*, pages 118–126. ACM.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI*, pages 1511–1517.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. 2021. Zero-shot visual question answering using knowledge graph. In *ISWC*, volume 12922 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z Pan, Wenting Song, et al. 2023a. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3317–3327.

Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z. Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023b. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *ISWC*, volume 14265 of *Lecture Notes in Computer Science*, pages 121–139. Springer.

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. 2022b. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. In *IJCKG*, pages 20–29. ACM.

Yunjun Gao, Xiaoze Liu, Junyang Wu, Tianyi Li, Pengfei Wang, and Lu Chen. 2022. Clusterea: Scalable entity alignment with stochastic training and normalized mini-batch similarities. In *KDD*, pages 421–431. ACM.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the*

8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), USA, February 2-7, 2018, pages 4832–4839. AAAI Press.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *NAACL*, pages 2284–2293.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4483–4491. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 6(2).

Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In *EMNLP*, pages 2723–2732.

Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. 2023. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In *WWW*, pages 2499–2508.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *COLING*, pages 2572–2584. International Committee on Computational Linguistics.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *AAAI*, pages 4257–4266. AAAI Press.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *The Semantic Web*, pages 459–474, Cham. Springer International Publishing.

Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, and Tat-Seng Chua. 2020. Exploring and evaluating attributes, values, and structures for entity alignment. In *EMNLP (1)*, pages 6355–6364. Association for Computational Linguistics.

Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021. Boosting the speed of entity alignment 10$\times$: Dual attention matching network with normalized hard sample mining. In *WWW*, pages 821–832.

Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020. MRAEA: an efficient and robust entity alignment approach for cross-lingual knowledge graph. In *WSDM*, pages 420–428.

Priya Radhakrishnan, Partha Talukdar, and Vasudeva Varma. 2018. Elden: Improved entity linking using densified knowledge graphs. In *NAACL*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Fabian M Suchanek, Serge Abiteboul, and Pierre Senellart. 2011. PARIS: probabilistic alignment of relations, instances, and schema. *PVLDB*, 5(3):157–168.

Zequn Sun, Muhao Chen, Wei Hu, Chengming Wang, Jian Dai, and Wei Zhang. 2020a. Knowledge association with hyperbolic knowledge graph embeddings. In *EMNLP*, pages 5704–5716.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *ISWC (1)*, volume 10587 of *Lecture Notes in Computer Science*, pages 628–644. Springer.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *IJCAI*, pages 4396–4402.

Zequn Sun, Jiacheng Huang, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *ISWC (1)*, volume 11778 of *Lecture Notes in Computer Science*, pages 612–629. Springer.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020b. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *AAAI*, pages 222–229. AAAI Press.

Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. 2019. Entity alignment between knowledge graphs using attribute embeddings. In *AAAI*, pages 297–304.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR (Poster)*. OpenReview.net.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective. MM '21, pages 2735–2743, New York, NY, USA. Association for Computing Machinery.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *EMNLP*, pages 349–357.

Derry Wijaya, Partha Pratim Talukdar, and Tom Mitchell. 2013. PIDGIN: ontology alignment using web text as interlingua. In *CIKM*, pages 589–598.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2020. Neighborhood matching network for entity alignment. In *ACL*, pages 6477–6487. Association for Computational Linguistics.

Kexuan Xin, Zequn Sun, Wen Hua, Wei Hu, and Xiaofang Zhou. 2022. Informed multi-context entity alignment. In *WSDM*, pages 1197–1205. ACM.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR (Poster)*.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *EMNLP/IJCNLP (1)*, pages 4430–4440. Association for Computational Linguistics.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view knowledge graph embedding for entity alignment. In *IJCAI*, pages 5429–5435. ijcai.org.

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. 2021. Deepvit: Towards deeper vision transformer. *CoRR*, abs/2103.11886.

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative entity alignment via joint knowledge embeddings. In *IJCAI*, pages 4258–4264.

Qiannan Zhu, Xiaofei Zhou, Jia Wu, Jianlong Tan, and Li Guo. 2019. Neighborhood-aware attentional representation for multilingual knowledge graphs. In *IJCAI*, pages 1943–1949. ijcai.org.