Leveraging Intra-modal and Inter-modal Interaction for Multi-Modal Entity Alignment

Zhiwei Hu[a,b], Víctor Gutiérrez-Basulto[c], Zhiliang Xiang[c], Ru Li[b,*], Jeff Z. Pan[d]

[a] *College of Information Science and Engineering, Shanxi Agricultural University, Jinzhong, 030801, Shanxi Province, China*
[b] *School of Computer and Information Technology, Shanxi University, Taiyuan, 030000, Shanxi Province, China*
[c] *School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 3AA, Wales, United Kingdom*
[d] *ILCC, School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, Scotland, United Kingdom*

Abstract

Multi-modal entity alignment (MMEA) aims to identify equivalent entity pairs across different multi-modal knowledge graphs (MMKGs). Existing approaches focus on how to better encode and aggregate information from different modalities. However, it is not trivial to leverage multi-modal knowledge in entity alignment due to the modal heterogeneity. In this paper, we propose a **M**ulti-Grained **I**nteraction framework for **M**ulti-Modal **E**ntity **A**lignment (**MIMEA**), which effectively realizes multi-granular interaction within the same modality or between different modalities. MIMEA is composed of four modules: i) a *Multi-modal Knowledge Embedding* module, which extracts modality-specific representations with multiple individual encoders; ii) a *Probability-guided Modal Fusion* module, which employs a probability guided approach to integrate uni-modal representations into joint-modal embeddings, while considering the interaction between uni-modal representations; iii) an *Optimal Transport Modal Alignment* module, which introduces an optimal transport mechanism to encourage the interaction between uni-modal and joint-modal embeddings; iv) a *Modal-adaptive Contrastive*

[*] Email: Zhiwei Hu (zhiweihu@whu.edu.cn), Víctor Gutiérrez-Basulto (gutierrezbasultov@cardiff.ac.uk), Zhiliang Xiang (xiangz6@cardiff.ac.uk), Ru Li (liru@sxu.edu.cn) and Jeff Z. Pan (j.z.pan@ed.ac.uk)

Learning module, which distinguishes the embeddings of equivalent entities from those of non-equivalent ones, for each modality. Extensive experiments conducted on two real-world datasets demonstrate the strong performance of MIMEA compared to the SoTA. Datasets and code are available at the following website: https://github.com/zhiweihu1103/MEA-MIMEA.

Keywords: Multi-Modal Knowledge Graph, Multi-Modal Entity Alignment, Knowledge Graph

1. Introduction

Knowledge graphs (KGs), such as DBpedia [1] and YAGO [2], employ a graph structure to organize real-world factual knowledge. They provide the backbone of various web-based applications like query answering [3, 4, 5] and search [6, 7]. Recently, several works have extended KGs with additional modeling capabilities, as required by different applications. *Multi-modal Knowledge Graphs (MMKGs)* extend traditional KGs with multi-modal information, *e.g.,* visual information. However, like traditional KGs, MMKGs suffer from incompleteness and low coverage [8]. Thus, the integration of independently developed MMKGs is paramount. A key task for MMKG integration is *multi-modal entity alignment (MMEA)*, which aims to identify equivalent entity pairs in different MMKGs by taking into account the structure of MMKGs, as well as the attribute and visual information of entities, see Figure 1. In this way, MMEA facilitates the exchange of knowledge among different MMKGs. MMEA shows strong potential in industrial intelligence scenarios, such as intelligent manufacturing and defect detection. In these settings, entities often correspond to real-world industrial objects (*e.g.,* components, products, or defects) that are described by heterogeneous information sources, including structural relations, sensor attributes, textual specifications, and visual inspection data. Aligning such multi-modal entities across independently developed industrial knowledge graphs can facilitate knowledge reuse, cross-system interoperability, and data-driven decision making [9, 10].

A wide variety of approaches to MMEA have been already introduced. Initial proposals [11, 12, 13, 14, 15] concentrated on the construction of distinct multi-modal fusion modules to integrate entity representations from multiple modalities into joint embeddings and then use aggregated embeddings to predict alignments. A shortcoming of these methods is that they only explore the use of diverse multi-modal representations to enhance the contextual embedding of entities, overlooking the capabilities of inter-modal representations to capture certain types of interactions. To overcome this, some works [16, 17, 18] use siamese networks, transformer mechanisms or contrastive learning strategies to enhance multi-modal knowledge by exploiting inter-modal interaction. However, existing frameworks for MMEA still suffer from serious shortcomings:

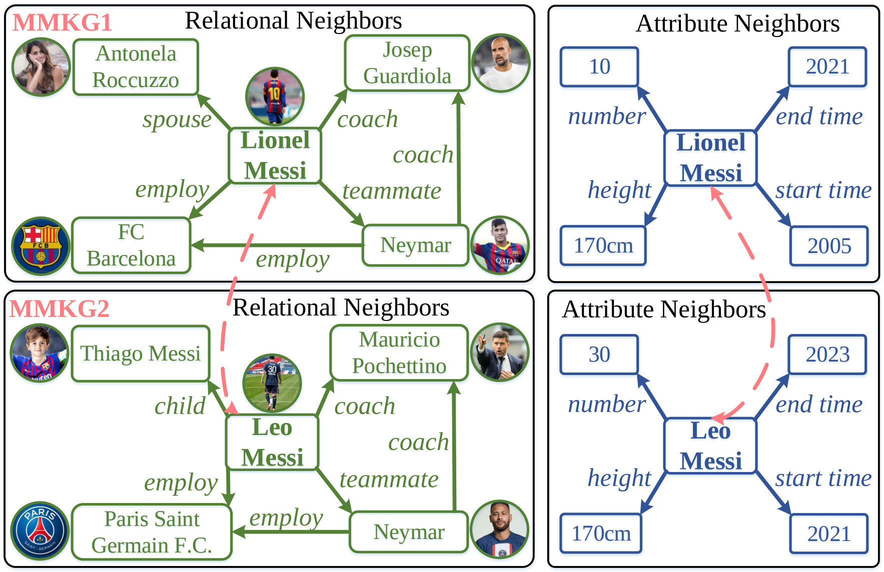


Figure 1: The MMEA task between MMKG1 and MMKG2, aligning the entities *Lionel Messi* and *Leo Messi*.

1. **Modality Distinctiveness.** Existing methods have difficulties to explicitly distinguish the importance of each modality. In fact, among all modalities, the structural modal knowledge is the most prevalent. For instance, the FB15K-DB15K dataset has a total of 714,720 structural triples, while it only has 1,624 relation categories, 341 attribute categories, and 26,281 images. Whether we look at the provided data ratios or the results of ablation experiments, it is evident that the structural modality provides a richer source of knowledge and therefore deserves more attention.

2. **Modality Interaction Diversity.** Existing models place more emphasis on the interaction between uni-modal embeddings while overlooking interactions between uni-modal and joint-modal embeddings, leading to a

lack of diversity in modality interactions. We advocate that, in practice, it is necessary to design mechanisms that better capture the interaction between uni-modal and joint-modal embeddings to fully harness the potential of all available modalities. Indeed, the interaction between the joint-modal and uni-modal representations enables simultaneous interactions with more than two modalities, covering the information gaps left by only looking at pairwise interactions.

To address the above two shortcomings, we propose the method **MIMEA**, a **M**ulti-Grained **I**nteraction framework for **M**ulti-Modal **E**ntity **A**lignment. Specifically, MIMEA includes the following four modules. The *Multi-modal Knowledge Embedding* module utilizes multiple individual encoders to obtain modality-specific representations for each entity. To effectively combine multi-modal knowledge, the *Probability-guided Modal Fusion* module takes structural knowledge as the core, and employs a probability distribution mechanism to integrate uni-modal information into joint-modal representations. Furthermore, we introduce an *Optimal Transport Modal Alignment* module to capture the interaction between uni-modal and joint-modal embeddings. The integration of the *Probability-guided Modal Fusion* and the *Optimal Transport Modal Alignment* modules realizes inter-modal interactions between uni-modal and joint-modal embeddings. Moreover, we introduce an intra-modal contrastive loss to distinguish the embeddings of equivalent entities from those of non-equivalent ones, for each modality. In summary, our main contributions are:

- We propose a framework to address the multi-modal entity alignment task by introducing multi-grained interaction mechanisms into the multi-modal knowledge representation process.

- We design mechanisms to explore intra-modal relationships and inter-modal interactions, ensuring that the aligned entities are semantically close.

- We conduct extensive experiments on two real-world datasets, showing the strong performance of MIMEA.

2. Related Work

2.1. Entity Alignment

Entity alignment (EA), which aims to identify equivalent entities across different knowledge graphs, is a fundamental data integration task. Existing research focuses on learning embeddings of entities by utilizing the structural information of KGs. Approaches to EA can be divided into two categories: *KGE-based methods* and *GNN-based methods*. KGE-based methods 'move' entity embeddings from different KGs into a unified latent space and measure the alignment by calculating the distance between entity embeddings, such as MTransE [19], JAPE [20], IPTransE [21], BootEA [22], RNM [23] and NeoEA [24]. Recently, GNN-based models have achieved remarkable performance in graph learning. Based on this, some works develop GNN-based frameworks for EA, such as KDCoE [25], AliNet [26], MuGNN [27], AttrGNN [28]. However, all the discussed methods ignore the multi-modal knowledge (especially the visual information) available in the knowledge graph.

2.2. Multi-Modal Entity Alignment

Recently various multimodal knowledge graphs have become available [13, 11]. Thus many works have investigated how to effectively incorporate visual knowledge into the entity alignment task. PoE [11] combines all multi-modal features into a single vector, and measures the trustworthiness of entity pairs by matching their underlying semantics. However, it cannot capture the potential interactions among different modalities. MMEA [12] integrates knowledge from different modalities into a joint representation and then calculates a similarity score between the holistic embeddings of aligned entities. EVA [13] introduces an iterative learning strategy to expand the set of training seeds. HMEA [14] encodes the multi-modal knowledge into the hyperbolic space, and uses aggregated embeddings to predict alignments. MSNEA [16] integrates visual features to guide the learning process of relation features and adaptively assigns attention weights to capture valuable attributes for alignment. MCLEA [18] explores intra-modal and inter-modal interactions via contrastive learning to reduce the gap between modalities. MEAformer [17] proposes a transformer-based model which can dynamically predict relativized mutual weights among modalities for each entity, encouraging the emergence of adaptive modality preferences. ACK-MMEA [29] designs a multi-modal attribute uniformization module to incorporate the consistent alignment knowledge. GEEA [30] studies embedding-based entity alignment from a perspective of generative models. It converts an entity from one knowledge graph to the other one, and generates new entities from random noise vectors. However, the aforementioned methods have the following two shortcomings: On one hand, the majority of methods, such as MMEA, EVA, HMEA, and MSNEA, have not been able to fully achieve multi-granular interactions within and across modalities. Consequently, they do not effectively integrate multimodal knowledge related to entities. On the other hand, even when some methods, like MCLEA and MEAformer, introduce mechanisms for intra-modality and inter-modality interactions, they are difficult to explicitly distinguish the importance of each modality, and also ignore the interaction between uni-modal and joint-modal embeddings, which results in a lack of diversity in modal interactions.

2.3. Optimal Transport

Optimal transport (OT) is a fundamental mathematical tool which aims to derive an optimal plan to transfer one distribution to another. OT has been used in many applications, such as, computer vision [31, 32], domain adaption [33, 34], and unsupervised learning [35, 36]. OTKGE [37] models the multi-modal fusion procedure as a transport plan moving different modal embeddings to a unified space by minimizing the Wasserstein distance between multi-modal distributions. MOTCat [38] proposes a multi-modal optimal transport-based co-attention transformer framework with global structure consistency for selecting informative patches. However, existing studies lack a comprehensive investigation of the correlations between uni-modal and joint-modal contexts. To the best of our knowledge, we are the first to adopt the optimal transport mechanism for MMEA task.

3. Preliminaries

Multi-modal Knowledge Graph. Let $\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{I}, \mathcal{V}$ respectively be finite sets of entities, relation types, attribute types, images, and values. A *multi-modal knowledge graph (MMKG)* $\mathcal{G}$ is defined as $\{\mathcal{P}, \mathcal{T}_r, \mathcal{T}_a\}$, where $\mathcal{P} = \{(e, i) \mid e \in \mathcal{E}, i \in \mathcal{I}\}$ is the set of *entity-image* pairs, $\mathcal{T}_r = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of *relational triples*, and $\mathcal{T}_a = \{(e, a, v) \mid e \in \mathcal{E}, a \in \mathcal{A}, v \in \mathcal{V}\}$ is the set of *attribute triples*.

Multi-modal Entity Alignment. The aim of the *multi-modal entity alignment (MMEA) task* is to identify pairs of entities in two multi-modal knowledge graphs which are equivalent. Concretely, given two MMKGs $\mathcal{G} = \{\mathcal{P}, \mathcal{T}_r, \mathcal{T}_a\}$ and $\mathcal{G}' = \{\mathcal{P}', \mathcal{T}_r', \mathcal{T}_a'\}$, we aim to find entity pairs $\mathcal{H} = \{(e_i, e_j) \mid e_i \in \mathcal{E}, e_j \in \mathcal{E}', e_i \equiv e_j\}$, where $\equiv$ represents the equivalence of two entities. Usually, we will select a small set of pre-aligned entity pairs $\mathcal{S}$ (seeds) for training, to learn entity representations in the two input MMKGs.

4. Framework

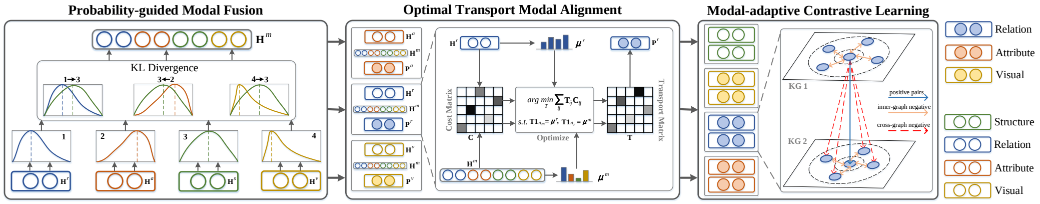


Figure 2: MIMEA's architecture, containing the modules: Probability-guided Modal Fusion, Optimal Transport Modal Alignment, and Modal-adaptive Contrastive Learning.

We now introduce the **MIMEA** framework (cf. Fig 2 for its architecture), which comprises four major components (cf. Sections 4.1-4.4).

4.1. Multi-modal Knowledge Embedding

We define entity embeddings for four modalities: structural, relation, attribute and visual. Structural embeddings are obtained based on the attribute and relational neighbors (described by attribute/relational triples) of an entity. Relation embeddings are derived from relation types, and they are expressed in the form of bag-of-words. Attribute embeddings are obtained analogously. Visual embeddings are derived from entity-image pairs.

Structural Embeddings. The graph attention network (GAT) [39] is an attention-based architecture which has been shown to effectively encode graph-like data. We thus leverage GAT to model the structural information of $\mathcal{G}$ and $\mathcal{G}'$. For the hidden state $\boldsymbol{h}_i \in \mathbb{R}^d$ (d represents the embedding dimension) of entity e_i, the aggregation of its one-hop neighbors $\mathcal{N}_i$ with self-loops is formulated as:

$$\boldsymbol{h}_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}_s \boldsymbol{h}_j\right) \tag{1}$$

where $\sigma(\cdot)$ denotes the nonlinear ReLU function; $\mathbf{W}_s \in \mathbb{R}^{d\times d}$ denotes a parameterized weight matrix [40, 18] — we restrict $\mathbf{W}_s$ to a diagonal matrix to reduce the number of computations; $\boldsymbol{h}_j$ is the hidden state of entity e_j; the attention weight α_{ij} measures the importance of entity e_j for entity e_i, formulated as:

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top[\mathbf{W}_s\boldsymbol{h}_i \parallel \mathbf{W}_s\boldsymbol{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^\top[\mathbf{W}_s\boldsymbol{h}_i \parallel \mathbf{W}_s\boldsymbol{h}_k]\right)\right)} \tag{2}$$

where $\mathbf{a} \in \mathbb{R}^{2d}$ is a learnable parameter, $\cdot^\top$ and $\parallel$ respectively represent the transposition and concatenation operations. To stabilize the learning process of self-attention, we introduce a multi-head strategy [40, 18, 17] to generate K independent representations based on the transformation of Equation 1. Then, we concatenate these features to obtain the structural embedding $\boldsymbol{h}_i^s$ of entity e_i as:

$$\boldsymbol{h}_i^s = \overset{K}{\underset{k=1}{\Big\|}} \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k\mathbf{W}_s^k\boldsymbol{h}_j\right) \tag{3}$$

where α_{ij}^k denotes the normalized attention coefficients computed by the k-th attention mechanism, and $\mathbf{W}_s^k$ is the corresponding input linear transformation's weight matrix. We use a two-layer GAT to aggregate the neighborhood information across multiple hops, and use the output of the final GAT layer as the structural embedding. The structural embedding of all entities is represented as $\mathbf{H}^s \in \mathbb{R}^{n\times d}$, where n represents the number of entities in the input dataset.

Relation and Attribute Embeddings. Note that the knowledge from attribute types is coarser than that of relational types. Thus, directly mixing the representations of relations and attributes using a GAT can easily lead to the problem of information contamination [13]. To alleviate this issue, we respectively regard the relations and attributes of entity e_i as bag-of-words features w_i^r and w_i^a. We further apply the multi-layer perceptrons MLP_r and MLP_a to respectively obtain the relation embedding $\boldsymbol{h}_i^r$ and attribute embedding $\boldsymbol{h}_i^a$, calculated as:

$$\boldsymbol{h}_i^r = \mathrm{MLP}_r(w_i^r), \quad \boldsymbol{h}_i^a = \mathrm{MLP}_a(w_i^a) \tag{4}$$

The relation and attribute embedding of all entities are respectively represented as $\mathbf{H}^r \in \mathbb{R}^{n\times d}$ and $\mathbf{H}^a \in \mathbb{R}^{n\times d}$.

Visual Embeddings. VGG [41] are usually pre-trained on large-scale image datasets and can extract useful features from images that are beneficial to different visual tasks. In practice, we feed the image v_i of entity e_i into the VGG-16 encoder Enc_v. We use the final layer output before logits as the visual feature, and finally apply a multi-layer perceptron MLP_v to obtain the visual embedding $\boldsymbol{h}_i^v$:

$$\boldsymbol{h}_i^v = \mathrm{MLP}_v(\mathrm{Enc}_v(v_i)) \tag{5}$$

The visual embedding of all entities is represented as $\mathbf{H}^v \in \mathbb{R}^{n \times d}$.

It is worth noting that the adoption of Bag-of-Words representations combined with lightweight MLPs for relation and attribute encoding is a deliberate design choice rather than a limitation of the proposed MIMEA framework. This setting follows common practice in recent entity alignment methods (*e.g.,* MCLEA [18] and MEAformer [17]), which employ similar encoders for structured textual modalities. Such a design enables a fair and controlled comparison by minimizing the influence of encoder capacity and allowing the effectiveness of the proposed multi-modal interaction mechanism to be isolated. Importantly, MIMEA itself is agnostic to the specific choice of encoders and can naturally accommodate more expressive textual and visual representations, such as contextualized language models (*i.e.,* BERT) or stronger visual backbones (*i.e.,* ViT-B/16), without requiring any architectural modification.

4.2. Probability-guided Modal Fusion

Different modalities concentrate on different types of knowledge. Thus, each modality contributes differently to the characterization of specific aspects of an entity. Typically, it is required to combine multiple modalities of knowledge to provide a more comprehensive understanding of an entity. For example, knowledge about the entity *Lionel Messi* includes the relational triple (*Lionel Messi, employ, FC Barcelona*) and a visual image (an image of Messi wearing a certain team's jersey). So, when evaluating the football club *Lionel Messi* plays for, the structural knowledge from the relational triple is more relevant than the knowledge from the image. However, when it comes to Messi's jersey number at a club, the triple (*Lionel Messi, employ, FC Barcelona*) does not contain relevant information, but a visual image of Messi wearing a *10* jersey can provide more appropriate clues. Holistically combining these two types of information will thus enable an accurate representation of the football club *Lionel Messi* plays for and his jersey number at that club. Therefore, an important challenge is *how to better integrate multi-modal knowledge to obtain effective fused representations in multi-modal contexts.*

A key source of knowledge in multi-modal knowledge graphs is the one provided by structural triples. The structural triples contain the relational triples and attribute triples, they can provide a more direct representation of the content of an entity and its relationship with other entities. For example, in the FB15K-DB15K dataset, there are a total of 714,720 structural triples, resulting in richer knowledge about the connections among entities. In contrast, the FB15K-DB15K dataset contains only 1,624 relational types and 341 attribute types, which means that the initialization vectors for the relation and attribute modalities will be bag-of-words vectors of length 1,624 and 341. Consequently, the representation of relation and attribute modalities of an entity lacks sufficient distinctiveness. Indeed, in subsequent ablation experiments we will show that the structural content has the most significant impact on the final performance of entity alignment. Therefore, using structural embeddings as a pivotal point, we introduce the *Probability-guided Modal Fusion (PMF)* module, which employs a probabilistic distribution to achieve initial interactions between relation embeddings and structural embeddings, attribute embeddings and structural embeddings, also visual embeddings and structural embeddings. It generates interactive weights in the first stage and aggregates different modal embeddings to obtain a joint-modal combined representation based on these weight coefficients. Specifically, the PMF module comprises the following three steps:

1. *Constructing Probability Distributions.* Given the structural embedding $\mathbf{H}^s$, relation embedding $\mathbf{H}^r$, attribute embedding $\mathbf{H}^a$, and visual embedding $\mathbf{H}^v$ of all entities, we represent each embedding using a probability density form based on the Beta probability distribution function. The Beta distribution has two shape hyperparameters α and β. Its *probability density function (PDF)* is defined as: $f_{(\alpha,\beta)}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathbf{B}(\alpha,\beta)}$, where $x \in [0,1]$ and $\mathbf{B}(\cdot)$ denotes the Beta function. To transform, for example, the structural embedding $\mathbf{H}^s$ into a Beta distribution, we proceed as follows: i) we first divide $\mathbf{H}^s$ into two equal parts $\boldsymbol{\alpha}^s$ and $\boldsymbol{\beta}^s$ according to the embedding dimension: $\mathbf{H}^s = \{[\boldsymbol{\alpha}^s, \boldsymbol{\beta}^s] \,|\, \boldsymbol{\alpha}^s \in \mathbb{R}^{n\times m}, \boldsymbol{\beta}^s \in \mathbb{R}^{n\times m}, m = \frac{d}{2}\}$. Then, we use each part as a shape parameter of the Beta distribution. ii) By combining the i-th element α_i^s in $\boldsymbol{\alpha}^s$ with the i-th element β_i^s in $\boldsymbol{\beta}^s$, we

will form the i-th Beta distribution. The combination of all elements will form m Beta distributions, represented as $\mathcal{D}^s = [(\alpha_1^s, \beta_1^s), \ldots, (\alpha_m^s, \beta_m^s)]$. We denote the PDF of the i-th Beta distribution in $\mathcal{D}^s$ as p_i^s. iii) We can analogously get the Beta distributions of the relation, attribute and visual embedding: $\mathcal{D}^r$, $\mathcal{D}^a$, $\mathcal{D}^v$, and the corresponding i-th Beta distributions: p_i^r, p_i^a, and p_i^v.

2. *Calculating Modal Weight Coefficients.* Given the relation and structural embedding's Beta distributions $\mathcal{D}^r$ with parameters $[(\alpha_1^r, \beta_1^r), ..., (\alpha_m^r, \beta_m^r)]$ and $\mathcal{D}^s$ with corresponding parameters $[(\alpha_1^s, \beta_1^s), ..., (\alpha_m^s, \beta_m^s)]$, we define the distance between the relation and structural embedding as the sum of the KL divergence between the two Beta distributions along each dimension:
$$\delta_{(r,s)} = \sum_{i=1}^{m} \mathrm{KL}(p_i^r, p_i^s) \tag{6}$$
Then, we convert the KL distance $\delta_{(r,s)}$ to a weight coefficient based on $w_{(r,s)} = \lambda(2 - \delta_{(r,s)})^2$, where λ represents the incremental rate, set empirically. Using the same method, we can obtain the weight coefficient $w_{(a,s)}$ between the attribute and structural embedding, and the weight coefficient $w_{(v,s)}$ between the visual and structural embedding.

3. *Fusing Different Modal Embeddings.* We add the three weight coefficients $w_{(r,s)}$, $w_{(a,s)}$, $w_{(v,s)}$ with the initialized value 1.0 (initially we assume that all modalities have the same weight coefficients) and normalize it to obtain the prior assumption: $W_{\mathrm{PMF}} = \mathit{softmax}([w_{(r,s)}, w_{(a,s)}, w_{(v,s)}] + 1.0)$. Then, we multiply these weight coefficients with the embedding representation of the corresponding modality, and concate the multiplied results to obtain the final fused modality representation $\mathbf{H}^m \in \mathbb{R}^{n \times 4d}$.

4.3. Optimal Transport Modal Alignment

The PMF deals with various modalities of knowledge by combining multimodal information. However, due to the introduced noise during the fusion process, an optimal representation of an entity cannot be based only on joint-modal information. For instance, if we are interested in the football club in which *Lionel Messi* plays, given the joint-modal embedding incorporating the relational triple (*Lionel Messi, employ, FC Barcelona*), the attribute triple (*Lionel Messi, number, 10*) and the visual image of Messi wearing the Barcelona jersey number 10, the knowledge provided by the attribute triple is regarded as noise, while the one provided by the relational triple is useful information. Therefore, in some cases, while using joint-modal embeddings, we need to retain the knowledge of individual embeddings for each modality to assess the extent to which a single modality represents an entity in a certain context. So, a natural question is *how to achieve better interaction between uni-modal and joint-modal embeddings to cover the information gap of single modalities and reduce the noise of joint-modal embeddings?*

Optimal transport (OT) aims to transport the density distribution of a group of elements to that of another group with minimal total cost. To consider the correlations of uni-modal and joint-modal representations, we can regard uni-modal as one group elements and joint-modal as another group elements. The expectation is that the uni-modal and joint-modal elements have an appropriate correlation with minimal total transportation cost. To achieve this, we first generate an intermediate transition matrix by aligning and optimizing uni-modal and joint-modal embeddings. Subsequently, by combining uni-modal information with the generated intermediate transition matrix, we obtain an enriched uni-modal embedding. The *Optimal Transport Modal Alignment (OTMA)* module consists of the following steps:

1. *Building the Transport Task.* We look, for instance, at how to obtain the intermediate modal embedding between the relation modal embedding $\mathbf{H}^r$ and the joint modal embedding $\mathbf{H}^m$. Optimal transport aims at computing a minimal cost transportation between a source distribution μ^r and a target distribution μ^m:

$$\mu^r = \sum_{i=1}^{n_r} q_i^r \varphi(x_i), \;\; \mu^m = \sum_{i=1}^{n_m} q_i^m \varphi(y_i) \tag{7}$$

   where μ^r and μ^m are defined on the probability space $\mathbf{H}^r$ and $\mathbf{H}^m$, $\varphi(\cdot)$ denotes the Dirac function, n_r and n_m are the number of samples, x_i and y_i are the i-th sample of $\mathbf{H}^r$ and $\mathbf{H}^m$ (in practice, to reduce the computational complexity, the number of selected samples will be lower than the embedding dimension), q_i^r and q_i^m are the probability mass of the i-th samples, satisfying the following conditions: $\sum_{i=1}^{n_r} q_i^r = \sum_{i=1}^{n_m} q_i^m = 1$, to simplify the calculations, we set $q_i^r = \frac{1}{n_r}$ and $q_i^m = \frac{1}{n_m}$. We define a cost matrix $\mathbf{C}$ with $\mathbf{C}_{ij}$ representing the distance (usually the cosine distance) between x_i and y_j.

2. *Optimal Transport Plan.* Based on distributions μ^r and μ^m, we can obtain all joint probability distributions $\Pi(\mu^r, \mu^m)$. Combining them with the

cost matrix $\mathbf{C}$, we can convert the optimal transport into the following form:

$$\mathcal{W}(\mu^r, \mu^m) = \underset{\mathbf{T}\in\Pi(\mu^r,\mu^m)}{min} \sum_{i=1}^{n_r}\sum_{j=1}^{n_m} \mathbf{T}_{ij}\mathbf{C}_{ij} \tag{8}$$

where $\Pi(\mu^r, \mu^m) = \{\mathbf{T} \in \mathbb{R}^{n_r \times n_m} | \mathbf{T}\mathbf{1}_{n_m} = \mu^r, \mathbf{T}^\top \mathbf{1}_{n_r} = \mu^m\}$, with $\mathbf{1}$ an all-one vector, $\mathbf{T}_{ij}$ the optimal amount of mass to move from x_i to y_j to obtain an overall minimum cost. We apply the Sinkhorn algorithm [42] to optimize Equation (8) to get the optimal transportation matrix $\mathbf{T}$.

3. *Translating Uni-Modal Embeddings.* We multiply the relation embedding $\mathbf{H}^r$ with the transportation matrix $\mathbf{T}$ to get the intermediate embedding $\mathbf{P}^r = \mathbf{H}^{r\top}\mathbf{T}$ between the relation-modal embedding $\mathbf{H}^r$ and the joint-modal embedding $\mathbf{H}^m$. The resulting embedding focuses on relational modal knowledge, but can also be aligned with joint modal embeddings at a small cost.

We can analogously obtain attribute and visual intermediate embeddings, denoted as $\mathbf{P}^a$ and $\mathbf{P}^v$, respectively. We found that there is no need to align the structural-modal embedding with the joint-modal embedding since the structural embedding in the joint embedding has the largest weight and therefore dominates the joint embedding.

4.4. Modal-adaptive Contrastive Learning

The OTMA module focuses on the interaction between uni-modal and joint-modal aspects of knowledge. However, both the OTMA and PMF modules overlook the interactions within a single modality. In many cases, for a given entity, there exist multiple associated pieces of information within a single modality. When predicting a specific attribute of an entity, typically only a subset of these related pieces of knowledge plays a decisive role. For instance, consider the entity *Lionel Messi*, which includes the relational triples: (*Lionel Messi, spouse, Antonela Roccuzzo*) and (*Lionel Messi, child, Thiago Messi*) related to family relationships and (*Lionel Messi, teammate, Neymar*) and (*Lionel Messi, coach, Josep Guardiola*) related to player attributes. Clearly, when describing Messi's family relationships, *Antonela Roccuzzo* and *Thiago Messi* are more important than *Neymar* and *Josep Guardiola*. However, when discussing Messi's football career, the situation is reversed. Therefore, it is preferable to make the embeddings of *Antonela*

Roccuzzo and *Thiago Messi* closer in the embedding space, while the embeddings of *Antonela Roccuzzo* and *Neymar* should be pushed farther apart. Based on these observations, an important challenge is *how to enforce embeddings to respect modal properties, while distinguishing the embedding of an entity from those of other entities, for each modality.*

Inspired by the contrastive learning mechanism [43, 18, 44], we devise a *Modal-adaptive Contrastive Learning (MCL)* module, which maps inner-graph aligned pairs to a proximate location, but also pushes the inner-graph and cross-graph unaligned pairs father apart. Specifically, MCL includes the following three parts:

- *Creating Positive and Negative Samples.* Following a 1-to-1 alignment constraint [18, 17], the entity pairs within the seed alignments $\mathcal{S}$ can be naturally regarded as positive samples, whereas any non-aligned pairs can be regarded as negative samples. Let (e_i^1, e_i^2) in $\mathcal{S}$ (with $e_i^1 \in \mathcal{G}$ and $e_i^2 \in \mathcal{G}'$) be the i-th aligned entity pair, the negative samples of e_i^1 are obtained from two sources: the inner-graph unaligned pairs from $\mathcal{G}$ and cross-graph unaligned pairs from $\mathcal{G}'$. More precisely, they are defined as $\mathcal{N}_i^1 = \{e_j^1 \mid \forall e_j^1 \in \mathcal{G}, j \neq i\}$ and $\mathcal{N}_i^2 = \{e_j^2 \mid \forall e_j^2 \in \mathcal{G}', j \neq i\}$. It should be noted that we use the in-batch negative sampling strategy [18, 17] to limit the negative sample scope within the mini-batch.

- *Contrastive Learning Loss.* For the constructed positive and negative examples, we perform contrastive learning under each modal condition. For instance, for the relational modality, we construct the contrastive learning loss $\mathcal{L}^r(e_i^1, e_i^2)$ of the positive pair (e_i^1, e_i^2) as:

$$-\log \frac{\theta(e_i^1, e_i^2)}{\theta(e_i^1, e_i^2) + \gamma \sum_{e_j^1 \in \mathcal{N}_i^1} \theta(e_i^1, e_j^1) + \sum_{e_j^2 \in \mathcal{N}_i^2} \theta(e_i^1, e_j^2)} \tag{9}$$

  where $\theta(x, y) = exp(f_r(x)^\top f_r(y)/\tau)$, $f_r(\cdot)$ is the relation encoder, τ is a temperature parameter, and γ is a hyper-parameter to control inner-graph alignment. The second and third terms in the denominator sum up inner-graph and cross-graph negative samples, respectively. We apply L2-normalisation to the input feature embeddings before computing the inner product [43, 18, 45]. Similarly, we can obtain the loss for the other direction as $\mathcal{L}^r(e_i^2, e_i^1)$. The final contrastive loss of the relational modality is the average of the losses in the two directions, expressed as: $\mathcal{L}^r = \frac{1}{2}[\mathcal{L}^r(e_i^1, e_i^2) + \mathcal{L}^r(e_i^2, e_i^1)]$.

- *Optimization Objective.* Using the same idea, we can obtain the contrastive loss of structural, attribute, visual and joint modalities, respectively expressed as $\mathcal{L}^s$, $\mathcal{L}^a$, $\mathcal{L}^v$, and $\mathcal{L}^m$. The overall loss is defined as:

  $$\mathcal{L} = \sum_{\ell \in \mathcal{M}} \phi^\ell \mathcal{L}^\ell,\ \mathcal{M} = \{s, r, a, v, m\} \tag{10}$$

  where ϕ^ℓ is the hyper-parameter that balances the importance of different modal losses. Similar to [46], we introduce a multi-task learning paradigm and then use homoscedastic uncertainty to weight each loss automatically during model training. Details of this strategy can be found in [46]. It should be noted that only the MCL module has loss values, and the PMF and OTMA modules do not have any loss content.

5. Experiments

To evaluate the effectiveness of MIMEA, we aim to explore the following research questions:

- **RQ1 (Effectiveness):** How does MIMEA perform compared to the SoTA?
- **RQ2 (Ablation studies):** How do different components of MIMEA contribute to its performance?
- **RQ3 (Complexity analysis):** What is the amount of computation and parameters used by MIMEA?
- **RQ4 (Parameter analysis):** How do hyper-parameters influence the performance of MIMEA?

Table 1: Statistics for different datasets. #Ent.: the number of entities, #Rel.: the number of relations, #Att.: the number of attribute, #Rel tri.: the number of relation triples, #Attr tri.: the number of attribute triples, #Image: the number of visual images.

Datasets	#Ent.	#Rel.	#Att.	#Rel tri.	#Attr tri.	#Image
FB15K	14,951	1,345	116	592,213	29,395	13,444
DB15K	12,842	279	225	89,197	48,080	12,837
YAGO15K	15,404	32	7	122,886	23,532	11,194

5.1. Experimental Setup

Datasets. We conduct experiments on two well known multimodal datasets, following previous studies [18, 17, 29, 30]: FB15K-DB15K and FB15K-YAGO15K, where FB15K, DB15K, and YAGO15K are extracted from the Freebase [47], DBpedia [48], and YAGO [49] knowledge bases, respectively. Note that a set of pre-aligned entity pairs are offered for guidance, according to the seed alignment ratio (20%, 50%, 80%). The FB15K-DB15K and FB15K-YAGO15K are proportionally split into training set and testing set. The statistics corresponding to FB15K, DB15K, and YAGO15K are shown in Table 1.

Hyperparameters. All experiments are carried out on a 32G Tesla V100 GPU, we use the AdamW as the optimizer to update the parameters. We fine-tune the hyper-parameters including the embedding size of modalities $d \in \{100, 200, 300, \mathbf{400}\}$, the learning rate of $lr \in \{$3e-4, 4e-4, **5e-4**, 6e-4$\}$, the incremental rate of the modal weight coefficient $\lambda \in \{\mathbf{0.1}, 0.2, 0.3, 0.4, 0.5\}$, the probability distribution function in PMF module including $\{\mathbf{Beta}, Cauchy, Gamma, Gumbel, Laplace\}$, the temperature parameter in contrastive loss $\tau \in \{0.01, 0.05, \mathbf{0.1}, 0.2, 0.4, 0.8, 1.0\ \}$, the inner-graph alignment ratio in contrastive loss $\gamma \in \{0.6, 0.7, \mathbf{0.8}, 0.9, 1.0\ \}$. The number of training epochs is 1000 with early-stopping and the batch size is 512. Similar to [13, 16, 18, 17], in the iterative training strategy, in the last 500 epochs of training, every $R = 5$ epochs, we select entity pairs that are mutual nearest neighbours in the vector space and add them to the buffer area. After the subsequent $M = 50$ epochs, if these entity pairs still exist, we add them to the training set. It should be noted that among the hyper-parameters, the parameters in bold are the configurations selected in our experimental report. We used the same parameter settings on all datasets: FB15K-DB15K and FB15K-YAGO15K datasets.

Evaluation Metrics. For every entity pair (e_1, e_2) in the test set, we obtain a ranking list for the possible alignment entities e_2. We use three evaluation metrics: mean reciprocal rank (MRR), and Hits@k($k \in \{1, 10\}$). MRR measures the inverse of the rank for the first correct answer, Hits@k represents the rate of correct entities ranked in the top-k answers. The higher the values of Hits@k and MRR, the better performance of the model.

Iterative Training. As in previous works [13, 16, 18, 17], we adopt a probation strategy for *iterative training*. Specifically, we constructed a buffer to temporarily store entity pairs that are close in the embedding space across different knowledge graphs. In every round R, we select entity pairs that meet the nearest distance criteria and add them to the buffer. If after M iterations, these entity pairs are still in the buffer, we will add them to the training set. This approach effectively serves as a data augmentation strategy during training, where the entity pairs in the buffer can be considered as pseudo-labels. In contrast, the training method that does not involve the aforementioned iterative process is referred to as *non-iterative training*.

Baselines. In the experiments, we used two training strategies: *non-iterative* and *iterative training*. For each training strategy, we used different baselines. For non-iterative trainin: PoE [11], HMEA [14], MMEA [12], EVA [13], MSNEA [16], MCLEA [18], MEAformer [17], ACK-MMEA [29], GEEA [30]. For iterative training: EVA [13], MSNEA [16], MCLEA [18], MEAformer [17]. The details of different baselines are described as follows:

- PoE [11] provides three multi-modal knowledge graphs enriched with numerical literals, images, and sameAs links, enabling research on link prediction and entity matching across KGs.
- HMEA [14] proposes a hyperbolic multi-modal entity alignment model that employs HGCNs for structural embeddings and DenseNet for visual embeddings, integrating both modalities in the hyperbolic space for improved alignment.
- MMEA [12] introduces a multi-modal entity alignment model that learns relational, visual, and numerical embeddings and integrates them through a multi-modal fusion module to improve alignment performance.
- EVA [13] leverages visual semantic representations with an attention-based fusion to align entities across KGs, and introduces an unsupervised setting by exploiting visual pivots.
- MSNEA [16] adopts a multi-modal Siamese network that exploits intermodal effects by enhancing relational features with visual cues and adaptively weighting attributes.
- MCLEA [18] employs contrastive learning to jointly model intra-modal and inter-modal interactions, generating discriminative cross-modal embeddings.
- MEAformer [17] designs a transformer-based entity alignment model with

meta-modality hybrid, which dynamically learns entity-level modality correlations for robust multi-modal fusion.

- ACK-MMEA [29] constructs attribute-consistent KGs through multi-modal attribute uniformization to mitigate contextual gaps, and applies a relation-aware GNN with a joint alignment loss.
- GEEA [30] considers a generative framework with a mutual variational autoencoder that enables both entity alignment and entity synthesis, overcoming the limitations of GAN-based approaches.

5.2. Main Results

To address **RQ1**, we conduct experiments on the non-iterative and iterative training settings, and on the number of selected pre-aligned seeds. The results are shown in Tables 2 and 3. To facilitate reproducibility, we explicitly summarize the final hyper-parameter settings used in all experiments. After tuning on the validation set, the incremental rate of the modal weight coefficient was fixed to $\lambda = 0.1$, the inner-graph alignment ratio in the contrastive loss was set to $\gamma = 0.8$, and the temperature parameter in the contrastive loss was set to $\tau = 0.1$. These configurations were consistently applied across all datasets, including FB15K-DB15K and FB15K-YAGO15K.

Performance Comparison. We can observe in the results that under both the non-iterative and iterative settings, MIMEA generally outperforms existing SoTA baselines by a large margin across all metrics. More precisely, we have the following observations. On the one hand, MIMEA achieves the best performance on the multi-modal entity alignment task. For example, in the non-iterative setting, on the FB15K-YAGO15K dataset, MIMEA achieves improvements of 8.9%, 2.4% and 0.5% on MRR compared to the best SoTA baselines when the given pre-aligned seeds are 20%, 50%, and 80%, respectively. Similar improvements are obtained on the FB15K-DB15K dataset. On the other hand, the iterative training strategy can significantly improve model performance of existing baselines and MIMEA. For example, on the FB15K-DB15K dataset when the given pre-aligned seeds are 20%, 50%, and 80%, depending on whether MIMEA uses the iterative training mechanism, there will be fluctuations of 10%, 2.2%, and 1.4% on MRR, respectively. This is primarily attributed to the generation of pseudo-entity alignments pairs during the iterative training process, which iteratively filters out potentially wrong entity pairs.

Table 2: Evaluation of different models in the non-iterative setting. Results marked with †, ‡ and ⋆ respectively come from [18] [17], and the corresponding paper. Best scores are in **bold**, the second best scores are underlined, and '–' indicates the results are not reported in previous work. All results for MIMEA are averaged over 5 random seeds. * indicates statistically significant improvement over the strongest baseline (GEEA) under a paired two-sided t-test ($p < 0.05$) across 5 random seeds.

Methods	**FB15K-DB15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
PoE†	0.170	0.126	0.251	0.533	0.464	0.658	0.721	0.666	0.820
HMEA†	–	0.127	0.369	–	0.262	0.581	–	0.417	0.786
MMEA†	0.357	0.265	0.541	0.512	0.417	0.703	0.685	0.590	0.869
EVA‡	0.283	0.199	0.448	0.422	0.334	0.589	0.563	0.484	0.696
MSNEA‡	0.175	0.114	0.296	0.388	0.288	0.590	0.613	0.518	0.779
MCLEA‡	0.393	0.295	0.582	0.637	0.555	0.784	0.790	0.735	0.890
MEAformer‡	0.518	0.417	0.715	0.698	0.619	0.843	0.820	0.765	0.916
ACK-MMEA⋆	0.387	0.304	0.549	0.624	0.560	0.736	0.752	0.682	0.874
GEEA⋆	0.450	0.343	0.661	0.723	0.651	0.852	0.836	0.787	**0.918**
MIMEA	**0.594**	**0.506**	**0.756**	**0.748**	**0.683**	**0.861**	**0.841**	**0.799**	0.914
±std	±0.004*	±0.005*	±0.003*	±0.003*	±0.004*	± 0.002*	±0.002*	±0.003*	±0.002

Methods	**FB15K-YAGO15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
PoE†	0.154	0.113	0.229	0.414	0.347	0.536	0.635	0.573	0.746
HMEA†	–	0.105	0.313	–	0.265	0.581	–	0.433	0.801
MMEA†	0.317	0.234	0.480	0.486	0.403	0.645	0.682	0.598	0.839
EVA‡	0.224	0.153	0.361	0.388	0.311	0.534	0.565	0.491	0.692
MSNEA‡	0.153	0.103	0.249	0.413	0.320	0.589	0.620	0.531	0.778
MCLEA‡	0.332	0.254	0.484	0.574	0.501	0.705	0.722	0.667	0.824
MEAformer‡	0.417	0.327	0.595	0.639	0.560	0.778	0.766	0.703	0.873
ACK-MMEA⋆	0.360	0.289	0.496	0.593	0.535	0.699	0.744	0.676	0.864
GEEA⋆	0.393	0.298	0.585	0.668	0.589	0.808	0.790	0.733	**0.890**
MIMEA	**0.506**	**0.417**	**0.671**	**0.692**	**0.622**	**0.818**	**0.795**	**0.741**	0.884
±std	±0.005*	±0.006*	±0.004*	±0.004*	±0.005*	±0.003*	±0.003*	±0.004*	±0.003

Table 3: Evaluation of different models under iterative setting. ‡ results come from [17]. † results from the corresponding papers. Best scores are highlighted in **bold**, the second best scores are underlined. All results for MIMEA are averaged over 5 random seeds. * indicates statistically significant improvement over the strongest baseline (MEAformer) under a paired two-sided t-test ($p < 0.05$) across 5 random seeds.

Methods	**FB15K-DB15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
EVA‡	0.318	0.231	0.488	0.449	0.364	0.606	0.573	0.491	0.711
MSNEA‡	0.232	0.149	0.392	0.459	0.358	0.656	0.651	0.565	0.810
MCLEA†	0.534	0.445	0.705	0.652	0.573	0.800	0.784	0.730	0.883
MEAformer‡	0.661	0.578	0.812	0.755	0.690	0.871	0.834	0.784	**0.921**
MIMEA	**0.694**	**0.622**	**0.824**	**0.770**	**0.716**	**0.872**	**0.855**	**0.821**	0.919
±std	±0.004*	±0.005*	±0.003*	±0.003*	±0.004*	±0.002*	±0.002*	±0.003*	±0.002*

Methods	**FB15K-YAGO15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
EVA‡	0.260	0.188	0.403	0.404	0.325	0.560	0.572	0.493	0.695
MSNEA‡	0.210	0.138	0.346	0.472	0.376	0.646	0.668	0.593	0.806
MCLEA†	0.474	0.388	0.641	0.616	0.543	0.759	0.715	0.653	0.835
MEAformer‡	0.529	0.444	0.692	0.682	0.612	0.808	0.783	0.724	0.880
MIMEA	**0.587**	**0.513**	**0.729**	**0.712**	**0.651**	**0.827**	**0.803**	**0.757**	**0.885**
±std	±0.006*	±0.004*	±0.005*	±0.004*	±0.006*	±0.003*	±0.003*	±0.004*	±0.003*

Statistical Significance Analysis. To assess the robustness and reliability of the observed performance gains, we conducted a statistical significance analysis on the key evaluation metrics. Specifically, all results of MIMEA are obtained by averaging over five runs with different random seeds, and we report the corresponding ±standard deviation. As shown in Tables 2 and 3, MIMEA consistently outperforms the strongest baseline methods across different alignment ratios and datasets under both non-iterative and iterative settings. To further verify that these improvements are not due to random variations, we performed paired two-sided t-tests between MIMEA and the strongest baseline for each setting (*i.e.,* GEEA in the non-iterative setting and MEAformer in the iterative setting). Improvements marked with "*" indicate statistical significance at the $p < 0.05$ level. The results demonstrate that the improvements achieved by MIMEA are statistically significant on almost all metrics, with particularly consistent gains in MRR and Hits@1, which are critical indicators of top-ranked alignment accuracy. In addition, the reported standard deviations are relatively small across all experimental configurations, suggesting that MIMEA maintains stable performance across different random initializations.

Impact of Number of Pre-aligned Seeds. We evaluate the sensitivity of MIMEA to the given number of pre-aligned seeds: 20%, 50%, and 80% [13, 16, 18, 17]. From the results, we can observe that MIMEA achieves the best performance on both the FB15K-DB15K and the FB15K-YAGO15K datasets in all metrics and proportions, confirming its robustness to the number of given pre-aligned seeds. For instance, in the iterative setting, on FB15K-YAGO15K, compared with the best-performing baseline MEAformer, for 20%, 50%, and 80%, the MRR metric is respectively improved by 5.8%, 3.0% and 2.0%. The higher improvement for 20% shows that MIMEA is well-suited for low-resource scenarios. This is mainly because, on the one hand, each modality can be explicitly given a differentiation weight according to the characteristics of such modality. Further, we take into account the interactions between uni-modal and joint-modal representations. On the other hand, intra-modal is able to differentiate uni-modal representations. The intra-modal and inter-modal multi-granularity interaction can indeed maximize the utility of having multi-modal knowledge.

5.3. Ablation Studies

We address **RQ2** from four perspectives, different modalities, including different variants, different encoders, impact of OTMA on different modalities, different distribution methods, different fusion strategies, and different pivotal modality. The results are shown in Tables 4, 6, and 7.

Impact of Modalities. The upper part of Table 4 shows the individual contribution of different modalities. We can observe that independent of the dataset or the number of pre-aligned seeds, the removal of different modalities has varying degrees of performance drop. The structural information has shown to be the main source, with its removal leading to the most significant drop (this is in line with previous findings [18]). This might be explained by the wealth of structural triples available in both datasets. On the other extreme, the performance gain brought by the visual modality is minimal.

Table 4: Ablation studies under different modals and different modules. Best scores are highlighted in **bold**. † and ‡ denote MIMEA variants using BERT-based relation and attribute encoders with ResNet-152 and ViT-B/16 visual backbones, respectively.

Settings	**FB15K-DB15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
w/o structure	0.094	0.044	0.183	0.134	0.066	0.264	0.220	0.117	0.447
w/o attribute	0.664	0.589	0.806	0.750	0.694	0.859	0.839	0.801	0.910
w/o relation	0.642	0.565	0.785	0.742	0.685	0.850	0.831	0.791	0.902
w/o visual	0.691	0.616	0.825	0.772	0.716	0.877	0.853	0.815	0.921
w/o PMF	0.595	0.507	0.757	0.747	0.682	0.862	0.841	0.797	0.914
w/o OTMA	0.576	0.486	0.748	0.724	0.648	0.860	0.845	0.797	0.923
w/o MCL	0.630	0.535	0.797	0.744	0.671	0.873	0.844	0.802	0.918
MIMEA	0.694	0.622	0.824	0.770	0.716	0.872	0.855	0.821	0.919
MIMEA†	0.713	0.639	0.836	0.786	0.731	0.881	0.868	0.832	0.927
MIMEA‡	**0.720**	**0.646**	**0.842**	**0.792**	**0.737**	**0.887**	**0.873**	**0.838**	**0.938**

Settings	**FB15K-YAGO15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
w/o structure	0.088	0.051	0.151	0.105	0.056	0.192	0.175	0.094	0.335
w/o attribute	0.553	0.479	0.700	0.684	0.620	0.806	0.774	0.718	0.871
w/o relation	0.507	0.429	0.660	0.661	0.587	0.801	0.776	0.717	0.875
w/o visual	0.568	0.492	0.717	0.699	0.633	0.822	0.791	0.736	0.886
w/o PMF	0.494	0.406	0.659	0.688	0.617	0.817	0.794	0.741	0.885
w/o OTMA	0.518	0.437	0.673	0.660	0.578	0.810	0.796	0.737	**0.897**
w/o MCL	0.535	0.449	0.690	0.680	0.612	0.795	0.780	0.723	0.878
MIMEA	0.587	0.513	0.729	0.712	0.651	0.827	0.803	0.757	0.885
MIMEA†	0.603	0.528	0.741	0.728	0.665	0.837	0.814	0.767	0.892
MIMEA‡	**0.610**	**0.534**	**0.746**	**0.734**	**0.673**	**0.842**	**0.819**	**0.772**	0.896

In fact, the removal of visual information can sometimes lead to achieve better results. The main reason is that the visual information provides limited additional knowledge. Only through the interaction with other modal information can bring certain performance improvement.

Table 5: Fine-grained ablation results of OTMA, in which w/o OTMA-A, w/o OTMA-R, and w/o OTMA-V indicate that OTMA is disabled for attribute, relation, and visual embeddings, respectively.

Settings	**FB15K-DB15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
w/o OTMA-A	0.635	0.556	0.792	0.758	0.702	0.873	0.851	0.809	0.927
w/o OTMA-R	0.598	0.512	0.764	0.742	0.684	0.865	0.846	0.801	0.924
w/o OTMA-V	0.662	0.588	0.812	0.764	0.709	0.878	0.853	0.816	0.929
MIMEA	0.694	0.622	0.824	0.770	0.716	0.872	0.855	0.821	0.919

Settings	**FB15K-YAGO15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
w/o OTMA-A	0.556	0.479	0.707	0.698	0.629	0.821	0.801	0.751	0.892
w/o OTMA-R	0.529	0.451	0.688	0.671	0.598	0.809	0.796	0.742	0.889
w/o OTMA-V	0.571	0.498	0.721	0.704	0.636	0.824	0.803	0.755	0.894
MIMEA	0.587	0.513	0.729	0.712	0.651	0.827	0.803	0.757	0.885

Impact of Modules. The lower part of Table 4 presents the results of the impact of each component of MIMEA on the performance. We can observe that by removing any module the performance dramatically degrades. This could be explained by the fact that different modules play different roles, realizing multi-granular modal information interaction. For example, the PMF module focuses on the interaction of uni-modal information (with the structural information as the core) and can ultimately form joint-modal representations. In contrast, the MCL module underscores the significance of intra-modal interactions for each modality. The MIMEA's modules are interrelated and form a complete data flow, so the absence of any one of them leads to a significant performance fluctuation.

Encoder Capacity and Encoder-independence. To verify whether the benefits of MIMEA are independent of encoder choice, we further evaluate MIMEA under stronger encoder configurations by introducing BERT-based relation/attribute encoders and replacing the visual backbone with ResNet-152 (MIMEA†) and ViT-B/16 (MIMEA‡), while keeping the remaining architecture and training protocol unchanged. Overall, MIMEA exhibits

Table 6: Evaluation of different models under different distribution methods and comparison with different fusion strategies, including Equal-weight Fusion (EWF), Attention-based Fusion (AF), Gated Fusion (GF). . Best scores are highlighted in **bold**.

Settings	**FB15K-DB15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
Beta	0.694	0.622	0.824	**0.770**	**0.716**	0.872	**0.855**	**0.821**	0.919
Cauchy	**0.697**	**0.625**	**0.825**	**0.770**	0.713	**0.877**	0.852	0.815	**0.921**
Gamma	0.691	0.621	0.822	0.769	0.713	0.874	0.851	0.818	0.915
Gumbel	0.690	0.619	0.821	0.769	0.714	0.872	0.851	0.816	0.917
Laplace	0.694	0.624	0.823	**0.770**	0.715	0.875	0.851	0.816	0.920
EWF	0.664	0.591	0.801	0.742	0.684	0.853	0.832	0.792	0.907
AF	0.683	0.611	0.816	0.758	0.702	0.865	0.846	0.807	0.914
GF	0.687	0.615	0.818	0.761	0.705	0.868	0.849	0.811	0.916
Settings	**FB15K-YAGO15K**								
	20%			**50%**			**80%**		
	MRR	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**	**MRR**	**H@1**	**H@10**
Beta	0.587	0.513	0.729	0.712	**0.651**	0.827	0.803	**0.757**	0.885
Cauchy	**0.592**	**0.516**	**0.734**	**0.715**	**0.651**	**0.833**	**0.806**	**0.757**	0.891
Gamma	0.575	0.496	0.724	0.708	0.646	0.828	0.802	0.751	**0.893**
Gumbel	0.573	0.493	0.722	0.708	0.645	0.827	0.801	0.749	0.890
Laplace	0.581	0.503	0.729	0.713	0.650	0.831	0.803	0.754	0.892
EWF	0.561	0.487	0.708	0.693	0.631	0.815	0.787	0.738	0.872
AF	0.574	0.499	0.721	0.704	0.642	0.824	0.796	0.747	0.880
GF	0.579	0.503	0.725	0.707	0.645	0.826	0.799	0.750	0.883

encoder-independent effectiveness: across all seed ratios (20%, 50%, 80%) and both datasets, the upgraded variants consistently improve upon the original MIMEA, *e.g.,* on FB15K-DB15K the MRR increases from 0.694/0.770/0.855 to 0.713/0.786/0.868 with ResNet-152 and further to 0.720/0.792/0.873 with ViT-B/16, and similar trends are observed on FB15K-YAGO15K (MRR: 0.587/0.712/0.803 → 0.603/0.728/0.814 → 0.610/0.734/0.819). Moreover, stronger encoders provide *complementary* yet *incremental* gains: switching from the original BoW+VGG setting to BERT+ResNet yields modest but consistent improvements, and replacing ResNet with ViT brings additional smaller gains, indicating diminishing returns as encoder capacity increases and as more pre-aligned seeds are available. Importantly, the relative performance ordering remains stable, suggesting that MIMEA does not rely on encoder superiority to achieve its advantage. These observations support that the performance gains primarily stem from MIMEA's multi-modal interaction and alignment modeling: even under stronger feature extractors, the improvements remain steady across data regimes, implying that MIMEA effectively.

Table 7: The MRR metric results of using different modal content as the central one in the PMF module. Best scores are highlighted in **bold**.

Methods	**FB15K-DB15K**			**FB15K-YAGO15K**		
	20%	**50%**	**80%**	**20%**	**50%**	**80%**
attribute	0.595	0.747	0.841	0.494	0.688	0.794
relation	0.576	0.724	0.845	0.518	0.660	0.796
visual	0.630	0.744	0.844	0.535	0.680	0.780
structural	**0.694**	**0.770**	**0.855**	**0.587**	**0.712**	**0.803**

Impact of OTMA on Different Modalities. To identify which uni-modal embeddings benefit the most from the proposed OTMA, we conduct a fine-grained ablation study by selectively disabling OTMA for one modality at a time, yielding three variants: w/o OTMA-A, w/o OTMA-R, and w/o OTMA-V as shown in Table 5. Overall, OTMA contributes most significantly to *relation* embeddings, especially under low supervision, and its impact gradually diminishes as more seeds become available. On FB15K-DB15K, removing OTMA from relation embeddings causes the largest degradation in MRR across all seed ratios (20%/50%/80%): 0.694/0.770/0.855 → 0.598/0.742/0.846, compared with smaller drops when disabling OTMA for attributes (MRR: 0.694/0.770/0.855 → 0.635/0.758/0.851) or visuals (MRR: 0.694/0.770/0.855 → 0.662/0.764/0.853). The same pattern is consistently observed on FB15K-YAGO15K. Importantly, this modality-specific sensitivity is most evident at 20% seeds, aligning with the observation in Table 4 that OTMA is particularly crucial in sparse-seed regimes. We attribute this to the fact that relation embeddings encode highly structured semantics and are more vulnerable to noise introduced by joint-modal fusion; OTMA mitigates this issue by explicitly aligning uni-modal relational representations with the joint-modal space via low-cost transport plans, thereby preserving discriminative relational cues while reducing fusion-induced interference. In contrast, visual signals are often complementary but less directly tied to graph topology, and attribute facts can be partially redundant or noisy, which explains why OTMA-V and OTMA-A contribute consistent yet smaller gains.

Impact of Distribution Methods. We investigate the choice of different probability distribution functions in the PMF module. Table 6 reports the results by replacing the Beta function in the PMF module with the Cauchy, Gamma, Gumbel, or Laplace functions. We observe that using different probability distribution functions has a relatively limited impact on MIMEA's performance, showing the robustness of the PMF module. This is explained by the fact that the weight coefficients obtained by each probability distribution function tend to be similar after subsequent gradient updates.

Impact of Fusion Strategies. We compare the proposed PMF with several commonly used deterministic fusion strategies, including equal-weight fusion, attention-based fusion, and gated fusion. The experimental results in the Tabel 6 show that adaptive fusion strategies, such as attention-based and gated fusion, consistently outperform naive equal-weight fusion, highlighting the importance of learning modality-specific contributions. However, both deterministic fusion methods remain inferior to PMF across all experimental settings and datasets. This performance gap can be attributed to the fundamental difference in how modality interactions are modeled. Attention-based and gated fusion mechanisms compute fusion weights directly from point embeddings, which makes them more sensitive to noise and modality imbalance. In contrast, PMF models each modality as a probabilistic distribution and derives fusion weights based on distributional divergence with respect to structural embeddings. This probabilistic design enables PMF to better capture uncertainty and semantic discrepancies between modalities, resulting in more stable and discriminative fused representations.

Impact of Different Pivotal Modality. In the PMF module, we use the structural modality as the central one for the interaction between uni-modal representations. To verify the adequateness of this choice, we select attribute, relation, and visual as the central ones. The experimental results are shown in Table 7. We can observe that by choosing the structural modality as the core we achieve the best results. The main reason is that the datasets contain rich knowledge of structural triples, which can provide abundant evidence. Recall that the performance loss caused by removing the visual modality in Table 4 is lower than that of removing the relation and attribute modalities, that is, the visual modality seems to be of little importance in the MMEA task. However, when using the visual modality as the core for uni-modal interaction, it can achieve better results than the relation and attribute modalities. A possible explanation is that the subsequent OTMA module directly assists the functioning of the visual modality, because from Table 4 we find that removing the OTMA module has the largest impact on performance.

Table 8: Amount of calculations (#FLOPs), parameters (#Params), and inference time (#Time) required by different models on different datasets.

Metrics	**Model**	**FB15K-DB15K**	**FB15K-YAGO15K**
#FLOPs	MCLEA	103.345G	112.872G
	MEAformer	203.100G	219.175G
	MIMEA	67.770G	74.018G
#Params	MCLEA	3.720M	3.720M
	MEAformer	3.461M	3.374M
	MIMEA	2.440M	2.440M
#Time	MCLEA	1.00×	1.00×
	MEAformer	1.97×	1.94×
	MIMEA	0.66×	0.66×

5.4. Complexity Analysis

To address **RQ3**, we analyze the model's complexity from three perspectives: *i.e.,* time complexity (#FLOPs), space complexity (#Params) and average inference time (#Time). Table 8 presents the corresponding time complexity, space complexity, and inference time results comparison of MIMEA and the best-performing MCLEA [18] and the MEAformer [17] model.

- The time complexity can be measured by the amount of model calculations, which refers to the number of floating-point operations performed during the inference process of the model, usually expressed in units of FLOPs (Floating-Point Operations Per Second). Specifically, all multiplication and addition operations involved in the forward inference process are considered fundamental floating-point operations. Formally, the

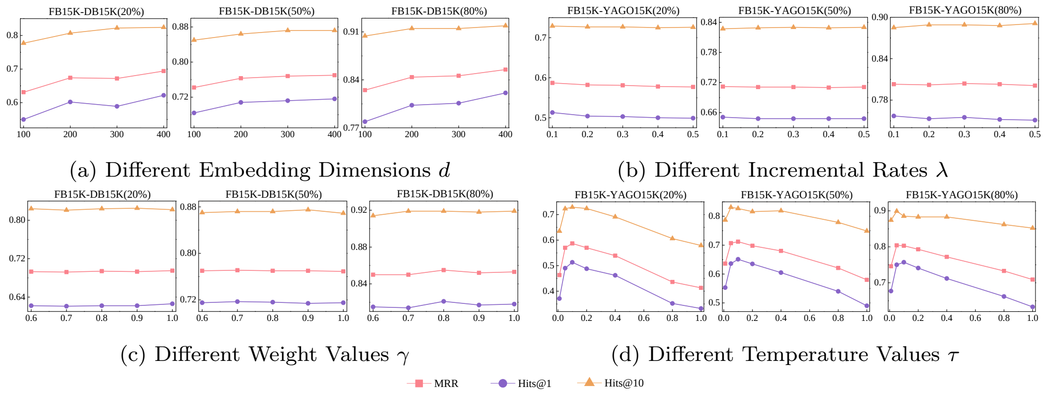
FLOPs is calculated as:

$$\text{FLOPs} = \sum_{i=1}^{N} (\text{Mul}_i + \text{Add}_i) \tag{11}$$

where N denotes the number of computational units involved in a single forward pass, and Mul_i and Add_i represent the numbers of multiplication and addition operations required by the i-th computational unit, respectively. The unit "Giga" (G) in FLOPs denotes a factor of 10^9, *i.e.,* one billion floating-point operations.

- The space complexity can be measured by the amount of model's parameters, which refers to the number of adjustable parameters that need to be learned in the model. These parameters are the weights and biases of the model that are adjusted through optimization algorithms such as gradient descent during the training process. The number of parameters is usually expressed in "Millions" (M).
- The average inference time is an important metric for evaluating inference efficiency, as it measures the actual wall-clock time required by the model to complete a single forward inference. Specifically, the average inference time is defined as the mean execution time over multiple inference runs under the same hardware and software environment. It reflects the practical deployment efficiency of the model and is affected by factors such as model architecture, hardware platform, and implementation optimization. In the Table 8, the #Time metric represents the relative inference time required by each model. Specifically, the inference time of MCLEA is used as the baseline and normalized to 1.00×, while the inference times of other models are reported as multiples of this baseline.

We can find that MIMEA simultaneously reduces the computational cost, the number of parameters, and inference time in comparison to the other two baselines. In particular, the amount of calculation needed by MIMEA is one-third of MEAformer's. Regard for inference time, the reported #Time values are derived from the relative computational complexity of the models under identical inference settings (*i.e.,* same hardware platform, batch size, and evaluation protocol). Since inference time is dominated by floating-point operations, the relative #Time values are proportional to the corresponding #FLOPs, providing a hardware-independent comparison of inference efficiency. To sum up, MIMEA can achieve the best performance while minimizing the model's computational load and video memory footprint.



(a) Different Embedding Dimensions d

(b) Different Incremental Rates λ

(c) Different Weight Values γ

(d) Different Temperature Values τ

Figure 3: The ablation studies results under different experimental conditions.

5.5. Parameter Analysis

To address **RQ4**, we investigate the effect of choosing different hyperparameters. The results are shown in Figure 3.

Different Embedding Dimensions d. Figure 3(a) presents results on the influence of the hidden sizes. We observe that the increase of the hidden sizes helps to capture more messages, which helps to improve the model performance. However, increasing the hidden size will increase the amount of video memory occupied. Thus, in practice, we might need to find a compromise between performance and computational complexity.

Different Incremental Rates λ. The coefficient λ in the PMF module is used to scale the calculation results of the KL divergence. Figure 3(b) shows that the value of λ does not have a decisive impact on the performance, showing that MIMEA's robustness in this respect.

Different Weight Values γ. The weight coefficient γ in Equation (9) is used to adjust the importance level of positive and negative samples between the inner-graph and cross-graph for a modality. Figure 3(c) shows the results for different weight values. We can observe that variation of the weight coefficient only brings subtle performance differences.

Different Temperature Values τ. The temperature coefficient adjusts the similarity distribution between contrastive embeddings. It affects the performance of contrastive learning models by expanding or shrinking the relative distance between different samples in the embedding space. Figure 3(d) shows that setting different temperature coefficients has significant impact on MIMEA. A higher temperature coefficient makes the difference between similarity scores smoother, making the samples harder to distinguish. Although lower temperatures might make it easier for the model to distinguish different samples, it will increase the model's sensitivity. So, in practice, we need to choose a temperature coefficient that ensures a balance between the stability and flexibility of the model.

6. Conclusion and future work

In this paper, we proposed MIMEA, a framework for multi-modal entity alignment that effectively leverages multi-modal knowledge with the exploitation of intra-modal and inter-modal interactions. The experimental results demonstrate the effectiveness of MIMEA. For future work, given that the structural information is the most significant and that in practice, structural knowledge is often incomplete, we could first employ knowledge graph completion techniques to fill in missing parts. In addition, it would be interesting to explore the application of multimodal entity alignment techniques in industrial intelligence domains, such as intelligent manufacturing and defect detection, where aligning heterogeneous multimodal descriptions of industrial entities across systems is a fundamental yet challenging problem.

Declaration of Competing Interests

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Generative AI in Scientific Writing

We declare that no generative artificial intelligence tools were used in the writing of this manuscript or in the generation of any figures or results.

CRediT authorship contribution statement

Zhiwei Hu: Conceptualization, Investigation, Methodology, Software, Writing - original draft. **Víctor Gutiérrez-Basulto**: Writing - Review &

Editing. **Zhiliang Xiang**: Writing - Review & Editing. **Ru Li**: Writing - Review & Editing, Supervision, Funding acquisition. **Jeff Z. Pan**: Writing - Review & Editing, Supervision, Funding acquisition.

Funding Sources

This work was supported by the National Natural Science Foundation of China (No.62376144), the Science and Technology Cooperation and Exchange Special Project of Shanxi Province (No.202204041101016), the Key Research and Development Project of Shanxi Province (No.202102020101008), the Independent Research Project of Key Laboratory of Minor Crop Germplasm Innovation and Molecular Breeding (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs (No.202404010930003-J08), the Open Fund of the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education (No.CICIP2025003), the Science and Technology Innovation Enhancement Program of Shanxi Agricultural University (No.CXGC2025101), the Research Start-up Fund for Introduced Talents of Shanxi Agricultural University (No.2026BQ64).

References

[1] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web 6 (2) (2015) 167–195.

[2] F. Mahdisoltani, J. Biega, F. M. Suchanek, YAGO3: A knowledge base from multilingual wikipedias, in: CIDR, www.cidrdb.org, Asilomar, the United States, 2015, pp. 1–12.

[3] Z. Zhang, J. Wang, J. Chen, S. Ji, F. Wu, Cone: Cone embeddings for multi-hop reasoning over knowledge graphs, in: NeurIPS, Curran Associates, online, 2021, pp. 19172–19183.

[4] Z. Hu, V. Gutiérrez-Basulto, Z. Xiang, X. Li, R. Li, J. Z. Pan, Type-aware embeddings for multi-hop reasoning over knowledge graphs, in: IJCAI, ijcai.org, Vienna, Austria, 2022, pp. 3078–3084.

[5] C. Nguyen, T. French, W. Liu, M. Stewart, Scone: Simplified cone embeddings with symbolic operators for complex logical queries, in: ACL, ACL, Toronto, Canada, 2023, pp. 11931–11946.

[6] D. Q. Nguyen, T. Vu, T. D. Nguyen, D. Q. Nguyen, D. Q. Phung, A capsule network-based embedding model for knowledge graph completion and search personalization., in: NAACL, NAACL-HLT, Minneapolis, the United States, 2019, pp. 2180–2189.

[7] Y. Gu, T. Zhou, G. Cheng, Z. Li, J. Z. Pan, Y. Qu, Relevance Search over Schema-Rich Knowledge Graphs, in: WSDM, ACM Press, Melbourne, Australia, 2019, pp. 114–122.

[8] N. Zheng, H. Xia, Z. Liang, Y. Du, MK-SGN: A spiking graph convolutional network with multimodal fusion and knowledge distillation for skeleton-based action recognition, Neurocomputing 662 (2026) 131796.

[9] Y. Wu, F. Liu, L. Wan, Z. Wang, Intelligent fault diagnostic model for industrial equipment based on multimodal knowledge graph, IEEE Sensors Journal 23 (21) (2023) 26269–26278.

[10] J. Xu, H. Xu, P. Liang, Z. Ma, D. Lu, Y. Hou, Textual and visual features alignment and attention interaction based multimodal entity alignment method for electric power operation and inspection, in: 2024 6th International Conference on Smart Power & Internet Energy Systems (SPIES), IEEE, 2024, pp. 195–200.

[11] Y. Liu, H. Li, A. García-Durán, M. Niepert, D. Oñoro-Rubio, D. S. Rosenblum, MMKG: multi-modal knowledge graphs, in: ESWC, Springer, Portorož, Slovenia, 2019, pp. 459–474.

[12] L. Chen, Z. Li, Y. Wang, T. Xu, Z. Wang, E. Chen, MMEA: entity alignment for multi-modal knowledge graph, in: KSEM, Springer, Hangzhou, China, 2020, pp. 134–147.

[13] F. Liu, M. Chen, D. Roth, N. Collier, Visual pivoting for (unsupervised) entity alignment, in: AAAI, AAAI Press, online, 2021, pp. 4257–4266.

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

[15] B. Cheng, J. Zhu, M. Guo, Multijaf: Multi-modal joint entity alignment framework for multi-modal knowledge graph, Neurocomputing 500 (2022) 581–591.

[16] L. Chen, Z. Li, T. Xu, H. Wu, Z. Wang, N. J. Yuan, E. Chen, Multi-modal siamese network for entity alignment, in: KDD, ACM, Washington, the United States, 2022, pp. 118–126.

[17] Z. Chen, J. Chen, W. Zhang, L. Guo, Y. Fang, Y. Huang, Y. Zhang, Y. Geng, J. Z. Pan, W. Song, H. Chen, Meaformer: Multi-modal entity alignment transformer for meta modality hybrid, in: MM, ACM, Ottawa, ON, Canada, 2023, pp. 3317–3327.

[18] Z. Lin, Z. Zhang, M. Wang, Y. Shi, X. Wu, Y. Zheng, Multi-modal contrastive representation learning for entity alignment, in: COLING, ACL, online, 2022, pp. 2572–2584.

[19] M. Chen, Y. Tian, M. Yang, C. Zaniolo, Multilingual knowledge graph embeddings for cross-lingual knowledge alignment, in: IJCAI, ijcai.org, Melbourne, Australia, 2017, pp. 1511–1517.

[20] Z. Sun, W. Hu, C. Li, Cross-lingual entity alignment via joint attribute-preserving embedding, in: ISWC, Springer, Vienna, Austria, 2017, pp. 628–644.

[21] H. Zhu, R. Xie, Z. Liu, M. Sun, Iterative entity alignment via joint knowledge embeddings, in: IJCAI, ijcai.org, Melbourne, Australia, 2017, pp. 4258–4264.

[22] Z. Sun, W. Hu, Q. Zhang, Y. Qu, Bootstrapping entity alignment with knowledge graph embedding, in: IJCAI, ijcai.org, Stockholm, Sweden, 2018, pp. 4396–4402.

[23] Y. Zhu, H. Liu, Z. Wu, Y. Du, Relation-aware neighborhood matching model for entity alignment, in: AAAI, AAAI Press, online, 2021, pp. 4749–4756.

[24] L. Guo, Q. Zhang, Z. Sun, M. Chen, W. Hu, H. Chen, Understanding and improving knowledge graph embedding for entity alignment, in: ICML, PMLR, Maryland, the United States, 2022, pp. 8145–8156.

[25] M. Chen, Y. Tian, K. Chang, S. Skiena, C. Zaniolo, Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment, in: IJCAI, ijcai.org, Stockholm, Sweden, 2018, pp. 3998–4004.

[26] Z. Sun, C. Wang, W. Hu, M. Chen, J. Dai, W. Zhang, Y. Qu, Knowledge graph alignment network with gated multi-hop neighborhood aggregation, in: AAAI, AAAI Press, California, the United States, 2020, pp. 222–229.

[27] Y. Cao, Z. Liu, C. Li, Z. Liu, J. Li, T. Chua, Multi-channel graph neural network for entity alignment, in: ACL, ACL, Florence, Italy, 2019, pp. 1452–1461.

[28] Z. Liu, Y. Cao, L. Pan, J. Li, T. Chua, Exploring and evaluating attributes, values, and structures for entity alignment, in: EMNLP, ACL, Zurich, Switzerland, 2020, pp. 6355–6364.

[29] Q. Li, S. Guo, Y. Luo, C. Ji, L. Wang, J. Sheng, J. Li, Attribute-consistent knowledge graph representation learning for multi-modal entity alignment, in: WWW, ACM, Washington, the United States, 2023, pp. 2499–2508.

[30] L. Guo, Z. Chen, J. Chen, Y. Fang, W. Zhang, H. Chen, Revisit and outstrip entity alignment: A perspective of generative models, in: ICLR, OpenReview.net, Vienna, Austria, 2024.

[31] N. Bonneel, M. van de Panne, S. Paris, W. Heidrich, Displacement interpolation using lagrangian mass transport, ACM Trans. Graph. 30 (6) (2011) 158.

[32] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, L. J. Guibas, Convolutional wasserstein distances: Efficient optimal transportation on geometric domains, ACM Trans. Graph. 34 (4) (2015) 1–11.

[33] X. Gu, Y. Yang, W. Zeng, J. Sun, Z. Xu, Keypoint-guided optimal transport with applications in heterogeneous domain adaptation, in: NeurIPS, Curran Associates, New Orleans, LA, USA, 2022, pp. 1–14.

[34] W. Chang, Y. Shi, H. Tuan, J. Wang, Unified optimal transport framework for universal domain adaptation, in: NeurIPS, Curran Associates, New Orleans, LA, USA, 2022, pp. 1–13.

[35] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, in: NeurIPS, Curran Associates, online, 2020, pp. 1–13.

[36] Y. M. Asano, M. Patrick, C. Rupprecht, A. Vedaldi, Labelling unlabelled videos from scratch with multi-modal self-supervision, in: NeurIPS, Curran Associates, online, 2020, pp. 1–12.

[37] Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, Q. Huang, OTKGE: multimodal knowledge graph embeddings via optimal transport, in: NeurIPS, NeurIPS Foundation, New Orleans, the United States, 2022, pp. 1–13.

[38] Y. Xu, H. Chen, Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction, in: ICCV, IEEE, Paris, France, 2023, pp. 21184–21194.

[39] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: ICLR, OpenReview.net, Vancouver, Canada, 2018, pp. 1–12.

[40] C. Li, Y. Cao, L. Hou, J. Shi, J. Li, T. Chua, Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model, in: EMNLP, ACL, Hong Kong, China, 2019, pp. 2723–2732.

[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: ICLR, OpenReview.net, Sainte-Maxime, France, 2015, pp. 1–14.

[42] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in: NeurIPS, Curran Associates, Lake Tahoe, the United States, 2013, pp. 2292–2300.

[43] M. Zolfaghari, Y. Zhu, P. V. Gehler, T. Brox, Crossclr: Cross-modal contrastive learning for multi-modal video representations, in: ICCV, IEEE, online, 2021, pp. 1430–1439.

[44] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, L. Wang, Graph contrastive learning with adaptive augmentation, in: WWW, ACM, Lisbon, Portugal, 2021, pp. 2069–2080.

[45] Y. Suh, B. Han, W. Kim, K. M. Lee, Stochastic class-based hard example mining for deep metric learning, in: CVPR, IEEE Computer Society, Long Beach, the United States, 2019, pp. 7251–7259.

[46] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: CVPR, IEEE Computer Society, Salt Lake City, the United States, 2018, pp. 7482–7491.

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

[48] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. G. Ives, Dbpedia: A nucleus for a web of open data, in: ISWC, Springer, Ingolstadt, Austria, 2007, pp. 722–735.

[49] T. Rebele, F. M. Suchanek, J. Hoffart, J. Biega, E. Kuzey, G. Weikum, YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames, in: ISWC, Springer, Yokohama, Japan, 2016, pp. 177–185.