Multi-modal Siamese Network for Entity Alignment

Liyi Chen
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
liyichencly@gmail.com

Zhi Li
Shenzhen International Graduate School, Tsinghua University
zhilizl@sz.tsinghua.edu.cn

Tong Xu*
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
tongxu@ustc.edu.cn

Han Wu
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
wuhanhan@mail.ustc.edu.cn

Zhefeng Wang
Huawei Cloud
wangzhefeng@huawei.com

Nicholas Jing Yuan
Huawei Cloud
nicholas.jing.yuan@gmail.com

Enhong Chen
University of Science and Technology of China
State Key Laboratory of Cognitive Intelligence
cheneh@ustc.edu.cn

ABSTRACT

The booming of multi-modal knowledge graphs (MMKGs) has raised the imperative demand for multi-modal entity alignment techniques, which facilitate the integration of multiple MMKGs from separate data sources. Unfortunately, prior arts harness multi-modal knowledge only via the heuristic merging of uni-modal feature embeddings. Therefore, inter-modal cues concealed in multi-modal knowledge could be largely ignored. To deal with that problem, in this paper, we propose a novel Multi-modal Siamese Network for Entity Alignment (MSNEA) to align entities in different MMKGs, in which multi-modal knowledge could be comprehensively leveraged by the exploitation of inter-modal effect. Specifically, we first devise a multi-modal knowledge embedding module to extract visual, relational, and attribute features of entities to generate holistic entity representations for distinct MMKGs. During this procedure, we employ inter-modal enhancement mechanisms to integrate visual features to guide relational feature learning and adaptively assign attention weights to capture valuable attributes for alignment. Afterwards, we design a multi-modal contrastive learning module to achieve inter-modal enhancement fusion with avoiding the overwhelming impact of weak modalities. Experimental results on two public datasets demonstrate that our proposed MSNEA provides state-of-the-art performance with a large margin compared with competitive baselines.

CCS CONCEPTS

• **Information systems** → Information integration; Multimedia and multimodal retrieval; Data mining.

KEYWORDS

Entity alignment, multi-modal learning, knowledge graph

ACM Reference Format:
Liyi Chen, Zhi Li, Tong Xu*, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022. Multi-modal Siamese Network for Entity Alignment. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539244

1 INTRODUCTION

Multi-modal knowledge graphs (MMKGs), which naturally organize real-world knowledge from visual, relational, and attribute perspectives, have drawn massive attention in various scenarios and motivated the development of numerous downstream applications [25, 37, 45]. Generally, multi-modal knowledge graphs are constructed from separate multi-modal corpora for diverse purposes. With the surge in demand for redundant multi-modal knowledge integration, multi-modal entity alignment [5, 18] has become one of the emerging tasks in this area.

In the literature, there are considerable efforts made in entity alignment to identify entities that denote the equivalent real-world concept from distinct knowledge graphs. Most existing methods directing towards conventional knowledge graphs are dedicated to

*Corresponding author.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539244

probing similarities in relation or graph structures. For instance, MTransE [6], IPTransE [41], and BootEA [27] follow translation-based embedding methods to mine semantics in multiple relations, while GCN-Align [30] and MuGNN [3] intend to model global graph structures. Nevertheless, they cannot probably fit the newly-emerged multi-modal knowledge graphs, which prompts recent research to concentrate on the exploitation of multi-modal knowledge for tackling entity alignment [5, 17, 18]. Although these methods have achieved promising performance, they still directly merge uni-modal feature embeddings, leaving the inter-modal effect in multi-modal entity alignment largely unexplored.

Indeed, the inter effect of different modality information serves a significant role in multi-modal representation and fusion [13, 29]. However, it is still an open issue with enormous challenges to incorporate the inter-modal effect in knowledge learning of multi-modal entity alignment. First, entities with analogous relations are more likely to be aligned, and other modalities can provide enhanced semantics to express the relations between entities [19, 34]. However, relational representation for entity alignment is often learned merely from fundamental graph structure and then directly merged, which ignores external information contained in other modalities. Thus, how to leverage the inter-modal effect to enrich relational representation remains pretty much open. Second, aligned entities possess attributes that are often sparse and heterogeneous, whereas all attributes are usually given the same degree of importance. This would introduce noise to render the entity alignment effect worse [5, 17]. Fortunately, introducing other modalities could bring extensive information to determine important attributes and make an enhanced inter-modal understanding of entity attributes. Therefore, how to use other modalities to capture valuable attributes for guaranteeing the effect of entity alignment becomes a significant challenge. Third, after learning inter-modal enhanced representations for multiple modalities, weak modalities may have an excessive influence on the overall modeling, thus reducing the inter-modal effect [5]. Along this line, how to achieve inter-modal enhancement fusion with avoiding the overwhelming impact of weak modalities is a critical issue.

To tackle these challenges, in this paper, we address the entity alignment problem for multi-modal knowledge graphs by proposing a Multi-modal Siamese Network for Entity Alignment (MSNEA) to incorporate inter-modal enhancement of entity representations. Specifically, we first learn visual, relational, and attribute features of entities to generate holistic entity representations for distinct knowledge graphs via a Multi-modal Knowledge Embedding (MKE) module. Particularly, considering the inter-modal effect, we develop a vision-guided relation learning mechanism and a vision-adaptive attribute learning mechanism, to achieve inter-modal enhancement entity representation in MKE module. On the one hand, we employ visual features to supplement abundant semantics in order to guide relational feature learning. On the other hand, we utilize visual features to adaptively assign attention weights to each attribute to capture valuable attributes for alignment. Afterwards, a Multi-modal Contrastive Learning (MCL) module is designed to avoid the overwhelming impact of weak modalities, to achieve inter-modal enhancement fusion for multi-modal entity alignment. Finally, we conduct extensive experiments on two public benchmarks. Experimental results clearly demonstrate the effectiveness and rationality of our proposed MSNEA. The main contributions of this paper are summarized as follows:

- In this paper, we present a novel study on inter-modal knowledge enhancement for multi-modal entity alignment task in the area of multi-modal knowledge graph.
- We propose an end-to-end framework named MSNEA, to address the multi-modal entity alignment task by introducing inter-modal enhancement mechanisms in multi-modal knowledge representation.
- Extensive experiments are conducted on two real-world datasets, and the results apparently indicate the effectiveness and rationality of MSNEA.

2 RELATED WORK

In this section, we briefly provide a review of the related work which could be grouped into three lines of literature: 1) *Entity Alignment*, 2) *Multi-modal Knowledge Graph*, 3) *Knowledge Graph Embedding*.

2.1 Entity Alignment

In recent years, entity alignment methods based on embedding techniques have been prevalent. These methods could be divided into two main categories, i.e., translation-based and GNNs-based methods. Translation-based entity alignment methods principally constrain the entity embeddings into a fixed distribution by translation-based knowledge graphs embedding methods. MTransE [6] learns transitions to map embedding vectors to their counterparts. Instead of training map embeddings, ITransE [41] and IPTransE [41] are iterative and parameter sharing methods. BootEA [27] also utilizes the iterative strategy to label new entity alignment as supervision. Then, in view of the degree difference between aligned entities, SEA [22] prevents entities with similar degrees from being aggregated into the same region. In the same year, due to the limited use of attributes, IMUSE [11] employs bivariate regression to merge relation and attribute alignment results. GNNs-based methods are originated with the intention to model global information of graphs via graph neural networks. GCN-Align [30] employs graph convolutional networks to model entities based on their neighborhood information. HMAN [36] combines multi-aspect information of entities by graph convolutional networks. MuGNN [3] designs multiple channels to reconcile structural differences. NAEA [42] incorporates neighborhood subgraph-level information of entities. RAGA [43] spreads entity information to the relations and then aggregates relation information back to entities. In general, these methods can be transferred to our model for further improvement, but cannot deal with multi-modal knowledge in real-world scenes.

2.2 Multi-modal Knowledge Graph

Although lots of multi-modal knowledge graphs have been constructed recently [17, 18], few attempts have been made towards entity alignment for multi-modal knowledge graphs. The first method PoE [18] deems the aligned link between entities as the relation of SameAs to perform link prediction. Then, MMEA [5] proposes a common space to fuse multi-modal representations by minimizing the distance between holistic embeddings of aligned entities. After that, in terms of visual advantages, EVA [17] leverages visual similarities to create an initial seed dictionary and applies iterative learning

to expand the set of training seeds. However, the aforementioned methods ignore the inter-modal effect and limit the exploitation of multi-modal knowledge. Besides, there are many studies putting emphasis on multi-modal knowledge graph embedding and applications. For multi-modal knowledge graph embedding, Xie et al. [34] use an attention mechanism to integrate image representations into an aggregated image-based representation. Pezeshkpour et al. [23] design encoders to embed multi-modal evidence types and decoders to generate multi-modal attributes. Mousselly-Sergieh et al. [19] define the translational energy of a relation triple. As for downstream applications, Zhu et al. [45] propose a multi-modal KB framework to handle an assortment of visual queries. Sun et al. [25] enhance recommender systems by conducting information propagation over multi-modal knowledge graphs. Yang et al. [37] integrate external multi-modal knowledge graph reasoning for dialogue systems. So far, many tasks associated with multi-modal knowledge graph have not been thoroughly investigated. Multi-modal knowledge graph is still attracting tremendous attention as a novel research area.

2.3 Knowledge Graph Embedding

Knowledge graph embedding aims to embed entities and relations into a low-dimensional vector space. These techniques designed for traditional knowledge graphs are mainly classified into three groups, i.e., translational distance, tensor factorization, and neural network methods. Translational distance models [2, 14, 31] learn relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. Typically, TransE [2] assumes that $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$, where $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$ are the representations of head entity, relation, and tail entity. However, it shows the limitation in dealing with complex relations, e.g, one-to-many or many-to-many. Therefore, many methods extend TransE to solve this issue. For example, TransH [31] uses translations on relation-specific hyperplanes. TransR [16] applies linear transformations to heterogeneous relation spaces. TransD [14] constructs a dynamic mapping matrix for each entity-relation pair by considering the diversity of entities and relations simultaneously. As for tensor factorization methods [21, 35], minimizing the squared loss amounts to factorization of a three-mode tensor represented by the knowledge graph. RESCAL [21] performs collective learning by the latent components of the factorization. DistMult [35] and HolE [20] are the extensions on RESCAL. DistMult [35] restricts the diagonal matrices and HolE [20] utilizes the circular correlation operation. Neural network methods [1, 4, 33, 40] employ neural network architectures to represent entities and relations. For instance, SME [1] conducts semantic matching. NTN [24] devises a relation-specific linear output layer. HC-GCN [33] aggregates all subgraphs along the temporal trajectory for representations. In this work, we adopt the representative translation-based embedding method, TransE, as the basis of relational knowledge representation.

3 METHODOLOGY

In this section, we first formally introduce the problem definition and the overview of our proposed MSNEA towards aligning the entities from two multi-modal knowledge graphs. Then, we elaborate technical details of two major modules in MSNEA.

Table 1: Mathematical notations in problem definition.

Notation	Explanation
G	The knowledge graph
E	The set of entities
R	The set of relations between entities
I	The set of entity images
A	The set of entity attributes
V	The set of values for entity attributes
P	The set of entity-image pairs
i	The image of an entity
T_R	The set of relation triples
h	The head entity in a relation triple
t	The tail entity in a relation triple
r	The relation in a relation triple
T_A	The set of attribute triples
a	The attribute of an entity in an attribute triple
v	The value for the attribute of an entity in an attribute triple
H	The set of alignment seeds

3.1 Problem Definition

3.1.1 Multi-modal Knowledge Graph. A multi-modal knowledge graph is formalized as $G = (E, R, I, A, V, P, T_R, T_A)$. Here, E, R, I, A, and V denote the sets of entities, relations, images, attributes, and values, respectively. $P = \{(e, i) \mid e \in E, i \in I\}$ is the set of entity-image pairs. $T_R = \{(h, r, t) \mid h, t \in E, r \in R\}$ refers to the set of relation triples. $T_A = \{(e, a, v) \mid e \in E, a \in A, v \in V\}$ means the set of attribute triples.

3.1.2 Multi-modal Entity Alignment. Given two multi-modal knowledge graphs $G = (E, R, I, A, V, P, T_R, T_A)$ and $G' = (E', R', I', A', V', P', T'_R, T'_A)$, the set of alignment seeds across two multi-modal knowledge graphs is defined as $H = \{(e, e') \mid e \in E, e' \in E', e \equiv e'\}$, where $\equiv$ represents the equivalence of two entities. The task of multi-modal entity alignment targets to match the counterpart entities e and e' describing the same concepts in the real world from distinct multi-modal knowledge graphs.

3.2 Framework Overview

In this paper, we propose a Multi-modal Siamese Network for Entity Alignment (MSNEA), to conquer the aforementioned challenges. As illustrated in Figure 1, our proposed MSNEA comprises two major components: 1) *Multi-modal Knowledge Embedding* (MKE) module to extract visual, relational, and attribute features with inter-modal enhancement mechanisms, to generate holistic entity representations; 2) *Multi-modal Contrastive Learning* (MCL) module to avoid the overwhelming impact of weak modalities, to achieve inter-modal enhancement fusion.

3.3 Multi-modal Knowledge Embedding

In multi-modal knowledge graphs, there are various modalities of knowledge to depict an entity, i.e., visual, relational, and attribute knowledge. In this module, we learn three modalities of feature representations with inter-modal enhancement mechanisms. Finally, holistic entity representations are generated.

3.3.1 Visual Modality. Multi-modal knowledge graph varies from a traditional knowledge graph mainly in that it integrates visual

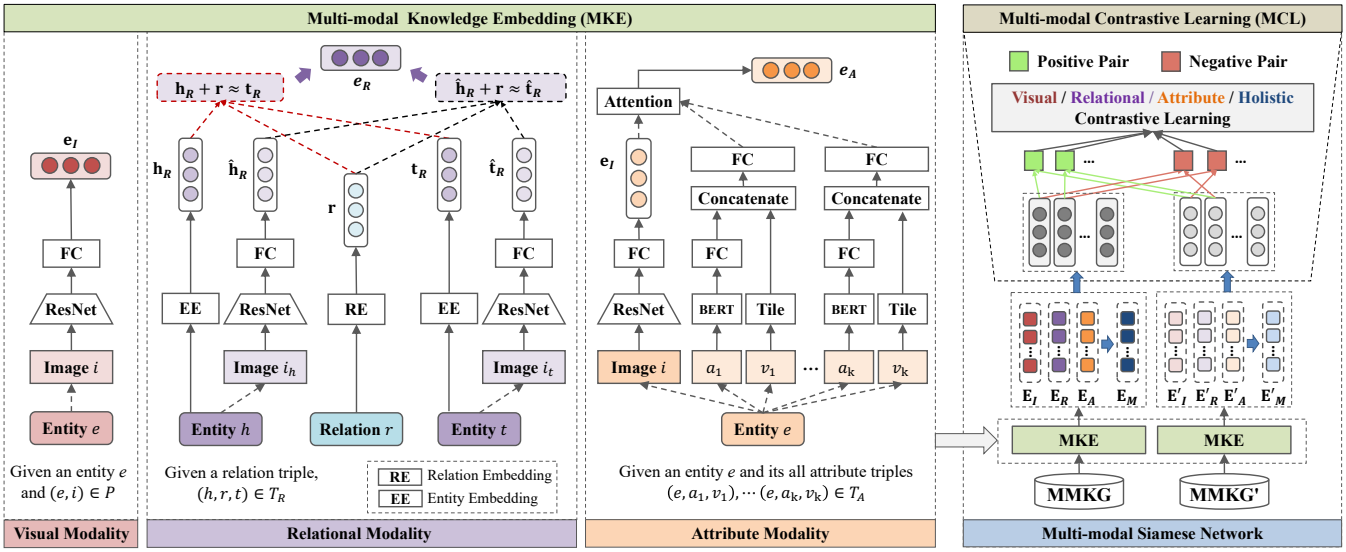

Figure 1: The framework overview of Multi-modal Siamese Network (MSNEA), which consists of two vital components, i.e., Multi-modal Knowledge Embedding (MKE) module and Multi-modal Contrastive Learning (MCL) module.

knowledge of entities. Visual knowledge can more intuitively reflect some entity characteristics such as appearance or behavior than other modal knowledge. Typically, entities referring to the same concept in the real world have more similarities in visual features. Therefore, visual knowledge improves the effect of entity alignment to some extent. In our model, image features are extracted by ResNet-50 [12]. Concretely, we drop the last fully connected layer and softmax layer to obtain the image embeddings of entities. Then, the linear transformation is made for image embeddings to get visual representations of entities. Given an entity-image pair $(e, i) \in P$ in multi-modal knowledge graphs, we generate a visual feature embedding of entity e as follows:

$$\mathbf{i}_e = ResNet(i), \tag{1}$$

$$\mathbf{e}_I = \mathbf{W}_1 \cdot \mathbf{i}_e + \mathbf{b}_1, \tag{2}$$

where $\mathbf{e}_I$ is the visual feature embedding of entity e.

3.3.2 Relational Modality. Relational knowledge forms traditional knowledge graphs and is also an essential component in multi-modal knowledge graphs. For the set of relation triples T_R in multi-modal knowledge graphs, we first expand T_R by swapping aligned entities in their involved relation triples [5, 39], as aligned entities hold identical relational facts in the real world. The swapping strategy makes aligned entities have closer representations in unified low-dimensional space. Then, we utilize the basic translation-based embedding model, TransE [2], to learn relational feature embeddings of entities. Given a relation triple $(h, r, t) \in T_R$, TransE motivates the embedding of the tail entity t to be close to the embedding of the head entity h plus the embedding of the relation r, i.e., $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$. In this way, TransE can preserve the structural information of entities in the embedding space, which reveals entities sharing similar neighbors should have close representations in the embedding space. The scoring function is defined as follows:

$$f_r(h, r, t) = \|\mathbf{h}_R + \mathbf{r} - \mathbf{t}_R\|_2^2, \tag{3}$$

where $\mathbf{h}_R$ is the relational feature embedding of head entity h, $\mathbf{r}$ is the embedding of relation r, $\mathbf{t}_R$ is the relational feature embedding of tail entity t, and $\|\cdot\|_2$ is the L_2-norm operation. Afterwards, we adopt the margin-based loss function [32] to differentiate positive examples and negative examples in the following:

$$\mathcal{L}_r = \sum_{\tau \in T_R} \sum_{\tau^- \in T_R^-} \max\left(0, \gamma_r + f_r(\tau) - f_r(\tau^-)\right), \tag{4}$$

where γ_r is the margin hyperparameter, and T_R^- is the set of negative examples. Here, negative examples are created by replacing either the head or tail entity with a random entity to corrupt the facts.

3.3.3 Vision-guided Relation Learning. Usually, it is highly probable that two entities exhibiting similar relations with other entities would be aligned. As a result, adequate comprehension and utilization of relational knowledge are important. Previous research [19, 34] manifests that entity features of other modalities have a similar pattern to relational features. Taking TransE [2] as an example, there exists $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ in a relation triple $(h, r, t) \in T_R$. And feature embeddings of other modalities also show this pattern, i.e., head entity embedding plus relation embedding approximately equal to tail entity embedding. In particular, visual knowledge contains rich semantics which can be used to guide the learning of relations between entities. Consequently, we incorporate semantic information in visual knowledge to generate an enhanced relational representation. With regard to the relation triple $(h, r, t) \in T_R$ and entity-image pairs $(h, i_h), (t, i_t) \in P$, we first adopt a linear transformation to project image features of head and tail entity derived from ResNet-50 into the corresponding space:

$$\hat{\mathbf{h}}_R = \mathbf{W}_2 \cdot \mathbf{i}_h + \mathbf{b}_2, \tag{5}$$

$$\hat{\mathbf{t}}_R = \mathbf{W}_2 \cdot \mathbf{i}_t + \mathbf{b}_2. \tag{6}$$

Following that, we employ the score function and the loss function in the same manner as below:

$$f_i(h, r, t) = \left\| \hat{\mathbf{h}}_R + \mathbf{r} - \hat{\mathbf{t}}_R \right\|_2^2, \tag{7}$$

$$\mathcal{L}_i = \sum_{\tau \in T_R} \sum_{\tau^- \in T_R^-} \max\left(0, \gamma_i + f_i(\tau) - f_i(\tau^-)\right), \tag{8}$$

where $\mathbf{r}$ is the embedding of relation r, and γ_i is the margin hyperparameter. Through the above vision-guided relation learning, we can learn a better relation embedding $\mathbf{r}$ to finally enhance relational feature representations $\mathbf{h}_R$ and $\mathbf{t}_R$.

3.3.4 Attribute Modality. In multi-modal knowledge graphs, attribute knowledge provides the attribute names and numerical values ascribed to entities. Analogous to the set of relation triples T_R, we adopt the swapping strategy to process the set of attribute triples T_A. Given an entity $e \in E$ and its all attribute triples $(e, a_1, v_1), (e, a_2, v_2), \ldots, (e, a_k, v_k) \in T_A$, we need to generate embeddings of these attributes and values. Specifically, for each attribute $a = \{w_1, w_2, \ldots\}$, the word sequence is inputted into BERT [8] to obtain word embeddings. Then, we average their embeddings and add a linear layer, to embed the attribute as $\mathbf{a}$:

$$\mathbf{a} = \mathbf{W}_3 \cdot Avg(BERT(w_1, w_2, \ldots)) + \mathbf{b}_3, \tag{9}$$

where $Avg(\cdot)$ denotes the average operation. Meanwhile, we apply the sigmoid function to normalize each numerical value v and tile the value to form the embedding $\mathbf{v}$ [28]. Next, we concatenate embeddings of attribute a and corresponding value v to represent the attribute triple. After the concatenation, another linear layer is appended to reduce the dimension as follows:

$$\mathbf{s} = \mathbf{W}_4 \cdot \mathbf{a} \| \mathbf{v} + \mathbf{b}_4, \tag{10}$$

where $\|$ is the concatenation operation. Now, we can obtain all attribute embeddings $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k$ belonging to entity e.

3.3.5 Vision-adaptive Attribute Learning. Aligned entities tend to have certain identical attributes. However, entity attributes are often sparse and heterogeneous. Some studies [17, 36] have shown that due to the sparsity of attribute quantity and the heterogeneity of attribute names, absorbing attributes into entity alignment task may pollute entity representations as noise, leading to unsatisfactory results. Moreover, it is unreasonable to assign the same weight to all attributes, as it would exacerbate the noise situation. In this case, we need to choose valuable attributes for multi-modal entity alignment task. In view of abundant semantics contained in visual knowledge, it is intuitive to utilize visual features to assign attention weights to each attribute triple, which allows adaptively capturing valuable attributes for alignment. We calculate the sum of these weighted attribute embeddings to represent the attribute feature of entity e in the following:

$$\alpha_j = \frac{\exp\left(\mathbf{e}_I^T \mathbf{s}_j\right)}{\sum_{c=1}^{k} \exp\left(\mathbf{e}_I^T \mathbf{s}_c\right)}, \tag{11}$$

$$\mathbf{e}_A = \sum_{j=1}^{k} \alpha_j \mathbf{s}_j, \tag{12}$$

where α_j denotes the attention weight assigned to $\mathbf{s}_j$, and $\mathbf{e}_A$ is the attribute feature embedding of entity e.

3.3.6 Holistic Representation. Depending on the above knowledge embedding with inter-modal enhancement mechanisms, we can learn three types of feature representations of entities in multi-modal knowledge graphs, i.e., relational, visual, and attribute feature embeddings. Given an entity $e \in E$, and its relational, visual, and attribute feature embeddings $\mathbf{e}_R \in \mathbf{E}_R$, $\mathbf{e}_I \in \mathbf{E}_I$, and $\mathbf{e}_A \in \mathbf{E}_A$, we concatenate three types of feature embeddings to generate a holistic entity representation:

$$\mathbf{e}_M = \mathbf{e}_R \| \mathbf{e}_I \| \mathbf{e}_A, \tag{13}$$

where $\mathbf{e}_M \in \mathbf{E}_M$ is the holistic entity embedding of entity e, and $\|$ is the concatenation operation.

3.4 Multi-modal Contrastive Learning

In the process of generating a holistic entity representation, weak modalities may have an excessive influence on the overall modeling, thus reducing the inter-modal effect [5, 17]. To achieve inter-modal enhancement fusion with avoiding the overwhelming impact of weak modalities, we design a multi-modal contrastive learning module from multiple perspectives of holistic entity representations and each uni-modal feature representations.

To be specific, we compute the cosine similarity of entity embeddings from different knowledge graphs as d and take the form of the contrastive loss [10] which is commonly used in Siamese networks. It has the ability to effectively deal with the relationship between paired data in Siamese neural networks. We minimize the loss so that positive entity pairs are encoded to similar representations whereas negative entity pairs are encoded to dissimilar representations. Here, positive entity pairs refer to two aligned entities, and negative entity pairs refer to two unaligned entities. The contrastive loss is formulated below:

$$\begin{aligned} \mathcal{L}_{cl}(E, E') = \frac{1}{2N} \sum_{n=1}^{N} & (1 - y_n) d^2(e_n, e'_n) \\ & + y_n \max(\gamma_{cl} - d(e_n, e'_n), 0)^2, \end{aligned} \tag{14}$$

where y_n is the label of entity pair (1 for positive examples, else 0), γ_{cl} is the margin hyperparameter, and $e_n \in E$ and $e'_n \in E'$ are entities in knowledge graph G and G'. Currently, most prior work [5, 18] solely minimizes the distance between holistic entity representations. Nevertheless, we deem that feature contrast in each modality for fusion facilitates preserving specific alignment information of each modality. Therefore, we adopt the contrastive learning from multiple perspectives of holistic entity representations and each uni-modal feature representations as follows:

$$\mathcal{L}_{mcl} = \mathcal{L}_{cl}(\mathbf{E}_M, \mathbf{E}'_M), \tag{15}$$

$$\mathcal{L}_{ucl} = \mathcal{L}_{cl}(\mathbf{E}_I, \mathbf{E}'_I) + \mathcal{L}_{cl}(\mathbf{E}_R, \mathbf{E}'_R) + \mathcal{L}_{cl}(\mathbf{E}_A, \mathbf{E}'_A). \tag{16}$$

Here, $\mathcal{L}_{mcl}$ and $\mathcal{L}_{ucl}$ are the contrastive learning losses for holistic entity representations and each uni-modal feature representations. $\mathbf{E}_I$, $\mathbf{E}_R$ and $\mathbf{E}_A$ are sets of visual, relational and attribute feature representations of entities in E, and $\mathbf{E}'_I$, $\mathbf{E}'_R$ and $\mathbf{E}'_A$ are sets of visual, relational and attribute feature representations of entities in E'.

Eventually, in training stage, we design the overall objective function in the following:

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_i + \mathcal{L}_{ucl} + \mathcal{L}_{mcl}, \tag{17}$$

and minimize $\mathcal{L}$ to update parameters in our model through the backpropagation.

4 EXPERIMENT

In this section, we evaluate MSNEA by conducting extensive experiments on two real-world datasets. First, we describe the experimental setup in detail. Second, we introduce the selected baselines for comparison. Finally, we present the experimental results to demonstrate the effectiveness and rationality of our proposed MSNEA.

4.1 Experimental Setup

4.1.1 Datasets. In the experiments, we use two public datasets FB15K-DB15K and FB15K-YG15K, which are most representative in multi-modal entity alignment task [5, 18]. The statistics of two real-world datasets are shown in Table 2. FB15K is a benchmark dataset widely used in link prediction task. Entities from DBpedia and YAGO aligned with FB15K are extracted through the SameAs links contained in DBpedia and YAGO dumps. They are utilized to build DB15K and YG15K datasets. Since the degree of a node relates to the probability of an entity appearing in a subsampled version of a knowledge graph, other entities highly connected to the aligned entities are also included in DB15K and YG15K. Furthermore, three multi-modal knowledge graphs are populated with numerical attributes and images of entities.

Table 2: Statistics of three multi-modal knowledge graphs.

Dataset	Entity	Relation Triple	Attribute Triple	Image	Seed
FB15K	14,951	592,213	29,395	13,444	-
DB15K	12,842	89,197	48,080	12,837	12,846
YG15K	15,404	122,886	23,532	11,194	11,199

4.1.2 Evaluation Metrics. Following related work [5], we adopt cosine similarity to measure the alignment probability between entities from different knowledge graphs and select Hits@*n*, MRR, and MR as metrics to evaluate all the models. Based on similarity calculating, Hits@*n* is the proportion of correctly aligned entities ranked in the top *n* list, MRR represents the mean reciprocal rank of correctly aligned entities, and MR denotes the mean rank of correctly aligned entities. Higher Hits@*n* and MRR suggest the better performance of the method, whereas lower MR indicates it.

4.1.3 Model Configurations. We initialize all weight matrices with Xavier Normal initializer [9]. ResNet-50 is pre-trained on ImageNet [7]. And we select the uncased base version of BERT [1] pre-trained on BookCorpus [44] and Wikipedia [2]. For fast training, we freeze the parameters in ResNet-50 and BERT in the following stage. Similar to MMEA [5] and PoE [18], we set the embedding of missing images as a zero vector to avoid the influence on other modalities. The dimensions of $\mathbf{e}_R$, $\mathbf{e}_I$, and $\mathbf{e}_A$ are 100, and the dimensions of $\mathbf{r}$, $\mathbf{a}$, $\mathbf{v}$, and $\mathbf{s}$ are also 100. Besides, in the loss functions, we set γ_r as 1, γ_i as 1 and γ_{cl} as 2. We adopt the mini-batch method with the batch size of 5000. We set the learning rate as 0.001 and minimize the loss function by Adam optimizer [15]. The max epochs are 200 and 400 on FB15K-DB15K and FB15K-YG15K datasets. We select the model with the best performance on the validation set to test. Our model is implemented with the deep learning framework PyTorch [3]. The experiments are conducted on a server with two Intel Xeon Silver 4214R CPUs @ 2.40GHz, four NVIDIA GeForce RTX 3090 GPUs, and 256 GB RAM memory. The code for MSNEA is available at https://github.com/liyichen-cly/MSNEA.

4.2 Compared Methods

In order to verify the effectiveness of MSNEA, we select several representative and competitive methods as baselines for comparison. They can be classified into two categories: traditional and multi-modal entity alignment methods.

Traditional Methods:

- **MTransE** [6] embeds different knowledge graphs in a separated embedding space and provides transitions for entities to map the aligned entities.
- **IPTransE** [41] iteratively adds newly aligned entity pairs into a set for soft alignment and adopts a parameter sharing strategy for different knowledge graphs.
- **GCN-Align** [30] combines structural and attribute information via graph convolutional networks.
- **BootEA** [27] iteratively labels likely alignment as training data and reduces error accumulation during iterations by an alignment editing method.
- **SEA** [22] uses both labeled entities and the abundant unlabeled entity information for the alignment and perceives the difference in degree of points with adversarial training.
- **IMUSE** [11] employs a bivariate regression model to learn the respective weights of relation and attribute similarities for a result combination.
- **HyperKA** [26] proposes a hyperbolic relational graph neural network to explore low-dimensional hyperbolic embeddings for entity alignment.
- **RAGA** [43] adopts the self-attention mechanism to aggregate relation information back to entities and proposes a global alignment algorithm to make one-to-one entity alignments with a fine-grained similarity matrix.
- **RAC** [38] devises an unsupervised contrastive loss to contrast different views of entity representations and augment supervision signals by exploiting the vast unlabeled data.

Multi-modal Methods:

- **PoE** [18] judges the SameAs links between aligned entities by assigning high probability to true triples and low probability to false triples. Then, it defines overall probability distribution as the product of all uni-modal experts.
- **MMEA** [5] first generates the entity representations of relational, visual, and numerical knowledge, and then migrates

[1] https://github.com/huggingface/transformers
[2] https://en.wikipedia.org
[3] https://pytorch.org/

Table 3: The performance of entity alignment methods with 20% alignment seeds on FB15K-DB15K and FB15K-YG15K.

Method	FB15K-DB15K					FB15K-YG15K				
	Hits@1	Hits@5	Hits@10	MR	MRR	Hits@1	Hits@5	Hits@10	MR	MRR
MTransE	0.359	1.414	2.492	1239.465	0.014	0.308	0.988	1.783	1183.251	0.011
IPTransE	3.985	11.226	17.277	387.512	0.086	3.079	9.505	14.443	522.235	0.070
GCN-Align	4.311	10.956	15.548	810.648	0.082	2.270	7.209	10.736	1109.845	0.053
BootEA	32.319	49.877	57.948	205.532	0.410	23.384	37.417	44.548	272.120	0.307
SEA	16.974	33.464	42.512	191.903	0.255	14.084	28.694	37.147	207.236	0.218
IMUSE	17.602	34.677	43.523	182.843	0.264	8.094	19.241	25.654	397.571	0.142
HyperKA	13.653	24.948	30.724	712.154	0.195	16.863	29.545	34.882	738.034	0.232
RAGA	6.101	11.482	14.965	1518.991	0.092	5.547	10.268	12.801	1337.759	0.082
RAC	20.285	35.954	43.205	453.313	0.281	15.066	28.103	34.494	501.795	0.216
PoE	12.0	-	25.6	-	0.167	10.9	-	24.1	-	0.154
MMEA	26.482	45.133	54.107	124.807	0.357	23.391	39.764	47.999	147.441	0.317
EVA	55.590	66.644	71.587	139.995	0.609	10.257	21.663	27.790	616.789	0.164
MSNEA	**65.268**	**76.847**	**81.214**	**54.025**	**0.708**	**44.288**	**62.554**	**69.831**	**85.074**	**0.529**

multi-modal knowledge embeddings from separate spaces to a common space for fusion.

- **EVA** [17] highlights the visual similarity of entities. It fuses multi-modal information into a joint embedding with an attention-based modality weighting scheme and allows the alignment model to automatically adjust modality weights.

For all baselines, we report the original results from the literature or run public implementations from GitHub. The results of MTransE, IPTransE, GCN-Align, SEA, IMUSE, PoE, and MMEA are reported by Chen et al. [5]. And the results of BootEA [4], HyperKA [5], RAGA [6], RAC [7], and EVA [8] are our experimental results with default settings.

4.3 Experimental Results

4.3.1 Performance Comparison. To demonstrate the effectiveness of our proposed model, we compare MSNEA with several state-of-the-art methods on the entity alignment task for multi-modal knowledge graphs. Table 3 shows the performance of all methods with 20% alignment seeds for training on FB15K-DB15K and FB15K-YG15K datasets. From the overview, MSNEA achieves the best performance on multi-modal entity alignment task. Specifically, we have the following observations. First, MSNEA significantly outperforms these baselines in terms of all evaluation metrics. On both datasets, Hits@1 and MRR are at least improved by 17.4% and 16.2%, and MR is at least decreased by 42.3%. This apparently demonstrates the effectiveness of our proposed model. Second, superior results are presented by multi-modal methods instead of traditional methods in most cases. When compared with traditional methods, MSNEA achieves at least 89.3% and 72.3% improvement in Hits@1 and MRR. But actually, all multi-modal methods only adopt simple models such as TransE and GCN to represent relational knowledge and already achieve promising results. This further implies the importance of exploiting multi-modal knowledge for entity alignment task. Third, MSNEA is more effective in leveraging multi-modal knowledge for multi-modal entity alignment. MSNEA gains 53.268%, 38.786%, and 9.678% absolute improvement in Hits@1 on FB15K-DB15K as compared to PoE, MMEA, and EVA, respectively. This demonstrates the effectiveness of inter-modal enhancement mechanisms. Besides, we notice a sharp decline in the performance of EVA on FB15K-YG15K. The possible reason is that its modality weighting scheme falls short of fusing multi-modal knowledge well, resulting in a strong influence by weak modality.

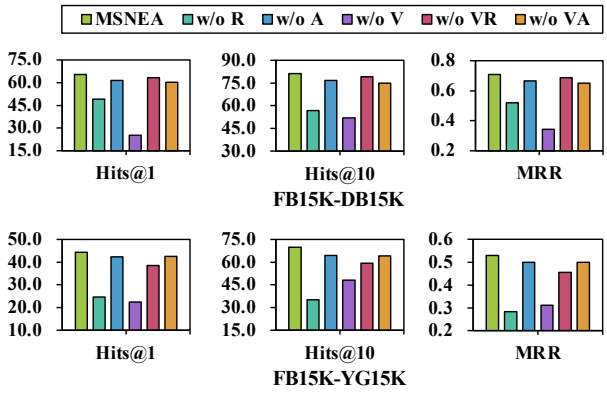

Figure 2: Ablation study on two real-world datasets.

4.3.2 Ablation Study. For investigating the impact of each part in our proposed MSNEA, we design two groups of variants for ablation study: 1) MSNEA without a modality, including relation, attribute, and vision, i.e., w/o R, w/o A, and w/o V; 2) MSNEA without the vision-guided relation learning and without the vision-adaptive attribute learning, i.e., w/o VR and w/o VA. The experimental results are illustrated in Figure 2. From the first group of variants, we can observe that three modalities all make contributions to entity alignment. However, incorporating attribute modality shows relatively less improvement on both datasets. This may be a consequence of

[4]https://github.com/nju-websoft/OpenEA
[5]https://github.com/nju-websoft/HyperKA
[6]https://github.com/zhurboo/RAGA
[7]https://github.com/DexterZeng/RAC
[8]https://github.com/cambridgeltl/eva

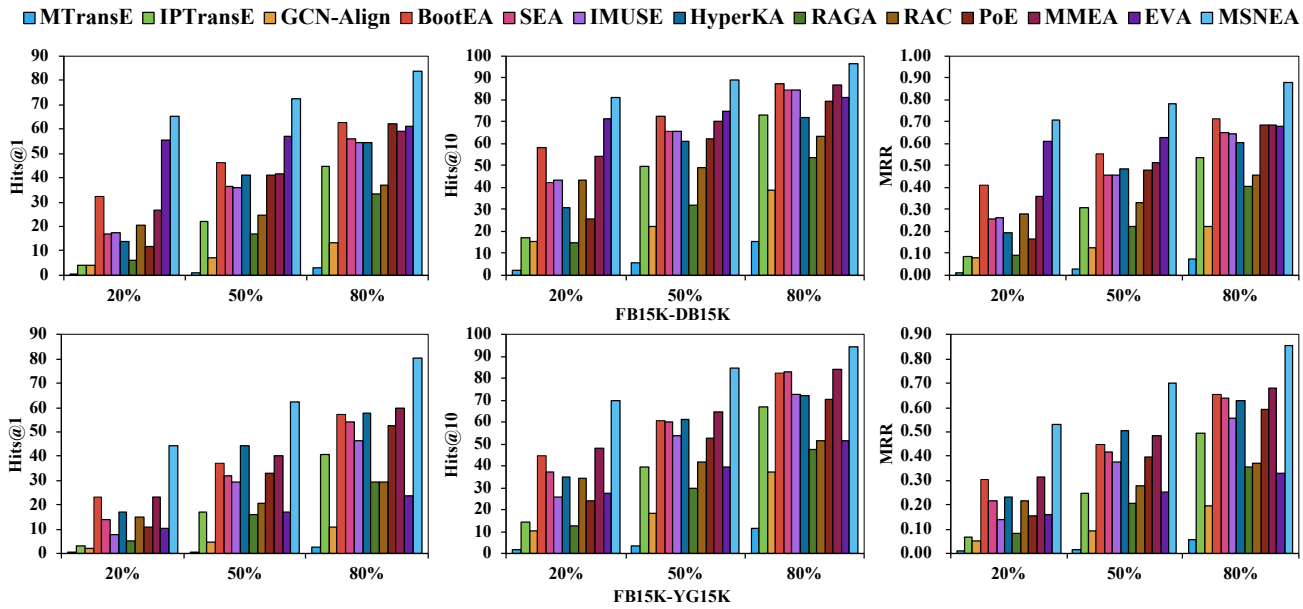

Figure 3: Comparison results with different proportions of alignment seeds on FB15K-DB15K and FB15K-YG15K.

Table 4: The verification of the contrastive learning approach which is designed from multiple perspectives.

Method	FB15K-DB15K			FB15K-YG15K		
	Hits@1	Hits@10	MR	Hits@1	Hits@10	MR
MSNEA	**65.268**	**81.214**	**54.025**	**44.288**	**69.831**	**85.074**
w/o VC	64.459	80.159	60.047	27.989	54.355	157.628
w/o RC	64.257	80.119	62.234	39.027	60.095	151.722
w/o AC	45.751	56.758	405.063	41.015	64.181	114.302
w/o MC	64.627	80.59	55.401	43.522	65.682	117.339

the noise caused by attributes as mentioned before. Moreover, we can find that the effectiveness of each modality on the two datasets is different. According to Hits@10 and MRR, w/o R performs worse than w/o V on FB15K-YG15K while not on FB15K-DB15K. Although FB15K-YG15K contains less useful visual knowledge, MSNEA still shows stable performance, which indicates that visual knowledge does not dominate the final results of MSNEA. Regarding the second group of variants, w/o VR and w/o VA simultaneously surpass MSNEA, which validates the effectiveness of vision-guided relation learning and vision-adaptive attribute learning mechanisms. Besides, it is worth noting that w/o A outperforms w/o VA in most cases. This not only explains that setting an inappropriate strategy to utilize attribute knowledge may have a negative influence [36], but also proves the effectiveness of vision-adaptive attribute learning mechanism again.

4.3.3 Analysis on Contrastive Learning. In MCL module, we develop a novel contrastive learning approach from multiple perspectives, i.e., holistic entity representations and each uni-modal feature representations. To confirm the effectiveness and rationality of MCL module, we concretely remove each contrastive loss in our MSNEA, i.e., $\mathcal{L}_{cl}(\mathbf{E}_I, \mathbf{E}'_I)$, $\mathcal{L}_{cl}(\mathbf{E}_R, \mathbf{E}'_R)$, $\mathcal{L}_{cl}(\mathbf{E}_A, \mathbf{E}'_A)$, and $\mathcal{L}_{cl}(\mathbf{E}_M, \mathbf{E}'_M)$. We denote these variants as w/o VC, w/o RC, w/o AC, and w/o MC, respectively. Table 4 reports the results of above variants on two datasets. On the whole, the performance of these variants is inconsistent on two datasets. Dropping the attribute contrastive learning increases MR from 54.025 to 405.063 on FB15K-DB15K dataset. Relatively speaking, other contrastive learning losses have less influence. Nevertheless, on FB15K-YG15K dataset, visual contrastive learning develops the most critical effect. Compared with MSNEA, Hits@1 of w/o VC is reduced by 36.8%. This reveals that MSNEA can avoid the excessive influence of different weak modalities. Consequently, the combination of multi-perspective contrastive learning is proved to be rational and effective.

4.3.4 Seed Sensitivity. For the sake of assessing the sensitivity of these entity alignment methods to pre-aligned entities, we divide 20%, 50%, and 80% alignment seeds as training sets in line with existing studies [5, 18]. Figure 3 shows the comparison results of all models with different proportions of alignment seeds on two datasets. From the overview, we can find that MSNEA substantially achieves the best results on both datasets in all metrics and proportions. In concrete terms, many methods manifest poor performance with only 20% training seeds. With the increase of the proportion of training seeds, the effects of some methods, e.g., IPTransE and PoE, gradually become better and have a clear gap compared to the condition with limited alignment seeds. This is because these methods depend more on pre-aligned entities. Instead, MTransE and GCN-Align still have poor results due to the simplicity of their network structures. By comparison, MSNEA does not have a strong dependency on pre-aligned entities and can perform well even with a limited number of pre-aligned entities for training.

5 CONCLUSION

In this paper, we proposed a novel Multi-modal Siamese Network for Entity Alignment (MSNEA), to comprehensively leverage multi-modal knowledge with the exploitation of inter-modal effect. Specifically, we first devised the multi-modal knowledge embedding module to extract visual, relational, and attribute features of entities, to generate holistic entity representations. In view of inter-modal effect, we integrated visual features to guide relational feature learning and adaptively assign attention weights to capture valuable attributes for alignment, to achieve inter-modal enhancement representation. Then, we developed the multi-modal contrastive learning module to contrast positive entity pairs with negative entity pairs from multiple perspectives of holistic entity representations and each uni-modal feature representations, which avoids the overwhelming impact of weak modalities and achieves inter-modal enhancement fusion. Extensive experiments were conducted on two public datasets, where the experimental results demonstrated the effectiveness and rationality of MSNEA.

6 ACKNOWLEDGMENTS

This research was partially supported by grants from the National Natural Science Foundation of China (No.62072423) and the USTC Research Funds of the Double First-Class Initiative (No.YD2150002009). Also, this research was partially supported by the Huawei-USTC Joint Innovation Program. We sincerely acknowledge all the people who gave the support and help.

REFERENCES

[1] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2014. A semantic matching energy function for learning with multi-relational data. *Machine Learning* (2014).
[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proc. of NeurIPS*.
[3] Yixin Cao, Zhiyuan Liu, Chengjiang Li, Juanzi Li, and Tat-Seng Chua. 2019. Multi-Channel Graph Neural Network for Entity Alignment. In *Proc. of ACL*.
[4] Liyi Chen, Zhi Li, Weidong He, Gong Cheng, Tong Xu, Nicholas Jing Yuan, and Enhong Chen. 2022. Entity Summarization via Exploiting Description Complementarity and Salience. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
[5] Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020. MMEA: Entity Alignment for Multi-modal Knowledge Graph. In *Proc. of KSEM*.
[6] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual Knowledge Graph Embeddings for Cross-Lingual Knowledge Alignment. In *Proc. of IJCAI*.
[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.
[9] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of AISTATS*.
[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proc. of CVPR*.
[11] Fuzhen He, Zhixu Li, Yang Qiang, An Liu, Guanfeng Liu, Pengpeng Zhao, Lei Zhao, Min Zhang, and Zhigang Chen. 2019. Unsupervised entity alignment using attribute triples and relation triples. In *Proc. of DASFAA*.
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*.
[13] Weidong He, Zhi Li, Dongcai Lu, Enhong Chen, Tong Xu, Baoxing Huai, and Jing Yuan. 2020. Multimodal Dialogue Systems via Capturing Context-aware Dependencies of Semantic Elements. In *Proc. of ACM MM*.
[14] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proc. of ACL*.
[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.
[16] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proc. of AAAI*.
[17] Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual Pivoting for (Unsupervised) Entity Alignment. In *Proc. of AAAI*.
[18] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. MMKG: multi-modal knowledge graphs. In *Proc. of ESWC*.
[19] Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*.
[20] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proc. of AAAI*.
[21] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proc. of ICML*.
[22] Shichao Pei, Lu Yu, Robert Hoehndorf, and Xiangliang Zhang. 2019. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *Proc. of WWW*.
[23] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. In *Proc. of EMNLP*.
[24] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Proc. of NeurIPS*.
[25] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proc. of CIKM*.
[26] Zequn Sun, Muhao Chen, Wei Hu, Chengming Wang, Jian Dai, and Wei Zhang. 2020. Knowledge Association with Hyperbolic Knowledge Graph Embeddings. In *Proc. of EMNLP*.
[27] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In *Proc. of IJCAI*.
[28] Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro A. Szekely. 2021. Representing Numbers in NLP: a Survey and a Vision. In *Proc. of NAACL*.
[29] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. 2020. Deep multimodal fusion by channel exchanging. *Proc. of NeurIPS*.
[30] Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proc. of EMNLP*.
[31] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proc. of AAAI*.
[32] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. 2017. Sampling matters in deep embedding learning. In *Proc. of ICCV*.
[33] Shiwei Wu, Joya Chen, Tong Xu, Liyi Chen, Lingfei Wu, Yao Hu, and Enhong Chen. 2021. Linking the Characters: Video-oriented Social Graph Generation via Hierarchical-cumulative GCN. In *Proc. of ACM MM*.
[34] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied Knowledge Representation Learning. In *Proc. of IJCAI*.
[35] 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 *Proc. of ICLR*.
[36] Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning Cross-Lingual Entities with Multi-Aspect Information. In *Proc. of EMNLP*.
[37] Shiquan Yang, Rui Zhang, Sarah Erfani, and Jey Han Lau. 2021. UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases into End-to-End Task-Oriented Dialogue Systems. In *Proc. of IJCAI*.
[38] Weixin Zeng, Xiang Zhao, Jiuyang Tang, and Changjun Fan. 2021. Reinforced Active Entity Alignment. In *Proc. of CIKM*.
[39] Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view Knowledge Graph Embedding for Entity Alignment. In *Proc. of IJCAI*.
[40] Guanqi Zhu, Hanqing Tao, Han Wu, Liyi Chen, Ye Liu, Qi Liu, and Enhong Chen. 2022. Text Classification via Learning Semantic Dependency and Association. In *Proc. of IJCNN*.
[41] Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative Entity Alignment via Joint Knowledge Embeddings. In *Proc. of IJCAI*.
[42] Qiannan Zhu, Xiaofei Zhou, Jia Wu, Jianlong Tan, and Li Guo. 2019. Neighborhood-Aware Attentional Representation for Multilingual Knowledge Graphs. In *Proc. of IJCAI*.
[43] Renbo Zhu, Meng Ma, and Ping Wang. 2021. RAGA: Relation-Aware Graph Attention Networks for Global Entity Alignment. In *Proc. of KDD*.
[44] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proc. of ICCV*.
[45] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base system for answering visual queries. *Computer Science* (2015).