Attribute-Consistent Knowledge Graph Representation Learning for Multi-Modal Entity Alignment

Qian Li
School of Computer Science and Engineering, BDBC, Beihang University
Beijing, China
liqian@act.buaa.edu.cn

Shu Guo
National Computer Network Emergency Response Technical Team/Coordination Center of China
Beijing, China
guoshu@cert.org.cn

Yangyifei Luo
School of Computer Science and Engineering, BDBC, Beihang University
Beijing, China
luoyangyifei@buaa.edu.cn

Cheng Ji
School of Computer Science and Engineering, BDBC, Beihang University
Beijing, China
jicheng@act.buaa.edu.cn

Lihong Wang
National Computer Network Emergency Response Technical Team/Coordination Center of China
Beijing, China
wlh@cert.org.cn

Jiawei Sheng
Institute of Information Engineering, Chinese Academy of Sciences, School of Cyber Security, UCAS
Beijing, China
shengjiawei@iie.ac.cn

Jianxin Li[*]
SCSE, Beihang University, Zhongguancun Lab
Beijing, China
lijx@act.buaa.edu.cn

ABSTRACT

The multi-modal entity alignment (MMEA) aims to find all equivalent entity pairs between multi-modal knowledge graphs (MMKGs). Rich attributes and neighboring entities are valuable for the alignment task, but existing works ignore contextual gap problems that the aligned entities have different numbers of attributes on specific modality when learning entity representations. In this paper, we propose a novel attribute-consistent knowledge graph representation learning framework for MMEA (ACK-MMEA) to compensate the contextual gaps through incorporating consistent alignment knowledge. Attribute-consistent KGs (ACKGs) are first constructed via multi-modal attribute uniformization with merge and generate operators so that each entity has one and only one uniform feature in each modality. The ACKGs are then fed into a relation-aware graph neural network with random dropouts, to obtain aggregated relation representations and robust entity representations. In order to evaluate the ACK-MMEA facilitated for entity alignment, we specially design a joint alignment loss for both entity and attribute evaluation. Extensive experiments conducted on two benchmark datasets show that our approach achieves excellent performance compared to its competitors.

[*]Corresponding author

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; **Natural language processing**; • **Theory of computation** → **Semantics and reasoning**.

KEYWORDS

Entity alignment, Multi-modal knowledge graph representation

1 INTRODUCTION

Knowledge graphs (KGs) have become a popular data structure for representing factual knowledge in form of RDF triples. Recently, there is a growing trend to incorporate multi-modal information into KGs, i.e., Multi-Modal Knowledge Graphs (MMKGs), which support various cross-modal tasks, e.g., recommendation systems [19, 28] and question answering systems [11, 27]. However, MMKGs often suffer from low coverage and incompleteness. To improve the coverage of these MMKGs, a viable approach termed as multi-modal entity alignment (MMEA) is proposed to identify the equivalent entity pairs (i.e., alignment seeds) in different MMKGs, by integrating the attribute information of text and image. In this way, MMKGs can obtain useful knowledge from other KG.

Although the rich attributes and neighboring entities in MMKGs provide valuable pieces of evidence for MMEA [14], the inevitable heterogeneity of MMKGs makes it difficult to learn and fuse knowledge representations from distinct modalities. A series of effective
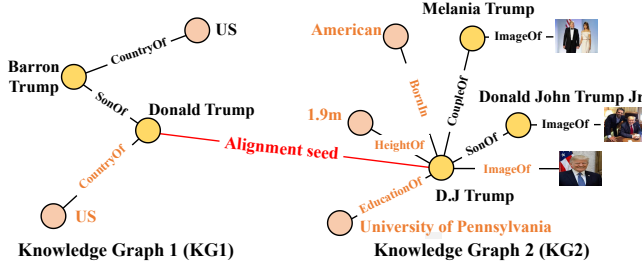
Figure 1: An example of the MMEA task between KG1 and KG2. The yellow and orange circles are the entity node and attribute nodes, respectively.

methods have been developed to conquer these challenges, and the detailed description is in Appendix A. The PoE method [14] composited representations of entities by concatenating all modality features, which could not capture the potential interactions among heterogeneous modalities and therefore limited the performance of MMEA. Later works [3, 7] designed multi-modal fusion modules to properly integrate attributes and entities, in order to better predict alignments according to aggregated embeddings. All of these methods would learn entity representations by harnessing their whole associated attributes and neighboring entities. Nevertheless, they ignore the contextual gaps between entity pairs and in turn constrain the effectiveness of the entity alignment.

The contextual gap, which means entities may associate with different numbers of attributes or even lack some modalities, is inevitable due to information redundancy or absence. Such inconsistencies between the equivalent entity pairs make alignments error-prone. Figure 1 illustrates a toy example of the contextual gap in MMEA. (a) Difference in the number of attributes. The entity *Donald Trump* in KG1 associates with only one text attribute *US*. With such limited contextual information, it is not easy to determine the identity of entity *Donald Trump*. In contrast, the identity of entity *D.J Trump* in KG2 will be more specific as he contains richer text attributes, which causes that similarity of text attributes (between *US* and *American*) is diluted by existing aggregation-based approaches. The contextual gaps caused by different numbers of text attributes make it hard to obtain the attribute-consistent alignment knowledge and judge that both actually refer to the same real-world identity. (b) Lack of modal attribute. Missing attributes also leads to contextual gap problems since unique attributes are neglected for alignment on existing fusion baselines, as such missing image attribute of *Donald Trump* makes *D.J Trump* difficult to align with it.

To overcome the above challenges, we propose a novel **M**ulti-**M**odal **E**ntity **A**lignment framework based on **A**ttribute-**C**onsistent **K**nowledge graph representation learning , termed as ACK-MMEA[1]. Specifically, for the attribute information, we design a multi-modal attribute uniformization method to obtain attribute-consistent KGs (ACKGs) with one uniform attribute for each modality of all entities. That means in the ACKGs every entity will possess only one attribute for each modality (i.e., one text and one image attribute). To generate such an attribute-consistent MMKG, we devise merge

and generate operators for each entity to deal with the attribute redundancy and absence respectively. The former is to compress associated multiple attributes into one with an attention-based approach to filter out the noise attributes, while the latter is to expand a new attribute if the entity has no attribute in specific modality. We further devise a ConsistGNN to enforce consistent attribute aggregation for entity representation. Given a relation triple in the above ACKGs, ConsistGNN first obtains an aggregated relation representation by simultaneously integrating relational features under every modality (i.e., entity/text/image). Then entity representations are obtained using a relation-aware entity encoder. As the newly constructed ACKGs may introduce noises, random dropouts on neighbors are employed to produce more robust representations. Finally, we design a joint alignment loss for entity and attribute evaluation. Experimental results show that our approach can achieve excellent performance on two MMEA benchmark datasets. Our contributions are summarized as follows:

- We propose a novel multi-modal entity alignment framework to incorporate the consistent alignment knowledge through leveraging an attribute-consistent knowledge graph representation learning method. To the best of our knowledge, this is the first work to tackle the contextual gap problems in the MMEA task.
- We design a multi-modal attribute uniformization method using merge and generate operators to derive an attribute-consistent MMKG, and a ConsistGNN model to aggregate consistent information.
- Experimental results indicate the framework achieves state-of-the-art performance on two public MMEA datasets.

2 PRELIMINARIES

We first provide the definitions of multi-modal knowledge graph (MMKG) and multi-modal entity alignment (MMEA) as follows.

Definition of MMKG. A multi-modal knowledge graph, denoted as $KG = (\mathcal{E}, \mathcal{R}, \mathcal{A})$, is composed of relations between entities and associated attributes. Specifically, $\mathcal{E}, \mathcal{R}, \mathcal{A}$ are the sets of entities, relations, and multi-modal attributes, respectively, with size of n_E, n_R, n_A. We suppose that a KG has two kinds of attributes, i.e., the text attributes $\mathcal{T}$ and image attributes $\mathcal{I}$.

In this paper, we aims to resolve the problem of contextual gaps for the entity alignment task on MMKG. In general, the entity alignment task includes cross-language entity alignment and multi-source entity alignment [30]. Following previous MMEA studies [3, 7, 13], this paper focuses on the latter one.

Definition of MMEA Task. The multi-modal entity alignment task [3, 7, 13] is to identify whether a pair of entities in two multi-modal knowledge graphs is equivalent or not. Concretely, given two multi-modal knowledge graphs KG_1 and KG_2 with a pair of entity alignment seed (v, v'), where v and v' are entities in KG_1 and KG_2, the multi-modal entity alignment aims to identify whether they are equivalent. The main procedure is to learn entity representations in two multi-modal knowledge graphs and calculate the similarity between alignment seed (v, v'). The set of entity alignment seeds is $\mathcal{S} = \{(v, v') \mid v \in \mathcal{E}, v' \in \mathcal{E}', v \equiv v'\}$.
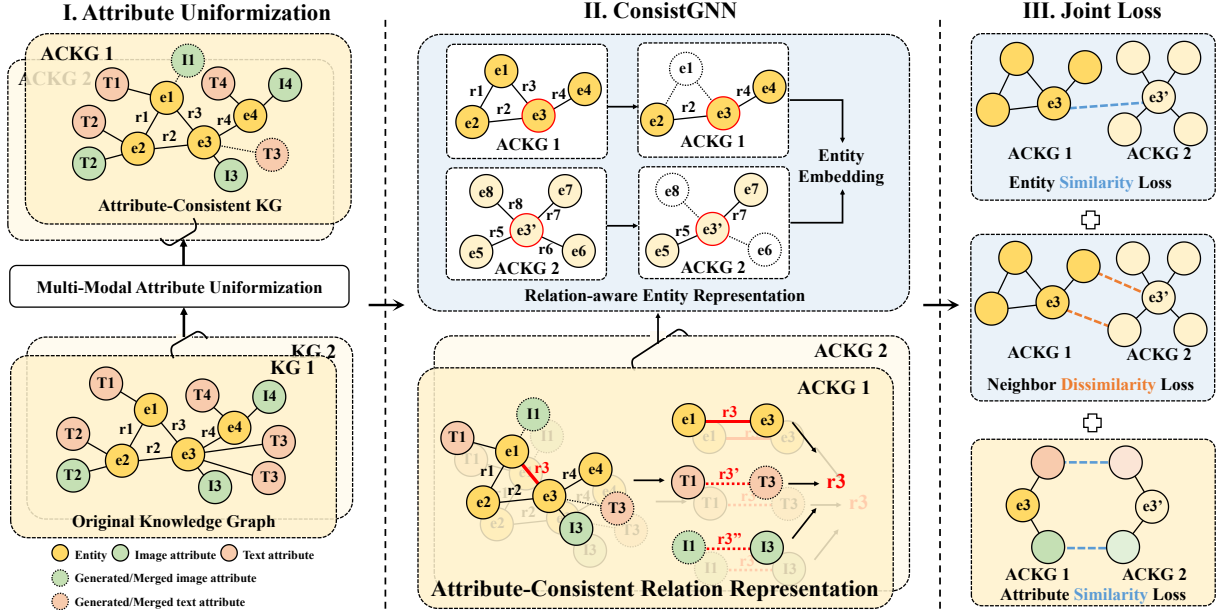
Figure 2: The framework of ACK-MMEA. (I) Two new attribute-consistent MMKGs ($ACKG_1$ and $ACKG_2$) are generated by performing the multi-modal attribute uniformization on the original ones (KG_1 and KG_2). (II) ConsistGNN: Relation-aware GNN with dropouts is to aggregate consistent attributes and learn robust representations of entities. (III) Joint alignment loss with three objectives is used for parameter optimization.

3 FRAMEWORK

This section introduces our proposed framework ACK-MMEA. As shown in Figure 2, ACK-MMEA consists of the following three modules: attribute uniformization, ConsistGNN and joint alignment loss. Firstly, the attribute uniformization module generates consistent attributes for each entity, wherein each entity has one attribute for every modality, respectively. Then, the ConsistGNN maps entities, attributes, and relations to a common representation space and learns heterogeneous relation as well as robust entity representations by aggregating consistent multi-modal attributes. Finally, the joint alignment loss combines the losses of entity similarity, attribute similarity, and neighbor dissimilarity to comprehensively evaluate the attribute-consistent MMKG.

3.1 Multi-Modal Attribute Uniformization

To better tackle the inconsistency issue of attributes in MMKG, we divide the original multi-modal knowledge graph KG into multiple knowledge graphs under each modality.

An MMKG contains text and image attributes, which can be divided into three knowledge graphs $\{KG_E, KG_T, KG_I\}$, corresponding to the entity, text and image graphs, respectively. For each divided knowledge graph KG_X, where $X \in \{E, T, I\}$, the representations of nodes $\mathbf{E}_X$ (entities, texts, images) are :

$$\mathbf{E}_X = \mathbf{F}_X \cdot \mathbf{W}_X, \tag{1}$$

where $\mathbf{F}_X$ are the initial representations of nodes. For each modality of nodes, (a) in KG_E, entity $\mathbf{F}_E \in \mathbb{R}^{n_E \times d_E}$ is initialized by the TransE model [1], with d_E as its dimension; (b) in KG_T, text attribute $\mathbf{F}_T \in \mathbb{R}^{n_T \times d_T}$ is initialized by the BERT [21], with d_T as

its dimension; (c) in KG_I, image attribute $\mathbf{F}_I \in \mathbb{R}^{n_I \times d_I}$ is initialized by the VGG16 [18], with d_I as its dimension. $\mathbf{W}_E \in \mathbb{R}^{d_E \times d}$, $\mathbf{W}_T \in \mathbb{R}^{d_T \times d}$, $\mathbf{W}_I \in \mathbb{R}^{d_I \times d}$ are learnable transformation matrix, mapping the initial representations of different types of nodes into a common d-dimensional space.

It is noticed that attributes of entities in the original MMKG are inconsistent, as shown in Figure 3. The inconsistency means that the entities in each pair have different number of attributes for a specific modality as discussed in Section 1. Such contextual gaps require that the model has the ability to select and generate attributes that contribute to the MMEA task. However, it is hard to determine which and how many attributes to use for the entity alignment, since the contextual gap has no particular structural pattern and different entities have different severity of problems.

To this end, we want to seek generic solutions to contextual gap problem in this paper. We implement the following two uniformization operators on the original MMKG to map entities, attributes, and their relations to a common representation space as well as ensure consistency among them simultaneously. (a) Merge Operator. We aggregate all attributes of each modality into one through an attention-based mechanism, to represent the uniform feature of the entity in the specific modality that is helpful to the MMEA task. (b) Generate Operator. We propose to use the neighbors' attributes to generate the missing attribute. The combination of merge and generate operators alleviates the problem of contextual gap.

Merge Operator. For MMKGs, there may be multiple attributes $\{\mathbf{e}_{v,A,i}\}_{i=1}^{n_v}$ of entity $v \in \mathcal{E}$ in $\mathbf{E}_A$, where $A \in \{T, I\}$, and n_v is the number of attribute in $\mathbf{h}_{v,A}$. In order to aggregate information across the attributes in the same modality, we perform a merge operator
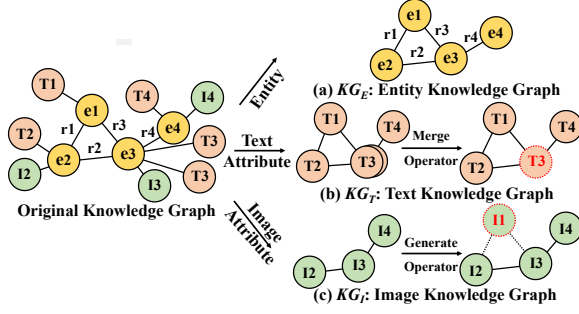
Figure 3: Schematic diagram of the multi-modal attribute uniformization. (a) is the entity knowledge graph only including entity. (b) and (c) are the text and image knowledge graphs, connected relying on the relation of (a). We use the merge and generate operators to make the text and image attributes uniformization.

for each modality. The operator is implemented by a learnable graph attention [15] which can discard redundant contextual information with lower weights, making the final consistent attribute helpful to the MMEA task. Thus we can obtain an aggregated attribute $\mathbf{E}_A^0$ which compresses redundant contextual information:

$$\mathbf{E}_A^{(0)}[v] = \sigma\left(\sum_{i=1}^{n_v} \alpha_i \cdot \mathbf{W}_M \mathbf{e}_{v,A,i}\right), \qquad (2)$$

where α_i is the learned attention weight for i-th attribute, $\mathbf{W}_M \in \mathbb{R}^{d \times d}$ is a learnable parameter, σ is ReLU($\cdot$) function.

Generate Operator. For the MMKG, many entities miss their attributes of the specific modality. It is observed that the neighbors' attributes in the same modality usually provide helpful information to generate the attribute of the target entity. Intuitively, the image attribute of the entity "Donald Trump" is similar to their children's, which means their representations are close. Therefore, to compensate the contextual gap caused by missing attributes, we generate the attribute $\mathbf{E}_A^{(0)}$ with an average aggregation of only the first-order neighbor attributes.

$$\mathbf{E}_A^{(0)}[v] = \sigma\left(\mathbf{W}_G \cdot \text{MEAN}\left(\{\mathbf{e}_{u,A} | u \in \mathcal{N}(v)\}\right)\right), \qquad (3)$$

where $\mathbf{W}_G \in \mathbb{R}^{d \times d}$ is a learnable transformation matrix, $\mathcal{N}(v)$ is the first-order neighbor set of the entity v, σ is the ReLU($\cdot$) function, and $\mathbf{e}_{u,A} = \mathbf{E}_A[u]$.

In this way, the contextual gaps can be relieved, as the attributes of entities in KGs are consistent via the two uniformization operators. Such attribute-consistent KGs (ACKGs) would be helpful to balanced attribute integration during entity representation learning, leading to more accurate alignments.

3.2 ConsistGNN

We further design a GNN model named ConsistGNN to derive relation and entity representations respectively based on attribute-consistent relation representation encoder and relation-aware entity representation encoder, enabling consistent modality aggregation on attribute-consistent KGs.

In ACKGs, there is only one attribute for every entity under each modality. It makes sense that the relations between the same modality of attributes can be somewhat analogous to those between entities. Thus, we define the representations of entity relations $\mathbf{R}^{(0)}$ and attribute relations $\mathbf{R}_A^{(0)}$ are calculated as follows respectively:

$$\mathbf{R}^{(0)} = \mathbf{R} \cdot \mathbf{W}_0, \mathbf{R}_A^{(0)} = \mathbf{R}_A \cdot \mathbf{W}_{0,A}, \qquad (4)$$

where $\mathbf{R}, \mathbf{R}_A \in \mathbb{R}^{n_r \times d_E}$ are the initial representations of relations, which are calculated from TransE. Specifically, for two connected entities (h, t), the attribute relation $\mathbf{r}_{ht,A} = |\mathbf{e}_{t,A} - \mathbf{e}_{h,A}| \in \mathbf{R}_A$ is calculated by tail attribute $\mathbf{e}_{t,A}$ and head attribute $\mathbf{e}_{h,A}$. $\mathbf{W}_0, \mathbf{W}_{0,A} \in \mathbb{R}^{d_E \times d}$ are learnable transformation matrix for mapping initial relation representations into the common space.

To obtain entity representations containing the consistent attribute information, we first initialize the entity representations of the ACKGs as follows:

$$\mathbf{E}_E^{(0)} = \sigma(\mathbf{E}_E \mathbf{W}_{c,E} + \sum_{A \in \{T,I\}} \mathbf{E}_A^{(0)} \mathbf{W}_{c,A}), \qquad (5)$$

where $\mathbf{W}_{c,E}, \mathbf{W}_{c,A} \in \mathbb{R}^{d \times d}$ are learnable transformation matrices. We then feed the initial entity and relation representations into ConsistGNN equipped with attribute-consistent relation representation encoder and relation-aware entity representation encoder. Specifically, we calculate the representations of nodes $\mathbf{E}^{(l)} = \{\mathbf{E}_E^{(l)}, \mathbf{E}_A^{(l)}\}$ and relations $\mathbf{R}^{(l)}$ in the l-th layer as follows:

$$\mathbf{E}^{(l)}, \mathbf{R}^{(l)} = \text{ConsistGNN}\left(\mathbf{E}^{(l-1)}, \mathbf{R}^{(l-1)}\right). \qquad (6)$$

Attribute-Consistent Relation Representation. The attribute uniformization obtains unique identity of each modality for every entity. The relation of two entities is close to the attribute relations of theirs. Thus, we propose to use the consistent attribute information and utilize the attribute relation for entity relation learning. Specifically, we propose an attribute-consistent relation representation encoder for utilizing the multi-modal attribute information, where the relation of two attributes is represented by the combination of themselves:

$$\mathbf{R}^{(l)}[u,v] = \text{ReLU}(\mathbf{W}_{E-E}^{(l)} \mathbf{r}_{uv}^{(l-1)}$$
$$+ \sum_{A \in \{T,I\}} \mathbf{W}_{E-A}^{(l)} [\mathbf{e}_{u,A}^{(l-1)} || \mathbf{e}_{v,A}^{(l-1)}]), \qquad (7)$$

where $\mathbf{r}_{uv}^{(l-1)} = \mathbf{R}^{(l-1)}[u,v]$, and ReLU($\cdot$) is the activation function.

Relation-aware Entity Representation. To make the entity representation have fault tolerance ability for the generated ACKGs, we adopt random dropouts [29] on neighbors to improve the robustness of entity representation, which assumes that missing part of entities does not affect the semantic meaning of ACKG. For each entity v, we randomly discard certain portion of neighboring entities along with the relations connected to themselves. The neighbor set of each entity v is therefore updated as:

$$\mathcal{N}(v) \leftarrow \text{drop}\left(\mathcal{N}(v); \rho\right)$$
$$= \{u_i | u_i \in \mathcal{N}(v), p(i) = 1, p(i) \sim \text{Ber}(1-\rho)\}, \qquad (8)$$

where ρ is the random dropouts rate. drop($\cdot$) is the neighbors random dropouts function, and Ber($\cdot$) is the Bernoulli distribution. The best random dropouts rate is 0.35. Afterwards, we update the

entity representation with the neighbors random dropouts knowledge graph, utilizing the relation and attribute information for each entity. The entity representations in the l-th layer are

$$\mathbf{E}_E^{(l)}[v] = \mathbf{W}_{h,E}[\mathbf{e}_{v,E}^{(l-1)} || \frac{1}{D_v} \sum_{u \in \mathcal{N}(v)} [\mathbf{e}_{u,E}^{(l-1)} || \mathbf{r}_{uv}^{(l)}]], \qquad (9)$$

where D_v is the degree of entity v. The updates of attribute is the special case of entity updates. The attribute representations in the l-th layer $E_A^l[v]$ are the integration of attribute representations $E_A^{l-1}[v]$ and entity representations $E_E^{l-1}[v]$ in $(l-1)$-th layer.

3.3 Joint Alignment Loss

We design a joint alignment loss function, which utilizes the alignment losses of entity similarity, attribute similarity and neighbor dissimilarity to evaluate the attribute-consistent MMKG for multi-perspective assessment.

Aligned Entity Similarity. The entity similarity constraint loss is as follows:

$$\mathcal{L}_i^{EA} = \text{sim}(\mathbf{e}_i, \mathbf{e}_i') - \text{sim}(\mathbf{e}_i, \overline{\mathbf{e}}_i') - \text{sim}(\overline{\mathbf{e}}_i, \mathbf{e}_i'), \qquad (10)$$

where $(\mathbf{e}_i, \mathbf{e}_i')$ are the final representations of aligned seeds (v_i, v_i') of KG_1 and KG_2. $\overline{\mathbf{e}}_i$ and $\overline{\mathbf{e}}_i'$ are the negative samples of the seeds. $\text{sim}(\cdot, \cdot)$ is the cosine distance.

Aligned Attribute Similarity. In addition, in order to make the relation between the same type of attributes similar to the two adjacent entities, we design the attribute similarity constraint loss as follows:

$$\mathcal{L}_i^{attr} = \sum_{A \in \{T, I\}} \text{sim}(\mathbf{e}_{i,A}, \mathbf{e}_{i,A}'). \qquad (11)$$

Aligned Neighbor Dissimilarity. Furthermore, for more precise alignment of candidate entities, the entity v_i' in KG_2 should be characterized similar to entity v_i in KG_1, and the neighbors $\mathcal{N}(v_i')$ of v_i' in KG_2 should be characterized dissimilar to entity v_i in KG_1. Thus, we introduce the neighbor dissimilarity constraint loss, inspired by [9, 25]:

$$\mathcal{L}_i^{cont} = -\log \frac{\exp(\text{sim}(\mathbf{e}_i, \mathbf{e}_i')/\tau)}{\sum_{v_j' \in \mathcal{N}(v_i')} \exp(\text{sim}(\mathbf{e}_i, \mathbf{e}_j'))/\tau)}, \qquad (12)$$

where v_j' is the neighbors of entity v_i', and τ is the temperature coefficient.

The total alignment loss $\mathcal{L}$ is the weighted sum of three loss functions:

$$\mathcal{L} = \lambda_1 \mathcal{L}^{EA} + \lambda_2 \mathcal{L}^{attr} + \lambda_3 \mathcal{L}^{cont}, \qquad (13)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the learnable hyper-parameters for joint alignment loss.

4 EXPERIMENT

4.1 Dataset

We conducted experiments on FB15K-DB15K and FB15K-YAGO15K datasets [14], which are the two most popular datasets in MMEA. FB15K-DB15K is the entity alignment dataset[2] of FB15K and DB15K multi-modal knowledge graph, including 12,846 alignment seeds.

---

[2] https://github.com/mniepert/mmkb

FB15K-YAGO15K is the entity alignment dataset of FB15K and YAGO15K knowledge graphs, including 11,199 alignment seeds. As previous works [3, 7] recommended, we divided the two data sets into training and testing sets at 2:8, 5:5, and 8:2, respectively. We report the official MRR, Hits@1, and Hits@10 metrics for evaluation on different proportions of alignment seeds. For more details, please refer to the Appendix B.

4.2 Comparision Methods

We compare our method with six EA methods. They originally aggregate the text attribute and relation information. Here, we introduce the image attributes initialized by VGG16 for entity representation with the same aggregation manner of text attributes: (1) **TransE** [1] assumes that the entity embedding v should be close to the attribute embedding a plus their relation r. (2) **IPTransE** [32] is a translation-based method to jointly optimize entities and relations representation in knowledge graphs with an iterative and parameter sharing strategy. (3) **GCN-align** [23] transfers entities and attributes of per language to a common representation space through GCN. (4) **SEA** [16] leverages labeled and unlabeled entities through adversarial training, and combines the image attributes. (5) **IMUSE** [8] uses a bivariate regression to merge the relations and multiple attributes. (6) **AttrGNN** [15] divides KG into multiple subgraphs, effectively modeling various types of attributes.

Furthermore, we compare our method with four MMEA methods, which also do not introduce name attributes and focus on how to utilize the multi-modal attributes: (7) **PoE** [14] utilizes the image features and measures the credibility by matching the semantics of the entities to mining the relations. (8) **PoE-rni** [14] uses the relation, numeric literals and images attributes of PoE with the best performance. (9) **Chen et al.** [3] design a fusion module to integrate multi-modal attributes. (10) **HEA** [7] characterizes MMKG in hyperbolic space.

In contrast to these methods, our method generates a new attribute-consistent MMKG to uniform attribute information and learns more robust entity representations via ConsistGNN with a joint loss.

4.3 Implementation Details

For all baselines, we adopt the best hyper-parameters reported in their literature. For the EA baselines (1-6), we reproduce the performance through adding image attributes. For the MMEA baselines (7-10), we copy the existing results reported in the literature [3, 7, 14].

Our model is implemented based on PyTorch, an open-source deep learning framework. The BERT version is bert-base-uncased in huggingface[3] for text attributes initialization and VGG version is VGG16[4] for image attributes initialization. The GNN (GCN and GAT) layer is 2, the training epoch is 200, the L2 regularization value is 0.0001, and the margin gramma value is 1.0. For hyper-parameters, the best random dropping rate ρ is 0.35 and temperature coefficient τ is 0.5, and coefficients $\lambda_1, \lambda_2, \lambda_3$ are 5, 3 and 2. For the learning rate, we adopt the method of grid search with a step size of 0.001. The optimal learning rate is 0.001. All hyper-parameter settings are tuned on the validation data by the grid search with 5 trials. Refer to Appendix C for more details. All experiments were conducted

---

[3] https://github.com/huggingface/transformers
[4] https://github.com/machrisaa/tensorflow-vgg

Table 1: Main experiments on FB15K-DB15K (top) and FB15K-YAGO15K (bottom) with different proportions of entity alignment seeds. The best results are highlighted in bold, and the underlined values are the second best result. The "↑" means the improvement compared to the second best result, and "-" means that the results are not available.

Methods	FB15K-DB15K (20%)			FB15K-DB15K (50%)			FB15K-DB15K (80%)		
	MRR (%)	Hits@1 (%)	Hits@10 (%)	MRR (%)	Hits@1 (%)	Hits@10 (%)	MRR (%)	Hits@1 (%)	Hits@10 (%)
TransE [1]	13.4	7.8	24.0	30.6	23.0	44.6	50.7	42.6	65.9
IPTransE [32]	9.4	6.5	21.5	28.3	21.0	42.1	46.9	40.3	62.7
GCN-align [23]	8.7	5.3	17.4	29.3	22.6	43.5	47.2	41.4	63.5
SEA [16]	25.5	17.0	42.5	47.0	37.3	65.7	50.5	51.2	78.4
IMUSE [8]	26.4	17.6	43.5	40.0	30.9	57.6	55.1	45.7	72.6
AttrGNN [15]	34.3	25.2	53.5	<u>54.7</u>	<u>47.3</u>	<u>72.1</u>	70.3	<u>67.1</u>	83.9
PoE [14]	17.0	12.6	25.1	53.3	46.4	65.8	<u>72.1</u>	66.6	82.0
PoE-rni [14]	28.3	23.2	39.0	44.2	38.0	55.7	55.8	50.2	64.1
Chen et al. [3]	<u>35.7</u>	<u>26.5</u>	<u>54.1</u>	51.2	41.7	70.3	68.5	59.0	<u>86.9</u>
HEA [7]	-	12.7	36.9	-	26.2	58.1	-	41.7	78.6
ACK-MMEA (ours)	**38.7** (↑3.0)	**30.4** (↑3.9)	**54.9** (↑0.8)	**62.4** (↑7.7)	**56.0** (↑8.7)	**73.6** (↑1.5)	**75.2** (↑3.1)	**68.2** (↑1.1)	**87.4** (↑0.5)
Methods	FB15K-YAGO15K (20%)			FB15K-YAGO15K (50%)			FB15K-YAGO15K (80%)		
	MRR(%)	Hits@1 (%)	Hits@10 (%)	MRR (%)	Hits@1 (%)	Hits@10 (%)	MRR (%)	Hits@1 (%)	Hits@10 (%)
TransE [1]	11.2	6.4	20.3	26.2	19.7	38.2	46.3	39.2	59.5
IPTransE [32]	8.4	4.7	16.9	24.8	20.1	36.9	45.8	40.1	60.2
GCN-align [23]	15.3	8.1	23.5	29.4	23.5	42.4	47.7	40.6	64.3
SEA [16]	21.8	14.1	37.1	38.8	29.4	57.7	60.5	51.4	77.3
IMUSE [8]	14.2	8.1	25.7	46.9	39.8	60.1	58.1	51.2	70.7
AttrGNN [15]	31.8	22.4	39.5	46.2	38.0	63.9	67.1	<u>59.9</u>	78.7
PoE [14]	15.4	11.3	22.9	41.4	34.7	53.6	63.5	57.3	74.6
PoE-rni [14]	<u>33.4</u>	<u>25.0</u>	<u>49.5</u>	<u>49.8</u>	<u>41.1</u>	<u>66.9</u>	57.2	49.2	70.5
Chen et al. [3]	31.7	23.4	48.0	48.6	40.3	64.5	<u>68.2</u>	59.8	<u>83.9</u>
HEA [7]	-	10.5	31.3	-	26.5	58.1	-	43.3	80.1
ACK-MMEA (ours)	**36.0** (↑2.6)	**28.9** (↑3.9)	**49.6** (↑0.1)	**59.3** (↑9.5)	**53.5** (↑12.4)	**69.9** (↑3.0)	**74.4** (↑6.2)	**67.6** (↑7.7)	**86.4** (↑2.5)

on a server with one GPU (Tesla V100). The time analysis of our method is shown in Appendix D.

4.4 Main Results

To verify the effectiveness of our ACK-MMEA, we report overall average results in Table 1. It shows performance comparisons on FB15K-DB15K and FB15K-YAGO15K datasets with different splits on training/testing data of alignment seeds, i.e., 2:8, 5:5, and 8:2.

From the table, we can observe that: 1) Our attribute-consistent model outperforms all the baselines of both EA and MMEA methods, in terms of three metrics on both datasets. Specifically, our model improves $3.0\% - 7.7\%$ (4% on average) on FB15K-DB15K and $2.6\% - 9.5\%$ (6% on average) on FB15K-YAGO15K in terms of MRR for all proportions of training data, respectively. It demonstrates that our model is robust to different proportions of training resource, achieving reliable performance on multi-modal entity alignment. 2) Compared to EA baselines (1-4), especially for MRR and Hits@1, our model improves 5% and 9% up on average on FB15K-DB15K and FB15K-YAGO15K, tending to achieve more significant improvements. It demonstrates that effectiveness of multi-modal consistent-attribute uniformization for incorporating consistent alignment knowledge. 3) Compared to MMEA baselines (5-8), our model designs a ConsistGNN model on new attribute-consistent MMKGs, the average gains of our model regarding MRR, Hits@1

and Hits@10 are 5%, 5%, and 1%, respectively. The reason is that our method incorporates the consistent multi-modal attributes and robust relation-aware entity information. 4) In terms of three proportions of training data on both datasets, our model improves 4.5% on average and 8% on average on FB15K-DB15K and FB15K-YAGO15K for the Hits@1 metric, which means the proportion that only prediction label is equal to the global label. It demonstrates that our method is more accurate compared to baselines, which can provide more correct predictions when only one outcome can be predicted. All the observations demonstrate the effectiveness of the ACK-MMEA framework.

4.5 Discussions for Model Variants

To investigate the effectiveness of each module in ACK-MMEA, we conduct variant experiments, showcasing the results in Table 2 and Figure 4. The "↓" means the value of performance degradation compared to the ACK-MMEA.

From the Table 2, we can observe that: 1) The impact of the attribute uniformization tends to be more significant on using original attributes. We believe the reason is that the consistent attributes captures more clues for entity alignment. 2) By replacing the ConsistGNN to GCN, GAT or without random dropouts on neighbors, or random replacement on neighbors, the performance decreased significantly. It demonstrates that the ConsistGNN captures more

Table 2: Variant experiments on FB15K-DB15K (80%) and FB15K-YOGA15K (80%). "w/o" means removing corresponding module from the complete model. "repl." means replacing corresponding module with the other module. The "↓" means the value of performance degradation compared to the ACK-MMEA.

Variants	FB15K-DB15K (80%)			
	MRR (%)	Hits@1 (%)	Hits@10 (%)	△ Avg (%)
ACK-MMEA (ours)	**75.2**	**68.2**	**87.4**	-
w/o attribute uniformization	73.2 (↓2.0)	63.9 (↓4.3)	82.8 (↓4.6)	↓3.6
w/o attribute uniformization (Merge Operator)	74.0 (↓1.2)	65.1 (↓3.1)	83.5 (↓3.9)	↓2.7
w/o attribute uniformization (Generate Operator)	74.2 (↓1.0)	65.6 (↓2.6)	84.1 (↓3.3)	↓2.3
w/o text attribute	73.4 (↓1.8)	66.2 (↓2.0)	85.9 (↓1.5)	↓1.7
w/o image attribute	72.5 (↓2.7)	65.8 (↓2.4)	86.2 (↓1.2)	↓2.1
repl. GCN	73.5 (↓1.7)	66.6 (↓1.6)	82.7 (↓4.7)	↓2.6
repl. GAT	74.1 (↓1.1)	65.5 (↓2.7)	83.1 (↓4.3)	↓2.7
w/o random dropouts	72.7 (↓2.5)	66.0 (↓2.2)	84.6 (↓2.8)	↓2.5
repl. random replacement	71.2 (↓4.0)	64.9 (↓3.3)	83.8 (↓3.6)	↓3.6
w/o attribute similarity loss	74.0 (↓1.2)	68.1 (↓0.1)	85.7 (↓1.7)	↓1.0
w/o neighbor dissimilarity loss	73.6 (↓1.6)	67.7 (↓0.5)	86.9 (↓0.5)	↓0.8

Variants	FB15K-YOGA15K (80%)			
	MRR (%)	Hits@1 (%)	Hits@10 (%)	△ Avg (%)
ACK-MMEA (ours)	**74.4**	**67.6**	**86.4**	-
w/o attribute uniformization	73.6 (↓0.8)	64.2 (↓3.4)	84.3 (↓2.1)	↓2.1
w/o attribute uniformization (Merge Operator)	73.5 (↓0.9)	65.3 (↓2.3)	83.9 (↓2.5)	↓1.9
w/o attribute uniformization (Generate Operator)	74.1 (↓0.3)	66.0 (↓1.6)	84.5 (↓1.9)	↓1.2
w/o text attribute	73.9 (↓0.5)	65.8 (↓1.8)	84.7 (↓1.7)	↓1.3
w/o image attribute	73.6 (↓0.8)	65.7 (↓1.9)	84.6 (↓1.8)	↓1.5
repl. GCN	73.0 (↓1.4)	66.3 (↓1.3)	84.3 (↓2.1)	↓1.6
repl. GAT	73.8 (↓0.6)	65.9 (↓1.7)	83.9 (↓2.5)	↓1.6
w/o random dropouts	72.3 (↓2.1)	65.7 (↓1.9)	84.1 (↓2.3)	↓2.1
repl. random replacement	70.9 (↓3.5)	64.2 (↓3.4)	83.1 (↓3.3)	↓3.4
w/o attribute similarity loss	72.8 (↓1.6)	66.7 (↓0.9)	85.3 (↓1.1)	↓1.2
w/o neighbor dissimilarity loss	73.2 (↓1.2)	67.0 (↓0.6)	86.1 (↓0.3)	↓0.7

effective consistent-attribute and relation information. 3) The impacts of the attribute similarity and neighbor dissimilarity loss tend to be significant. Since the consistent attributes tackle the contextual gap, and the neighbor loss guides our model to learn robust representations. 4) When we remove all image attributes as "w/o image attribute", our method drops 2.1% and 1.5% on average on FB15K-DB15K and FB15K-YAGO15K. The performance decreases 3.6% and 2.1% on average when we remove attribute uniformization module as "w/o attribute uniformization". It demonstrates that image attributes can improve model performance and our method utilizes image attributes effectively through capturing more alignment knowledge. All the observations demonstrate the effectiveness of each component in our model.

To further investigate the impact of multi-modal attributes on all compared methods, we report the results by deleting different modality of attributes, as shown in Figure 4. From the figure, we can observe that: 1) The variants without the text or image attributes significantly decline on all evaluation metrics, which demonstrates that the multi-modal attributes are necessary and effective for the entity alignment task. 2) Our model is less affected by deleting all

multi-modal attributes. The reason we think is that the random dropouts on neighbors and the neighbor dissimilarity loss are beneficial to obtaining better entity representations. 3) Compared to other baseline methods, our model derives better results both in the case of using all multi-modal attributes or abandoning some of them. It demonstrates our model makes full use of existing multi-modal attributes, and consistent attributes are effective for the multi-modal entity alignment task. 4) When we delete all attributes as "Del. all attributes", our method drops 1%-5% in terms of MRR, and performs best compared to all baselines. It demonstrates that our model makes the entity representation having fault tolerance through the relation-aware entity representation and more precise alignment of candidate entities by the dissimilarity constraint loss. 5) When we delete all image attributes as "Del. image attributes", which means that the original multi-modal knowledge graph transferred into a KG, our method is better than other baselines. It demonstrates that our model incorporates consistent alignment knowledge by the attribute-consistent relation representation encoder and the relation-aware entity representation encoder. All the observations
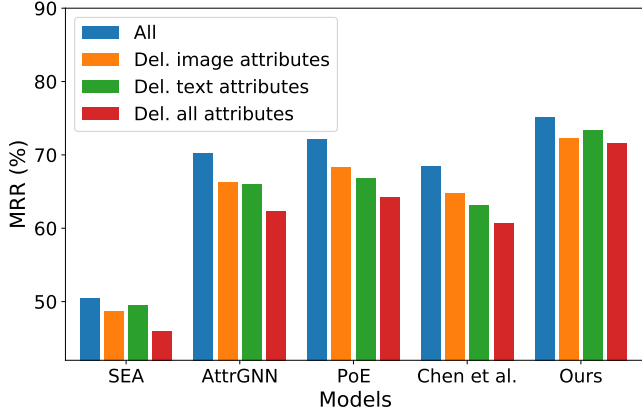
Figure 4: Results of deleting attributes on FB15K-DB15K (80%). "Del." means deleting the corresponding attribute.

Figure 5: Analysis of the dropping rate ρ.

Figure 6: Impact of differences in attribute number.

demonstrate that the effectiveness of the constructed attribute-consistent MMKG and the ConsistGNN.

4.6 Impact of Dropping Rate

We investigate the impact of the random dropouts rate on neighbors. Figure 5 shows the metric values with various hyper-parameter setting of ρ on the FB15K-DB15K (80%). As the dropping rate increases, the MRR, Hits@1 and Hits@10 gradually increase and then falling after an optimal value. The peak performance of the model is when the dropping rate of neighbors reaches 35%, reflecting the effectiveness to learn the robust entity representation. It demonstrates the capacity of the random dropouts for improving the fault tolerance ability and enforcing consistent attribute aggregation.

4.7 Impact of Attribute Number

We investigate the impact of different degrees of the contextual gap between alignment seeds. To do so, we choose entity alignment seeds with the same number of image attributes, and vary the gaps of text attribute number of entity pairs in $[0, 24]$. Figure 6 shows the performance of different models in the case of varied attribute gaps on the FB15K-DB15K (80%). From the figure, we can observe that: 1) With the increase of the gaps on alignment seeds, the performance of all methods gradually decreases. The main reason is that the bigger the gaps between entity seeds, the more difficult it is to match entities. This again confirms our intuition claimed in the introduction. 2) Compared to baseline methods, the performance of our model decreased slowly, demonstrating the superiority of our method in tackling the contextual gap issue. All the observations demonstrate that our method can reduce the impact of the contextual gap.

5 CONCLUSION

This paper proposes a novel multi-modal entity alignment framework, namely ACK-MMEA. It generates an attribute-consistent MMKG with each entity containing only one attribute of each modality by the multi-modal attribute uniformization. We further propose the ConsistGNN to integrate the consistent multi-modal attributes and obtain aggregated rela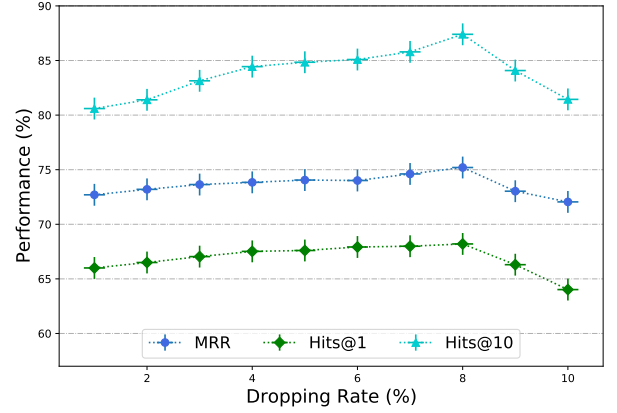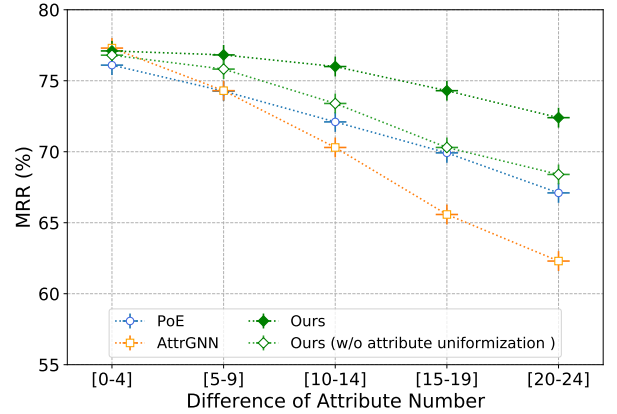tion representations and robust entity representations. To evaluate the attribute-consistent MMKG, we design the joint alignment loss with three objectives. Our work overcomes the contextual gaps between entity pairs, caused by the information redundancy and absence of the attribute. The empirical experiments demonstrate that our method tackles the contextual gap problem. However, the operator of attribute generation will introduce noise data. In future work, we will study how to avoid the influence of noise data on the MMEA task.

ACKNOWLEDGMENTS

REFERENCES

[1] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* 2787–2795. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

[2] Yixin Cao, Zhiyuan Liu, Chengjiang Li, Zhiyuan Liu, Juanzi Li, and Tat-Seng Chua. 2019. Multi-Channel Graph Neural Network for Entity Alignment. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers.* Association for Computational Linguistics, 1452–1461. https://doi.org/10.18653/v1/p19-1140

[3] Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020. MMEA: Entity Alignment for Multi-modal Knowledge Graph. In *Knowledge Science, Engineering and Management - 13th International Conference, KSEM 2020, Hangzhou, China, August 28-30, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12274).* Springer, 134–147. https://doi.org/10.1007/978-3-030-55130-8_12

[4] 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 *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022.* ACM, 118–126. https://doi.org/10.1145/3534678.3539244

[5] Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2018. Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.* ijcai.org, 3998–4004. https://doi.org/10.24963/ijcai.2018/556

[6] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017.* ijcai.org, 1511–1517. https://doi.org/10.24963/ijcai.2017/209

[7] Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing* 461 (2021), 598–607. https://doi.org/10.1016/j.neucom.2021.03.132

[8] Fuzhen He, Zhixu Li, Qiang Yang, An Liu, Guanfeng Liu, Pengpeng Zhao, Lei Zhao, Min Zhang, and Zhigang Chen. 2019. Unsupervised Entity Alignment Using Attribute Triples and Relation Triples. In *Database Systems for Advanced Applications - 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22-25, 2019, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 11446).* Springer, 367–382. https://doi.org/10.1007/978-3-030-18576-3_22

[9] Cheng Ji, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Qingyun Sun, and Philip S. Yu. 2023. Unbiased and Efficient Self-Supervised Incremental Contrastive Learning. *CoRR* abs/2301.12104 (2023). https://doi.org/10.48550/arXiv.2301.12104 arXiv:2301.12104

[10] Jin Jiang, Mohan Li, and Zhaoquan Gu. 2021. A Survey on Translating Embedding based Entity Alignment in Knowledge Graphs. In *Sixth IEEE International Conference on Data Science in Cyberspace, DSC 2021, Shenzhen, China, October 9-11, 2021.* IEEE, 187–194. https://doi.org/10.1109/DSC53577.2021.00033

[11] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. 2022. Reasoning over Different Types of Knowledge Graphs: Static, Temporal and Multi-Modal. *CoRR* abs/2212.05767 (2022). https://doi.org/10.48550/arXiv.2212.05767 arXiv:2212.05767

[12] Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal Contrastive Representation Learning for Entity Alignment. *CoRR* abs/2209.00891 (2022). https://doi.org/10.48550/arXiv.2209.00891 arXiv:2209.00891

[13] Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual Pivoting for (Unsupervised) Entity Alignment. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021.* AAAI Press, 4257–4266. https://ojs.aaai.org/index.php/AAAI/article/view/16550

[14] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11503).* Springer, 459–474. https://doi.org/10.1007/978-3-030-21348-0_30

[15] Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, and Tat-Seng Chua. 2020. Exploring and Evaluating Attributes, Values, and Structures for Entity Alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020.* Association for Computational Linguistics, 6355–6364. https://doi.org/10.18653/v1/2020.emnlp-main.515

[16] Shichao Pei, Lu Yu, Robert Hoehndorf, and Xiangliang Zhang. 2019. Semi-Supervised Entity Alignment via Knowledge Graph Embedding with Awareness of Degree Difference. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019.* ACM, 3130–3136. https://doi.org/10.1145/3308558.3313646

[17] Xiaofei Shi and Yanghua Xiao. 2019. Modeling Multi-mapping Relations for Precise Cross-lingual Entity Alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019.* Association for Computational Linguistics, 813–822. https://doi.org/10.18653/v1/D19-1075

[18] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.* http://arxiv.org/abs/1409.1556

[19] 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 *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020.* ACM, 1405–1414. https://doi.org/10.1145/3340531.3411947

[20] Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge Graph Alignment Network with Gated Multi-Hop Neighborhood Aggregation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020.* 222–229. https://aaai.org/ojs/index.php/AAAI/article/view/5354

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA.* 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[22] Zeshi Wang, Mohan Li, and Zhaoquan Gu. 2021. A Review of Entity Alignment based on Graph Convolutional Neural Network. In *Sixth IEEE International Conference on Data Science in Cyberspace, DSC 2021, Shenzhen, China, October 9-11, 2021.* IEEE, 144–151. https://doi.org/10.1109/DSC53577.2021.00027

[23] Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018.* 349–357. https://doi.org/10.18653/v1/d18-1032

[24] Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing.* 349–357.

[25] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised Graph Learning for Recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021.* ACM, 726–735. https://doi.org/10.1145/3404835.3462862

[26] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2020. Neighborhood Matching Network for Entity Alignment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020.* Association for Computational Linguistics, 6477–6487. https://doi.org/10.18653/v1/2020.acl-main.578

[27] Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022. LingYi: Medical Conversational Question Answering System based on Multi-modal Knowledge Graphs. *CoRR* abs/2204.09220 (2022). https://doi.org/10.48550/arXiv.2204.09220 arXiv:2204.09220

[28] Guohai Xu, Hehong Chen, Feng-Lin Li, Fu Sun, Yunzhou Shi, Zhixiong Zeng, Wei Zhou, Zhongzhou Zhao, and Ji Zhang. 2021. AliMe MKG: A Multi-modal Knowledge Graph for Live-streaming E-commerce. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021.* ACM, 4808–4812. https://doi.org/10.1145/3459637.3481983

[29] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph Contrastive Learning with Augmentations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.* https://proceedings.neurips.cc/paper/2020/hash/3fe230348e9a12c13120749e3f9fa4cd-Abstract.html

[30] Kaisheng Zeng, Chengjiang Li, Lei Hou, Juanzi Li, and Ling Feng. 2021. A comprehensive survey of entity alignment for knowledge graphs. *AI Open* 2 (2021), 1–13. https://doi.org/10.1016/j.aiopen.2021.02.002

[31] Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view Knowledge Graph Embedding for Entity Alignment. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019.* ijcai.org, 5429–5435. https://doi.org/10.24963/ijcai.2019/754

[32] Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative Entity Alignment via Joint Knowledge Embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017.* ijcai.org, 4258–4264. https://doi.org/10.24963/ijcai.2017/595

[33] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-Modal Knowledge Graph Construction and Application: A Survey. *CoRR* abs/2202.05786 (2022). arXiv:2202.05786 https://arxiv.org/abs/2202.05786

A   RELATED WORK

A.1   Entity Alignment

Entity alignment (EA) technology mainly includes early embedding-based methods [1, 6] and recently popular GNN-based methods [2, 17]. The focus of various GNN-based methods is the aggregation way of attributes [2, 15], relations [5, 24], and neighbor features [20, 26]. Specifically, the attribute-aware methods [2, 5, 31] aggregate multi-type attributes or combine multiple models to encode entities for learning the entity embedding from multiple perspectives. AttrGNN [15] divides KG into subgraphs for attributes aggregation, effectively modeling various types of attribute triples. The above methods demonstrate the effectiveness of aggregating attributes for EA. Nevertheless, all of these methods ignore the inconsistency of attributes, as well as the image attributes.

A.2   Multi-Modal Entity Alignment

Furthermore, because of the multi-modal nature of KGs in real-world, there are several works [10, 22, 33] beginning to focus on the MMEA technology. Similar to EA on single-modal KG, many tasks on MMKG [13, 19] provide the possibility of the fusion of multi-modal attributes and relations. As the first work on the MMEA, PoE [14] characterized each entity as a single vector wherein all modality features of entities are concatenated. However, it cannot capture the potential interactions among heterogeneous modalities, limiting its capacity for performing accurate entity alignments. Later, Chen et al. [3] proposed a multi-modal knowledge embedding method to discriminatively generate knowledge representations of different types of knowledge, and then designed a multi-modal fusion module to integrate them. Guo et al. [7] developed hyperbolic multi-modal entity alignment (HEA) approach to combine both attribute and entity representations in the hyperbolic space and used aggregated embeddings to predict alignments. MCLEA [12] and MSNEA [4] reduce the gaps between modalities for each entity as well as utilize name embeddings. Nevertheless, the above existing methods ignore contextual gaps between entity pairs and in turn may constrain the effectiveness of alignment.

B   DATASETS AND EVALUATION METRICS

B.1   Datasets

In our experiments, we use two multi-modal datasets which are built in [14], namely FB15K-DB15K and FB15K-YAGO15K. FB15K is a representative subset extracted from the Freebase knowledge base. Aiming to maintain an approximate entity number of FB15K, DB15K from DBpedia and YAGO15K from YAGO are mainly selected based on the entities aligned with FB15K. Table 3 depicts the statistics of multi-modal datasets.

Table 3: Statistics for the datasets. (Rel.: Relation, Attr.: Attribute, Rel. T.: Relational Triple, Attr. T.: Attributes triple.)

Dataset	#Entity	#Rel.	#Attr.	#Rel. T.	#Attr. T.	#Images
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

B.2   Evaluation Metrics

We utilize cosine similarity to calculate the similarity between two entities and employ Hits@n, and MRR as metrics to evaluate all the models. Hits@n means the rate correct entities rank in the top n according to similarity computing. MRR denotes the mean reciprocal rank of correct entities. The higher values of Hits@n and MRR explain the better performance of the method.

C   HYPER-PARAMETERS

To enable replication and foster research, we report our hyper-parameter settings in Table 4. Note that all the hyper-parameter settings are tuned on the validation set by the grid search with 5 trials. We adopt bert-large-uncased in huggingface as our encoder, whose layer number is 24 and the embedding size is 1024. The best values of hyper-parameters $\lambda_1, \lambda_2, \lambda_3$ are 5, 3 and 2. Specifically, if the model does not decrease the loss function of the validation set for 100 consecutive turns, the operation is stopped. All baseline models use the same data set partitioning to ensure fairness. To ensure fairness, all baselines use the same entity representation dimension, which is set to 128 dimensions. All experiments were conducted on a server with one GPU (Tesla V100).

D   EMPIRICAL RUNTIME ANALYSIS

The time complexity of the proposed framework is acceptable. Table 5 shows the time costs of the training of our method and three good performance baseline methods on FB15K-DB15K and FB15K-YAGO15K. For fairness, we only use one GPU, including AttrGNN. Thus, we train the three channels of AttrGNN in turn, which makes the model take the longest training time. Our method generates a attribute-consistent knowledge graph, which enhances the time cost. However, we design a random dropouts mechanism to save time. Overall, the time complexity of our approach can be on par with other efficient approaches.

Table 4: Hyper-parameter settings of model.

Hyper-parameters	FB15K-DB15K	FB15K-YAGO15K
Batch size	512	512
Train epoch	200	200
Learning rate	0.001	0.001
Temperature	0.5	0.5
Negative Sample Number	15	15
Weight Decay	0.01	0.01
Random Dropouts Rate	0.35	0.35
λ_1	5	5
λ_2	3	3
λ_3	2	2

Table 5: Average Training time (s) on the FB15K-DB15K and FB15K-YAGO15K datasets.

Methods	AttrGNN	PoE	PoE-rni	ACK-MMEA
FB15K-DB15K	396	164	162	165
FB15K-YAGO15K	389	155	151	156