Multi-Modal Entities Matter: Benchmarking Multi-Modal Entity Alignment

Guanchen Xiao[1], Weixin Zeng[1]*, Shiqi Zhang[1], Mingrui Lao[1], Xiang Zhao[1]
[1]National University of Defense Technology
*Corresponding Author
Correspondence: 921873590@qq.com

Abstract

Multi-modal entity alignment (MMEA) is a long-standing task that aims to discover identical entities between different multi-modal knowledge graphs (MMKGs). However, most of the existing MMEA datasets consider the multi-modal data as the attributes of textual entities, while neglecting the correlations among the multi-modal data and do not fit in the real-world scenarios well. In response, in this work, we establish a novel yet practical MMEA dataset, i.e. NMMEA, which models multi-modal data (e.g., images) equally as textual entities in the MMKG. Due to the introduction of multi-modal data, NMMEA poses new challenges to existing MMEA solutions, i.e., heterogeneous structural representation learning and cross-modal alignment inference. Hence, we put forward a simple yet effective solution, CrossEA, which can effectively learn the structural information of entities by considering both intra-modal and cross-modal relations, and further infer the similarity of different types of entity pairs. Extensive experiments validate the significance of NMMEA, where CrossEA can achieve superior performance in contrast to competitive methods on the proposed dataset.

1 Introduction

Multi-modal knowledge graph (MMKG) (Liu et al., 2019) is a large-scale semantic network of text, image, audio of information in the real world, which includes entities and concepts of different modalities as nodes, and various semantic relations as edges. MMKGs have attracted wide attentions in various scenarios and promoted the development of many downstream applications including information retrieval (Sun et al., 2020; Zeng et al., 2023b) and question answering (Zhao et al., 2019; Zhu et al., 2024).

Current MMKGs can be typically categorized into A-MMKGs (Attribute Multi-Modal Knowledge Graphs) and N-MMKGs (Node Multi-Modal Knowledge Graphs) (Zhu et al., 2024), while the former takes the multi-modal data as the specific attribute values of entities or concepts (Liu et al., 2019), the latter directly treats multi-modal data as entities in knowledge graphs representing entities [1] (Li et al., 2020; Chen et al., 2013). The difference is also illustrated in Figure 1. Compared with A-MMKGs, N-MMKGs can better model the relations between the multi-modal data, for instance "liveIn" and "competitorOf" between visual entities and "bornIn" between visual and textual entities in Figure 1. In addition, when constructing MMKGs from real-world data, most entities do not inherently possess multiple modalities of data, which may cause attribute missing issue in A-MMKGs (Zhang et al., 2023; Li et al., 2023). In contrast, N-MMKGs, with multi-modal data inherently as entities, can avoid such an attribute missing issue.

In general, MMKGs are constructed from independent multi-modal corpora for different purposes, and hence, there consequently may be entities representing the same real-world objects in different MMKGs. Thus, it calls for the study of multi-modal entity alignment (MMEA) that aims to integrate the same information in different multi-modal knowledge sources, which has become one of the emerging tasks over recent years (Chen et al., 2020; Cheng et al., 2022). For instance, as shown in Figure 1, the entity "Joe Biden" in a MMKG can be aligned to entity "Joseph Robinette Biden Jr" in the other MMKG for they represent the same person in real world. [2]

Noteworthily, all of existing MMEA benchmarks and solutions are built upon A-MMKGs (Zhu et al.,

[1]Note that in this work, we only consider the visual and textual modalities, where the entities are referred to as "textual" and "visual" entities, respectively. The studied contents can be extended to more modalities.

[2]The definition of N-MMKGs and multi-modal entity alignment task on N-MMKGs can be found in Appendix A

Proceedings of the 31st International Conference on Computational Linguistics, pages 8714–8724
January 19–24, 2025.

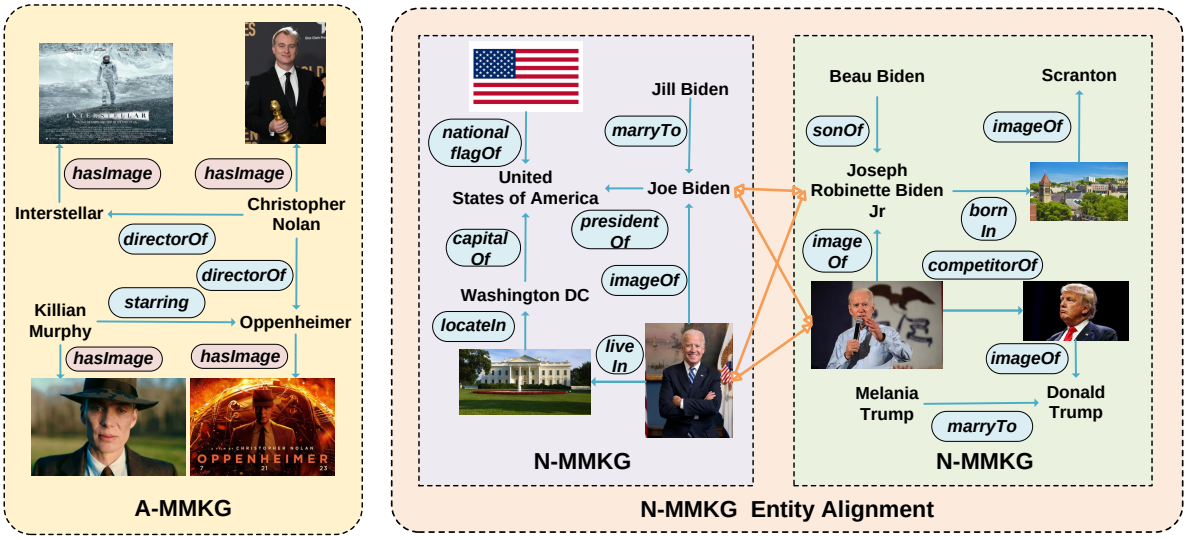


Figure 1: An example to illustrate the difference between A-MMKG and N-MMKG, and to showcase entity alignment between N-MMKGs. In A-MMKG, images serve as attribute values for the "hasImage" relation associated with a "textual" entity. Conversely, in N-MMKG, images exist as "visual" entities and are treated similar to "textual" entities. Thus, the alignment involves the intra-modal matching, e.g., "textual-textual" and "visual-visual", as well as the cross-modal matching, e.g., "textual-visual".

2024), and they adopt the "representation learning then alignment inference" paradigm to obtain the alignment results (Zhao et al., 2023; Zeng et al., 2023a), where they first learn and combine the representations of different modalities and then infer the equivalence between two "textual" entities (Lin et al., 2022; Chen et al., 2023). However, these methods are not directly applicable to N-MMKGs where the entities are in different modalities. In fact, such multi-modal entities introduce new challenges to alignment: (1) **Heterogeneous structural representation learning.** The multi-modal entities intrinsically make N-MMKG a heterogeneous graph, consisting of both intra-modal and cross-modal edges. Hence, current homogeneous graph representation learning methods may fail to effectively capture such differences. (2) **Cross-modal alignment inference.** In addition to comparing two entities in the same modality (i.e., intra-modal alignment), the alignment in N-MMKGs requires the comparison of cross-modal entities (i.e., cross-modal alignment). A major challenge with this type of problem is how to make the alignment similarities within intra-modal and cross-modal comparable.

In response, in this work, we aim to fill in these gaps by building an N-MMKG oriented MMEA dataset, NMMEA, which is sourced from a large image-text pairs dataset Laion-400m (Schuhmann et al., 2021). We carefully process and formalize Laion-400m using the large language model CLIP (Radford et al., 2021) to fit the requirement of N-MMEA. The resulting NMMEA contains 29,607 entities total in two MMKGs with two modalities and 20,832 aligned entity pairs. Furthermore, in order to address the challenges posed by NMMEA, we put forward a baseline model CrossEA, with heterogeneous relational path modeling and cross-modal inference components. We evaluate CrossEA on NMMEA against state-of-the-art MMEA algorithms, which demonstrate the challenges brought by NMMEA and the effectiveness of CrossEA.

Contribution. In summary, our contributions can be summarized as: (1) To the best of our knowledge, this work is the first attempt to study the MMEA task on N-MMKGs, and we construct a N-MMKG alignment dataset NMMEA to inspire following research on this task; (2) We propose an N-MMKG oriented MMEA method CrossEA to offer a preliminary solution to the challenges posed by NMMEA; (3) We conduct experiments on NMMEA using CrossEA and existing MMEA methods, which demonstrate the effectiveness of both NMMEA and CrossEA.

2 Related Work

Multi-modal knowledge graphs (MMKGs). During the construction of N-MMKG, which incor-

A-MMKG EA Datasets	A-MMKG	Ent.		Rel.			Att.		R.Triples	A.Triples
FB15K-DB15K	FB15K	14951		1345			116		592213	29395
	DB15K	14777		279			225		89197	48080
FB15K-YAGO15K	FB15K	14951		1345			116		592213	29395
	YAGO15K	15283		32			7		122886	23532
N-MMKG EA Datasets	N-MMKG	Ent.T	Ent.V	Rel.TT	Rel.TV	Rel.VV	Att.T	Att.V	R.Triples	A.Triples
NMMEA	NMMKG1	7758	7163	729	47	67	73	6	163142	34803
	NMMKG2	7597	7089	489	34	98	98	6	109834	31526

Table 1: Dataset statistics. Ent.T, Ent.V refer to the number of textual entities and visual entities, Rel.TT, Rel.TV, Rel.VV stand for the number of "textual-textual", "textual-visual" and "visual-visual" relations.

porates multi-modal data as entities, the methodology necessitates integrating a component for identifying and extracting visual entities present within images. NEIL (Chen et al., 2013) initially employs pre-trained classifiers to assign a solitary label to each image, followed by the extraction of visual relations through heuristic principles based on the spatial arrangement of the detected objects. The system GAIA (Li et al., 2020) employs object recognition coupled with detailed categorization to discern subtle concepts within news content. GAIA serves as a foundation for RESIN (Wen et al., 2021), which further specializes in extracting visual news occurrences and recognizing pertinent visual entities and concepts, functioning as arguments, from more confined resources such as individual news articles. Subsequently, MMEKG (Ma et al., 2022) enhances certain components of this process and scales it up to handle the extraction of universal events across billions of data points.

Multi-modal entity alignment (MMEA). Traditional knowledge graph entity alignment tasks have already been studied in great depth and breadth (Zhao et al., 2023; Zeng et al., 2021, 2022). Compared with traditional KGEA methods, MMEA methods typically involve integrating visual and text modalities to enhance KG-based entity alignment. Previous works, such as PoE (Liu et al., 2019), represent entities as single vectors, concatenating features from multiple modalities. HEA (Guo et al., 2021) combines attribute and entity representations in hyperbolic space, utilizing aggregated embeddings for alignment predictions. Methods like MCLEA (Lin et al., 2022) enhance intra-modal learning with contrastive methods, while MEAformer (Chen et al., 2023) improves modality fusion through hybrid frameworks. DESAlign (Wang et al., 2024) address the over-smoothing caused by semantic inconsistency and interpolating missing semantics using existing modalities. Despite their contributions, most of these methods are not suitable for the entity alignment task on N-MMKG where each entity has only one modality.

3 Construction Of NMMEA

NMMEA is sourced from Laion-400m (Schuhmann et al., 2021), which is an image-text pair dataset from random web pages crawling between 2014 and 2021.

N-MMKG construction. We first build N-MMKGs based on Laion-400m using the multi-modal graph construction framework GAIA (Li et al., 2020) [3]. We first selected 20,000 image-text pairs related to "countries", "locations" and "people" from Laion-400m. After constructing $NMMKG_1$ using GAIA on these image-text pairs, we obtained 7,163 visual entities and 7,758 textual entities. Next, we retrieved and selected 40,000 image-text pairs related to the entities in $NMMKG_1$ from Laion-400m, and similarly used GAIA to construct $NMMKG_2$, resulting in 16,581 visual entities and 16,927 textual entities.

Since it is easier to find matching entities between two knowledge graphs in the textual modality compared to the visual modality, we prioritize annotating the seed entity pairs in the textual modality of the two knowledge graphs first. Then, through the "imageOf" relation in the knowledge graphs, we identify the corresponding visual entities within each graph, thereby obtaining seed entity pairs for the other modalities. Afterwards, we will search for visual entity pairs that are not matched to textual entities via "imageOf" by comparing similar visual entities as a supplement. We use the entities in $NMMKG_1$ as target entities and search for corresponding entities in $NMMKG_2$ to form seed entity pairs. Finally, we get $NMMKG_2$ with 7089 visual entities and 7597 textual entities

[3] Details of the N-MMKGs construction based on GAIA can be found in the appendix B.

by the above matching method, the total number of entity alignment pairs is 20832.

Dataset analysis. The statistics of the dataset can be found in Table 1. We conducted a statistical analysis of the modality distribution of entities, relations and the aligned entity pairs in NMMEA.

Entities distribution Unlike the A-MMKG EA datasets, entities in NMMEA are divided into textual and visual entities, both accounting for approximately 50%. The large number of visual entities introduces a multitude of cross-modal and visual-visual relations that do not exist in existing datasets, making the knowledge graph more complex and heterogeneous. Visual entities also bring various types of aligned entity pairs, which complicates the alignment tasks.

Relations distribution The relations in the NMMEA can be categorized into three types: textual-textual, textual-visual, and visual-visual. Among these, textual-textual relations are the most numerous, while cross-modal relations are the least common. Statistics show that the "imageOf" relation accounts for 83.8% of cross-modal relations, with other relations such as "photoOf", "locatedIn" and "bornIn" also appearing frequently in cross-modal contexts. In visual-visual type relations, the "sameAs", "contain", "nearBy" and "similar" relations make up approximately 92.4%. Other relations like "beneath", "bornIn" and "dressedIn" also appear in certain specific triplets. It is evident that some relations can appear across different modalities, further highlighting the heterogeneity and complexity of the NMMEA.

EA pairs distribution NMMEA has three kinds of EA pairs, makes the number of EA pairs much larger than existing datasets. Among the 20832 aligned entity pairs, "textual-textual", "visual-visual" and "textual-visual" entity pairs account for about 40.07%, 19.97%, and 39.96%, respectively. It can be concluded that the proposed dataset exhibits a high degree of heterogeneity, effectively mirroring the complex scenarios encountered in actual MMEA applications.

Our aspiration is that these datasets will facilitate the development of more sophisticated MMEA models, and thereby provide a clearer trajectory for advancing MMEA research.

4 A Simple But Effective Method

In this section, we propose a simple but effective method CrossEA. It follows the "representation learning then alignment inference" paradigm adopted by existing works, while designing a method for learning structural feature representations based on the aggregation of features from neighbor nodes along modality meta-paths, as well as an alignment inference method that separately predicts intra-modal and cross-modal alignments. The overall framework of the model is shown in Figure 2.

4.1 Representation Learning

Entity feature encoding module. To generate visual and textual embeddings of entities, we employ RESNET (He et al., 2016) and BERT (Devlin et al., 2019) to extract features from all visual and textual entities, using the output of their last layer as the representation.

Structural feature encoding module. Due to the heterogeneous nature of NMMKG, using current representation learning algorithms that treat all nodes equally may not be able to effectively capture the structural information.

Hence, we propose to classify the paths of relations originating from different target entities according to the modalities, thus yielding a set of distinct modality paths that encompass all the modality types of relation paths appearing in the graph, referred to as "**modality meta-paths**".

Aggregation of neighbor nodes in each hop along single modality meta-path Within each modality meta-path, the modalities of the neighboring nodes at each hop are the same. We determine the weight of the neighboring nodes by calculating their similarity to the target node respectively. By weighted averaging the representations of nodes at the current hop, we obtain the structural feature representation of neighbors at each hop:

$$emb^{k\ l}_{s_i} = \sum_{j \in P^l} \alpha^l_j emb^{lj}, \tag{1}$$

where $emb^k_{s_i}$ is the l-th hop structural embedding of target entity i within certain modality meta-path k, P^l is the set of l-th neighbor nodes of target node, emb^{lj} stands for the j-th node modality feature representation among l-th hop neighbors, and the α^l_j is weight of the j-th node which can be

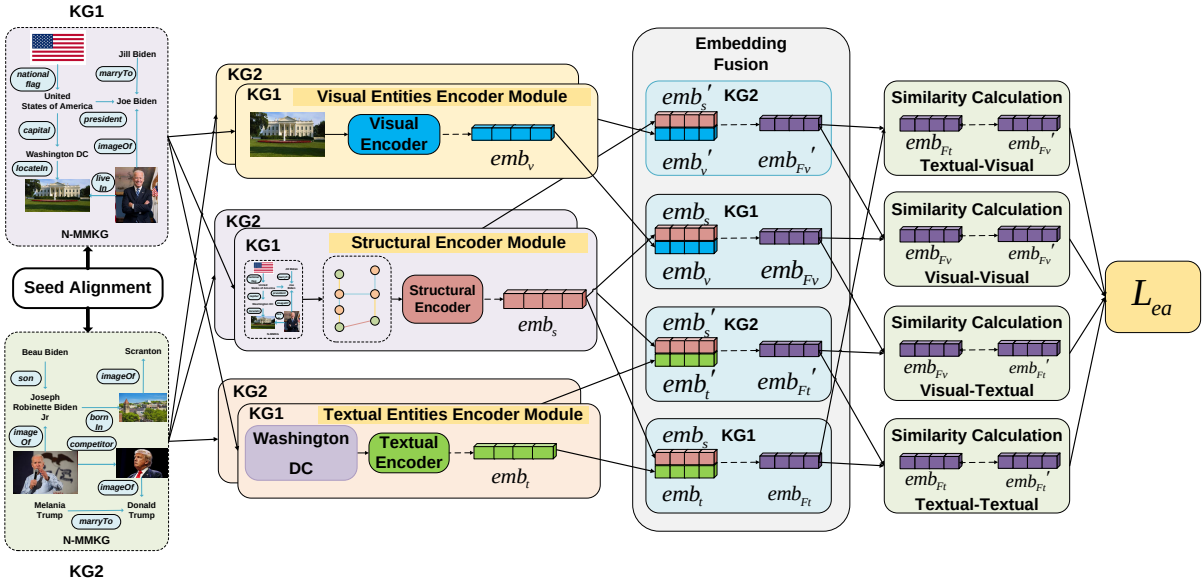


Figure 2: Overview of CrossEA. Given two N-MMKGs, CrossEA first constructs entities feature encoders and structural feature encoders to learn two kinds of embedding of each entity. The two embeddings of entities are fed into an embedding fusion module to get the representation of each entity of different modality. At last, a similarity calculation module is constructed to calculate the similarity between two entities in different N-MMKG.

calculated by:

$$\alpha_j^l = \frac{exp(sim_j^l)}{\sum_{j \in P^l} exp(sim_j^l)}, \tag{2}$$

where sim_j^l can be calculated by:

$$sim_j^l = tanh(emb^{iT} \cdot [W emb^{lj} + b]), \tag{3}$$

where emb^i is the modality feature of target entity (visual feature or textual feature).

Aggregation of all hops along single modality meta-path After obtaining the structural feature representation for each hop, we assign weights to the structural feature representations of different hops based on the proximity of the neighbors. Specifically, the calculation process of the structural feature representation of the specific target entity in the modal element path k is as follows:

$$\begin{aligned} emb_{s_i}^k &= \alpha(emb_{s_i}^{k\ 1}) + \alpha(1-\alpha)(emb_{s_i}^{k\ 2}) \\ &+ \ldots + \alpha(1-\alpha)^{(l-2)}(emb_{s_i}^{k\ (l-1)}) \\ &+ (1-\alpha)^{(l-1)}(emb_{s_i}^{k\ l}) \qquad (\alpha > 0.5), \end{aligned} \tag{4}$$

where α is a trainable weight of the first-hop neighbors, l is the highest order in the modal meta-path. The formula adheres to the principle that closer neighbors have greater weight than distant ones, with the total weight summing to 1.

Aggregation of all modality meta-paths from the target entity Next, we calculate the structural feature representations for all target entities within the same modality meta-path and average these vectors to represent the meta-path. By determining the similarity between the meta-path and the target entities, we measure how well each entity's structural information is conveyed. This similarity score serves as a weighting factor when combining different modality meta-paths. The specific calculation process is as follows:

$$emb^k = \frac{1}{N} \sum_{j=1}^{N} emb_{s_j}^k, \tag{5}$$

$$Sim_i^k = tanh(emb^{kT} \cdot [W emb_{s_i}{}^k + b]), \tag{6}$$

$$\beta_i^k = \frac{exp(Sim_i^k)}{\sum_{k \in K} exp(Sim_i^k)}, \tag{7}$$

$$emb_s^i = \sum_{k \in K} \beta_i^k emb_{s_i}{}^k, \tag{8}$$

where N is the number of initial entities within modality meta-path k, K is the set of modality meta-paths which contain the target entity i as initial entity. emb_s^i represent the structural feature representation of the target entity i.

4.2 Alignment Inference

After weighting and averaging the modality and structural features of entities, we obtain their rep-

resentations. To resolve the non-comparability between cross-modal and intra-modal similarities, we negate the representation vectors of visual entities when calculating similarities. This makes cross-modal similarities negative and intra-modal similarities positive. During training, we use the absolute values of the similarities and aim to increase the similarity for matching pairs. In testing, we apply different thresholds to positive and negative similarities to evaluate cross-modal and intra-modal matches separately. Entities are considered a match if the absolute value of their similarity exceeds the respective threshold. The specific similarity denoted as $Sim_{xy}(emb_{F_x}, emb^{'}_{F_y}) \quad (x, y \in t, v)$:

$$
\begin{aligned}
&Sim_{xy}(emb_{F_x}, emb^{'}_{F_y}) \\
&\quad = \frac{(-1)^n emb_{F_x} \cdot (-1)^m emb^{'}_{F_y}}{||emb_{F_x}||_2 ||emb^{'}_{F_y}||_2} \\
&\qquad n = \begin{cases} 1 & x = v \\ 2 & x = t \end{cases} \quad m = \begin{cases} 1 & y = v \\ 2 & y = t. \end{cases}
\end{aligned} \tag{9}
$$

4.3 Model Training

To ensure matching entities are closely positioned within the vector space, we construct a negative example set E'_s by replacing one entity in each pair within seed entities pairs set E_s. The model's training process revolves around optimizing a margin-based ranking loss function, which aims to minimize the distances between matching entities effectively. The margin-based ranking loss function is as followed:

$$
\begin{aligned}
L = \sum_{(i,j) \in E_s} \sum_{(i',j') \in E'_s} &[|Sim(emb^i_{F_x}, emb^j_{F_x})| \\
&+ \eta - |Sim(emb^{i}_{F_x}{}^{'}, emb^{j}_{F_x}{}^{'})|]_+,
\end{aligned} \tag{10}
$$

where $[X]_+ = max\{0, x\}$, η is the margin hyper-parameter separating positive and negative instances.

5 EXPERIMENTAL STUDY

In this section, we conduct extensive experimental studies to verify the effectiveness of our proposed method CrossEA.

5.1 Experimental Setting

Dataset. We conduct extensive experiments on our new proposed dataset NMMEA and use three different training data proportions 10%, 30% and 50%. The statistic of the NMMEA is summarized in Table 1.

Baselines. A-MMKG entity alignment methods:

(1) **POE** (Liu et al., 2019), which represent entities as single vectors, concatenating features from multiple modalities.

(2) **ACK-MMKG** (Li et al., 2023), which compensate the context gaps through incorporating consistent alignment knowledge.

(3) **MCLEA** (Lin et al., 2022), which enhance intra-modal learning with contrastive methods.

(4) **MEAformer** (Chen et al., 2023), which improves modality fusion through hybrid frameworks.

(5) **DESAlign** (Wang et al., 2024), which address the over-smoothing caused by semantic inconsistency and interpolating missing semantics using existing modalities.

CLIP (Radford et al., 2021) model can integrate image and text into a vector space to calculate similarity. For MMEA task based on N-MMKG, the alignment of visual entities and textual entities is required, alignment methods based on CLIP are worth trying. Since the knowledge graph contains structural information, we use GAT (Velickovic et al., 2018) to learn the structural information in the graph and integrate it into CLIP to strengthen its entity alignment ability.

CLIP based MMEA methods:

(1) **CLIP-MMEA**: The visual, textual encoder and similarity calculation methods of CLIP are applied to multi-modal entity alignment.

(2) **SE-CLIP-MMEA**: A similarity of structure is calculated by using the representation of entity structural information, and the final similarity is obtained by weighted averaging the structural similarity and the similarity calculated by CLIP.

(3) **SE-CLIP-MMEA+**: The representation of entity structural information in the graph is integrated into the visual representation and textual representation encoded by CLIP, and then the similarity of entity pairs is calculated.

Evaluation settings. When dealing with entity alignment task on N-MMKG, an entity in one N-MMKG might be aligned to several entities (same or different modality), therefore, commonly used metrics such as $Hits@k$ are no longer suitable for this type of entity alignment task. For this reason, we have chosen $Precision$, $Recall$, and $F1$ score as the evaluation metrics for our experiments.

Models		NMMEA-10%			NMMEA-30%			NMMEA-50%		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
A-MMKG Entity Alignment Methods	POE	0.564	0.378	0.453	0.612	0.364	0.456	0.657	0.394	0.493
	ACK-MMKG	0.763	0.537	0.631	0.792	0.541	0.643	0.810	0.562	0.664
	MCLEA	0.787	0.583	0.670	0.798	0.603	0.687	0.818	0.591	0.686
	MEAformer	0.806	0.613	0.696	0.823	0.618	0.706	0.828	0.624	0.712
	DESAlign	0.813	0.637	0.715	0.831	0.635	0.720	0.834	0.638	0.723
CLIP Based Methods	CLIP-MMEA	0.727	0.539	0.619	0.733	0.612	0.667	0.758	0.630	0.688
	SE-CLIP-MMEA	0.807	0.624	0.704	0.826	0.643	0.723	0.839	0.638	0.725
	SE-CLIP-MMEA+	0.802	0.628	0.704	0.834	0.639	0.724	0.841	0.647	0.731
Ours	CrossEA	**0.826**	**0.717**	**0.766**	**0.839**	**0.743**	**0.788**	**0.843**	**0.793**	**0.817**

Table 2: Main entity alignment results on different training data proportions of NMMEA.

Models	NMMEA-30%		
	Precision	Recall	F1
POE-md	0.621	0.453	0.524
ACK-MMKG-md	0.793	0.637	0.706
MCLEA-md	0.802	0.672	0.731
MEAformer-md	**0.836**	0.694	0.758
DESAlign-md	0.829	0.703	0.761
CLIP-MMEA-md	0.751	0.687	0.718
SE-CLIP-MME-md	0.819	0.714	0.763
SE-CLIP-MMEA+-md	0.833	**0.720**	**0.772**

Table 3: Results of A-MMKG entity alignment methods and CLIP-based methods with alignment inference method mentioned in Section 4.

5.2 Main Results

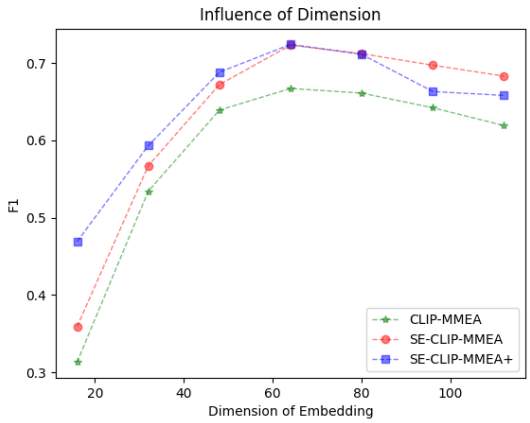
Comparison with A-MMKG entity algnment methods.

As can be seen in Table 2, compared to our model, the methods based on A-MMKG are much inferior to CrossEA. Moreover, the larger the proportion of the training set, the greater the gap of $Recall$, reaching a maximum decrease of 0.155. We hypothesize that this is because the methods based on A-MMKG often omit cross-modal aligned entity pairs during prediction, leading to the lower $Recall$. In order to test our hypothesis, we use the alignment inference method mentioned in Section 4 on each A-MMKG entity alignment method, and test the improved methods on NMMEA, as shown in Table 3. It can be seen from the experimental results that $Recall$ is improved to a large extent, which illustrates the significant impact that the incomparability between cross-modal and intra-modal entity similarities can have on the alignment effect. At the same time, it can be seen that the improved experimental results still have a significant gap compared with the results of CrossEA, which indicates that the heterogeneity structure of N-MMKGs also affects the effect of the A-MMKG entity alignment methods to a large extent.

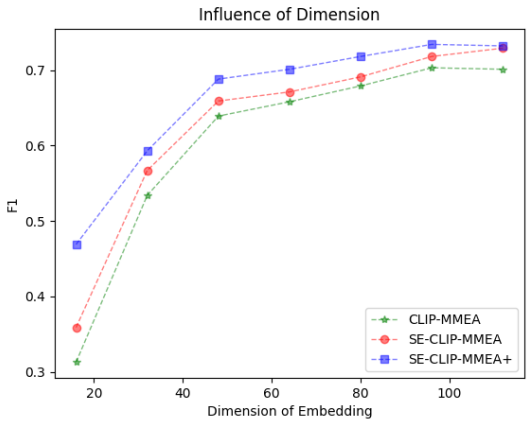
Comparison with CLIP based methods. We test three CLIP-based entity alignment methods on NMMEA, and the experimental results show that the SE-CLIP-MMEA+ has some improvement over the A-MMKG entity alignment methods on the N-MMKG dataset. Specifically, when the training data proportion is 30%, compared with the SOTA method DESAlign, the $F1$ is improved by 0.004, and with DESAlign-md, the $F1$ is improved by 0.011. Experimental results indicate that, for tasks requiring cross-modal entity alignment, introducing CLIP, which are designed for image-text matching, can yield considerable improvements.

However, compared with CrossEA, CLIP-based methods still have certain drawbacks when applied to multi-modal entity alignment tasks. In addition to the fact that it cannot handle the heterogeneity structure of N-MMKGs, their performance in entity alignment tends to degrade on N-MMKGs with a large number of entities. This is because CLIP model is essentially designed for image-text matching, with its pre-training process conducted in batches. However, when confronted with the task of entity alignment, the challenge lies in evaluating the similarity between a single entity from one graph and all entities in another graph. This requirement drastically expands the scope of similarity computations during model fine-tuning, far exceeding the scale encountered during pre-training. Such a substantial inflation in computational scale leads to significant adjustments in model parameters, which in turn negatively impact experimental results.

To validate this limitation, we reduced the dataset size by a factor of 10, resulting in a graph with 1,500 entities, and recorded the results of the CLIP-based methods at different embedding dimen-



(a) Results in NMMEA



(b) Results in reduced NMMEA

Figure 3: $F1$ of the CLIP-based methods under different embedding dimensions in NMMEA and the reduced dataset

Models	FB15K-DB15K	FB15K-YAGO15K	NMMEA
	Hits@1	Hits@1	Hits@1
POE	0.666	0.573	0.543
ACK-MMKG	0.682	0.676	0.657
MCLEA	0.730	0.667	0.642
MEAformer	0.762	0.703	0.695
DESAlign	**0.805**	**0.728**	**0.714**

Table 4: Results of A-MMKGs entity alignment methods on different datasets.

Models	NMMEA-30%		
	Precision	Recall	F1
CrossEA	**0.839**	**0.743**	**0.788**
CrossEA w/o SL	0.753	0.633	0.688
CrossEA w/o AIM	0.814	0.649	0.722

Table 5: Results of ablation study.

sions on both datasets, as shown in the Figure 3. The best $F1$ on smaller-scale datasets (0.737) is better than those on larger-scale datasets (0.724). Furthermore, when the vector dimension reaches a certain level on the larger datasets, there is a substantial decline in the experimental results.

5.3 Comparison of Datasets

In order to analysis the difficulty of NMMEA, we test existing A-MMKG entity alignment methods on three datasets, the results are shown in Table 4. For unifying evaluation indicators, $Hits@1$ was used here to evaluate experimental results. It can be seen that compared with the results on A-MMKG entity alignment datasets (FB15K-DB15K, FB15K-YAGO15K), the results on NMMEA have different degrees of decline. It is evident that the modality diversity of entities, relations, and entity pairs in NMMEA gives it a more complex structure, which significantly impacts the model during both representation learning and alignment inference.

5.4 Ablation Study

To evaluate the effectiveness of components in CrossEA, we conduct the ablation study. Firstly, we remove the structural feature learning module mentioned in Section 4.1, and replace it by GAT. It can be seen in Table 5 that the $F1$ of CrossEA w/o SL declines by 0.100 which prove the significance of overcoming the heterogeneous in structural representation learning.

Secondly, we remove the alignment inference method mentioned in Section 4.2, the $Recall$ of CrossEA w/o AIM declines by 0.094 which is very much larger than the decline of $Precision$. This demonstrates that the alignment inference method we propose can effectively handle situations where an entity may correspond to entities in multiple modalities.

5.5 Significance Testing

To conduct significance tests on the experimental results and further verify the effectiveness of CrossEA, we conducted experiments for 10 times for both CrossEA and the second best solution DESAlign. The specific results are shown in table 6. From the table, we calculated the p-value using the F1 scores obtained from these 10 experiments, which is $2.126 \times 10^{-}17$, much lower than 0.05. This indicates that, with 95% confidence, CrossEA shows a significant improvement over DESAlign in terms of performance.

	F1									
CrossEA	0.768	0.763	0.759	0.769	0.766	0.763	0.769	0.773	0.764	0.768
DESAlign	0.714	0.718	0.717	0.712	0.711	0.714	0.713	0.721	0.714	0.715

Table 6: Results of significance testing

6 Conclusion

This paper introduces a novel benchmark—NMMEA, which more accurately reflects the complex multi-modal entities and heterogeneous relations of MMKG in realistic applications. In conjunction with this, we propose an efficient and practical method CrossEA tailored to address these complex challenges. Our findings not only highlight the intricacy of the new benchmark but also attest to the efficacy of the suggested solution.

Moving forwards, for the point of more advanced methods, to address the limitations of traditional MMEA methods in effectively extracting structural information, prioritizing the development of more advanced models becomes crucial. This entails exploring new MMEA architectures capable of handling highly heterogeneous structures more efficiently, as well as incorporating sophisticated GNN techniques to delve deeper into the complex structural associations among multi-modal entities within N-MMKGs. These enhancements aim to facilitate the creation of more comprehensive and performance-enhanced MMEA solutions.

Limitations

In this section, we faithfully discuss the limitations that we would like to improve in future work.

Firstly, although the proposed CrossEA achieved good results on the dataset, there were still some incorrect judgments for entity pairs, such as the visual entity "teamFranklin" being aligned with both the textual entity "teamFranklin" and the textual entity "FranklinSchool." This situation arises because CrossEA uses the same optimal similarity threshold for all entity pairs to predict alignments, leading to some non-aligned entity pairs having similarity scores higher than the threshold. In future work, we will investigate a method to dynamically set similarity thresholds for each entity pair to improve this issue.

Secondly, the method CrossEA we designed may encounter the issue of redundant high-order neighbor node features when computing structural characteristics due to overly long relational paths in larger and more complex knowledge graphs. This is a direction for improvement in our future work.

Thirdly, in this paper, We only use data from two modalities, images and text, as entities in the knowledge graph for entity alignment, focusing on these two modalities as a starting point to study the new multi-modal entity alignment task. In future work, we will explore the alignment task on multi-modal knowledge graphs that incorporate additional modalities such as audio, video, heatmaps, etc.

Acknowledgements

This work was partially supported by National Key R&D Program of China (No. 2022YFB3103600), NSFC (Nos. U23A20296, 62272469, 62302513), and The Science and Technology Innovation Program of Hunan Province (No. 2023RC1007).

References

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*, volume 12274 of *Lecture Notes in Computer Science*, pages 134–147. Springer.

Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: extracting visual knowledge from web data. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1409–1416. IEEE Computer Society.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z. Pan, Wenting Song, and Huajun Chen. 2023. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 3317–3327. ACM.

Bo Cheng, Jia Zhu, and Meimei Guo. 2022. Multijaf: Multi-modal joint entity alignment framework for multi-modal knowledge graph. *Neurocomputing*, 500:581–591.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing*, 461:598–607.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare R. Voss, Daniel Napierski, and Marjorie Freedman. 2020. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 77–86. Association for Computational Linguistics.

Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. 2023. Attribute-consistent knowledge graph representation learning for multi-modal entity alignment. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2499–2508. ACM.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2572–2584. International Committee on Computational Linguistics.

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*, volume 11503 of *Lecture Notes in Computer Science*, pages 459–474. Springer.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, and Jing Shao. 2022. MMEKG: multi-modal event knowledge graph towards universal representation across modalities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 231–239. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114.

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*, pages 1405–1414. ACM.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yuanyi Wang, Haifeng Sun, Jiabo Wang, Jingyu Wang, Wei Tang, Qi Qi, Shaoling Sun, and Jianxin Liao. 2024. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. *CoRR*, abs/2401.17859.

Haoyang Wen, Ying Lin, Tuan Manh Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Ren Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 133–143. Association for Computational Linguistics.

Weixin Zeng, Xiang Zhao, Xinyi Li, Jiuyang Tang, and Wei Wang. 2022. On entity alignment at scale. *VLDB J.*, 31(5):1009–1033.

Weixin Zeng, Xiang Zhao, Zhen Tan, Jiuyang Tang, and Xueqi Cheng. 2023a. Matching knowledge graphs in entity embedding spaces: An experimental study. *IEEE Trans. Knowl. Data Eng.*, 35(12):12770–12784.

Weixin Zeng, Xiang Zhao, Jiuyang Tang, Xuemin Lin, and Paul Groth. 2021. Reinforcement learning-based collective entity alignment with adaptive features. *ACM Trans. Inf. Syst.*, 39(3):26:1–26:31.

Yawen Zeng, Qin Jin, Tengfei Bao, and Wenfeng Li. 2023b. Multi-modal knowledge hypergraph for diverse image retrieval. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 3376–3383. AAAI Press.

Yichi Zhang, Zhuo Chen, and Wen Zhang. 2023. MACO: A modality adversarial and contrastive framework for modality-missing multi-modal knowledge graph completion. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part I*, volume 14302 of *Lecture Notes in Computer Science*, pages 123–134. Springer.

Xiang Zhao, Weixin Zeng, and Jiuyang Tang. 2023. *Entity Alignment - Concepts, Recent Advances and Novel Approaches*. Springer.

Zhengwei Zhao, Xiaodong Wang, Xiaowei Xu, and Qing Wang. 2019. Multi-modal question answering system driven by domain knowledge graph. In *5th International Conference on Big Data Computing and Communications, BIGCOM 2019, QingDao, China, August 9-11, 2019*, pages 43–47. IEEE.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2024. Multi-modal knowledge graph construction and application: A survey. *IEEE Trans. Knowl. Data Eng.*, 36(2):715–735.

A Appendix

N-MMKG. N-MMKG takes multi-modal data (images in this paper) as entities, which denoted as $G = \{E, R, A, V, T_R, T_A\}$, where E, R, A, V represent the set of entities, relations, attributes and attribute values respectively, $T_R = (E_T \cup E_V) \times R \times (E_T \cup E_V)$ is the set of relation triples, E_T is the set of textual entities and E_V is the set of visual entities.

Multi-model Entity Alignment Based on N-MMKG. Given two N-MMKGs, $KG_1 = (E_1, R_1, A_1, V_1, T_{R_1}, T_{A_1})$ and $KG_2 = (E_2, R_2, A_2, V_2, T_{R_2}, T_{A_2})$, multi-model entity alignment based on N-MMKG aims to obtain the identical entity set $S = \{(e_{i_T}, e_{i_V}) \cup (e_{j_T}, e_{k_T}) \cup (e_{j_V}, e_{k_V}) \parallel e_{i_T}, e_{j_T}, e_{k_T} \in E_T, e_{i_V}, e_{j_V}, e_{k_V} \in E_V\}$ represent the same real-world entity. The three different types of entity pairs ((e_{i_T}, e_{i_V}), (e_{j_T}, e_{k_T}), (e_{j_V}, e_{k_V})) represent the cross-modal and intra-modal alignments that exist between the two entities.

B Appendix

N-MMKGs construction method based on GAIA. The N-MMKGs construction method based on GAIA includes three branches: textual knowledge extraction, visual knowledge extraction and MMKG fusion. In the textual knowledge extraction branch, an LSTM-CRF model is used for the coarse-grained extraction of entities, relations, and events. Entities are then clustered using entity linking and coreference resolution to group identical entities. For entities that cannot be linked to a background knowledge base, heuristic rules are applied to form NIL clusters. Additionally, an attention-based fine-grained type classification model is developed to determine the fine-grained types of entities. Finally, a weight score is assigned to each entity in a document to better filter information.

In the visual knowledge extraction branch, convolutional neural networks (CNNs) are employed to extract features from images, and deep learning models are used for semantic segmentation of the images. The results of semantic segmentation are then utilized to identify elements such as scenes, objects, and activities within the images.

In the MMKG fusion branch, for each text-extracted entity, GAIA uses an ELMo model on the text and sentence, compares it with surrounding images' CNN feature maps to generate a relevance score and heatmap. For relevant images, a bounding box is derived from the heatmap, compared with known visual entities, and assigned to the best match. After linking textual and visual entities via NIL clustering, a N-MMKG is constructed.