# Pseudo-Label Calibration Semi-supervised Multi-Modal Entity Alignment

**Luyao Wang, Pengnian Qi, Xigang Bao, Chunlai Zhou, Biao Qin**[*]

School of Information, Renmin University of China
{wangluyao123, pengnianqi, baoxigang, czhou, qinbiao} @ruc.edu.cn

## Abstract

Multi-modal entity alignment (MMEA) aims to identify equivalent entities between two multi-modal knowledge graphs for integration. Unfortunately, prior arts have attempted to improve the interaction and fusion of multi-modal information, which have overlooked the influence of modal-specific noise and the usage of labeled and unlabeled data in semi-supervised settings. In this work, we introduce a Pseudo-label Calibration Multi-modal Entity Alignment (PCMEA) in a semi-supervised way. Specifically, in order to generate holistic entity representations, we first devise various embedding modules and attention mechanisms to extract visual, structural, relational, and attribute features. Different from the prior direct fusion methods, we next propose to exploit mutual information maximization to filter the modal-specific noise and to augment modal-invariant commonality. Then, we combine pseudo-label calibration with momentum-based contrastive learning to make full use of the labeled and unlabeled data, which improves the quality of pseudo-label and pulls aligned entities closer. Finally, extensive experiments on two MMEA datasets demonstrate the effectiveness of our PCMEA, which yields state-of-the-art performance.

## Introduction

Multi-modal knowledge graphs (MMKGs) have drawn massive attention in various scenarios and motivated numerous downstream applications (Sun et al. 2020a; Ding et al. 2022; Shao et al. 2023). In MMKGs, knowledge is often summarized in various forms, such as relation triples, attribute triples, and images. Generally, MMKGs are constructed for specific purposes, leading to separate MMKGs with different descriptions for identical concepts. To improve the completeness of MMKGs, multi-modal entity alignment (MMEA) is an emerging tasks that link entities referring to the same real-world concept.

Figure 1 illustrates a toy example in MMEA. Commonly, aligned entities share similarities in attributes, relations, topology, or visual information. Thus, recent works (Chen et al. 2020; Lin et al. 2022) design interaction and fusion methods to integrate multi-modal embeddings. Nevertheless, 1) direct interaction and fusion introduce richer information and modal-specific noise. For instance, the entity */m/0r6c4* in FreeBase and the entity *Mountain_View,_California* in Dbpedia have similar relation (e.g. "Time zone" vs. "timeZone") and attribute (e.g. "population_number" vs. "populationTotal"), all of which favor the alignment of the two entities. But these entities may be incorrectly aligned due to significant visual differences. 2) The exploration of optimal embedding methods for each modality is often neglected. For example, "longitude" and "wgs84_pos#long" represent the full name and abbreviation, and the bag-of-words method might maximize similarity. Conversely, embedding by the pre-trained language model might be better for the case where "birthplace" and "people_born_here" are different phrases with the same meaning. In addition, most existing methods exploit embedding-based approaches relying heavily on human labeling. However, 3) the design of the training strategy with limited labels has been overlooked in MMEA. Many methods only use existing labels for supervised learning, neither fully utilizing unlabeled data nor preventing model bias.

**Contributions:** To address the abovementioned problems, we introduce PCMEA, a Pseudo-label Calibration based semi-supervised MMEA framework. It has three main components: PCMEA first utilizes diverse encoders and attention mechanisms to obtain modality-specific representations for each entity. To exploit complementarities across modalities and filter out model-specific noise, PCMEA then employs mutual information-enhanced cross-modality alignment methods, which can enrich intra-modal interaction and avoid the influence of noise. To leverage labeled and unlabeled data, PCMEA finally develops momentum-based contrastive learning with pseudo-label calibration, which can reduce error propagation and help to align entities. Experimental results show that our approach achieves state-of-the-art performance on two MMEA benchmark datasets.

## Related Work

**Entity alignment (EA) and multi-modal entity alignment.** With developments in knowledge graph representation learning, embedding-based entity alignment has emerged. Those embedding-based methods commonly have two steps: 1)KG embedding module encodes the entities into vectors according to the semantic or structural information; 2)entity alignment (Sun et al. 2020b) module captures the

[*]Corresponding author.

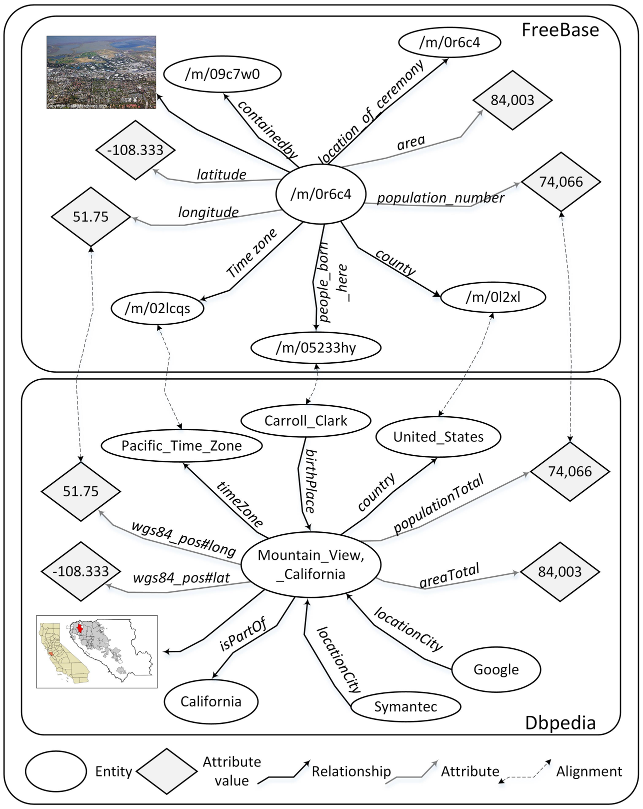


Figure 1: An example of multi-modal entity alignment. The oval shape represents entities, and the diamond shape represents attribute values. The dotted line indicates the relation or attribute of the alignment of two aligned entities.

correspondence of embedding vectors with seed alignment.

Recently, lots of multi-modal knowledge graphs have been constructed. Information from different independent modalities can complement each other. Nevertheless, their corresponding representations, reflected in separate spaces, can not directly merge in a shared space. A fusion module in MMEA (Chen et al. 2020) is proposed to migrate knowledge embeddings from multiple modalities for robustness. A novel method in MSNEA (Chen et al. 2022a) uses visual features to guide other modalities learning, promoting inter-modal enhanced entity representation. In addition to the modal embedding perspective, an intra-modal contrastive loss in MCLEA (Lin et al. 2022) is utilized to distinguish the embeddings of equivalent entities from other entities for each modality.

Different from previous methods, our proposed PCMEA not only strengthens the complementarity between different modal representations using mutual information and attention mechanism but also devises a more effective contrastive learning strategy to pull aligned entities closer.

**Semi-supervised and unsupervised entity alignment.** Recently, some EA methods based on representation learning have been prevalent due to their accuracy. However, their success often relies on annotated data, which means higher labor costs. For situations with few or no aligned seeds, several semi-supervised or unsupervised EA approaches have been proposed. Methods like CEAFF (Zeng et al. 2021b) and RLEA (Guo et al. 2022) transform the alignment process into sequence decision task. RAC (Zeng et al. 2021a) conducts reinforced active learning by selecting entities to manually label with minimal labeling efforts and exploit vast unlabeled data. Meanwhile, self-supervised EA methods like EVA (Liu et al. 2021) and ICLEA (Zeng et al. 2022) create pre-aligned entity pairs by leveraging visual similarity or contrastive learning. However, those semi-supervised or self-supervised methods usually suffer from error accumulation, leading to performance bottlenecks. The main reason is no guarantee of the accuracy of the pre-decision or pre-alignments. Therefore, in PCMEA, we calibrate the pseudo-labels before incorporating them into the supervised contrastive learning framework, decreasing error accumulation.

# Methodology

## Problem Definition and Notations

*Definition 1:Multi-modal Knowledge Graph.* A multi-modal knowledge graph is formalized as $G = (E, R, I, A, V, T_R, T_A)$. Here, $E, R, I, A$, and $V$ denote the sets of entities, relations, images, attributes, and values, respectively. $T_R = \{(h, r, t)|h, t \in E, r \in R\}$ refers to the set of relation triples. $T_A = \{(e, a, v)|e \in E, a \in A, v \in V\}$ denotes the set of attribute triples.

*Definition 2:Multi-modal Entity alignment.* Given two multi-modal knowledge graphs $G$ and $G'$, $G$=$(E, R, I, A, V, T_R, T_A)$ and $G'$=$(E', R', I', A', V', T'_R, T'_A)$, the set of alignment seeds across two multi-modal knowledge graphs is defined as $H = \{(e, e')|e \in E, e' \in E', e \equiv e'\}$, where $\equiv$ represents the equivalence of two entities. The task of multi-modal entity alignment targets to match the counterpart entities $e$ and $e'$, which describe the same concepts in the real world from distinct multi-modal knowledge graphs.

## Framework Overview

In this paper, we introduce a semi-supervised multi-modal entity alignment framework called PCMEA to solve the challenges above. Our proposed PCMEA comprises three components: *Attention-guided Multi-modal Embedding* to extract visual, relation, attribute, and structure features with diverse encoders and attention mechanisms; *Mutual Information Enhanced Cross-modality Alignment* to encourage cross-modality knowledge transfer and to filter modality-invariant noise; *Contrastive Learning with Pseudo-Label Calibration* method to help align entities with a few label supervision.

## Attention-guided Multi-modal Embedding

In multi-modal knowledge graphs, there are various modalities of knowledge to depict an entity, i.e., neighborhood structure, relations, attributes, and images. Each modality is processed using different encoders depending on the nature of the signal. Furthermore, uni-modal embeddings are fused with weighted concatenation to form the joint embedding.

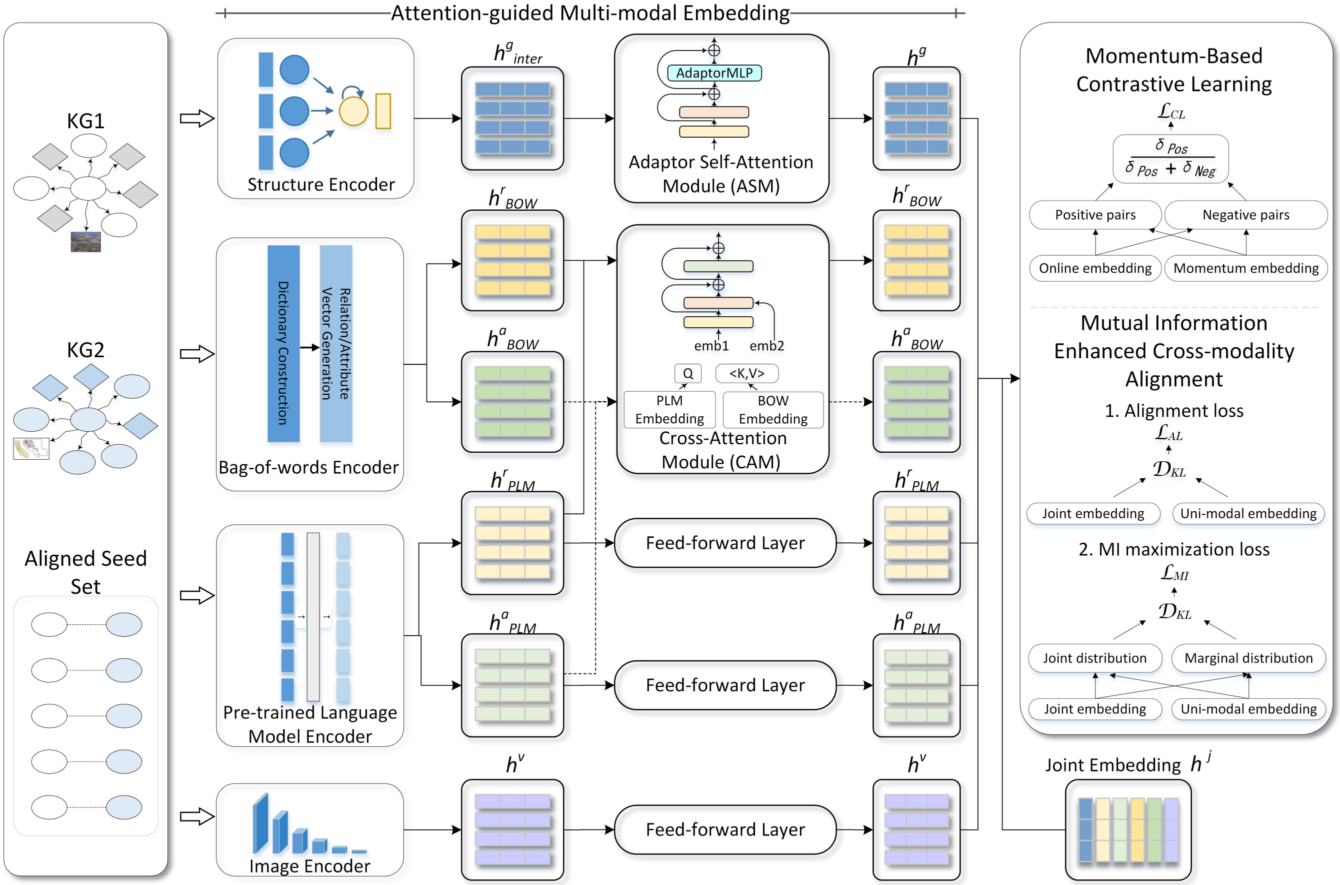


Figure 2: The overall architecture of PCMEA, which combines heterogeneous multi-modal attention-guided embedding and learns through MI maximization enhanced alignment loss (consists of alignment loss $\mathcal{L}_{AL}$ and MI maximization loss $\mathcal{L}_{MI}$), and momentum-based contrastive loss $\mathcal{L}_{CL}$.

**Self-attention Augmented Structure Embedding.** We utilize Graph Attention Networks (GAT) (Veličković et al. 2018) to model the neighborhood structure information of entities as real-valued vectors, since GAT aggregates neighborhood information with the attention mechanism and focuses on the most relevant neighbors. In practice, we apply a two-layer GAT model to embed the neighborhood information. We use the output of the last layer as the intermediate representation of neighbor structure embedding $h^g_{inter}$.

Because diverse modalities have different statistical properties, which are distributed cross feature spaces, we apply an adaptor self-attention module (ASM) to narrow the semantic gap. The main components of ASM are a multi-head self-attention (MHSA) layer and a plug-and-play bottleneck layer called AdaptorMLP (Chen et al. 2022b). The processes of $h^g_{inter}$ and $h^g$ are formulated as:

$$h^g_{inter} = W_g \cdot GAT_2(g_i) + b_g \tag{1}$$

$$h^g = ASM(h^g_{inter}) \tag{2}$$

where $W_g$ and $b_g$ are learnable parameters.

**Semantic Information Guided Relation and Attribute Embedding.** In KGs, the relation triples (attribute triples) of the corresponding entities have similarities in characters or semantics. Thus, we adopt two approaches to embed the relations (attributes) triples. On one hand, we represent the relations (attributions) of entities $e_i$ as bag-of-words features and feed them into a feed-forward layer to obtain the relation embedding $h^r_{BOW}$ (attribute embedding $h^a_{BOW}$). On the other hand, we expand the relation (attribute) triples $t_m$ into word sequences $s_m$ and input $s_m$ into a pre-trained language model, which can understand the meaning of sentences. In our work, we apply T5 (Raffel et al. 2020) and Roberta (Liu et al. 2019b) to encode relation and attribute triples, respectively. After a feed-forward layer, the semantic representation of relation (attribute) triples can be obtained.

$$h^m_{BOW} = W_m \cdot BOW(t_m) + b_m, m \in \{r, a\} \tag{3}$$

$$h^m_{PLM} = PLM(s_m), m \in \{r, a\} \tag{4}$$

$$h^m_{PLM} = W'_m \cdot h^m_{PLM} + b'_m, m \in \{r, a\} \tag{5}$$

where $W_m$, $W'_m$, $b_m$, and $b'_m$ are learnable parameters.

The character and semantic information of relation triples and attribute triples can be mutually complemented. Therefore, we adopt a cross-attention module and use semantic information to guide further learning of character information. The cross-attention module (CAM) is very similar to multi-head self-attention, with the difference that semantic embeddings $h^m_{PLM}$ are used as query inputs and character embeddings $h^m_{BOW}$ are used as the input of keys and values. The structure of CAM is shown in Figure 2.

$$h^m_{BOW} = CAM(h^m_{BOW}, h^m_{PLM}), m \in \{r, a\} \tag{6}$$

**Visual Information Embedding.** For visual information embedding, we follow (Lin et al. 2022) and adopt the pre-trained visual model (PVM) to learn visual representation. For consistency, we use the embeddings generated by (Lin et al. 2022). The visual representation is sent through a feed-forward layer to get final visual embedding $h^v$.

$$h^v = W_v \cdot PVM(v_i) + b_v \tag{7}$$

where $W_v$ and $b_v$ are learnable parameters.

**Joint Embedding** Inspired by recent works (Lin et al. 2022; Chen et al. 2022a, 2023), we implement a simple weighted concatenation by integrating the multi-modal features into a joint representation $h^j$:

$$h^j = \underset{m}{Concat}\left[softmax(\alpha^m) \cdot h^m\right] \tag{8}$$

where $m$ denotes the modal type and $m \in \{g, r_{BOW}, r_{PLM}, a_{BOW}, a_{PLM}, v\}$, which means that $h^m \in \{h^g, h^r_{BOW}, h^r_{PLM}, h^a_{BOW}, h^a_{PLM}, h^v\}$. $\alpha^m$ is the trainable attention weight for the modality of $m$. $Concat$ means concatenation operation.

## Mutual Information Enhanced Cross-modality Alignment

Different modality information enriches the entity representation from different perspectives. The joint embedding generated from fusion mechanism is more comprehensive than uni-modal embedding. Thus, aligning uni-modal embedding with the joint embedding can transfer the knowledge from the joint embedding back to uni-modal ones, resulting in better uni-modal representation. Concretely, we minimize the Align Loss ($\mathcal{L}_{AL}$) to reduce the difference between uni-modality and joint modality and to realize the knowledge transfer. For aligned pair $(e^i_1, e^i_2)$:

$$\begin{aligned}\mathcal{L}^m_{AL} = -\sum_m E_{i\in B}\big[&D_{KL}(\mathbb{Q}_{h^j}(e^i_1, e^i_2)||\mathbb{Q}_{h^m}(e^i_1, e^i_2)) \\ &+D_{KL}(\mathbb{Q}_{h^j}(e^i_2, e^i_1)||\mathbb{Q}_{h^m}(e^i_2, e^i_1))\big]\end{aligned} \tag{9}$$

where $D_{KL}(\cdot)$ represents the Kullback-Leibler Divergence, $\mathbb{Q}_{h^j}(e^i_1, e^i_2) = h^j_{e^i_1} \otimes h^j_{e^i_2}$ and $h^j_{e^i_1}$ is the joint embedding of $e^i_1$. $\mathbb{Q}_{h^m}(e^i_1, e^i_2)$ is calculated in an analogous way using the uni-modal embedding $h^m$, $m \in \{g, r_{BOW}, r_{PLM}, a_{BOW}, a_{PLM}, v\}$.

However, direct alignment of cross-modality only encourages integrating information from different modalities while mixing the noise from each modality irrelevant to our task. Thus, we use the mutual information estimator MINE (Belghazi et al. 2018) to enhance mutual information, which can be utilized to mine the modal-invariant information between different modalities and filter out modality-specific random noise (Qi and Qin 2023; Bao et al. 2023). Specifically, we maximize the MI between joint embedding $h^j$ and uni-modal embedding $h^m$:

$$\begin{aligned}I(\widehat{h^j, h^m}) &= max\ I(h^j; h^m) \\ &= max\ D_{KL}(\mathbb{P}_{h^j h^m} \mid\mid \mathbb{P}_{h^j} \otimes \mathbb{P}_{h^m}) \\ &= sup\ \mathbb{E}_{\mathbb{P}_{h^j h^m}}[\Phi] - log(\mathbb{E}_{\mathbb{P}_{h^j} \otimes \mathbb{P}_{h^m}})[e^{\Phi}]\end{aligned} \tag{10}$$

where $\mathbb{P}_{h^j h^m}$ represents joint distribution, $\mathbb{P}_{h^j}$ and $\mathbb{P}_{h^m}$ are marginal distributions of joint embedding $h^j$ and uni-modal embedding $h^m$, respectively. $sup$ represents supremum function and $\Phi$ is a simple nonlinear layer. Specifically, in the multi-modal entity alignment task, the loss for MI maximization is:

$$\mathcal{L}_{MI} = -\sum_m I(\widehat{h^j, h^m}) \tag{11}$$

where $h^m \in \{h^g, h^r_{PLM}, h^a_{PLM}, h^v\}$ is uni-modal entity representation, $h^j$ denotes the joint-modal embedding.

## Contrastive Learning with Pseudo-Label Calibration

In this section, contrastive learning with pseudo-label calibration strategy is designed for semi-supervised EA. It mainly consists of two parts: (1) *Pseudo-label calibration* provides more reliable pseudo-aligned entity pairs to expand the training data set and decrease error propagation. (2) *Momentum-based contrastive learning mechanism* can be more effective in pulling the aligned pairs closer and pushing unaligned pairs away.

**Pseudo-label Calibration.** Ensemble learning always plays a crucial role in improving prediction performance and overcoming the model homogenization issues of single model. Accordingly, we devise a pseudo-label calibration strategy that improves the confidence of pseudo-aligned pairs via the modal ensemble. Specially, we introduce a dynamic prediction dictionary, where the prediction of the present epoch will be stored. After $\omega$ epochs, the new prediction will first be compared with the previously stored results. For simplicity, we set $\omega$ to 2. If the two results are identical, the sample, as well as the predicted sample, will be classified as pseudo-label. Otherwise, the new prediction will be used to update the dictionary.

In addition, we introduce a data reordering method to accelerate model learning and optimize the quality of pseudo-label generation. In the early stage of model training, we rearrange the order of the labeled data, which is used to speed up the convergence of the model and capture crucial features. We first calculate the cosine similarity based on the entity joint representation $h^j$ and then put together data items with higher cosine similarity. Therefore, we put the more similar data in one mini-batch that the model can fit the general features and make the model initially distinguishable.

**Momentum-Based Contrastive Learning.** Recent studies (Zeng et al. 2022; Liu et al. 2022) have shown the popularity of contrastive learning in entity alignment. The aligned seeds can be naturally regarded as positive samples, whereas any non-aligned pairs can be considered as negative samples due to the convention of 1-to-1 alignment constraint.

| Seeds | Model | FB15K-DB15K | | | | FB15K-YAGO15K | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $Hits$@1 | $Hits$@5 | $Hits$@10 | $MRR$ | $Hits$@1 | $Hits$@5 | $Hits$@10 | $MRR$ |
| 20% | PoE | 0.1260 | - | 0.2510 | 0.1700 | 0.2500 | - | 0.4950 | 0.3340 |
| | MMEA | 0.2648 | 0.4513 | 0.5411 | 0.3570 | 0.2339 | 0.3976 | 0.4800 | 0.3170 |
| | EVA | 0.1340 | - | 0.3380 | 0.2010 | 0.0980 | - | 0.2760 | 0.1580 |
| | ACK-MMEA | 0.3040 | - | 0.5490 | 0.3870 | 0.2890 | - | 0.4960 | 0.3600 |
| | MSNEA | 0.6527 | 0.7685 | 0.8121 | 0.7080 | 0.4429 | 0.6255 | 0.6983 | 0.5290 |
| | MCLEA | 0.4450 | - | 0.7050 | 0.5340 | 0.3880 | - | 0.6410 | 0.4740 |
| | MultiJAF | 0.4800 | 0.5760 | 0.6010 | 0.5230 | 0.4630 | 0.6580 | 0.7310 | 0.5540 |
| | MEAformer | 0.4340 | - | 0.7280 | 0.5340 | 0.3250 | - | 0.5980 | 0.4160 |
| | MEAformer(+) | 0.5780 | - | 0.8120 | 0.6610 | 0.4440 | - | 0.6920 | 0.5290 |
| | PCMEA(ours) | **0.6763** | **0.8214** | **0.8872** | **0.7280** | **0.5896** | **0.7518** | **0.8347** | **0.6460** |
| 50% | PoE | 0.4640 | - | 0.6580 | 0.5330 | 0.4110 | - | 0.6690 | 0.4980 |
| | MMEA | 0.4165 | 0.6210 | 0.7035 | 0.5120 | 0.4026 | 0.5723 | 0.6451 | 0.4860 |
| | EVA | 0.2230 | - | 0.4710 | 0.3070 | 0.2400 | - | 0.4770 | 0.3210 |
| | ACK-MMEA | 0.5600 | - | 0.7360 | 0.6240 | 0.5350 | - | 0.6990 | 0.5930 |
| | MCLEA | 0.5730 | - | 0.8000 | 0.6520 | 0.5430 | - | 0.7590 | 0.6160 |
| | MEAformer | 0.6250 | - | 0.8470 | 0.7040 | 0.5600 | - | 0.7800 | 0.6400 |
| | MEAformer(+) | 0.6900 | - | 0.8710 | 0.7550 | 0.6120 | - | 0.8080 | 0.6820 |
| | PCMEA(ours) | **0.7375** | **0.8614** | **0.9154** | **0.7810** | **0.6702** | **0.8151** | **0.8857** | **0.7210** |
| 80% | PoE | 0.6660 | - | 0.8200 | 0.7210 | 0.4920 | - | 0.7050 | 0.5720 |
| | MMEA | 0.5903 | 0.8041 | 0.8687 | 0.6850 | 0.5976 | 0.7849 | 0.8389 | 0.6820 |
| | EVA | 0.3700 | - | 0.5850 | 0.4440 | 0.3940 | - | 0.6130 | 0.4710 |
| | ACK-MMEA | 0.6820 | - | 0.8740 | 0.7520 | 0.6760 | - | 0.8640 | 0.7440 |
| | MCLEA | 0.7300 | - | 0.8830 | 0.7840 | 0.6530 | - | 0.8350 | 0.7150 |
| | MEAformer | 0.7730 | - | 0.9180 | 0.8250 | 0.7050 | - | 0.8740 | 0.7680 |
| | MEAformer(+) | 0.7840 | - | 0.9210 | 0.8340 | 0.7240 | - | 0.8800 | 0.7830 |
| | PCMEA(ours) | **0.8204** | **0.9232** | **0.9644** | **0.8580** | **0.7556** | **0.8819** | **0.9424** | **0.8020** |

Table 1: Main experiments on FB15K-DB15K and FB15K-YAGO15K with different proportions of entity alignment seeds. The best results are highlighted in bold and the second best results are underlined. The "-" denotes that the results are not available, and the "+" means the iterative results.

Formally, for the $i$-th entity $e_1^i \in E_1$ of mini-batch $B$, the positive set is defined as $P^i = \{e_2^i | e_2^i \in E_2\}$, where $(e_1^i, e_2^i)$ is an aligned pair. The negative set includes two parts, inner-graph unaligned pairs from the source KG $G_1$ and cross-graph unaligned pairs from the target KG $G_2$, defined as $N_1^i = \{e_1^i | \forall e_1^j \in E_1, i \neq j\}$ and $N_2^i = \{e_2^j | \forall e_2^j \in E_2, i \neq j\}$, respectively. Since contrastive learning and pseudo-label calibration are performed simultaneously, the scale of the alignment seeds is gradually expanded as training progressively proceeds, and the corresponding positive and negative sets are dynamically updated.

Motivated by MoCo (He et al. 2020) and Fast-MoCo (Ci et al. 2022), momentum-based contrastive learning methods adopt an asymmetric forward path, and the two encoded samples from two paths (online path and target path) form a pair for contrastive learning, which has been proven to be effective in many scenarios. In this work, we propose to apply momentum-based contrastive learning to the field of semi-supervised entity alignment. On the online path, online entity representation is generated by the online encoder. On the target path, momentum entity representation is generated by a slowly moving momentum encoder. Thus, for the $i$-th entity $e_1^i \in E_1$ of mini-batch $B$, its representation is generated by the online encoder, and the representations of its positive set $P^i$ and negative set $N^i$ are generated by momentum encoder. While the online encoder's parameter $\theta_{online}$ is instantly updated with the back-propagation, the target encoder's parameter $\theta_{target}$ is asynchronously updated with momentum by:

$$\theta_{target} \leftarrow \kappa \cdot \theta_{target} + (1 - \kappa) \cdot \theta_{online}, \kappa \in [0, 1) \quad (12)$$

To be specific, we define the alignment probability distribution $q_m(e_1^i, e_2^i)$ of the modality $m$ for each positive pair $(e_1^i, e_2^i)$ as:

$$q_m(e_1^i, e_2^i) = \frac{\delta_m(e_1^i, e_2^i)}{\delta_m(e_1^i, e_2^i) + \sum\limits_{e_1^j \in N_1^i} \delta_m(e_1^i, e_1^j) + \sum\limits_{e_2^j \in N_2^i} \delta_m(e_1^i, e_2^j)} \quad (13)$$

where $\delta_m(u, v) = exp(f_m(u)^T g_m(v)/\tau)$, $f_m(\cdot)$ and $g_m(\cdot)$ are the online encoder and the momentum encoder of the modality $m$, respectively. $T$ denotes transpose operation and $\tau$ is a temperature parameter. Notably, the distribution is directional and asymmetric for each input; the distribution for another direction is thus defined similarly as $q_m(e_2^i, e_1^i)$. The loss function of contrastive learning can be calculated by:

$$\mathcal{L}_{CL}^m = -E_{i \in B} log \Big[\frac{1}{2}(q_m(e_1^i, e_2^i) + q_m(e_2^i, e_1^i))\Big] \quad (14)$$

We employ stage-wise momentum-based contrastive learning loss on uni-modal and joint representation. Specifically, in the first stage, only the online network is trained and updated. In the second stage, the momentum network is initialized and trained simultaneously with the online network. The time of changing the training strategy is $ts$, and the time span for updating the momentum network is $\rho$ epochs.

## Optimization Objective

The overall loss of the PCMEA is given below,

$$\mathcal{L} = \sum_{m \in M1} \mathcal{L}_{AL}^{m} + \mathcal{L}_{MI} + \sum_{m \in M2} \mathcal{L}_{CL}^{m} \tag{15}$$

where $M1 = \{g, r_{BOW}, r_{PLM}, a_{BOW}, a_{PLM}, v\}, M2 = M1 \cup \{h^j\}$, and $h^j$ denotes the joint embedding in Eq. (8).

# Experiments

## Experimental Settings

Two cross-KG EA datasets are adopted for evaluation, including FB15K-DB15K and FB15K-YAGO15K, which are the most representative datasets in MMEA task (Chen et al. 2020, 2022a; Lin et al. 2022). FB15K is one of the most widely used data sets in the field of link prediction. Entities from DBpedia and YAGO aligned with FB15K are extracted through the SameAs links, which are utilized to build DB15K and YAGO15K datasets.

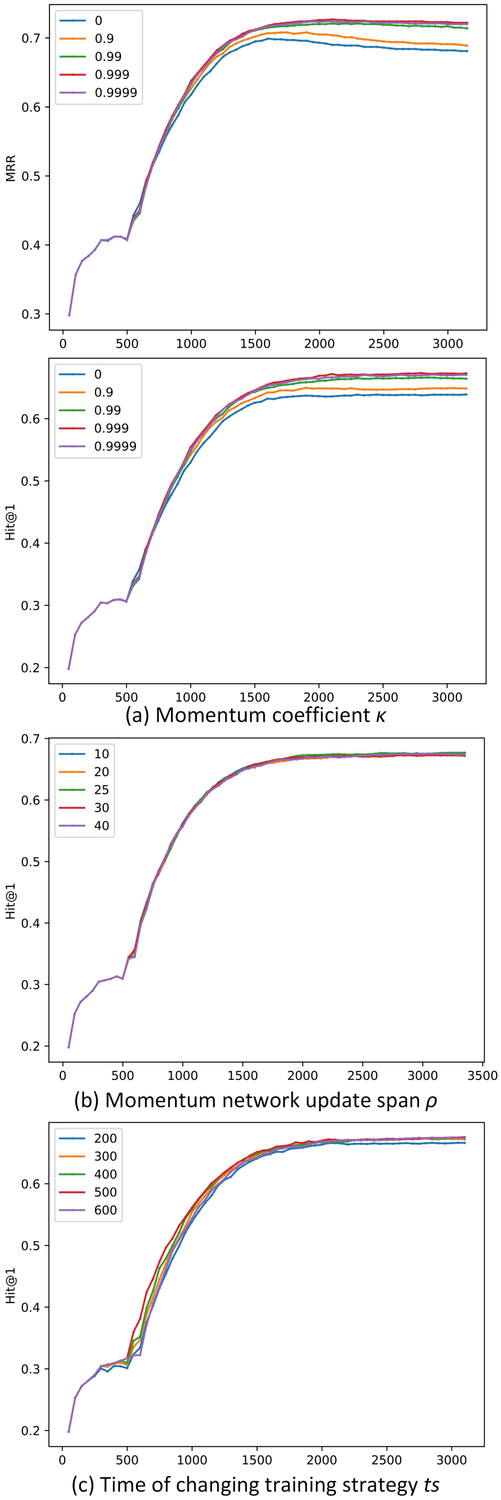
Our baseline model is MCLEA (Lin et al. 2022) and we also compare our method against other 7 state-of-the-art multi-modal EA methods, which can be classified into three categories: 1) traditional multi-modal EA methods, including PoE (Liu et al. 2019a), MMEA (Chen et al. 2020), and EVA (Liu et al. 2021); 2) multi-modal EA method based on pre-trained language model, such as ACK-MMEA (Li et al. 2023); 3) multi-modal EA method based on contrastive learning, including MSNEA (Chen et al. 2022a), MultiJAF (Cheng, Zhu, and Guo 2022) and MEAFormer (Chen et al. 2023). For all baselines, we report the original results from their literature.

Our model is implemented based on Pytorch, an open-source deep learning framework. The pre-trained language models (Bert (Kenton and Toutanova 2019), T5 (Raffel et al. 2020), RoBerta (Liu et al. 2019b), Albert (Lan et al. 2019) ChatGLM-6B (Du et al. 2022) and LLaMA-7B (Touvron et al. 2023)) are downloaded from Hugging Face[1] and all of them are base version. All experiments were conducted on a server with two GPUs (NVIDIA-SMI 3090).

## Results

To verify the effectiveness of our method, 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 Table 1, we can observe that: 1) Our model outperforms all the baseline of MMEA methods in terms of all metrics on both datasets. Particularly, our model brings about 9.04%-23.13% (16.21% on average) improvement on



(a) Momentum coefficient $\kappa$

(b) Momentum network update span $\rho$

(c) Time of changing training strategy $ts$

Figure 3: Study on (a) momentum coefficient, (b) momentum network update span, (c) time of changing training strategy.

[1] https://huggingface.co/

FB15K-DB15K and 10.26%-20.16% (14.38% on average) on FB15K-YAGO15K in terms of $Hits$@1 for all proportions of training data over baseline MCLEA. The superiority of our method demonstrates that the proposed structure and training strategy substantially boosts performance in the semi-supervised settings. 2) Our model shows a clear improvement over the traditional multi-modal EA method, which reveals that a more rational multi-modal information fusion method, as well as an appropriate training strategy, can make full use of the available data. 3) Our model is significantly more effective than other EA methods based on PLM, suggesting that PLM is useful for embedding relations and attributes but needs to be coupled with the neighborhood structure of the entity. 4) Moreover, our method clearly surpasses the state-of-the-art baseline by 2.36% in $Hits$@1 and 2% in $MRR$ on FB15K-DB15K and significantly outperforms the best baseline by 12.66% in $Hits$@1 and 9.2% in $MRR$ on FB15K-YAGO15K based on 20% aligned seeds. The above results also indicate that our modified contrastive learning strategy and model structure are superior.

| **Variants** | **FB15K-DB15K** | | |
|---|---|---|---|
| | $Hits$@1 | $Hits$@10 | $MRR$ |
| **Ours** | **0.6763** | **0.8872** | **0.7280** |
| w/o Image | 0.6510 | 0.8740 | 0.7070 |
| w/o Bag-of-words | 0.5303 | 0.7582 | 0.5860 |
| w/o PLM | 0.6502 | 0.8746 | 0.7050 |
| w/o Cross-attention | 0.6521 | 0.8719 | 0.7070 |
| w/o MI Loss | 0.6718 | 0.8841 | 0.7230 |
| w/o Align Loss | 0.6674 | 0.8818 | 0.7200 |
| w/o CL Loss | 0.3610 | 0.7204 | 0.4320 |
| w/o LC1 | 0.6647 | 0.8809 | 0.7190 |
| w/o LC2 | 0.6727 | 0.8753 | 0.7220 |
| w/o LC1&LC2 | 0.6588 | 0.8706 | 0.7120 |
| Attribute Embedding using different PLM | | | |
| Bert | 0.6682 | 0.8803 | 0.7190 |
| **Roberta** | **0.6763** | **0.8872** | **0.7280** |
| Albert | 0.6634 | 0.8811 | 0.7180 |
| T5 | 0.6718 | 0.8811 | 0.7230 |
| ChatGLM-6B | 0.6650 | 0.8788 | 0.7170 |
| LLaMA-7B | 0.6746 | 0.8787 | 0.7230 |
| Relation Embedding using different PLM | | | |
| Bert | 0.6706 | 0.8853 | 0.7230 |
| Roberta | 0.6693 | **0.8952** | 0.7260 |
| Albert | 0.6672 | 0.8828 | 0.7200 |
| **T5** | 0.6763 | 0.8872 | **0.7280** |
| ChatGLM-6B | 0.6726 | 0.8825 | 0.7230 |
| LLaMA-7B | **0.6768** | 0.8829 | 0.7270 |

Table 2: Variant experiments on FB15K-DB15K (20%). w/o means removing the corresponding module from the complete model. LC1 and LC2 mean dynamic prediction dictionary and the reordering method in pseudo-label calibration strategy, respectively. $In$ represents the maximum sequence length input to LLMs.

## Ablation Study

To investigate the effectiveness of each module in PCMEA, we perform variant experiments, whose results are shown in Table 2. From Table 2, we notice that: 1) The impact of the bag-of-words method (BOW) tends to be more significant than PLM on encoding relations and attributes. When combining PLM and BOW, the cross-attention mechanism must be used to bring out the power of PLM. We believe this is because the semantic differences between the vectors produced by the two encoders are obvious, and the attention mechanism can reduce the semantic gap. 2) The removal of the contrastive learning loss (CL loss) has the greatest impact with respect to the removal of the other two loss functions, since CL loss directly clusters together similar entities, while the others transfer information between different modalities. 3) Removing pseudo-label calibration drops 0.36%-1.75% in $Hits$@1, showing that improving pseudo-labeling quality is necessary for semi-supervised contrastive learning and can contribute to model performance. 4) We analyze the influence of embedding models by replacing different pre-trained language models. Specially, we test four PLMs (Bert et al.) and two large language models (ChatGLM-6B and LLaMA-7B). The results show that different embedding models affect entity alignment to a certain extent, and stronger model can improve performance. 5) The variants without image modality decline on all metrics, which hints that the multi-modal information and rational utilization are necessary for the EA task.

**Impact of hyper-parameters.** We conduct hyper-parameter studies using FB15K-DB15K, showing results in Figure 3. The main hyper-parameters in our method are momentum coefficient $\kappa$, momentum network update span $\rho$, and time of changing training strategy $ts$. For momentum coefficient $\kappa$, a proper large $\kappa$ (e.g. 0.999) bring better stability and accuracy in $Hits$@1 and $MRR$, illustrating momentum-based contrast learning is more effective than just contrast learning. Varying time span $\rho$ shows little difference. For time $ts$ of changing training strategy, time $ts$ obvious effects the surge in $Hits$@1 after changing the strategy, with $ts = 500$ allowing faster convergence. Besides, $ts$ barely affects post-convergence performance.

## Conclusion

In this work, we propose a semi-supervised pseudo-label calibration multi-modal entity alignment framework named PCMEA. It utilizes various embedding methods and attention mechanisms to obtain multi-modal entity representation. Instead of direct interaction and fusion of multi-modal embedding, we apply mutual information maximization to filter out task-independent noise and transfer cross-modality information. To boost the quality of pseudo-label and contrastive learning, we combine pseudo-label calibration with momentum-based contrastive learning, which helps pull aligned pairs closer and improve alignment performance. Experimental results show that PCMEA can consistently outperform prior state-of-the-art methods, producing high-quality alignment performance even under 20% labeled data settings.

## Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No. 61772534, and partially supported by Public Computing Cloud, Renmin University of China.

## References

Bao, X.; Wang, S.; Qi, P.; and Qin, B. 2023. Wukong-CMNER: A Large-Scale Chinese Multimodal NER Dataset with Images Modality. In *International Conference on Database Systems for Advanced Applications*, 582–596. Springer.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *International conference on machine learning*, 531–540. PMLR.

Chen, L.; Li, Z.; Wang, Y.; Xu, T.; Wang, Z.; and Chen, E. 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 13*, 134–147. Springer.

Chen, L.; Li, Z.; Xu, T.; Wu, H.; Wang, Z.; Yuan, N. J.; and Chen, E. 2022a. Multi-modal siamese network for entity alignment. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 118–126.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022b. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Chen, Z.; Chen, J.; Zhang, W.; Guo, L.; Fang, Y.; Huang, Y.; Zhang, Y.; Geng, Y.; Pan, J. Z.; Song, W.; et al. 2023. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3317–3327.

Cheng, B.; Zhu, J.; and Guo, M. 2022. MultiJAF: Multi-modal joint entity alignment framework for multi-modal knowledge graph. *Neurocomputing*, 500: 581–591.

Ci, Y.; Lin, C.; Bai, L.; and Ouyang, W. 2022. Fast-MoCo: Boost momentum-based contrastive learning with combinatorial patches. In *European Conference on Computer Vision*, 290–306. Springer.

Ding, Y.; Yu, J.; Liu, B.; Hu, Y.; Cui, M.; and Wu, Q. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5089–5098.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Guo, L.; Han, Y.; Zhang, Q.; and Chen, H. 2022. Deep Reinforcement Learning for Entity Alignment. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2754–2765.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

Li, Q.; Guo, S.; Luo, Y.; Ji, C.; Wang, L.; Sheng, J.; and Li, J. 2023. Attribute-Consistent Knowledge Graph Representation Learning for Multi-Modal Entity Alignment. In *Proceedings of the ACM Web Conference 2023*, 2499–2508.

Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; and Zheng, Y. 2022. Multi-modal Contrastive Representation Learning for Entity Alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2572–2584.

Liu, F.; Chen, M.; Roth, D.; and Collier, N. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4257–4266.

Liu, X.; Hong, H.; Wang, X.; Chen, Z.; Kharlamov, E.; Dong, Y.; and Tang, J. 2022. Selfkg: self-supervised entity alignment in knowledge graphs. In *Proceedings of the ACM Web Conference 2022*, 860–870.

Liu, Y.; Li, H.; Garcia-Duran, A.; Niepert, M.; Onoro-Rubio, D.; and Rosenblum, D. S. 2019a. MMKG: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, 459–474. Springer.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Qi, P.; and Qin, B. 2023. SSMI: Semantic Similarity and Mutual Information Maximization Based Enhancement for Chinese NER. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13474–13482.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Shao, Z.; Yu, Z.; Wang, M.; and Yu, J. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14974–14983.

Sun, R.; Cao, X.; Zhao, Y.; Wan, J.; Zhou, K.; Zhang, F.; Wang, Z.; and Zheng, K. 2020a. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 1405–1414.

Sun, Z.; Zhang, Q.; Hu, W.; Wang, C.; Chen, M.; Akrami, F.; and Li, C. 2020b. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 13(12).

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Zeng, K.; Dong, Z.; Hou, L.; Cao, Y.; Hu, M.; Yu, J.; Lv, X.; Cao, L.; Wang, X.; Liu, H.; et al. 2022. Interactive contrastive learning for self-supervised entity alignment. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2465–2475.

Zeng, W.; Zhao, X.; Tang, J.; and Fan, C. 2021a. Reinforced active entity alignment. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2477–2486.

Zeng, W.; Zhao, X.; Tang, J.; Lin, X.; and Groth, P. 2021b. Reinforcement learning–based collective entity alignment with adaptive features. *ACM Transactions on Information Systems (TOIS)*, 39(3): 1–31.