CateEA: Enhancing Entity Alignment via Implicit Category Supervision

Guandong Feng[1,2], Tao Ren[1,2]*, Jun Hu[1,2], Dandan Wang[1,2]
[1]State Key Laboratory of Intelligent Game, Institute of Software
Chinese Academy of Sciences, Beijing, China,
[2]University of Chinese Academy of Sciences, Beiiing, China,
{guandong2022,taoren22,hujun,dandan}@iscas.ac.cn

Abstract

Entity Alignment (EA) is essential for integrating Knowledge Graphs (KGs) by matching equivalent entities across diverse KGs. With the rise of multi-modal KGs, which emerged to better depict real-world KGs by integrating visual, textual and structured data, Multi-Modal Entity Alignment (MMEA) has become crucial in enhancing EA. However, existing MMEA methods often neglect the inherent semantic category information of entities, limiting alignment precision and robustness. To address this, we propose *Cate*gory-enhanced *E*ntity *A*lignment (CateEA), which combines implicit entity category information into multi-modal representations. By generating pseudo-category labels from entity embeddings and integrating them into a multi-task learning framework, CateEA captures latent category semantics, enhancing entity representations. CateEA allows for adaptive adjustments of similarity measures, leading to improved alignment precision and robustness in multi-modal contexts. Experiments on benchmark datasets, such as FB15K-DB15K/YAGO15K, demonstrate that CateEA outperforms state-of-the-art methods in various settings.[1]

1 Introduction

Knowledge Graphs (KGs) have become fundamental infrastructures underpinning a wide array of artificial intelligence applications, including but not limited to question-answering systems (Cui et al., 2019), recommendation engines (Zhang et al., 2016), and semantic search (Xiong et al., 2017). Due to the presence of overlapped entities in KGs from different data sources, integrating knowledge through these common entities is essential for completing KGs. Entity Alignment (EA) (Chen et al., 2016) emerges as a promising technology to identify and align these entities, facilitating knowledge integration across diverse KGs.

While traditional entity alignment primarily relies on structured data (Li et al., 2019) and textual descriptions, recent advances in Multi-Modal Entity Alignment (MMEA) leverage diverse data modalities, such as images, text, and structured information, to capture richer semantic information and provide a more comprehensive understanding of entities. MMEA has evolved from early feature fusion techniques, which combined visual, textual, and structural data (Chen et al., 2020), to contrastive learning approaches that reduce cross-modal discrepancies (Lin et al., 2022). Recently, adaptive integration methods have emerged, introducing dynamic strategies to handle data inconsistencies and ambiguities (Chen et al., 2023a). However, these methods mostly employ standard similarity measures, such as cosine or Euclidean distance, on fused multi-modal features to judge alignment. This similarity measure could fail to distinguish appearance-similar but semantic-different entities based solely on explicit features.

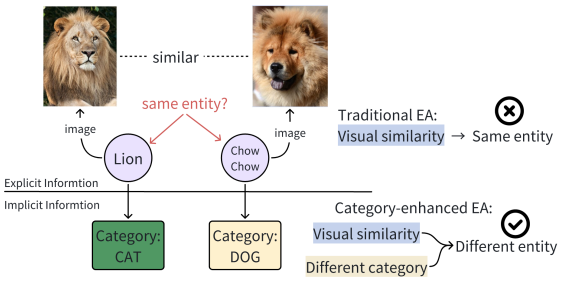


Figure 1: Possible misjudgement of different entities due to directly comparing embedding similarities.

As illustrated in Figure 1, entities like "Chow Chow" and "Lion" may be misaligned due to their visual similarities, despite belonging to distinct categories. This misalignment primarily results from

*Corresponding authors: Tao Ren and Dandan Wang
[1]The source code is available at https://github.com/Melkor0007/CateEA.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5975–5986
January 19–24, 2025.

the exclusive focus of existing methods on explicit multi-modal feature similarities in the fused embedding space, *overlooking the implicit category relationships among entities* involved in the contextual semantics of multi-modal data.

To address the issue, we propose a novel Category-enhanced Entity Alignment (CateEA) framework that integrates implicit category information into multi-modal entity alignment through multi-task learning. Specifically, CateEA clusters entity embeddings to generate pseudo-category labels, which guide the training process by incorporating category-aware classification into the alignment task. This process enables entity embeddings to capture category-specific information, refining alignment through enhanced semantic representation. During testing, entities are classified into categories, and similarity scores are adjusted based on category proximity, further improving alignment accuracy. Comprehensive experiments on widely used benchmark datasets (FB15K-DB15K/YAGO15K, DBP15K ZH/JA/FR-EN) highlight the advantages of CateEA compared to state-of-the-art methods. The main contributions are summarized as follows:

- We introduce a novel entity alignment framework that exploits implicit entity category information within multi-modal data to enhance the semantic discriminating ability of entity alignments.
- We design a multi-task learning strategy that incorporates pseudo-category labels obtained from embedding clusters into both alignment and auxiliary classification tasks to capture latent semantic structures of entities, producing category-enhanced entity representations.
- We conduct extensive experiments on popular benchmark datasets to demonstrate the superiority of CateEA over state-of-the-art methods, along with various ablation studies to validate the efficacy of CateEA.

2 Related Work

We categorize related work into KG representation learning and multi-modal entity alignment, highlighting the progression from traditional embedding methods to advanced multi-modal integration strategies.

2.1 KG Representation Learning

Representation learning-based methods, including translation models and GCN-based approaches, have proven effective in capturing semantic information of KGs. TransE (Bordes et al., 2013) and its variants, such as TransH (Wang et al., 2014), TransEdge (Sun et al., 2019) and TransR (Lin et al., 2015), as well as GCN-based models like MuGNN(Cao et al., 2019) JAPE (Sun et al., 2017), GCN-Align (Wang et al., 2018) and ClusterEA (Gao et al., 2022) focus primarily on leveraging the structural information of knowledge graphs with GCN (Kipf and Welling, 2016) to enhance entity embeddings, providing simplicity, scalability, and improved alignment accuracy. However, these approaches often overlook the rich multi-modal and semantic information embedded within entities, which limits their ability to fully capture the complex relationships and diverse contexts in real-world data. To address these limitations, our proposed CateEA framework integrates implicit entity category information into the alignment process, combining multi-modal data with latent semantic structures.

2.2 Multi-Modal Entity Alignment

Recent advancements in MMEA mainly lie in three aspects: multi-modal feature fusion, inter-modal contrastive learning, and adaptive modality integration. Feature fusion methods, such as PoE (Liu et al., 2019), MMEA (Chen et al., 2020) and HMEA (Guo et al., 2021), embed visual, textual, and structural data into a unified representation to enhance alignment accuracies. To improve discrimination ability of feature fusion methods, inter-modal contrastive learning, e.g., MSNEA (Chen et al., 2022) and MCLEA (Lin et al., 2022), distinguishes positive and negative samples across modalities to reduce cross-modal gaps and enhance the interaction between visual, relational, and attribute features, but could face challenges with incomplete or ambiguous modality information. Thus, adaptive modality integration is proposed, i.e., MeaFormer (Chen et al., 2023a) employs a dynamic meta-modality hybrid strategy with transformer-based architectures to adaptively fuse features and enhance robustness against noisy and missing data, and UMAEA (Chen et al., 2023b) introduces an uncertainty-aware alignment mechanism that manages modality inconsistencies and visual ambiguities to maintain alignment accuracy.

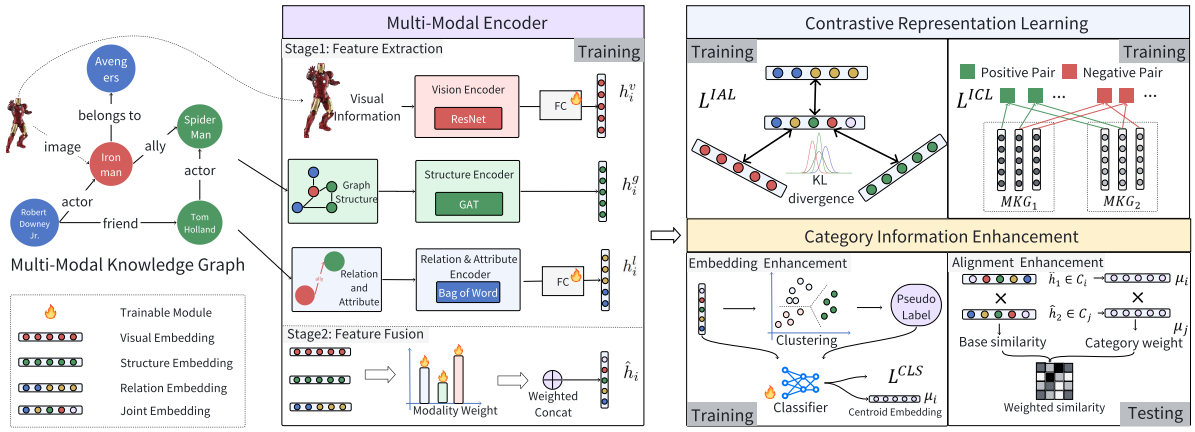


Figure 2: Overview of the CateEA framework consisting of Multi-Modal Encoder that extracts and fuses multi-modal features, Contrastive Representation Learning that improves multi-modal representation via contrastive learning, and Category Information Enhancement that enhances entity alignment via implicit category supervision.

ACK-MMEA (Li et al., 2023) addresses modality inconsistencies in multi-modal entity alignment by creating attribute-consistent representations.

Despite the significant progress, existing methods still focus on judging entity similarity primarily through explicit features, which could fail to distinguish appearance-similar but semantic-different entities. Facing this, CateEA enhances alignment accuracy by leveraging intrinsic semantic category.

3 Method

CateEA mainly consists of three key modules: Multi-Modal Encoder (MME), Contrastive Representation Learning (CRL), and Category Information Enhancement (CIE) which include Category Embedding Enhancement (CEE) and Category Alignment Enhancement (CAE).

The MME effectively integrates various modalities into a joint embedding, CRL refines the embeddings by distinguishing positive and negative pairs, and CIE further optimizes the embeddings by incorporating implicit category supervision.

The overall framework of CateEA is shown in Figure 2.

3.1 Problem Definition

A knowledge graph $G = (E, R, T)$ is a directed graph that includes an entity set E, a relation set R, and a set of triples $T \subseteq E \times R \times E$. Given a source knowledge graph $G_1 = (E_1, R_1, T_1)$ and a target knowledge graph $G_2 = (E_2, R_2, T_2)$, along with a potential entity pair seed set $S = \{(u, v) \mid u \in G_1, v \in G_2, u = v\}$, where u and v represent equivalent entities referring to the same real-world object or concept, the goal of entity alignment is to discover equivalent entity pairs between the source and target KGs, which can be seen as an extension of the seed set S.

Based on the expansion of relationships between the source and target knowledge graphs, multi-modal knowledge is integrated with textual knowledge. Similar to traditional knowledge graphs, a multi-modal knowledge graph can be formally defined as $MKG = (E, R, A, V, T)$, where A represents the set of entity attributes, and V represents the set of entity images. The task of entity alignment in multi-modal knowledge graphs can be further seen as an extension of traditional entity alignment (Zhu et al., 2022), specifically identifying equivalent entity pairs between $MKG_1 = (E_1, R_1, A_1, V_1, T_1)$ and $MKG_2 = (E_2, R_2, A_2, V_2, T_2)$.

3.2 Multi-Modal Encoder

For different modalities of information (text, structure, image, etc.) in KG, the MME adopts different encoders to extract embedding features, after which an attention mechanism is employed to assign weights and integrate these embeddings, yielding the joint embedding of the entity.

3.2.1 Structure Embedding

Graph attention network (GAT) (Veličković et al., 2017) is a typical neural network that is good at extracting features from structured data. Therefore, we use GAT to model the structural information of the knowledge graph, as shown in the "Structure Encoder" in Figure 2. Specifically, the hidden state $h_i \in \mathbb{R}^d$ of an entity e_i (where d is the size of the hidden layer) is formalized by aggregating the

one-hop neighbors N_i (including self-loops) of the entity e_i as:

$$h_i = \sigma\Big(\sum_{j \in N_i} \alpha_{ij} h_j\Big), \tag{1}$$

where $\sigma_(\cdot)$ denotes the ReLU nonlinearity, and α_{ij} is the importance of e_i to e_j, computed by self-attention. Multi-head attention is performed and by concatenating these features, we obtain the structural embedding of entity e_i:

$$h_i^g = \bigoplus_{k=1}^{K} \sigma\Big(\sum_{j \in N_i} \alpha_{ij}^k h_j\Big), \tag{2}$$

where α_{ij}^k is the normalized attention coefficient obtained from the k-th attention mechanism. In practice, we apply a two-layer GAT model to aggregate information, and the output of the last GAT layer is taken as structural embedding.

3.2.2 Relation and Attribute Embedding

We follow (Yang et al., 2019) and represent the relation r, attribute a, and name n of entity e_i as bag-of-words features and input them into a linear transformation to obtain the embedding as follows,

$$h_i^l = \mathbf{W}_l \mathbf{u}_i^l + \mathbf{b}_l, \quad l \in \{r, a, n\}. \tag{3}$$

Specifically, the name feature is obtained by averaging the pre-trained GloVe (Pennington et al., 2014) vectors of name strings.

3.2.3 Visual Embedding

To be consistent with previous work, we adopt a pre-trained visual model, ResNet-152 (He et al., 2016), to learn visual embedding. We input the images of entities e_i into the pre-trained visual model (PVM) and use the logits from the last fully connected layer before the softmax as the visual features. These features are passed through a linear transformation to obtain the visual embedding:

$$h_i^v = \mathbf{W}_v \cdot \mathrm{PVM}(v_i) + \mathbf{b}_v. \tag{4}$$

3.2.4 Joint Embedding

The embeddings from different modalities are concatenated with attention weights to obtain the joint embedding of the entity:

$$\hat{h}_i = \bigoplus_{m \in \mathcal{M}} \left[\frac{\exp(w_m)}{\sum_{j \in \mathcal{M}} \exp(w_j)} h_i^m\right], \tag{5}$$

where $\mathcal{M} = \{g, r, a, n, v\}$, and w_m is the trainable attention weight for modality m.

3.3 Contrastive Representation Learning

After obtaining the joint embedding of the entity and the embedding of each modality, we design CRL to encourage the embeddings of the same entity to be closer while pushing the embeddings of different entities further apart.

The contrastive learning is conducted in two dimensions (Lin et al., 2022): using Intra-Modal Contrastive Loss (ICL) and Inter-Modal Alignment Loss (IAL) to construct the interaction between intra-modality and inter-modality embeddings. The CRL helps to refine both joint and individual modality embeddings, ensuring that embeddings of similar entities maintain minimal distance.

For each entity pair (e_i^1, e_i^2) in seed set S, its negative entity set is defined as $\mathcal{N}_i^{neg} = \{e_j^1 \mid \forall e_j^1 \in E_1, j \neq i\} \cup \{e_j^2 \mid \forall e_j^2 \in E_2, j \neq i\}$. The alignment probability is defined as:

$$q_m(e_i^1, e_i^2) = \frac{\delta_m(e_i^1, e_i^2)}{\delta_m(e_i^1, e_i^2) + \sum\limits_{e_j \in \mathcal{N}_i^{neg}} \big(\delta_m(e_j, e_i^1) + \delta_m(e_i^2, e_j)\big)}, \tag{6}$$

where $\delta_m(u, v) = \exp\left(\frac{f_m(u)^T f_m(v)}{\tau_1}\right)$, $f_m(\cdot)$ is the encoder of the modality m, and τ_1 is a temperature parameter.

IAL aligns the joint and individual modality embeddings using the Kullback–Leibler divergence, with the goal of making the output distributions of different modalities as consistent as possible, thereby facilitating cross-modal interactions:

$$\begin{aligned}\mathcal{L}_m^{\mathrm{IAL}} = \mathbb{E}_{i \in \mathcal{B}} \frac{1}{2} [&\mathrm{KL}(q_o'(e_1^i, e_2^i) \parallel q_m'(e_1^i, e_2^i)) \\ &+ \mathrm{KL}(q_o'(e_2^i, e_1^i) \parallel q_m'(e_2^i, e_1^i))],\end{aligned} \tag{7}$$

where $q_o'(e_1^i, e_2^i)$, $q_o'(e_2^i, e_1^i)$, $q_m'(e_1^i, e_2^i)$, and $q_m'(e_2^i, e_1^i)$ denote the output predictions for both directions of the joint embedding and the uni-modal embedding of modality m, respectively, similar to Eq. 6 but with another temperature parameter τ_2.

ICL selects aligned entities as positive samples and other entities as negative samples, facilitating the proximity of semantically similar entities within the same knowledge graph, thereby forming a more compact and distinct representation suitable for cross-modal matching. ICL can be formulated

as:

$$\mathcal{L}_m^{ICL} = -\mathbb{E}_{i\in\mathcal{B}} \log \left[\frac{1}{2} \left(q_m(e_i^1, e_i^2) + q_m(e_i^2, e_i^1) \right) \right]. \tag{8}$$

3.4 Category Information Enhancement

3.4.1 Category Embedding Enhancement

Similar to the idea of DeepCluster (Caron et al., 2018), the CEE component utilizes K-Medoids (Rdusseeun and Kaufman, 1987) to cluster the joint embeddings generated by multi-modal encoders, which selects actual data points as cluster centers and enhances robustness to noises and outliers, as it captures complex, nonspherical category structures often found in multi-modal data, providing more representative category labels. These labels are crucial for the auxiliary classification task, improving the model's ability to reflect true semantic differences and enhancing alignment accuracy.

Once the clustering results are obtained, they are used as pseudo-labels for the entity's category. The joint embedding of the entity is then fed into a classifier which predicts the category to which the entity belongs. The classification loss is added to the alignment loss and propagated back to the multi-modal encoder module during training, thus allowing the model to perform the tasks simultaneously. Specifically, the classification task is optimized using a two-layer feedforward neural network, aiming at minimizing the cross-entropy loss:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}), \tag{9}$$

where C is the number of categories, N represents the total number of samples in the batch, $y_{i,c}$ is the true label of sample i in category c. If the sample i belongs to category c, then $y_{i,c} = 1$, otherwise $y_{i,c} = 0$.

Unlike traditional methods that depend on labeled supervision, CateEA does not simply rely on annotated entity information. Instead, it uses the clustering results as pseudo-labels for learning, thereby alleviating the reliance on manual annotation and improving the model's robustness.

3.4.2 Category Alignment Enhancement

Mainstream MMEA methods compute similarities between joint embeddings of entities, producing a $N \times N$ similarity matrix that treats all entities uniformly, ignoring their semantic and category differences. They often result in suboptimal alignment, especially with complex or ambiguous multi-modal data. To address these limitations, the CAE component adjusts similarity scores using category labels generated during training, allowing the model to better capture semantic distances and dynamically refine entity relationships, thus enhancing alignment accuracy and robustness.

For each category C_i, compute its embedding centroid μ_i, which is defined as the mean of the embeddings of all entities within that category:

$$\mu_i = \frac{1}{|C_i|} \sum_{x\in C_i} \hat{h}_x, \tag{10}$$

where $\hat{h}_x$ represents the joint embedding of entity x and $|C_i|$ is the number of entities in category C_i.

To determine the semantic proximity between different categories, we begin by calculating the Euclidean distance between the centroids of the categories. Let d_{ij} denote the distance between the centroids of categories C_i and C_j:

$$d_{ij} = \|\mu_i - \mu_j\|, \tag{11}$$

where μ_i and μ_j represent the centroids of categories C_i and C_j, respectively.

Given these distances, we transform them into similarity measures, with smaller distances indicating higher similarity. An inverse distance function is then applied:

$$s_{ij} = \frac{1}{1 + d_{ij}}, \tag{12}$$

which ensures that closer centroids result in higher similarity scores, aligning with the intuition that semantically similar categories are more likely to align with each other.

To further refine similarity scores into weights that reflect category proximity, we adopt a normalization approach. The final category weight w_{ij}, which indicates the relative closeness of category C_j to category C_i, is computed as:

$$w_{ij} = \frac{\exp(s_{ij})}{\sum_k \exp(s_{ik})}. \tag{13}$$

This normalization step ensures that the weights sum up to one, providing a probabilistic interpretation of the relative influence of each category.

During alignment, these weights are used to adjust the similarity between entities of different categories. Specifically, for a pair of entities x_i and y_j,

Table 1: Non-Iterative results on three bilingual datasets.

Models	$\text{DBP15K}_{\text{ZH-EN}}$			$\text{DBP15K}_{\text{JA-EN}}$			$\text{DBP15K}_{\text{FR-EN}}$		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
w/o SF									
EVA	.680	.910	.762	.673	.908	.757	.683	.923	.767
MSNEA	.601	.830	.684	.535	.775	.617	.543	.801	.630
MCLEA	.715	.923	.788	.715	.909	.785	.711	.909	.782
MEAformer	.771	.951	.835	.764	.959	.837	.770	.961	.841
CateEA (Ours)	**.776**	**.955**	**.839**	**.772**	**.963**	**.840**	**.786**	**.972**	**.855**
w/ SF									
EVA	.929	.986	.951	.946	.997	.976	.962	.996	.978
MSNEA	.887	.961	.913	.938	.983	.955	.933	.983	.953
MCLEA	.926	.983	.946	.961	.994	.973	.987	.999	.992
MEAformer	**.948**	.993	**.965**	**.977**	**.999**	.986	.991	1.00	.995
CateEA (Ours)	.945	**.993**	.964	.972	.998	**.987**	**.991**	**1.00**	**.995**

Table 2: Iterative results on three bilingual datasets.

Models	$\text{DBP15K}_{\text{ZH-EN}}$			$\text{DBP15K}_{\text{JA-EN}}$			$\text{DBP15K}_{\text{FR-EN}}$		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
w/o SF									
EVA	.746	.910	.807	.741	.918	.805	.767	.939	.831
MSNEA	.643	.865	.719	.572	.853	.660	.584	.841	.671
MCLEA	.811	.954	.865	.806	.953	.861	.811	.954	.865
MEAformer	**.847**	.970	**.892**	.842	.974	.892	.845	.976	.894
CateEA (Ours)	.839	**.971**	.887	**.851**	**.978**	**.899**	**.862**	**.985**	**.908**
w/ SF									
EVA	.956	.993	.969	.979	.995	.987	.995	.999	.997
MSNEA	.896	.969	.922	.942	.971	.958	.971	.998	.982
MCLEA	.964	.996	.977	.995	1.00	.992	.995	1.00	.997
MEAformer	.973	.998	.983	.991	**1.00**	**.995**	.996	1.00	.998
CateEA (Ours)	**.974**	**.998**	**.984**	**.992**	.999	.993	**.997**	**1.00**	**.998**

where x_i belongs to category C_i and y_j belongs to category C_j, their final similarity is defined as:

$$\text{Similarity}(x_i, y_j) = w_{ij} \cdot \text{BaseSimilarity}(x_i, y_j), \tag{14}$$

where $\text{BaseSimilarity}(x_i, y_j)$ is the original similarity of embeddings, such as cosine similarity. The pseudo-algorithm is shown in Appendix A.3.

4 Experiments

4.1 Experiment Settings

We present below the datasets, model configurations, baseline methods, iterative training strategy, and evaluation metrics used in our experiments.

4.1.1 Datasets

Our experiments are conducted on five popular MMEA datasets, including two monolingual cross-graph multi-modal entity datasets, i.e., FB15K-DB15K/YAGO15K (Liu et al., 2019), and three bilingual datasets, i.e., ZH-EN/JA-EN/FR-EN versions of DBP15K (Liu et al., 2021). These datasets combine KGs from different sources and leverage multi-modal information to assist the entity alignment task. Dataset details are shown in Appendix A.1.

It is worth noting that not all entities have corresponding images. For entities without images, random vectors are assigned as visual features. Following previous work, we use 20%, 50%, and 80% aligned entity pairs as the seed set for FB15K-DB15K/YAGO15K, and 30% for DBP15K.

4.1.2 Baselines

We compare CateEA with five multi-modal entity alignment methods: MMEA(Chen et al., 2020), MSNEA(Chen et al., 2022), EVA(Liu et al., 2021), MCLEA(Lin et al., 2022), Meaformer(Chen et al., 2023a). Previous studies have demonstrated that surface forms (SF, entity names) significantly impact the performance of entity alignment. To ensure consistency with previous methods, on bilingual datasets, we use both with and without surface forms, while on monolingual datasets, surface

forms are excluded.

4.1.3 Model Configuration

The hidden layer size of each GAT layer is 300, while the embedding size of other modules is 400. We use the AdamW optimizer with a learning rate of 5×10^{-4} to update the parameters. The total number of training epochs is 1000, with early stopping applied, and the batch size is 512. Number of categories is set to 10. The hyperparameters τ_1 and τ_2 are set to 0.1 and 0.4, respectively. For visual embeddings, we use the preprocessed image features provided by (Liu et al., 2021), with ResNet-152 as the pre-trained backbone network.

4.1.4 Iterative Training

To mitigate the shortage of training data, we employ a bidirectional iterative strategy (Liu et al., 2021). Specifically, every K_e epochs ($K_e = 5$), cross-KG entity pairs that are mutual nearest neighbors in the vector space are identified and added to a candidate list $\mathcal{N}^{cd}$. An entity pair from $\mathcal{N}^{cd}$ is incorporated into the training set if it remains mutual nearest neighbors for K_s consecutive rounds ($K_s = 10$). This approach progressively enhances the training set by introducing new aligned pairs during each iteration.

4.1.5 Evaluation Metrics

The experimental results are evaluated using two metrics: MRR (Mean Reciprocal Rank) and Hits@N. Both metrics assess the ranking performance of the model. A higher value for these metrics indicates better performance. Details of these metrics are shown in the Appendix A.2.

4.2 Main Results

The results on bilingual datasets are displayed in Table 1 (non-iterative) and Table 2 (iterative), while the results on monolingual datasets are presented in Table 3 (non-iterative) and Table 4 (iterative).

CateEA is compared with various entity alignment methods, using different proportions of seed sets. Particularly, CateEA outperforms the baselines with notable improvements across various settings. Specifically, on FB15K-DB15K, we achieve an H@1 increase ranging from 3.4% to 7.6%, and on FB15K-YAGO15K, H@1 improves by up to 7.5% over the baselines. Moreover, consistent gains are observed in H@10 and MRR, with improvements ranging from 1.0% to 8.0% across different settings. CateEA shows significant performance improvements especially in the 20% seed set scenario, where the gains are most pronounced. By leveraging implicit category information, CateEA excels in situations with limited labeled data, demonstrating its effectiveness in enhancing alignment accuracy, particularly in low resource environments where traditional methods struggle.

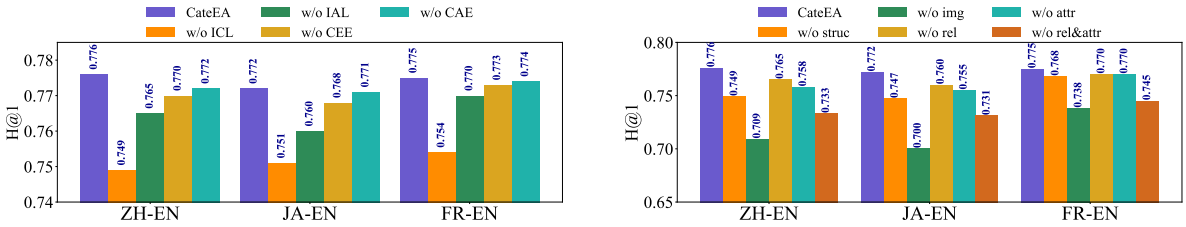
In summary, CateEA demonstrates excellent performance across different settings, showing strong robustness and generalization ability, especially in low seed set scenarios. The experimental results fully validate the effectiveness and superiority of CateEA, demonstrating the value of incorporating the implicit entity category strategy.

Table 3: Non-iterative results on two monolingual datasets, where Seed% is the seed set proportion.

Seed %	Models	FB15K-DB15K H@1	FB15K-DB15K H@10	FB15K-DB15K MRR	FB15K-YAGO15K H@1	FB15K-YAGO15K H@10	FB15K-YAGO15K MRR
20%	MMEA	.265	.541	.357	.234	.480	.317
	EVA	.199	.448	.283	.153	.361	.224
	MSNEA	.114	.296	.175	.103	.249	.153
	MCLEA	.295	.582	.393	.254	.484	.332
	MEAformer	.417	.715	.518	.327	.595	.417
	CateEA	.493	.759	.584	.402	.675	.497
	improv.	+7.6%	+4.4%	+6.6%	+7.5%	+8.0%	+8.0%
50%	MMEA	.417	.703	.512	.403	.645	.486
	EVA	.334	.589	.422	.311	.534	.388
	MSNEA	.288	.590	.388	.320	.589	.413
	MCLEA	.555	.784	.637	.501	.705	.574
	MEAformer	.619	.843	.698	.560	.778	.639
	CateEA	.674	.874	.745	.608	.829	.686
	improv.	+5.5%	+3.1%	+4.7%	+4.8%	+5.1%	+4.7%
80%	MMEA	.590	.869	.685	.598	.839	.682
	EVA	.484	.696	.563	.491	.692	.565
	MSNEA	.518	.779	.613	.531	.778	.620
	MCLEA	.735	.890	.790	.667	.824	.722
	MEAformer	.765	.916	.820	.703	.873	.766
	CateEA	.799	.933	.849	.742	.912	.805
	improv.	+3.4%	+1.7%	+2.9%	+3.9%	+3.9%	+3.9%

Table 4: Iterative results on two monolingual datasets.

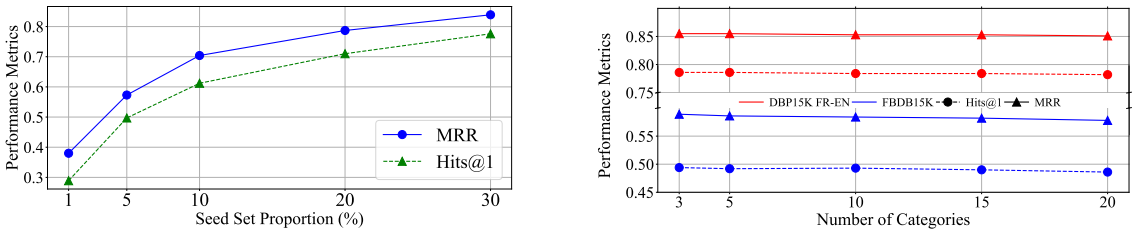
Seed %	Models	FB15K-DB15K H@1	FB15K-DB15K H@10	FB15K-DB15K MRR	FB15K-YAGO15K H@1	FB15K-YAGO15K H@10	FB15K-YAGO15K MRR
20%	EVA	.231	.488	.318	.188	.403	.260
	MSNEA	.149	.392	.232	.138	.346	.210
	MCLEA	.395	.656	.487	.322	.546	.400
	MEAformer	.578	.812	.661	.444	.692	.529
	CateEA	.599	.822	.675	.518	.745	.594
	improv.	+2.1%	+1.0%	+1.4%	+7.4%	+5.3%	+6.5%
50%	EVA	.364	.606	.449	.325	.560	.404
	MSNEA	.358	.656	.459	.376	.646	.472
	MCLEA	.620	.832	.696	.563	.751	.631
	MEAformer	.690	.871	.755	.612	.808	.682
	CateEA	.705	.882	.764	.640	.844	.708
	improv.	+1.5%	+1.1%	+0.9%	+2.8%	+3.6%	+2.6%
80%	EVA	.491	.711	.573	.493	.695	.572
	MSNEA	.565	.810	.651	.593	.806	.668
	MCLEA	.741	.900	.802	.681	.837	.737
	MEAformer	.784	.921	.834	.724	.880	.783
	CateEA	.806	.934	.853	.746	.915	.807
	improv.	+2.2%	+1.3%	+1.9%	+2.2%	+3.5%	+2.4%



(a) Component analysis on DBP15K.

(b) Modality analysis on DBP15K.

Figure 3: Performance analysis on the components and modalities of CateEA in DBP15K.



(a) Impact of the seed set proportion.

(b) Impact of the number of categories.

Figure 4: Impact of the seed set proportion and the number of categories.

4.3 Ablation Study

4.3.1 Effectiveness of Each Component

To validate the effectiveness of each component of our model, we conduct an ablation study comparing the complete model (CateEA) with versions missing different modules across three subsets of the DBP15K dataset (ZH-EN, JA-EN, FR-EN). As shown in Figure 3a, the complete model consistently achieves the highest H@1 scores across all subsets, and removing any module results in performance degradation, underscoring the contribution of each module to the overall performance.

4.3.2 Influence of Each Modality

To investigate the impact of different modal information on model performance, we conduct ablation experiments on the DBP15K dataset. As depicted in Figure 3b, removing any feature causes a performance drop, with the removal of image information having the greatest impact, emphasizing its crucial role in multi-modal entity alignment. The removal of relationship and attribute information individually results in minor performance degradation, but their simultaneous removal significantly reduces model performance, highlighting their complementary and essential roles in the alignment process. Overall, the complete model performs best across all sub-datasets, validating the necessity of multi-feature fusion for enhancing multi-modal entity alignment effectiveness.

4.4 Parameter Analysis

4.4.1 Seed Set Proportion

To assess the robustness of CateEA under low-resource conditions, we analyze the impact of seed set proportions on alignment performance on DBP15K(ZH-EN). This evaluation aims to determine how well the model can maintain alignment accuracy when faced with limited labeled data, a common challenge in real-world applications. By varying the seed set proportion, we observe the model's ability to adapt and perform reliably with minimal supervision..

Figure 4a illustrates that CateEA maintains strong performance with minimal labeled seeds, achieving reasonable alignment accuracy even with seed proportions as low as 1%, as indicated by MRR and Hits@1. This highlights the robustness and ability of CateEA to effectively leverage limited seed data, demonstrating its adaptability and effectiveness in handling multi-modal information and semantic structures, making it suitable for real-world applications with sparse labeled data.

4.4.2 Number of Categories

To investigate the effect of the number of categories on model performance, we conduct experiments on DBP15K(FR-EN) with various category settings ($C = 3, 5, 10, 15, 20$). As shown in Figure 4b, MRR and Hits@1 slightly decrease as the number of categories increases. While a finer granular-

ity of categories allows for a more detailed representation of entity category information, it also introduces noise, reduces the distances between categories, increases model complexity, and can cause category imbalance. These combined factors contribute to making the alignment process less robust, weakening the classifier's generalization ability, and slightly degrading the alignment performance. Therefore, selecting an appropriate number of categories is crucial to balancing the utilization of fine-grained information and maintaining optimal alignment performance.

5 Conclusion

We propose CateEA, a knowledge graph entity alignment method leveraging implicit entity category information from multi-modal data. By introducing a classification task as an additional training objective and using category-driven clustering results, CateEA captures richer category-level semantics. Experimental results show that CateEA outperforms state-of-the-art methods with notable gains: on FB15K-DB15K, H@1 increases by 3.4%–7.6%, and on FB15K-YAGO15K, H@1 improves by up to 7.5%. These results highlight the effectiveness of CateEA for multi-modal knowledge graph entity alignment.

6 Limitations

Despite its promising results, CateEA still faces several limitations. The clustering process, which is crucial for extracting category information, can heavily influence alignment quality if not sufficiently optimized. Additionally, the current approach is restricted to static knowledge graphs, leaving temporal and event-centric applications unexplored. Another practical concern is increased computational overhead due to the additional classification step, leading to longer running times compared to baseline methods. This issue can be further exacerbated by larger graph sizes, where the increasing number of nodes raises training time, potentially compromising CateEA's scalability. Future work will focus on refining clustering method, incorporating temporal aspects, and improving overall efficiency.

7 Ethics Statement

To the best of our knowledge, this work does not involve any discrimination, social bias, or private data. All the datasets are constructed from open-source knowledge graphs such as Wikidata, YAGO, and DBpedia. Therefore, we believe that our study complies with the Ethics Policy.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Yixin Cao, Zhiyuan Liu, Chengjiang Li, Juanzi Li, and Tat-Seng Chua. 2019. Multi-channel graph neural network for entity alignment. *arXiv preprint arXiv:1908.09898*.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149.

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 13*, pages 134–147. Springer.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022. Multi-modal siamese network for entity alignment. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 118–126.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2016. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *arXiv preprint arXiv:1611.03954*.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z Pan, Wenting Song, et al. 2023a. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3317–3327.

Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023b. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *International Semantic Web Conference*, pages 121–139. Springer.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2019. Kbqa: learning question answering over qa corpora and knowledge bases. *arXiv preprint arXiv:1903.02419*.

Yunjun Gao, Xiaoze Liu, Junyang Wu, Tianyi Li, Pengfei Wang, and Lu Chen. 2022. Clusterea: Scalable entity alignment with stochastic training and normalized mini-batch similarities. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 421–431.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. 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*, pages 2499–2508.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. *arXiv preprint arXiv:2209.00891*.

Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4257–4266.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-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 16*, pages 459–474. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

LKPJ Rdusseeun and P Kaufman. 1987. Clustering by means of medoids. In *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland*, volume 31.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*, pages 628–644. Springer.

Zequn Sun, Jiacheng Huang, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, pages 612–629. Springer.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

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*, pages 349–357.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. *arXiv preprint arXiv:1910.06575*.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362.

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. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735.

A Appendix

A.1 Datasets

The detailed statistics of the dataset are presented in Table 5, which includes the number of entities (#Ent.), relations (#Rel.), attributes (#Attr.), relation triples (#Rel tr.), attribute triples (#Attr tr.), images (#Image), and seed entities (#Seed.). It is important to note that not all entities have corresponding images or matching counterparts in the target knowledge graph.

A.2 Metric Detail

Definition of Hits@N metric:

$$\text{Hits@N} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\mathbb{I}(\text{rank}_i \leq N)$$

where $|Q|$ is the total number of test samples. rank_i is the rank of the correct answer for the i-th query in the list returned by the model. For each query, if the correct answer's rank is within the top N, it is considered a hit; otherwise, it is not. Finally, the average of all query hits is taken to get the Hits@N.

Definition of MRR metric:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

For each query, the inverse rank of the correct answer is calculated, and the average of all the inverse ranks is then taken.

A.3 Category Information Enhancement Algorithm

Algorithm 1 Enhanced Multi-Modal Entity Alignment.

Input: Joint embeddings $\{\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_N\}$, and the number of categories C
Output: Adjusted similarity scores
for $i = 1$ *to* N **do**
  Cluster $\hat{h}_i$ using K-Medoids to obtain labels $\{C_1, C_2, \ldots, C_C\}$
end
for $i = 1$ *to* N **do**
  $L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c})$,
  $L = L_{align} + \lambda L_{cls}$
end
for $i = 1$ *to* C **do**
  $\mu_i = \frac{1}{|C_i|}\sum_{x\in C_i}\hat{h}_x$
end
for $i, j = 1$ *to* C **do**
  $d_{ij} = \|\mu_i - \mu_j\|$,
  $s_{ij} = \frac{1}{1+d_{ij}}$,
  $w_{ij} = \frac{\exp(s_{ij})}{\sum_k \exp(s_{ik})}$
end
for *each pair* (x_i, y_j) **do**
  $\text{Sim}(x_i, y_j) = w_{ij} \cdot \text{BaseSim}(x_i, y_j)$
end

A.4 Performance over Epochs

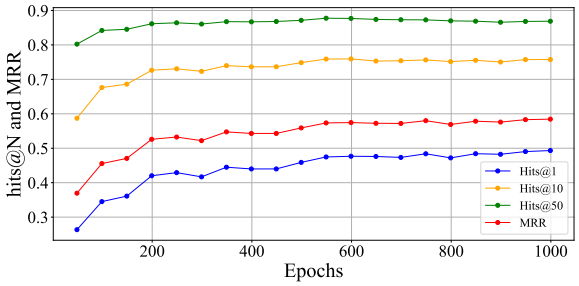
To thoroughly evaluate the learning effectiveness, convergence and robustness of CateEA, we examine the performance changes of CateEA over training epochs on the FBDB15K dataset. From Figure 5, it is observed that Hits@1, Hits@10, and Hits@50 gradually increase and stabilize as the training progresses, indicating that the overall alignment capability of CateEA is consistently

Table 5: Dataset Statistics.

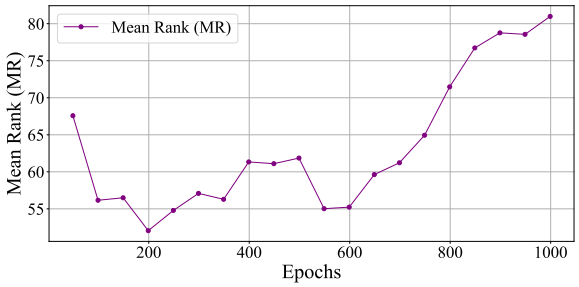
Dataset	KG	#Ent.	#Rel.	#Attr.	#Rel tr.	#Attr tr.	#Image	#Seed.
$\text{DBP15K}_{\text{ZH-EN}}$	ZH	19,388	1,701	8,111	70,414	248,035	15,912	15,000
	EN	19,572	1,323	7,173	95,142	343,218	14,125	15,000
$\text{DBP15K}_{\text{JA-EN}}$	JA	19,814	1,299	5,882	77,214	248,991	12,739	15,000
	EN	19,780	1,153	6,066	93,484	320,616	13,741	15,000
$\text{DBP15K}_{\text{FR-EN}}$	FR	19,661	903	4,547	105,998	273,825	14,174	15,000
	EN	19,993	1,208	6,422	115,722	351,094	13,858	15,000
FB15K-DB15K	FB15K	14,951	1,345	116	592,213	29,395	13,444	12,846
	DB15K	12,842	279	225	89,197	48,080	12,837	12,846
FB15K-YAGO15K	FB15K	14,951	1,345	116	592,213	29,395	13,444	11,199
	YAGO15K	15,404	32	7	122,886	23,532	11,194	11,199

improving. The Mean Rank (MR) shows some fluctuations, especially an upward trend in the later training stages, possibly due to poor ranking performance on certain samples. The MRR continues to rise and stabilize, demonstrating the enhanced ability of CateEA to find the target entity among the top few candidates.

The concurrent increase in MR and MRR indicates that, despite occasional errors in some edge cases, overall performance continues to improve, demonstrating the robustness of CateEA.



(a) HitsN and MRR over epochs.



(b) Mean rank over epochs.

Figure 5: Performance evaluation over epochs.

A.5 Case Study

To assess the impact of Implicit Category Supervision (ICS) during the testing phase, we take a case study (as shown in Figure 6) on the DBP15K JA-EN dataset, listing the predicted target entities of CateEA with and without ICS in Table 6. The introduction of the ICS substantially improves the alignment accuracy of source entities. For instance, in cases like '*Slovenian PrvaLiga*' and '*Premier League of Bosnia and Herzegovina*', although both entities belong to the football league category, they are from different countries (Slovenia and Bosnia-Herzegovina). The ICS provides more refined discrimination, allowing CateEA to capture subtle differences between leagues from different nations and match them correctly. On the other hand, in scenarios with entities like '*Hachisuka Narihiro*' (a historical figure) and '*Masahiro Chono*' (a wrestler), which are from different categories, the ICS enhances the category information, reducing cross-category mismatches and improving the identification and alignment of entities from diverse categories.

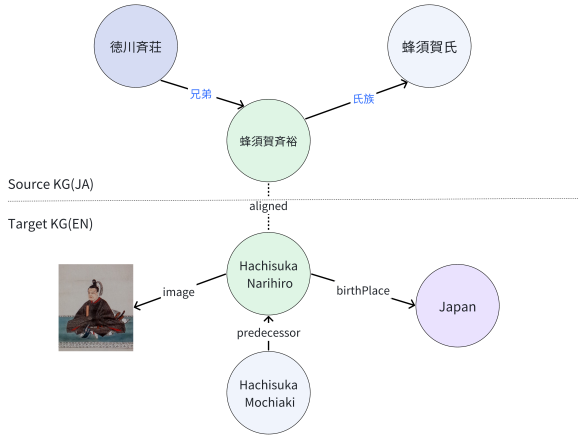


Figure 6: A case of source and target sub-KGs from the DBP15K JA-EN dataset.

Table 6: Example of predicted entities with and without category information supervision (ICS).

Source Entity ID	Target Entity	Prediction w. ICS	Prediction w/o ICS
176	Slovenian PrvaLiga	Slovenian PrvaLiga	Premier League of Bosnia and Herzegovina
893	Hachisuka Narihiro	Hachisuka Narihiro	Masahiro Chono
22753	Sone Arasuke	Sone Arasuke	Inukai Tsuyoshi
8089	Circuit de Monaco	Circuit de Monaco	Autodromo Enzo e Dino Ferrari
25225	& (Ayumi Hamasaki EP)	& (Ayumi Hamasaki EP)	A Complete: All Singles
8853	Hàm Nghi	Hàm Nghi	T Đc
529	2010 Winter Olympics	2010 Winter Olympics	2008 Summer Olympics