Exploring the Impacts of Feature Fusion Strategy in Multi-modal Entity Alignment

Chenxiao Li, Jingwei Cheng*, Qiang Tong, Fu Zhang,
Northeastern University, China
chenxiaoli_joe@163.com, {chengjingwei,tongqiang,zhangfu}@mail.neu.edu.cn

Abstract

Multi-modal entity alignment aims to identify equivalent entities between two different multi-modal knowledge graphs, which consist of structural triples and images associated with entities. Unfortunately, prior works fuse the multi-modal knowledge of all entities only via solely one single fusion strategy. Therefore, the impact of the fusion strategy on individual entities could be largely ignored. To solve this challenge, we propose AMF^2SEA, an adaptive multi-modal feature fusion strategy for entity alignment, which dynamically selects the optimal entity-level feature fusion strategy. Additionally, we build a new dataset based on DBP15K, which includes a full set of entity images from multiple inconsistent web sources, making it more representative of the real world. Experimental results demonstrate that our model achieves state-of-the-art (SOTA) performance compared to models using the same modality on DBP15K and its variants with richer image sources and styles. Our code and data are available at https://github.com/ChenxiaoLi-Joe/AMFFSEA.

1 Introduction

Multi-modal knowledge graphs (MMKGs) organize real-world knowledge across modalities such as text and vision, have drawn massive attention in various scenarios and supported numerous AI applications (Zhu et al., 2015; Yang et al., 2021). Due to the increasing need for comprehensive multi-modal knowledge integration, multi-modal entity alignment (MMEA) (Chen et al., 2020; Liu et al., 2019) has emerged as a significant task in this field.

Several previous MMEA works have shown that the inclusion of visual modality in modeling helps to improve the performance of entity alignment. For instance, Lin et al. (2022) obtain discriminative entity representations based on contrastive learning for entity alignment. Chen et al. (2022) employ inter-modal enhancement mechanisms to integrate visual features to guide relational feature learning. Chen et al. (2023a) dynamically predict the mutual correlation coefficients among modalities for entity-level feature aggregation. Chen et al. (2023b) proactively complete missing modality information to alleviate the impact of incompleteness on the alignment process.

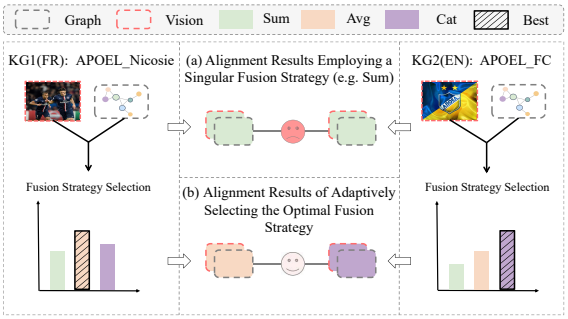

Figure 1: (a) Alignment results employing a single fusion strategy and (b) alignment results of adaptively selecting the optimal strategy from multiple fusion strategies.

However, these methods mainly utilize unified joint representations from different modalities, and the impact of fusion strategies on an entity level has not been fully explored. Indeed, as shown in Figure 1, due to the variability of image styles, a single fusion strategy may lead to model overfitting or fail to adequately capture the semantic information of the images. In addition, the incompleteness of visual data and the high correlation between images and entities that violates real retrieval scenarios present significant challenges for multi-modal entity alignment. As reported by Liu et al. (2021), approximately 15–50% of entities in the commonly used benchmark DBP15K (Sun et al., 2017) lack images. The only existing images all originate from DBpedia and do not reflect real-world scenarios.

*Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 7809–7818
January 19–24, 2025.

To tackle these challenges, in this paper, we address the entity alignment problem for multi-modal knowledge graphs by proposing an Adaptive Multi-modal Feature Fusion Strategy for Entity Alignment (AMF^2SEA) to accurately obtain the semantic information of images of different styles. Specifically, we first enhance the structural information by incorporating filtered visual information and assign weights based on the importance of each modality. We then select and apply the optimal fusion strategy to improve the performance of multi-modal entity alignment.

In summary, our main contributions are threefold:

- We propose a new MMEA method called AMF^2SEA, which adaptively selects the optimal feature fusion strategy on an entity level. To the best of our knowledge, we are the first to investigate the impact of feature fusion strategies on individual entities.
- We build a new dataset based on DBP15K (Sun et al., 2017), which includes a full set of entity images from multiple inconsistent web sources. This dataset mirrors real-world scenarios by showcasing the diversity of image sources and styles, and highlighting that images retrieved based on entities are not always directly associated with those entities.
- Extensive experiments on multiple datasets, including the one proposed above, demonstrate the robustness and effectiveness of AMF^2SEA, which significantly outperforms several state-of-the-art baseline methods. Experimental results indicate that existing entity alignment methods are affected by the aforementioned issues, and our model effectively addresses these problems.

2 Related Work

Generally, the related work can be classified into two perspectives, i.e., text-based entity alignment and multi-modal entity alignment. In addition, we present insights on datasets for multi-modal entity alignment.

2.1 Text-based Entity Alignment

Embedding-based approaches for entity alignment (EA) can be generally divided into two categories: that only utilized graph structures and that used additional side information of entities (Zhang et al., 2020, 2021). By encoding entities and relations of each language in a separated embedding space, MTransE (Chen et al., 2016) provides transitions for each embedding vector to its cross-lingual counterparts in other spaces, while preserving the functionalities of monolingual embeddings. IPTransE (Zhu et al.) jointly encodes both entities and relations of various KGs into a unified low-dimensional semantic space according to a small seed set of aligned entities. BootEA (Sun et al.) iteratively labels likely entity alignment as training data for learning alignment-oriented KG embeddings. GCN-Align (Wang et al., 2018) trains GCNs to embed entities of each language into a unified vector space. JAPE (Sun et al., 2017) jointly embeds the structures of two KBs into a unified vector space and further refines it by leveraging attribute correlations in the KBs. AttrE (Trisedya et al., 2019) uses a transitivity rule to further enrich the number of attributes of an entity to enhance the attribute character embedding. MultiKE (Zhang et al., 2019) unifies multiple views of entities to learn embeddings for entity alignment. To exploit the literal descriptions of entities expressed in different languages, HMAN (Yang et al., 2019) integrates GCN-based and BERT-based modules to boost performance. UEA (Zhao et al., 2022) offers an unsupervised framework that performs entity alignment in the open world. Although some of the above approaches can achieve high accuracy on EA, the fusion strategy of visual context has not been explored yet.

2.2 Multi-modal Entity Alignment

Nevertheless, the above-mentioned methods focus on the textual facts with few multi-modal sources. Actually, in addition to text and structured data, visual and auditory data, such as pictures, videos and audio, can also be the data sources. In recent years, the incorporation of visual modalities for entity alignment in knowledge graphs has garnered increasing attention within academic communities, driven by advancements in multi-modal learning. EVA (Liu et al., 2021) provides a completely unsupervised solution by leveraging the visual similarity of entities to create an initial seed dictionary (visual pivots). MCLEA (Lin et al., 2022) utilizes an entity alignment model based on multi-modal contrastive learning to obtain effective joint representations for multi-modal entity alignment. MSNEA (Chen et al., 2022) introduces inter-modal enhancement mechanisms in multi-modal knowledge represen-

tation. Masked-MMEA (Shi et al., 2022) exploits image classification techniques and entity types to remove potentially visual noises via generating entity mask vectors in the learning and inference processes. MEAformer (Chen et al., 2023a) provides a multi-modal entity alignment transformer approach for meta modality hybrid, which dynamically predicts the mutual correlation coefficients among modalities for entity-level feature aggregation. UMAEA (Chen et al., 2023b) proactively completes missing modality information to alleviate the impact of incompleteness on the alignment process.

2.3 Datasets for Multi-modal Entity Alignment

Mainstream datasets typically retrieve images of entities from DBpedia. Existing research has shown that visual information significantly enhances multi-modal entity alignment. However, we notice that all of them are based on an ideal assumption that images are strongly associated with entities and are always available.

In real-world scenarios, the process of data construction is often more complex. Consequently, we focus on two more pragmatic and demanding issues: (i) Irrelevant images are frequently introduced during image acquisition. (ii) Images from web sources exhibit more styles, which may be challenging for the model to understand multi-modal semantics. To tackle these challenges, we propose a multi-modal variant of DBP15K (Sun et al., 2017), sourced entirely from web sources. Details of the dataset and its construction will be provided in Section 4.4.4.

3 Methodology

3.1 Problem Formulation

Multi-modal Knowledge Graph. A multi-modal knowledge graph is formalized as $G = (E, R, T, I, P)$. Here, E, R, T, and I are the sets of entities, relations, triples, and images, respectively. $P = \{(e, i) \mid e \in E, i \in I\}$ is the set of entity-image pairs.

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

3.2 Framework Overview

In this paper, we propose an Adaptive Multi-modal Feature Fusion Strategy for Entity Alignment (AMF^2SEA), to conquer the aforementioned challenges. Our proposed AMF^2SEA comprises four major components: 1) Multi-modal Knowledge Embedding module to extract structural and visual features, to generate holistic entity representations; 2) Image Noise Filter to mitigate image noise by using image classification techniques and entity types; 3) Adaptive Multi-modal Feature Fusion Strategy module to dynamically generate the entity-level weight for each modality and adaptively select the optimal feature fusion strategy for each entity; 4) Alignment Learning and Inference module uses HAL loss (Liu et al., 2020) and cosine similarity matrices to select the entity with the highest similarity score as a match. The framework overview is illustrated in Figure 2, and its primary components will be detailed in the following sections.

3.3 Multi-modal Knowledge Embedding

Due to the integration of information from at least two modalities in MMKGs, and in order to better analyze the impact of visual information on MMEA, we only model two modalities, namely structural and visual information.

Structural Embedding. As a typical neural network, GCN (Kipf and Welling, 2016) is utilized to model the structural information. Given as input the adjacency matrix $\mathbf{A}$ of a KG and a randomly initialized feature matrix $\mathbf{H}^{(0)}$ of its entities, a multi-layer GCN sequentially updates entity representations from the i-th layer to the $(i+1)$-th layer using the following propagation rule:

$$\mathbf{H}^{(i+1)} = \phi\left(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(i)}\mathbf{W}^{(i+1)}\right), \quad (1)$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ and $\mathbf{I}$ is an identity matrix, $\hat{\mathbf{D}}$ is the diagonal degree matrix of $\hat{\mathbf{A}}$, $\mathbf{W}^{(i+1)}$ denotes learnable parameters in the $(i+1)$-th layer and ϕ is the activation function ReLU.

Visual Embedding. We choose ResNet-152 (He et al., 2016) pre-trained on the ImageNet (Deng et al., 2009) recognition task as the initial image classifier. We then fine-tuned it on DBP15K to

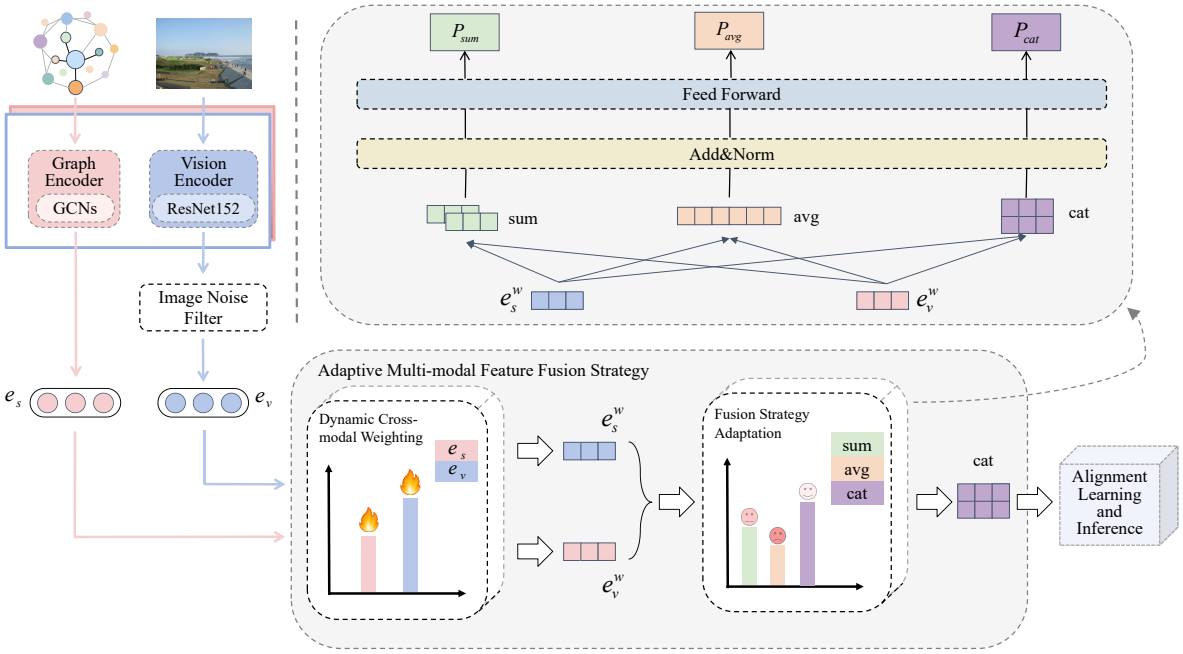

Figure 2: The overall framework and implementation details of AMF^2SEA.

extract image features, resulting in the visual embedding $\mathbf{e}_v^{early}$ as follows:

$$\mathbf{e}_v^{early} = \mathbf{W}_v \cdot \text{ResNet}(i) + \mathbf{b}_v, \tag{2}$$

where $\mathbf{W}_v$ is a projection matrix and $\mathbf{b}_v$ is a bias vector.

3.4 Image Noise Filter

When the visual information contains noise (e.g., irrelevant or distracting elements), it can lead to inconsistencies between different information modalities. In the face of such information, models succumb to overfitting the modality noise and exhibit performance oscillations or declines. This indicates that additional multi-modal data negatively impacts entity alignment and leads to even worse results than when no visual modality information is used.

To mitigate image noise, we introduce the image noise filter with special masks provided by Shi et al. (2022) to effectively identify and remove noise from visual images. Specifically, we employ an entity masking vector $\mathbf{M}$, where $\mathbf{M}_{e_i}$ denotes the masking value of the i-th entity e_i. If the image of entity e_i is determined to contain potential noise, then $\mathbf{M}_{e_i}$ is set to 0, indicating that e_i is masked and its image is filtered during training or testing. Otherwise, we set $\mathbf{M}_{e_i} = 1$. We initialize M with all zeros and iteratively update it. Given a conflict threshold λ, for each entity $e \in E$, we feed its corresponding image to the classifier, resetting the masking value of e to 1 if the conflict between its prediction and the actual class is not greater than λ. Inspired by OntoEA (Xiang et al., 2021), we use a class conflict dictionary (CCD) to store inter-class conflicts. Given two classes a and b, we calculate their conflict degree as $C[a, b]$. Finally, the obtained mask $\mathbf{M}_{e_i}$ is multiplied element-wise with the image feature vector of e_i to obtain the final usable image feature.

$$\mathbf{e}_{v_i} = \mathbf{M}_{e_i} \cdot \mathbf{e}_{v_i}^{early}, \tag{3}$$

where $\mathbf{e}_{v_i}$ is the visual embedding of the i-th entity after processing.

3.5 Adaptive Multi-modal Feature Fusion Strategy

This section describes the detailed architecture of the adaptive multi-modal feature fusion strategy for aligning multi-modal entities between MMKGs.

Dynamic Cross-modal Weighting. We propose a multi-modal weight calculation mechanism aiming at enhancing interaction between the structural and visual modalities. First, two independent linear layers are initialized, each receiving embeddings of the same shape from different modalities. Subsequently, each embedding undergoes a linear transformation to map the original feature dimensions to a single output dimension, resulting in a scalar value representing the score for each sample in both the structural and visual modalities. Finally, the

output score vectors are processed through a softmax function, transforming them into normalized probability distributions. The weights $\mathbf{w}_s$ for the structural modality and $\mathbf{w}_v$ for the visual modality are calculated as follows:

$$\mathbf{w} = \text{softmax}\left(\mathbf{W} \cdot \mathbf{e} + \mathbf{b}\right), \tag{4}$$

where $\mathbf{e}$ is defined as the structural embedding and the visual embedding. $\mathbf{W}$ is a learned projection matrix, and $\mathbf{b}$ is a bias vector.

Fusion Strategy Adaptation. To reach our goal, we develop an adaptive feature fusion strategy that enables entities to dynamically select the optimal fusion method. Specifically, a predefined multi-modal weight calculation mechanism is used to compute the weighted structural embedding $\mathbf{e}_s^w$ and visual embedding $\mathbf{e}_v^w$. Subsequently, the weighted structural and visual embeddings are concatenated along the feature dimension to obtain a combined embedding. Finally, we utilize a linear layer with three output nodes to predict the fusion strategies, corresponding to the three potential fusion methods.

$$\mathbf{e}_{comb} = \begin{cases} \mathbf{e}_{comb}^{sum} : \frac{\mathbf{e}_s^w}{\mathbf{e}_s^w + \mathbf{e}_v^w}\mathbf{e}_s + \frac{\mathbf{e}_v^w}{\mathbf{e}_s^w + \mathbf{e}_v^w}\mathbf{e}_v, \\ \mathbf{e}_{comb}^{avg} : \sum_{i\in\{s,v\}}\left(\frac{\cos(\mathbf{e}_i, \bar{\mathbf{e}})}{\sum_{j\in\{s,v\}}\cos(\mathbf{e}_j, \bar{\mathbf{e}})}\right)\mathbf{e}_i, \\ \mathbf{e}_{comb}^{cat} : \frac{\mathbf{e}_s^w}{\mathbf{e}_s^w + \mathbf{e}_v^w}\mathbf{e}_s \oplus \frac{\mathbf{e}_v^w}{\mathbf{e}_s^w + \mathbf{e}_v^w}\mathbf{e}_v, \end{cases} \tag{5}$$

where the symbol $\oplus$ denotes concatenation of embeddings. Meanwhile, $\bar{e} = \frac{1}{2}(\mathbf{e}_s + \mathbf{e}_v)$ assigns weights to modality-specific entity embeddings, allowing the model to emphasize important modalities through this combination.

During the model initialization stage, weight matrices and bias vectors are automatically created and initialized with random values. After mutual learning, the similarity between entities and the three joint embeddings is calculated to obtain initial scores. These scores are then transformed through a linear layer and normalized via a softmax function to yield probability distributions for each fusion strategy. Finally, an argmax operation is applied to the probability distributions to predict the most similar fusion strategy.

$$\mathbf{e}_{final} = \text{argmax}(\text{softmax}\left(\mathbf{e}_{comb}\right)), \tag{6}$$

where $\mathbf{e}_{final}$ is defined as the final multi-modal fusion representation used for entity alignment reasoning. $\mathbf{e}_{comb}$ contains the three fusion representations in (4): $\mathbf{e}_{comb}^{sum}$, $\mathbf{e}_{comb}^{avg}$, and $\mathbf{e}_{comb}^{cat}$.

3.6 Alignment Learning and Inference

This section presents details about alignment learning and inference. We integrate two KGs as one KG and learn both structural embeddings and visual embeddings of entities in a unified space. To better punish hard negatives and mitigate the hubness problem (Conneau et al., 2017), we choose HAL loss (Liu et al., 2020) as the objective function and apply it to obtain the loss of the structural modality $L_{(s)}$. Likewise, we compute $L_{(v)}$ for the visual modality. We compute cosine similarity matrices for the structural and visual modalities. Then we combine them by a weighted addition and a position mask to obtain the fused similarity matrix **Sim**, the *(i, j)* entry of **Sim**, is computed as:

$$\mathbf{Sim}_{ij} = \begin{cases} \text{if } \mathbf{pos}_{ij} = 1 : \\ \quad w \cdot \mathbf{Sim}_{ij}^{(s)} + (1-w) \cdot \mathbf{Sim}_{ij}^{(v)} \\ \text{otherwise :} \\ \quad \mathbf{Sim}_{ij}^{(s)} \end{cases} \tag{7}$$

where $\mathbf{Sim}_{ij}^{(s)}$ and $\mathbf{Sim}_{ij}^{(v)}$ are cosine similarity matrices for the structural and visual modalities, respectively. $\mathbf{pos}_{ij}$ is used to determine if their visual similarity should be considered.

After obtaining **Sim**, we further use cross-domain similarity local scaling (CSLS) (Conneau et al., 2017) to post-process it. Then for $e_i \in \bar{E}_s$, we retrieve the similarity scores of the i-th row in **Sim**, rank them in a descending order, and take the top ranked entity as the match.

4 Experiments

4.1 Experiment Setup

Datasets. Since image retrieval results are significantly affected by language differences, and bilingual datasets can more comprehensively evaluate the performance of multi-modal entity alignment methods in cross-language environments, thereby verifying their effectiveness in practical applications, we only use bilingual datasets in our experiments. DBP15K (Sun et al., 2017) contains three datasets built from the multilingual versions of DBpedia, including $DBP15K_{ZH-EN}$, $DBP15K_{JA-EN}$ and $DBP15K_{FR-EN}$. We adopt their multi-modal variants (Liu et al., 2021) with entity-matched images attached. Inspired by Shi et al. (2022), we query the classes of each entity with $rdf:type$ via a public SPARQL endpoint to retrieve entity types. We also obtain the subsumption relationships between classes which are

Methods	FR-EN			JA-EN			ZH-EN		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
MTransE(Chen et al., 2016)	0.224	0.556	0.335	0.279	0.575	0.349	0.308	0.614	0.364
IPransE(Zhu et al.)	0.333	0.685	0.451	0.367	0.693	0.474	0.406	0.735	0.516
JAPE(Sun et al., 2017)	0.324	0.667	0.430	0.363	0.685	0.476	0.412	0.745	0.490
GCN-Align(Wang et al., 2018)	0.373	0.745	0.532	0.399	0.745	0.546	0.413	0.744	0.549
SEA(Pei et al., 2019)	0.400	0.797	0.533	0.385	0.783	0.518	0.424	0.796	0.548
MuGNN(Cao et al., 2019)	0.495	0.870	0.621	0.501	0.857	0.621	0.494	0.844	0.611
HMAN(Guo et al., 2021)	0.543	0.867	-	0.565	0.866	-	0.537	0.834	-
AliNet(Sun et al., 2020)	0.552	0.852	0.657	0.549	0.831	0.645	0.539	0.826	0.628
MultiKE(Zhang et al., 2019)	0.639	0.712	0.665	0.393	0.489	0.426	0.509	0.576	0.532
EVA(Liu et al., 2021)	0.700	0.891	0.768	0.622	0.846	0.701	0.596	0.816	0.674
Masked-MMEA(Shi et al., 2022)	0.712 ±.005	0.901 ±.003	0.779 ±.004	0.627 ±.005	0.858 ±.005	0.711 ±.004	0.612 ±.006	0.837 ±.006	0.693 ±.005
AMF^{2}SEA$_{(\mathbf{Ours})}$	**0.767** ±.005	**0.914** ±.004	**0.818** ±.004	**0.696** ±.003	**0.871** ±.002	**0.757** ±.002	**0.691** ±.005	**0.879** ±.005	**0.751** ±.004

Table 1: Entity alignment results on DBP15K. For fair comparison, the results of HMAN are from its variant that only uses training data in DBP15K as alignment signals, and the results of EVA are reproduced by only utilizing structural and visual context, as the setting of AMF^2SEA. For Masked-MMEA and AMF^2SEA, $\frac{\text{Means}}{\pm\text{Stds.}}$ are shown. Best results are shown in bold.

explicitly defined by the $rdfs:subClassOf$ property in the DBpedia ontology.

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

Implementation Details. Following conventions, we use 30% of the aligned pairs for training and the remaining for evaluation. We employ a three-layer GCN (including the input layer) and set the dimensions of the input, hidden and output layers to 400, 400 and 200, respectively. We train our model for 600 epochs and adopt **AdamW** to update parameters. The learning rate is set to 5×10^{-4}. When calculating losses, we set $\alpha = 5$, $\beta = 10$ for $L^{(s)}$, and $\alpha = 15$, $\beta = 10$ for $L^{(v)}$.

4.2 Comparative Methods

To generally verify the effectiveness of adaptively selecting the optimal fusion strategy incorporating visual information at the entity level, we selected 11 prominent EA algorithms proposed in recent years as comparative methods.

Recent multi-modal approaches for entity alignment, such as MCLEA (Lin et al., 2022), MSNEA (Chen et al., 2022), MEAformer (Chen et al., 2023a), UMAEA (Chen et al., 2023b), etc. use three or more types of information including structural data, numerical/attribute triples, visual knowledge and surface names of entities to improve alignment performance. Our work focuses on probing the impact of fusion strategies involving visual information. Due to the differences between modalities, including data distribution, noise, and feature representation, introducing other modalities may introduce more variables, involve different problem definitions, and it is difficult to explain the semantic understanding of visual information. Since MMKGs integrate at least two modalities, and structural information is dominant in MMEA with the least noise compared to other modalities (Chen et al., 2020), we aimed to better control the experimental conditions to accurately evaluate the fusion strategies for visual information. Therefore, we only utilized structural and visual information in our experiments. For a fair comparison, we did not include other methods.

4.3 Overall Results

The results of the bilingual datasets are shown in Table 1. It is clear that our model has better performance than baselines across all the datasets under all the metrics. The superiority of AMF^2SEA confirms that the proposed adaptive multi-modal feature fusion strategy substantially promotes the performance.

4.4 Ablation Studies

In this section, we first control the number of images to ensure that the visual modality benefits multi-modal entity alignment. We then verify the effectiveness of the adaptive multi-modal feature fusion strategy by comparing it with variants of AMF^2SEA that employ only a single fusion strategy. We also use relations and attributes to fuse with visual information, confirming that structural information has the least noise and better controls experimental conditions to accurately evaluate the visual information fusion strategy, providing relevant insights. Finally, to increase the difficulty of the task, we propose a dataset with lower correlation between images and entities and richer image sources and styles. Extensive experiments verify the generalization and robustness of AMF^2SEA on real-world data.

4.4.1 Importance of Visual Modality

To examine the effects of visual modality on the entity alignment, we follow the class conflict ratio proposed by Shi et al. (2022) and choose different class conflict ratios: $\lambda \in \{0, 0.4, 0.67, 1\}$, in which $\lambda = 0$ corresponds to the strictest setting and $\lambda = 1$ is the no-masking setting where no entity images are filtered. As shown in Figure 3, the alignment accuracy improves significantly as the number of input images increases. This result indicates that even without using the image noise filter, our strategy outperforms Masked-MMEA (Shi et al., 2022). This phenomenon may be attributed to the richer visual information provided by more images, thereby enhancing the model's understanding and ability to capture complex relationships between entities.

However, when we introduce the image noise filter with special masks provided by Shi et al. (2022), the performance increases again, even with the same number of images as Masked-MMEA (Shi et al., 2022). This phenomenon may be attributed to the image noise filter removing images judged as noisy or not conducive to entity alignment. This also verifies that our model has stronger capabilities in multi-modal semantic understanding.

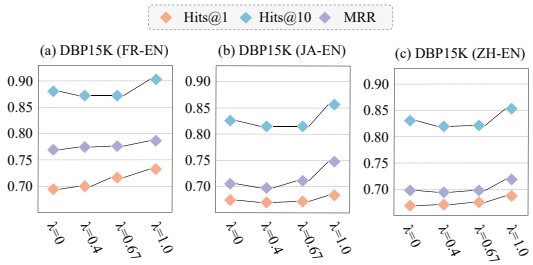

Figure 3: The alignment results on DBP15K as affected by the number of images and the strictness of the mask (λ).

4.4.2 Analysis of Fusion Strategy Adaptation

To assess the effectiveness of our model, we developed variants that exclusively employ a single fusion method and conducted a comparative analysis. The variants of AMF^2SEA employ a single fusion method to all entities, namely weighted sum (Sum), weighted average (Avg), and weighted concatenation (Cat). AMF^2SEA adaptively selects the optimal fusion strategy from three options at the entity level. The results are shown in Figure 4. The experimental results demonstrate that our model significantly outperforms the variant that relies solely on a single fusion strategy, particularly in terms of Hits@1 and MRR. These findings fully underscore the superiority and practical applicability of the adaptive feature fusion strategy in multi-modal entity alignment.

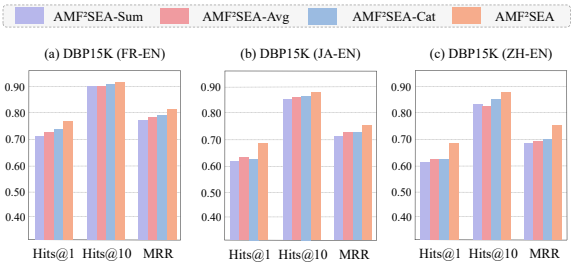

Figure 4: The results on DBP15K using both single and adaptive feature fusion strategies.

4.4.3 Analysis of Noise in Relation and Attribute

In this section, we verify that structural information has the least noise compared to other modalities

by applying independent fully connected layers to extract features of relations and attributes, and then fusing them with visual information. The results in Figure 5 indicate that their performance is inferior to the fusion of structural and visual information, particularly in the case of relation and visual information fusion. This finding confirms that relational and attribute data contain more noise than structural information.

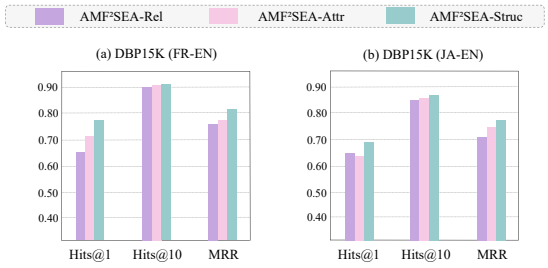

Figure 5: The results on DBP15K using relations and attributes with visual fusion respectively.

Furthermore, we explore the noise in relations and attributes in greater detail. We find that relations contain a significant amount of heterogeneous information. For instance, in a knowledge graph, relations such as $(Obama, President, United\ States)$ may have different interpretations in real-world scenarios, such as $place\ of\ birth$. However, the label is marked as $President$, which introduces noise. Reducing this noise is a critical issue. In addition, attribute information from various sources offers diverse descriptions of the same entity, known as attribute heterogeneity, which can impact the outcomes of entity alignment.

4.4.4 Data-oriented Generalization

To enrich image sources and image styles, we propose a multi-modal variant of DBP15K (Sun et al., 2017), sourced entirely from web sources. Specifically, we use names to extract images from their Uniform Resource Identifiers (URIs) via regular expressions. The statistics of image coverage are presented in Table 2.

Image coverd	FR-EN	JA-EN	ZH-EN
By DBpedia	70.69%	66.87%	77.09%
By web source	100%	100%	100%

Table 2: Statistics of image coverage.

Regarding classification criteria, we found that the common pre-trained models have defects in fine-grained semantic understanding. Therefore, we categorize entities into four base classes: $Person$, $Place$, $Organization$, and $Work$. This is different from the classification criteria of Shi et al. (2022).

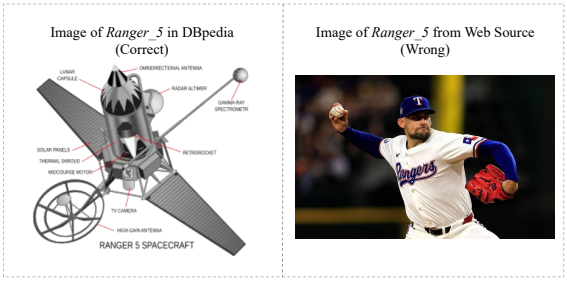

Figure 6: Results of searching for images of $Ranger\_5$ from web sources and DBpedia.

In our dataset, we identified numerous images that are not aligned with entities. For example, the entity Ranger_5 is a spaceship in the Ranger project and is classified as work. In DBP15K (Sun et al., 2017), this image is correct. However, a Bing search using the entity name retrieves an image of a player whose team name is Ranger. The example is shown in Figure 6. Therefore, before applying the image noise filter, this image would be incorrectly learned. However, after applying the image noise filter, the image can be properly cleaned. In addition, we discovered some images resembling text descriptions, a style not present in previous datasets. Performance analysis revealed that this style negatively impacts entity alignment and introduces additional noise.

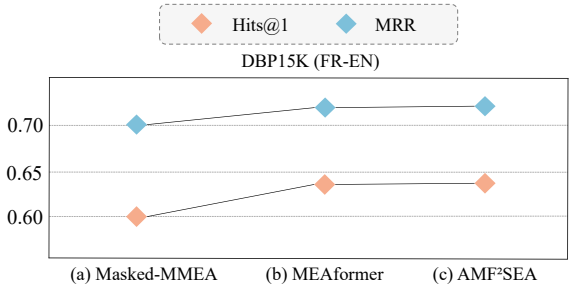

Figure 7: Analysis of performance based on data with rich image sources and styles.

For a fair comparison, we utilized a stripped-down version of MEAformer (Chen et al., 2023a) that incorporates only structural and visual information. Given that other models discussed in this

paper incorporate additional modalities to influence visual learning, their simplified versions are not suitable for comparison. The results are shown in Figure 7. The experimental results demonstrate that when the relevance between images and entities is not ensured and the images have more styles, models may learn noise. The results also demonstrate the strong adaptability and potential application value of AMF^2SEA in processing real-world data.

5 Conclusion

In this work, we have studied an Adaptive Multimodal Feature Fusion Strategy for Entity Alignment, which encourages the emergence of adaptive feature fusion strategy preferences. Extensive experiments on multiple real-world datasets demonstrated the robustness and effectiveness of our solution for multi-modal entity alignment, which outperformed several state-of-the-art baseline methods with a significant margin. In future work, we will continue to explore more fusion strategies that can properly capture the semantic information of images and add them into the framework while considering the efficiency issue.

Limitations

In this study, introducing additional or more complex fusion methods may incur significant computational overhead. This not only increases the model's training time but may also affect its efficiency in practical applications. Therefore, future research should explore the introduction of additional modalities and the adoption of more complex fusion methods to enhance the interaction between modalities, thereby improving model performance while maintaining computational efficiency.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments and suggestions, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057), and Sponsored by CAAIMindSpore Open Fund, developed on OpenI Community.

References

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*.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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*.

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.

Fangyu Liu, Rongtian Ye, Xun Wang, and Shuaipeng Li. 2020. Hal: Improved text-image matching by mitigating visual semantic hubs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11563–11571.

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.

Shichao Pei, Lu Yu, Robert Hoehndorf, and Xiangliang Zhang. 2019. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *The world wide web conference*, pages 3130–3136.

Yinghui Shi, Meng Wang, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. 2022. Probing the impacts of visual context in multimodal entity alignment. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 255–270. Springer.

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, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping entity alignment with knowledge graph embedding.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 222–229.

Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. 2019. Entity alignment between knowledge graphs using attribute embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 297–304.

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.

Yuejia Xiang, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Zhenxi Lin, and Yefeng Zheng. 2021. Ontoea: Ontology-guided entity alignment via joint knowledge graph embedding. *arXiv preprint arXiv:2105.07688*.

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*.

Shiquan Yang, Rui Zhang, Sarah M Erfani, and Jey Han Lau. 2021. Unimf: A unified framework to incorporate multimodal knowledge bases intoend-to-end task-oriented dialogue systems. In *IJCAI*, pages 3978–3984.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view knowledge graph embedding for entity alignment. *arXiv preprint arXiv:1906.02390*.

Yuxin Zhang, Bohan Li, Han Gao, Ye Ji, Han Yang, Meng Wang, and Weitong Chen. 2021. Fine-grained evaluation of knowledge graph embedding model in knowledge enhancement downstream tasks. *Big Data Research*, 25:100218.

Ziheng Zhang, Jiaoyan Chen, Xi Chen, Hualuo Liu, Yuejia Xiang, Bo Liu, and Yefeng Zheng. 2020. An industry evaluation of embedding-based entity alignment. *arXiv preprint arXiv:2010.11522*.

Xiang Zhao, Weixin Zeng, Jiuyang Tang, Xinyi Li, Minnan Luo, and Qinghua Zheng. 2022. Toward entity alignment in the open world: an unsupervised approach with confidence modeling. *Data Science and Engineering*, 7(1):16–29.

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Iterative entity alignment via joint knowledge embeddings.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*.