Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion

Xiang Chen, Ningyu Zhang*
Zhejiang University
AZFT Joint Lab for Knowledge Engine
Hangzhou Innovation Center
Hangzhou, Zhejiang, China
{xiang_chen,zhangningyu}@zju.edu.cn

Lei Li, Shumin Deng
Zhejiang University
AZFT Joint Lab for Knowledge Engine
Hangzhou Innovation Center
Hangzhou, Zhejiang, China
{leili21,231sm}@zju.edu.cn

Chuanqi Tan
Alibaba Group
Hangzhou, Zhejiang, China
chuanqi.tcq@alibaba-inc.com

Changliang Xu
State Key Laboratory of Media
Convergence Production Technology
and Systems
Beijing, China
xu@shuwen.com

Fei Huang, Luo Si
Alibaba Group
Hangzhou, Zhejiang, China
{f.huang,luo.si}@alibaba-inc.com

Huajun Chen*
Zhejiang University
AZFT Joint Lab for Knowledge Engine
Hangzhou Innovation Center
Hangzhou, Zhejiang, China
huajunsir@zju.edu.cn

ABSTRACT

Multimodal Knowledge Graphs (MKGs), which organize visual-text factual knowledge, have recently been successfully applied to tasks such as information retrieval, question answering, and recommendation system. Since most MKGs are far from complete, extensive knowledge graph completion studies have been proposed focusing on the multimodal entity, relation extraction and link prediction. However, different tasks and modalities require changes to the model architecture, and not all images/objects are relevant to text input, which hinders the applicability to diverse real-world scenarios. In this paper, we propose a hybrid transformer with multi-level fusion to address those issues. Specifically, we leverage a hybrid transformer architecture with unified input-output for diverse multimodal knowledge graph completion tasks. Moreover, we propose multi-level fusion, which integrates visual and text representation via coarse-grained prefix-guided interaction and fine-grained correlation-aware fusion modules. We conduct extensive experiments to validate that our MKGformer can obtain SOTA performance on four datasets of multimodal link prediction, multimodal RE, and multimodal NER[1].

CCS CONCEPTS

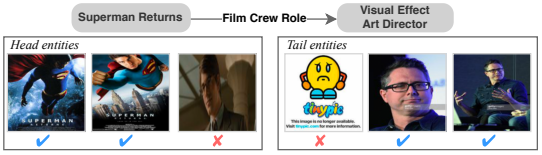
• **Information systems** → **Information extraction**; **Multimedia content creation**.

KEYWORDS

knowledge graph completion; multimodal; relation extraction; named entity recognition

ACM Reference Format:
Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3531992

1 INTRODUCTION



(a) Examples of Multimodal Link Prediction

(b) Examples of Multimodal Relation Extraction

Figure 1: Illustration of examples of multimodal knowledge graph facts. Imaged/objects with ✔ around the same entity have relevant visual features. In contrast, the other images/objects with ✗ are irrelevant with the corresponding entity.

Knowledge graphs (KGs) can provide back-end support for a variety of knowledge-intensive tasks in real-world applications, such as recommender systems [17, 57], information retrieval [10, 50]

*Corresponding author.
[1]Code is available in https://github.com/zjunlp/MKGformer.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531992

and time series forecasting [8]. Since KGs usually contain visual information, Multimodal Knowledge Graphs (MKGs) recently have attracted extensive attention in the community of multimedia, natural language processing and knowledge graph [24, 38]. However, most MKGs are far from complete due to the emerging entities and corresponding relations. Therefore, multimodal knowledge graph completion (MKGC), which aims to extract knowledge from the text and image to complete the missing facts in the KGs, has been proposed [5, 23, 56]. Concretely, visual data (images) can be regarded as complementary information used for MKGC tasks, such as multimodal link prediction, multimodal named entity recognition (MNER) and multimodal relation extraction (MRE).

For example, as shown in Figure 1, for the multimodal link prediction task, each entity possesses many associated images, which can enhance the entity representation for missing triple prediction; while for MNER and MRE, each short sentence contains a corresponding image to complement textual contexts for entity and relation extraction. Benefit from the development of multimodal representation learning [14], it is intuitive to fuse the heterogeneous features of KG entities and the visual information with similar semantics in unified embeddings.

To this end, Xie et al. [48] propose to integrate image features into the typical KG representation learning model for multimodal link prediction. Besides, Sergieh et al. [33] and Wang et al. [47] jointly encode and fuse the visual and structural knowledge for multimodal link prediction through simple concatenation and auto-encoder, respectively. On the other hand, Zheng et al. [59] present an efficient modality alignment strategy based on scene graph for the MRE task. Zhang et al. [55] fuse regional image features and textual features with extra co-attention layers for the MNER task.

Although previous studies for multimodal KGC have shown promising improvements compared with the unimodal methods, those approaches still suffer from several evident limitations as follows: (1) **Architecture universality**. Different MKGC tasks and modality representation demand changes to the model architecture. Specifically, different subtasks require task-specific, separately-parameterized fusion module on top of diverse encoder architectures (e.g., ResNet [16], Fast-RCNN [32] for visual encoder, and BERT [9], word2vec [29] for textual encoder). Therefore, a unified model should be derived to expand the application of the diverse subtasks of multimodal KGC more effectively and conveniently. (2) **Modality contradiction**. Most existing multimodal KGC models largely ignore the noise of incorporating irrelevant visual information, which may result in modality contradiction. To be specific, in most multimodal KG, each entity possesses many associated images; however, parts of images may be irrelevant to entities, and some images even contain a lot of background noise which may mislead entity representation. For example in Figure 1(a), each entity has many associated images, but the third image of the head entity has little relevance to the semantic meaning of "Superman Returns" for multimodal link prediction. Meanwhile, current SOTA methods for MNER and MRE tasks usually utilize valid visual objects by selecting top salient objects with the higher object classification scores, which may also introduce noise from irrelevant or redundant objects, such as the "Shirt0" and "Shirt1" objects in the Figure 1(b). In practice, irrelevant images/objects may directly exert adverse effects on multimodal KGC.

To overcome the above barriers, we propose **MKGformer**, a hybrid transformer for unified multimodal KGC, which implements the modeling of the multimodal features of the entity cross the **last few layers** of visual transformer and the textual transformer with **multi-level fusion**, namely **M-Encoder**. Previous works [7, 22] indicate that the pre-trained models (PLMs) can activate knowledge related to the input at the self-attention layer and Feed-Forward Network (FFN) layer in Transformer Encoder. Inspired by this, we consider the visual information as supplementary knowledge and propose multi-level fusion at the transformer architecture.

Specifically, we first present a coarse-grained **prefix-guided interaction module** at the self-attention part of M-Encoder to pre-reduce modal heterogeneity for the next step. Second, the **correlation-aware fusion module** is proposed in the FFN part of M-Encoder to obtain the fine-grained image-text representations which can alleviate the error sensitivity of irrelevant images/objects. In particular, apart from multimodal link prediction, MKGformer can be more generally applied to MRE and MNER tasks with a simple modification of task-specific head as shown in Figure 2(a). In a nutshell, the contributions of this paper can be summarized:

- To the best of our knowledge, our work is the first to propose a hybrid transformer framework that can be applied to multiple multimodal KGC tasks. Intuitively, leveraging a unified transformer framework with similar arithmetic units to encode text descriptions and images inside transformers naturally reduces the heterogeneity to model better multimodel entity representation.
- We propose multi-level fusion with coarse-grained prefix-guided interaction module and fine-grained correlation-aware fusion module in blocks of transformers to pre-reduce the modal heterogeneity and alleviate noise of irrelevant visual elements, respectively, which are empirically necessary for diverse MKGC tasks.
- We perform comprehensive experiments and extensive analysis on three tasks involving multimodal link prediction, MRE and MNER. Experimental results illustrate that our model can effectively and robustly model the multimodal representations of descriptive text and images and substantially outperform the current state-of-the-art (SOTA) models in standard supervised and low-resource settings.

2 RELATED WORKS

2.1 Multimodal Knowledge Graph Completion

Multimodal KGC has been widely studied in recent years, which leverages the associated images to represent relational knowledge better. Previous studies mainly focus on the following three tasks:

2.1.1 Multimodal Link Prediction. Existing methods for multimodal link prediction focus on encoding image features in KG embeddings. Xie et al. [48] extend TransE [3] to obtain visual representations that correspond to the KG entities and structural information of the KG separately. Sergieh et al. [33] and Wang et al. [47] further propose several fusion strategy to encoder the visual and structural features into unified embedding space. Recently, Wang et al. [46] study the noise from irrelevant images corresponding to entities

and designs a forget gate with an MRP metric to select valuable images for multimodal KGC.

2.1.2 Multimodal Relation Extraction. Recently, Zheng et al. [60] present a multimodal RE dataset with baseline models. The experimental results illustrate that utilizing multimodal information improves RE performance in social media texts. Zheng et al. [59] further revise the multimodal RE dataset and presents an efficient alignment strategy with scene graphs for textual and visual representations. Wan et al. [45] also present four multimodal datasets to handle the lack of multimodal social relation resources and propose a few-shot learning based approach to extracting social relations from both texts and face images.

2.1.3 Multimodal Named Entity Recognition. Zhang et al. [58], Lu et al. [25], Moon et al. [31] and Arshad et al. [1] propose to encode the textual information with RNN and model the whole image representation through CNN in the early stages. Recently, Yu et al. [53], Zhang et al. [55] propose to leverage regional image features to represent objects to exploit fine-grained semantic correspondences based on transformer and visual backbones since informative object features are more important than the whole images for MNER tasks. Sun et al. [37] propose a text-image relation propagation-based multimodal BERT, namely RpBERT, to reduce the interference from whole images. However, RpBERT only focuses on the irrelevance of the whole image but ignores the noise brought by irrelevant objects.

In conclusion, MKGC can handle the problem of extending a KG with missing triples, thus, have received significant attention. However, different tasks and modalities demand changes to the model architecture, hindering the applicability of diverse real-world scenarios. Therefore, we argue that a unified model should be derived to expand the application of the diverse tasks of multimodal KGC more effectively and conveniently.

2.2 Pre-trained Multimodal Representation

The pre-trained multimodal visual-language models have recently demonstrated great superiority in many multimodal tasks (e.g., image-text retrieval and visual question answering). The existing visual-linguistic pre-trained models can be summarized as two aspects: 1) **Architecture**. The single-stream structures include VL-BERT [36], VisualBERT [21], Unicoder-VL [20], and UNITER [6], where the image and textual embeddings are combined into a sequence and fed into transformer to obtain contextual representations. While two-stream structures separate visual and language processing into two streams with interacting through cross-modality or co-attentional transformer layers, which includes LXMERT [40] and ViLBERT [26]. 2) **Pretraining tasks**. The pretraining tasks of multimodal models usually involve masked language modeling (MLM), masked region classification (MRC), and image-text matching (ITM). However, the ITM task is based on the hypothesis that the text-image pairs in the caption datasets are highly related; however, there are much noise brought by irrelevant images/objects, thus not being completely satisfied with the ITM task. Further, most of the above models are pre-trained on the datasets of image caption, such as the Conceptual Captions [34] or COCO caption dataset [4] or visual question answering datasets. Thus, the target optimization objects of the above pre-trained multimodal models are less relevant to multimodal KGC tasks.

Therefore, directly applying these pre-trained multimodal methods to the multimodal KGC may not produce a good performance since multimodal KGC mainly focuses on leveraging visual information to enhance the text rather than on relying on information of both sides. Unlike previous methods that focus on learning pre-trained multimodal representation, we regard the image as supplementary information for knowledge graph completion and propose a hybrid transformer with multi-level fusion.

3 OUR APPROACH

In this section, we present the overall framework of MKGformer, which is a general framework that can be applied to widespread multimodal KGC tasks. To facilitate understanding, we introduce its detailed implementation, including the unified multimodal KGC framework in Section 3.1, the hybrid transformer architecture in Section 3.2 and the detailed introduction of M-Encoder in Section 3.3.

3.1 Unified Multimodal KGC Framework

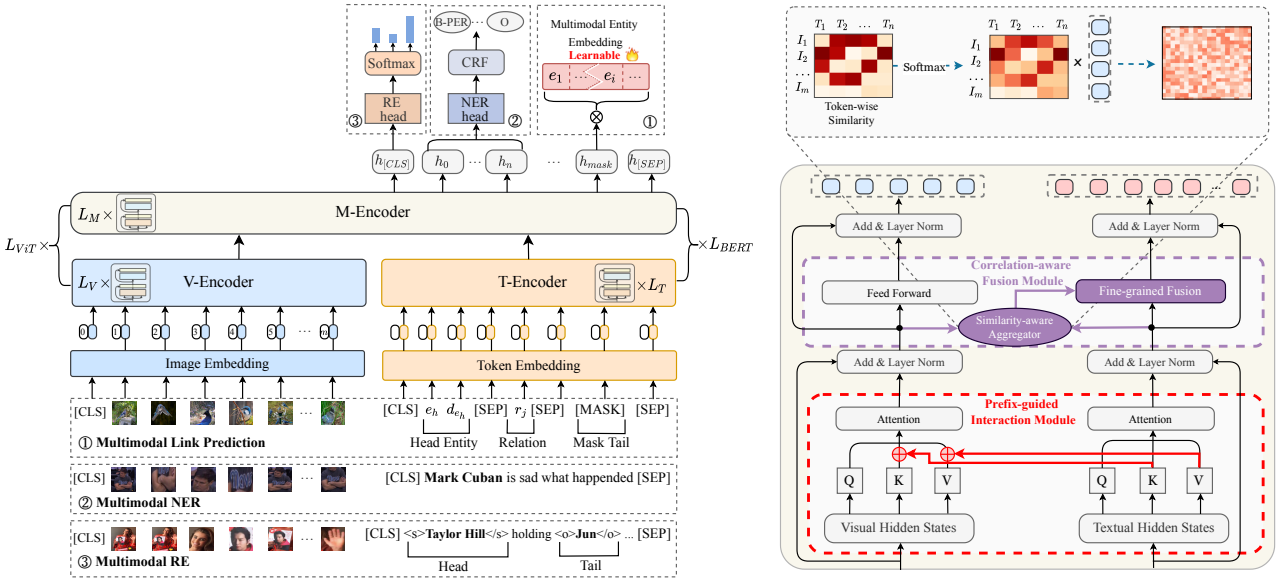
As shown in Figure 2(a), the unified multimodal KGC framework mainly includes hybrid transformer architecture and task-specific paradigm. Specifically, we adopt ViT and BERT as visual transformer and textual transformer models, respectively and conduct the modeling of the multimodel representations of the entity across the last L_M layers of transformers. We introduce its detailed implementation of task-specific paradigm in the following parts.

3.1.1 Applying to Multimodal Link Prediction. Multimodal Link Prediction is the most popular task for multimodal KGC, which focuses on predicting the tail entity given the head entity and the query relation, denoted by $(e_h, r, ?)$. And the answer is supposed to be always within the KG. In terms of the images I_h that related to the entity e_h, we propose to model the distribution over the tail entity e_t as $p(e_t|(e_h, r, I_h))$. As shown in Figure 2(a), to fully leverage the advantage of pre-trained models, we design the specific procedure for link prediction similar to masked language modeling of pre-trained language models (PLMs). We take first step to model the image-text incorporated entity representations and then predict the missing entity $(e_h, r, ?)$ over the multimodal entity representations.

Image-text Incorporated Entity Modeling. Unlike previous work simply concatenate or fuse based on particular visual and textual features of entities, we fully leverage the "masked language modeling" (MLM) ability of pre-trained transformers to model image-text incorporated multimodal representations of the entities in the knowledge graph. To be more specific, given an entity description $d_{e_i} = (w_1, ..., w_n)$ and its corresponding multiple images $I_{e_i} = \{I^1, I^2, ..., I^o\}$, we feed the patched images of the entity e_i into the visual side of hybrid transformer architecture and convert the textual side input sequence of hybrid transformer architecture as:

$$T_{e_i} = \texttt{[CLS]}d_{e_i}\ \texttt{is the description of [MASK][SEP].} \quad (1)$$

We extend the word embedding layer of BERT to treat each token embedding as corresponding multimodal representation E_{e_i} of i-th entity e_i. Then we train the MKGformer to predict the `[MASK]` over



(a) Unified Multimodal KGC Framework.

(b) Detailed M-Encoder.

Figure 2: Illustration of MKGformer for (a) Unified Multimodal KGC Framework and (b) Detailed M-Encoder.

the multimodal entity embedding E_{e_i} with cross entropy loss for classification:

$$\mathcal{L}_{link} = -log(p(e_i|(T_{e_i})), \tag{2}$$

Notably, we freeze the whole model except the newly added parameters of multimodal entity embedding. We argue that the modified input can guide MKGformer to incorporate the textual and visual information into multimodal entity embeddings attentively.

Missing Entity Prediction. Given a $triple = (e_h, r, e_t) \in \mathcal{G}$, KGC models predict the e_h or e_t in the head or tail batch. Similarly, we treat the link prediction as the MLM task, which uses entity e_h, entity description d_{e_h}, relation r and the entity images I_{e_h} to predict the masked tail entity over the multimodal entity embeddings described above. Specifically, we also process the multiple patched images of the entity e_h into the visual side of hybrid transformer architecture and convert this triple $(e_h, r, ?)$ to the input sequence of on the text side as follows:

$$T_{(e_h,r,?)} = \texttt{[CLS]}e_h\ d_{e_h}\texttt{[SEP]}r_j\texttt{[SEP][MASK][SEP]}. \tag{3}$$

Finally, we train the MKGformer to predict the `[MASK]` over the multimodal entity embedding E_{e_i} via binary cross-entropy loss for multilabel classification with the consideration that the prediction of e_t is not unique in link prediction.

3.1.2 Applying to MRE. Relation extraction aims at linking relation mentions from text to a canonical relation type in a knowledge graph. Given the text T and the corresponding image I, we aim to predict the relation between an entity pair (e_h, e_t) and outputs the distribution over relation types as $p(r|(T, I, e_h, e_t))$. Specifically, we take the representation of the special token `[CLS]` from the final output embedding of hybrid transformer architecture to compute the probability distribution over the class set $\mathcal{Y}$ with the softmax function $p(r|(I, T, e_h, e_t)) = \texttt{Softmax}(\mathbf{W}\boldsymbol{h}^{M_t}_{L_M\,\texttt{[CLS]}})$. $\boldsymbol{h}^{M_t}_{L_M} \in \mathbb{R}^{n\times d_T}$ denotes the final sequence representation of the L_M-th layer from textual side of M-Encoder in hybrid transformer architecture. The parameters of the model and $\mathbf{W}$ are fine-tuned by minimizing the cross-entropy loss over $p(r|(I, T, e_h, e_t))$ on the entire train sets.

3.1.3 Applying to MNER. MNER is the task of extracting named entities from text sequences and corresponding images. Given a token sequence $T = \{w_1, \ldots, w_n\}$ and its corresponding image I, we focus on modeling the distribution over sequence tags as $p(y|(T, I))$, where y is the label sequence of tags $y = \{y_1, \ldots, y_n\}$. We assign the procedure of MKGformer with CRF [18] function similar to previous multimodal NER tasks for the fair comparison. For a sequence of tags $y = \{y_1, \ldots, y_n\}$, we calculate the probability of the label sequence y over the pre-defined label set Y with the BIO tagging schema as described in [18].

3.2 Hybrid Transformer Architecture

The **hybrid transformer architecture** of MKGformer mainly includes three stacked modules: (1) the underlying textual encoder is designed to capture basic syntactic and lexical information from the input tokens, namely (**T-Encoder**), (2) the underlying visual encoder (**V-Encoder**), which is responsible for capturing basic visual features from the input patched images, and (3) the upper multimodal encoder (**M-Encoder**) is adopted to model image-text incorporated entity representations inside the underlying visual transformer and textual transformer. Besides, we denote the number of `V-Encoder` layers as L_V, the number of `T-Encoder` layers as L_T, and the number of `M-Encoder` layers as L_M, where $L_{ViT} = L_V + L_M$ and $L_{BERT} = L_T + L_M$.

Recap of the Transformer Architecture. Transformer [44] is now the workhorse architecture behind most SOTA models in CV and NLP, which is composed of L stacked blocks. While each block mainly includes two types of sub-layers: multi-head self-attention (MHA) and a fully connected feed-forward network (FFN). Layer Normalization (LN) and residual connection are also used in each layer. Given the input sequence vectors $\boldsymbol{x} \in \mathbb{R}^{n\times d}$, the conventional attention function maps $\boldsymbol{x}$ to queries $\boldsymbol{Q} \in \mathbb{R}^{n\times d}$ and key-value pairs $\boldsymbol{K} \in \mathbb{R}^{n\times d}, \boldsymbol{V} \in \mathbb{R}^{n\times d}$:

$$\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V}, \tag{4}$$

where n denotes the length of sequence. MHA performs the attention function in parallel over N_h heads, where each head is separately parameterized by $\boldsymbol{W}_q^{(i)}, \boldsymbol{W}_k^{(i)}, \boldsymbol{W}_v^{(i)} \in \mathbb{R}^{d\times d_h}$ to project inputs to queries, keys, and values. The role of MHA is to compute the weighted hidden states for each head, and then concatenates them as:

$$\begin{aligned} \text{MHA}(\boldsymbol{x}) &= [\text{head}_1; \cdots; \text{head}_\text{h}]\boldsymbol{W}_o, \\ \boldsymbol{Q}^{(i)}, \boldsymbol{K}^{(i)}, \boldsymbol{V}^{(i)} &= \boldsymbol{x}\boldsymbol{W}_q^{(i)}, \boldsymbol{x}\boldsymbol{W}_k^{(i)}, \boldsymbol{x}\boldsymbol{W}_v^{(i)} \\ \text{head}_\text{i} &= \text{Attn}(\boldsymbol{Q}^{(i)}, \boldsymbol{K}^{(i)}, \boldsymbol{V}^{(i)}), \end{aligned} \tag{5}$$

where $\boldsymbol{W}_o \in \mathbf{R}^{d\times d}$ and d denotes the dimension of hidden embeddings. $d_h = d/N_h$ is typically set in MHA. FFN is another vital component in transformer, typically consisting of two layers of linear transformations with a ReLU activation function as follows:

$$\text{FFN(x)} = \text{ReLU}(\boldsymbol{x}\boldsymbol{W}_1 + \mathbf{b}_1)\boldsymbol{W}_2 + \mathbf{b}_2, \tag{6}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d\times d_m}, \boldsymbol{W}_2 \in \mathbb{R}^{d_m\times d}$.

V-Encoder. We adopt the first L_V layers of ViT [11] pre-trained on ImageNet-1k from [42] as the visual encoder to extract image features. Given o images I_{e_i} of the entity e_i[2], we rescale each image to unified $H\times W$ pixels, and the i-th input image $I_i \in \mathbb{R}^{C\times H\times W}$ ($1 \le i < o$) is first reshaped into $u = HW/P^2$ flattened 2D patches, then pooled and projected as $X_{pc}^i \in \mathbb{R}^{u\times d_V}$, where the resolution of the input image is $H \times W$, C is the number of channels and d_V denotes the dimension of hidden states of ViT. We concatenate the patched embeddings of o images to get the visual sequence patch embeddings $X_{pc} \in \mathbb{R}^{m\times d_V}$, where $m = (u \times o)$.

$$\begin{aligned} X_0^V &= X_{pc} + V_{pos} \\ \bar{X}_l^V &= \textbf{MHA}(\textbf{LN}(X_0^V)) + X_{l-1}^V, l = 1...L_V \\ X_l^V &= \textbf{FFN}(\textbf{LN}(\bar{X}_l^V)) + \bar{X}_l^V, l = 1...L_V, \end{aligned} \tag{7}$$

where $V_{pos} \in \mathbb{R}^{m\times d_V}$ represents the corresponding position embedding layer, embedding, X_l^V is the hidden states of the l layer of visual encoder.

T-Encoder. We leverage the first L_T layers of BERT [9] as the text encoder, which also consists of L_T layers of MHA and FFN blocks similar to the visual encoder except that LN comes after MHA and FFN. To be specific, a token sequence $\{w_1, \ldots, w_n\}$ is embedded to $X_{wd} \in \mathbb{R}^{n\times d_T}$ with a word embedding matrix, and the textual representation is calculated as follows:

$$\begin{aligned} X_0^T &= X_{wd} + T_{pos} \\ \bar{X}_l^T &= \textbf{LN}(\textbf{MHA}(X_{l-1}^T)) + X_{l-1}^T, l = 1...L_T \\ X_l^T &= \textbf{LN}(\textbf{FFN}(\bar{X}_l^T)) + \bar{X}_l^T, l = 1...L_T, \end{aligned} \tag{8}$$

where $T_{pos} \in \mathbb{R}^{(n)\times d_T}$ denotes position embedding, X_l^T is the hidden states of the l layer for the output textual sequence.

M-Encoder. Multimodal KGC mainly faces the issues of heterogeneity and irrelevance between different modalities. Different from previous works leveraging extra co-attention layers to integrate modality information, we propose to model the multimodal features of the entity cross the last L_M layers of ViT and BERT with multi-level fusion, namely M-Encoder. To be specific, we present a **P**refix-**G**uided **I**nteraction module (**PGI**) at the self-attention block to pre-reduce the modality heterogeneity. We also propose a **C**orrelation-**A**ware **F**usion module (**CAF**) in the FFN layer to reduce the impact of noise caused by irrelevant image elements. Here, we omit the calculation of LN and residual connection for simplicity.

$$\begin{aligned} h_0^{M_t} &= X_{L_T}^T \\ h_0^{M_v} &= X_{L_V}^V \\ \tilde{h}_l^{M_t}, \tilde{h}_l^{M_v} &= \textbf{PGI}(h_{l-1}^{M_t}, h_{l-1}^{M_v}), l = 1...L_M \\ h_l^{M_t}, h_l^{M_v} &= \textbf{CAF}(\tilde{h}_{l-1}^{M_t}, \tilde{h}_{l-1}^{M_v}), l = 1...L_M. \end{aligned} \tag{9}$$

3.3 Insights of M-Encoder

3.3.1 Prefix-guided Interaction Module. Inspired by the success of textual prefix tuning [22] and corresponding analysis [15], we propose a prefix-guided interaction mechanism to pre-reduce the modality heterogeneity through the calculation of multi-head attention at every layer, which is performed on the hybrid keys and values. In particular, we redefine the computation of visual $\text{head}_i^{M_v}$ and textual $\text{head}_i^{M_t}$ in Eq. 5 as:

$$\begin{aligned} \text{head}^{M_t} &= \text{Attn}(\boldsymbol{x}^t\boldsymbol{W}_q^t, \boldsymbol{x}^t\boldsymbol{W}_k^t, \boldsymbol{x}^t\boldsymbol{W}_v^t), \\ \text{head}^{M_v} &= \text{Attn}(\boldsymbol{x}^v\boldsymbol{W}_q^v, [\boldsymbol{x}^v\boldsymbol{W}_k^v, \boldsymbol{x}^t\boldsymbol{W}_k^t], [\boldsymbol{x}^v\boldsymbol{W}_v^v, \boldsymbol{x}^t\boldsymbol{W}_v^t]), \end{aligned} \tag{10}$$

We also derive the variant formula of Eq. 10 and provide another perspective of prefix-guided interpolated attention:[3]

$$\begin{aligned} \text{head}^{M_v} &= \text{Attn}(\boldsymbol{x}^v\boldsymbol{W}_q^v, [\boldsymbol{x}^v\boldsymbol{W}_k^v, \boldsymbol{x}^t\boldsymbol{W}_k^t], [\boldsymbol{x}^v\boldsymbol{W}_v^v, \boldsymbol{x}^t\boldsymbol{W}_v^t]), \\ &= \text{softmax}\big(\boldsymbol{Q}_v[\boldsymbol{K}_v; \boldsymbol{K}_t]^\top\big)\begin{bmatrix}\boldsymbol{V}_v \\ \boldsymbol{V}_t\end{bmatrix} \\ &= (1-\lambda(\boldsymbol{x}^v))\text{softmax}(\boldsymbol{Q}_v\boldsymbol{K}_v{}^\top)\boldsymbol{V}_v + \lambda(\boldsymbol{x}^v)\text{softmax}(\boldsymbol{Q}_v\boldsymbol{K}_t{}^\top)\boldsymbol{V}_t \\ &= (1-\lambda(\boldsymbol{x}^v))\underbrace{\text{Attn}(\boldsymbol{Q}_v, \boldsymbol{K}_v, \boldsymbol{V}_v)}_{\text{standard attention}} + \lambda(\boldsymbol{x}^v)\underbrace{\text{Attn}(\boldsymbol{Q}_v, \boldsymbol{K}_t, \boldsymbol{V}_t)}_{\text{Cross-modal Interaction}}, \end{aligned} \tag{11}$$

$$\lambda(\boldsymbol{x}^v) = \frac{\sum_i \exp(\boldsymbol{Q}_v\boldsymbol{K}_t^\top)_i}{\sum_i \exp(\boldsymbol{Q}_v\boldsymbol{K}_t^\top)_i + \sum_j \exp(\boldsymbol{Q}_v\boldsymbol{K}_v^\top)_j}. \tag{12}$$

where $\lambda(\boldsymbol{x}^v)$ denotes the scalar for the sum of normalized attention weights on the textual key and value vectors.

[2]Here, we take the multimodal link prediction as example. As for multimodal NER and RE, we choose the top o salient objects according to the text.

[3]Without loss of generalization, we ignore the softmax scaling factor $\sqrt{d}$ for ease of representation.

Remark 1. *As shown in Eq. 11, the first term $Attn(\boldsymbol{Q}_v, \boldsymbol{K}_v, \boldsymbol{V}_v)$ is the standard attention in the visual side, whereas the second term represent the cross-modal interaction. Monolithic in the sense that the prefix-guided interaction mechanism down-weights the original visual attention probabilities by a scalar factor (i.e., $1-\lambda$) and re-distributes the remaining attention probability mass λ to attend to textual attention, which likes the linear interpolation. By applying this to the attention flow calculation over hidden visual states and hidden textual states, MKGformer learns coarse-grained modality fusion to pre-reduce the modality heterogeneity.*

3.3.2 Correlation-aware Fusion Module. To alleviate the adverse effects of noise, we apply a correlation-aware fusion module to conduct the token-wise cross-modal interaction (e.g., word-patch alignment) between the two modalities. Specifically, we denote m and n as the sequence length of the visual vectors $\boldsymbol{x}^v \in \mathbb{R}^{m\times d}$ and textual vectors $\boldsymbol{x}^t \in \mathbb{R}^{n\times d}$ respectively, which are the corresponding output features of the prefix-guided interaction module. For the textual tokens, we compute its similarity matrix of all visual tokens as follows:

$$\boldsymbol{S} = \boldsymbol{x}^t (\boldsymbol{x}^v)^\top. \tag{13}$$

We then conduct softmax function over similarity matrix $\boldsymbol{S}$ of i-th textual token and use the average token-wise aggregator over visual tokens in the image as follows:

$$\mathrm{Agg}_i(\boldsymbol{x}^v) = \mathrm{softmax}(\boldsymbol{S}_i)\boldsymbol{x}^v, (1 \le i < n) \tag{14}$$

$$\mathrm{Agg}(\boldsymbol{x}^v) = [\mathrm{Agg}_1(\boldsymbol{x}^v); ..., \mathrm{Agg}_n(\boldsymbol{x}^v)] \tag{15}$$

where Agg_i denotes the similarity-aware aggregated visual representation for i-th textual token. Inspired by the finding [13] that the FFN layer learns task-specific textual patterns, we propose to incorporate similarity-aware aggregated visual hidden states into textual hidden states in FFN layers and modify the calculation of the FFN process as:

$$\mathrm{FFN}(\boldsymbol{x}^t) = \mathrm{ReLU}(\boldsymbol{x}^t\boldsymbol{W}_1 + \mathbf{b}_1 + \mathrm{Agg}(\boldsymbol{x}^v)\boldsymbol{W}_3)\boldsymbol{W}_2 + \mathbf{b}_2, \tag{16}$$

where $\boldsymbol{W}_3 \in \mathbb{R}^{d\times d_m}$ represent the new added parameters for aggregated visual hidden states.

Remark 2. *Note that the token-wise similarity in Equation 13, 14 and 15 indicates that we would like to obtain the closest image patch for each textual token. By inserting the similarity-aware aggregated visual representation into the FFN calculation in the textual side, our MKGformer learns fine-grained alignment between image patches and textual tokens, which makes our modal more robust to the noise of irrelevant images of entities in KG.*

4 EXPERIMENTS

We next introduce the experimental settings of MKGformer in three tasks: multimodal link prediction, multimodal RE, and multimodal NER. Following results show that MKGformer can outperforms the other baselines in both standard supervised and few-shot settings.

4.1 Experimental Setup

4.1.1 Datasets. We adopt two publicly available datasets for multimodal link prediction, including: 1) **WN18-IMG:** WN18 [3] is a knowledge graph originally extracted from WordNet [30]. While WN18-IMG is an extended dataset of WN18 [3] with 10 images for each entity. **FB15K-237-IMG**: FB15K-237-IMG[4] [3] has 10 images for each entity and is a subset of the large-scale knowledge graph Freebase [2], which is a popular dataset in knowledge graph completion. Detailed statistics are shown in Table 3. For multimodal RE, we evaluate on **MNRE** [59], a manually-labeled dataset for multimodal neural relation extraction, where the texts and image posts are crawled from Twitter. For multimodal NER, we conduct experiments on public Twitter dataset **Twitter-2017** [25], which mainly include multimodal user posts published on Twitter during 2016-2017.

4.1.2 Compared Baselines. We compare our MKGformer with several baseline models for a comprehensive comparison to demonstrate the superiority of our MKGformer. Firstly, we choose the conventional text-based models for comparison to demonstrate the improvement brought by the visual information. Secondly, we also compare our MKGformer with *VisualBERT* [21] and *ViLBERT* [27], which are pre-trained visual-language model with single-stream structure and two-stream structure respectively. Besides, we further consider another group of previous SOTA multimodal approaches for multimodal knowledge graph completion models as follows:

Multimodal link prediction: 1) *IKRL* [48], which extends TransE to learn visual representations of entities and structural information of KG separately; 2) *TransAE* [47], which combines multimodal autoencoder with TransE to encode the visual and textural knowledge into the unified representation, and the hidden layer of the autoencoder is used as the representation of entities in the TransE model. 3) *RSME* [46], which designs a forget gate with an MRP metric to select valuable images for the multimodal KG embeddings learning. **Multimodal RE:** 1) *BERT+SG* is proposed in [59] for MRE, which concatenates the textual representation from BERT with visual features generated by scene graph (SG) tool [41]. 2) *MEGA* [59] designs the dual graph alignment of the correlation between entities and objects, which is the newest SOTA for MRE. **Multimodal NER:** 1) *AdapCoAtt-BERT-CRF* [58], which designs an adaptive co-attention network to induce word-aware visual representations for each word; 2) *UMT* [53], which extends Transformer to multi-modal version and incorporates the auxiliary entity span detection module; 3) *UMGF* [55], which proposes a unified multimodal graph fusion approach for MNER and achieves the newest SOTA for MNER.

4.1.3 Settings. Notably, we assign the layer of M-Encoder as $L_M = 3$ and conduct experiments with BERT_base and ViT-B/32 [9] for all experiments. We further conduct extensive experiments in the low-resource setting by running experiments over five randomly sampled $\mathcal{D}_{train}$ for each task and report the average results on test set. For the few-shot multimodal link prediction and MRE, we follow the settings of [12]. We adopt whole images corresponding to entities for multimodal link prediction tasks. As for the MRE and MNER tasks, we follow [55] to adopt the visual grounding toolkit [51] for extracting local visual objects with top m salience.

[4]Since the dataset FB15k has the inverse relation. Therefore, we adopt corresponding sub-datasets FB15k-237 to mitigate the problem of the reversible relation between. The multimodal datasets of link prediction can be acquired from https://github.com/wangmengsd/RSME, which is the public code of RSME.

Table 1: Results of the link prediction on FB15K-237-IMG and WN18-IMG. Note that the universal pre-trained vision-language model cannot be directly applied to the multimodal link prediction; thus, we follow KG-BERT to leverage the pre-trained VL model for multimodal link prediction. The best scores are in bold.

Model	FB15k-237-IMG				WN18-IMG			
	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓	Hits@1 ↑	Hits@3 ↑	Hits@10 ↑	MR ↓
Unimodal approach								
TransE [3]	0.198	0.376	0.441	323	0.040	0.745	0.923	357
DistMult [49]	0.199	0.301	0.446	512	0.335	0.876	0.940	655
ComplEx [43]	0.194	0.297	0.450	546	0.936	0.945	0.947	-
RotatE [39]	0.241	0.375	0.533	177	0.942	0.950	0.957	254
KG-BERT [52]	-	-	0.420	**153**	0.117	0.689	0.926	58
Multimodal approach								
VisualBERT_base [21]	0.217	0.324	0.439	592	0.179	0.437	0.654	122
ViLBERT_base [27]	0.233	0.335	0.457	483	0.223	0.552	0.761	131
IKRL(UNION) [48]	0.194	0.284	0.458	298	0.127	0.796	0.928	596
TransAE [47]	0.199	0.317	0.463	431	0.323	0.835	0.934	352
RSME (ViT-B/32+Forget) [46]	0.242	0.344	0.467	417	0.943	0.951	0.957	223
MKGformer	**0.407**	**0.483**	**0.573**	225	**0.944**	**0.961**	**0.972**	**28**

Table 2: Performance of low-resource setting (8-shot) FB15K-237-IMG for multimodal link prediction.

Model	FB15K-237-IMG	
	MR ↓	Hit@10 ↑
TransE [3]	5847	0.0925
DistMult [49]	5791	0.0059
ComplEx [43]	6451	0.0046
RotatE [39]	7365	0.0066
KG-BERT [52]	2023	0.0451
VisualBERT_base [21]	5983	0.0772
ViLBERT_base [27]	5754	0.0831
IKRL [48]	4913	0.0973
RSME(ViT-B/16+Forget) [46]	2654	0.1183
MKGformer	**2718**	**0.1540**

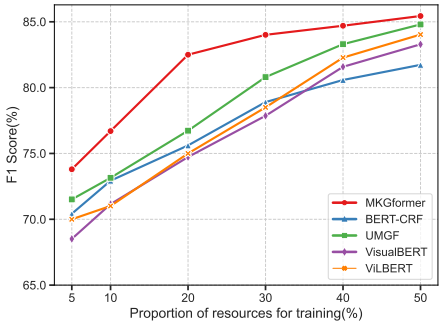


Figure 3: Performance of low-resource setting on Twitter-2017 dataset for multimodal NER task.

Table 3: Dataset statistics for Multimodal Link Prediction.

Dataset	#Rel.	#Ent.	#Train	#Dev	#Test
FB15k-237-IMG	237	14,541	272,115	17,535	20,466
WN18-IMG	18	40,943	141,442	5,000	5,000

4.2 Overall Performance

4.2.1 Multimodal link prediction. The experimental results in Table 1 show that incorporating the visual features is generally helpful for link prediction tasks, indicating the superiority of our MKGformer. Compared to the multimodal SOTA method RSME, MKGformer has an increase of 16.5% hits@1 scores and 10.6% hits@10 scores on FB15k-237-IMG. We further observe that VisualBERT and ViLBERT perform even worse than SOTA modal RSME. Note that the implementation of MKGformer is also relatively efficient and straightforward compared with previous knowledge graph embedding approaches [3] that iteratively query all entities.

4.2.2 Multimodal RE and NER. From the experimental results shown in Table 4, we can find that our MKGformer is superior to the newest SOTA models UMGF and MEGA, which improves 1.98% F1 scores for the Twitter-2017 dataset and 15.44% F1 scores for the MNRE dataset. We further modify the typical pre-trained vision-language model VisualBERT and ViLBERT with [CLS] classifier for the MRE task and CRF classifier for the MNER task to conduct experiments for comparison. We notice that VisualBERT and ViLBERT perform worse than our methods. We hold that the poor performance of the pre-trained multimodal model may be attributed to the fact that the pre-training datasets and objects have gaps in information extraction tasks. This finding also demonstrate that our MKGformer is more beneficial for multimodal MNER and MRE tasks.

4.3 Low-Resource Evaluation

Previous experiments illustrate that our methods achieve improvements in standard supervised settings. We further report the experimental results in low-resource scenarios compared with several baselines in Figure 3, Table 2 and Table 5.

Table 4: Performance comparison of the different competitive baseline approaches for multimodal RE and NER. "(CRF)" represents CRF is only for MNER dataset Twitter-2017.

Modality	Methods	MNRE			Twitter-2017		
		Precision	Recall	F1	Precision	Recall	F1
Text	CNN-BiLSTM-(CRF) [28]	60.18	46.32	52.35	80.00	78.76	79.37
	HBiLSTM-(CRF) [19]	60.22	47.13	52.87	82.69	78.16	80.37
	PCNN [54]	62.85	49.69	55.49	83.28	78.30	80.72
	BERT-(CRF)	63.85	55.79	59.55	83.32	83.57	83.44
	MTB [35]	64.46	57.81	60.86	83.88	83.22	83.55
Text+Image	AdapCoAtt-BERT-(CRF) [58]	64.67	57.98	61.14	85.13	83.20	84.10
	UMT [53]	62.83	61.32	62.56	85.28	85.34	85.31
	BERT_base+SG [59]	62.95	62.65	62.80	84.13	83.88	84.00
	VisualBERT_base [21]	56.34	58.28	57.29	84.06	85.39	84.72
	ViLBERT_base [27]	64.50	61.86	63.16	84.62	85.47	85.04
	UMGF [55]	64.38	66.23	65.29	86.54	84.50	85.51
	MEGA [59]	64.51	68.44	66.41	84.03	84.75	84.39
	MKGformer	**82.67**	**81.25**	**81.95**	**86.98**	**88.01**	**87.49**

Table 5: Results of compared models for MRE in the low-resource setting. We report the results of $F1$ score and adopt $K = 1, 5, 10, 20$ (# examples per class).

Dataset	Model	$K = 1$	$K = 5$	$K = 10$	$K = 20$
MNRE	BERT [9]	3.67	6.27	12.65	18.94
	VisualBERT [21]	2.93	4.91	6.10	14.96
	ViLBERT [27]	3.89	7.78	13.38	18.45
	MEGA [59]	9.84	11.71	15.39	20.26
	MKGformer	**12.04** (+2.2)	**18.54** (+6.83)	**21.09** (+5.7)	**40.93** (+20.67)

Table 6: Performance when the layer of M-Encoder varies, we take different levels of the activation states into computation, where 1 indicates the bottom layer and 12 indicates the top layer.

Methods	FB15k-237-IMG	MNRE	Twitter-2017
	Hits@10	F1	F1
(layer 12)	0.557	80.21	85.25
(layer 10-12)	0.573	81.95	87.49
(layer 7-12)	0.576	82.20	87.62
(All layers)	0.575	82.25	87.60

4.3.1 Q1: Does the pre-trained vision-language model successful in the low-resource multimodal KGC? . As shown in Figure 3, Table 2 and Table 5, VisualBERT and ViLBERT yield slightly improvements compared with typical unimodal baselines in low-resource settings while obtain worse results than previous multimodal SOTA methods. It reveals that applying these pre-trained multimodal methods to the multimodal KGC may not always achieve a good performance. This result may be attributed to the fact that the pre-training dataset and objects of the above visual-language models are less relevant to KGC tasks, which may bias the original language modeling capabilities and knowledge distribution of BERT.

4.3.2 Q2: Whether hybrid transformer framework data-efficient? Since pre-trained multimodal models do not show promising advantages in low-resource settings, we further analyze whether the hybrid transformer framework is data-efficient. We hold that leveraging a hybrid transformer framework with similar arithmetic units to fuse entity description and images inside transformers intuitively reduces heterogeneity, playing a more critical role in low-resource settings. Thus, we compare with previous multimodal KGC SOTA models in low-resource settings. The experimental results in Figure 3, Table 2 and Table 5 indicate that the performance of MKGformer still outperforms the other baselines, which further proving that our proposed method can more efficiently leverage multimodal data. This success may be attributed to the prefix-guided fusion module in MKGformer leveraging a form similar to linear interpolation to fuse features in the attention layer, thus, effectively pre-reducing modal heterogeneity.

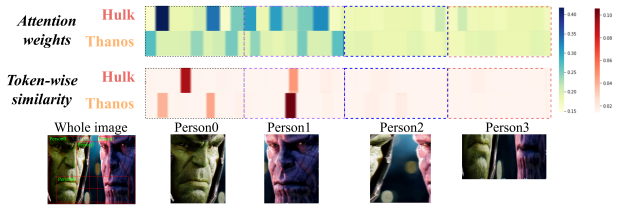
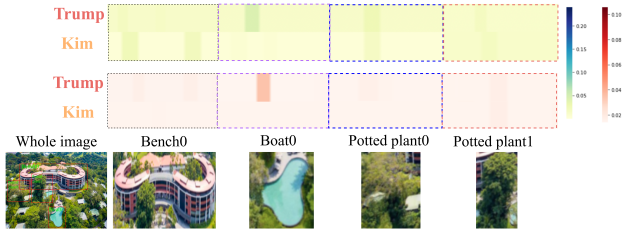
4.4 Sensitivity Analysis of M-Encoder Layers

We assign the M-Encoder with last three layers of ViT and BERT in the previous experiments. However, it is intuitive to investigate whether the performance of MKGformer is sensitive to the layers of the M-Encoder. Thus, we take the M-Encoder in different layers into computation for further analysis. As shown in Table 6, the performance of MKGformer on FB15k-237-IMG, MNRE and Twitter-2017 only achieve improvements of 0.2% Hits@10 scores, 0.3% F1 scores and 0.11% F1 scores by conducting M-Encoder with all 12 layers. Furthermore, the performance of assigning M-Encoder with only one layer also drops slightly compared with the original result (three layers M-Encoder), showing that M-Encoder applied to higher transformer layers can stimulate knowledge and perform modality fusion for downstream tasks more efficiently. It also reveals that our approach is not sensitive to the layers of the M-Encoder.

4.5 Ablation Study

4.5.1 Ablation Setting. The ablation study is further conducted to prove the effects of different modules in MKGformer: *1) (Independent)* indicate that we add extra three layers M-Encoder based on ViT and BERT rather than conduct fusion inside them; *2) w/o PGI* refers to the model without the prefix-guided interaction module; *3) w/o CAF* refers to the model without the whole correlation-aware fusion module. We report detailed experimental results in Figure 4 and observe that ablation models both show a performance decay,

Table 7: The first row shows the split of the relevance of image-text pairs, and the several middle rows indicate representative samples together with their entity-object attention and token-wise similarity in the test set of MNRE datasets, and the bottom five rows in the figure show predicted relation of different approaches on these test samples.

Relevant Image-text Pair			Irrelevant Image-text Pair	
Infinity War Director Confirms Hulk is NOT Afraid of Thanos .			History beckons as Trump - Kim summit kicks off in Singapore.	
				
Gold Relations: per/per/peer			per/per/peer	
BERT:	per / per /couple	✗	per / per /peer	✔
VisualBERT:	per / per /peer	✔	per / org /member_of	✗
MEGA:	per / per /peer	✔	per / loc /place_of_residence	✗
Ours:	per / per /peer	✔	per / per /peer	✔
Ours(w/o CAF):	per / per /peer	✔	per / per /member_of	✗

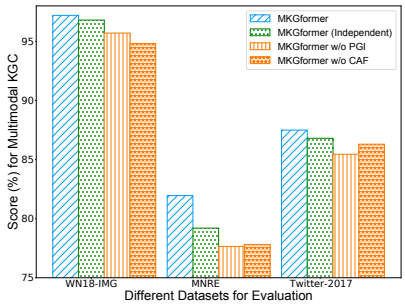


Figure 4: Ablation study results of MKGformer.

which demonstrates the effectiveness of each component of our approach.

4.5.2 Effectiveness of Internal Hybrid Fusion in Transformer. A specific feature of our method is that we conduct modal fusion inside the dual-stream transformer rather than adding a fusion layer outside the transformer like IKRL [48], UMGF [55] and MEGA [59]. To this end, we add extra three layers M-Encoder based on ViT and BERT to evaluate the impact of the internal fusion mechanism. We observe that the performance of *MKGformer (Independent)* drops on three sub-tasks of multimodal KGC, revealing the effectiveness of internal fusion in Transformer.

4.5.3 Importance of multi-level fusion. The highlights of our MKGformer is the multi-level fusion in M-Encoder with coarse-grained prefix-guided interaction module and fine-grained correlation-aware fusion module. We argue that these two parts can mutually reinforce each other: the heterogeneity reduced visual and textual features can help the correlation-aware module better understand fine-grained information. On the contrary, the prefix-guided interaction module in the next layer can reduce the modality heterogeneity more gently based on fine-grained fusion in the last layer. The results shown in Figure 4 demonstrate that multi-level fusion holds the most crucial role in achieving excellent performance. At the same time, the case analysis in Table 7 also reveals the impact of the correlation-aware module for alleviating error sensitivity.

4.6 Case Analysis for Image-text Relevance

To further analyze the robustness of our method for error sensitivity, we conduct a specific case analysis on the multimodal RE task as indicated in Table 7. We notice that VisualBERT, MEGA, and our method can recognize the relation for the relevant image-text pair. Through the visualization of the case, we can further notice: 1) The attention weights in the prefix-guided interaction module reveal that our model can capture the significant attention between relevant entities and objects. 2) The similarity matrix also shows that the entity representation from our model is more similar to the corresponding object patch. Moreover, in the situation that image represents the abstract semantic that is irrelevant to the text, only our method success in prediction due to MKGformer captures the more fine-grained multimodal features. It is worth noting that another two multimodal baselines fail in irrelevant image-text pairs while text-based BERT and ours still predict correctly. These observations reveal that irrelevant visual features may hurt the performance, while our model can learn more robust and fine-grained multimodal representation, which is essential for reducing error sensitivity.

5 CONCLUSION AND FUTURE WORK

In this paper, we present a hybrid Transformer network for multimodal knowledge graph completion, which presents M-Encoder

with multi-level fusion at the last several layers of ViT and BERT to conduct image-text incorporated entity modeling. To the best our knowledge, MKGformer is the first work leveraging unified transformer architecture to conduct various multimodal KGC tasks, involving multimodal link prediction, multimodal relation extraction, and multimodal named entity recognition. Concretely, we propose a prefix-guided interaction module at the self-attention layer to pre-reduce modality heterogeneity and further design a correlation-aware fusion module which realize token-wise fine-grained fusion at the FFN layer to mitigate noise from irrelevant images/objects. Extensive experimental results on four datasets demonstrate the effectiveness and robustness of our MKGformer.

In the future, we plan to 1) apply our approach to more image enhanced natural language processing and information retrieval tasks, such as multimodal event extraction, multimodal sentiment analysis, and multimodal entity retrieval; 2) apply the reverse version of our approach to boost visual representation with text for CV; 3) extend our approach to pre-training of multimodal KGC.

ACKNOWLEDGMENTS

We want to express gratitude to the anonymous reviewers for their hard work and kind comments. This work is funded by NSFC U19B2027/91846204, National Key R&D Program of China (Funding No.SQ2018YFC000004), Zhejiang Provincial Natural Science Foundation of China (No. LGG22F030011), Ningbo Natural Science Foundation (2021J190), and Yongjiang Talent Introduction Programme (2021A-156-G).

REFERENCES

[1] Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. *ArXiv preprint* abs/1904.01356 (2019). https://arxiv.org/abs/1904.01356

[2] Kurt Bollacker, Georg Gottlob, and Sergio Flesca. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *KDD*. 1247–1250.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. 2787–2795.

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR* abs/1504.00325 (2015). arXiv:1504.00325 http://arxiv.org/abs/1504.00325

[5] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. *CoRR* abs/2104.07650 (2021). arXiv:2104.07650 https://arxiv.org/abs/2104.07650

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX (Lecture Notes in Computer Science, Vol. 12375)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 104–120. https://doi.org/10.1007/978-3-030-58577-8_7

[7] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2021. Knowledge Neurons in Pretrained Transformers. *CoRR* abs/2104.08696 (2021). arXiv:2104.08696 https://arxiv.org/abs/2104.08696

[8] Shumin Deng, Ningyu Zhang, Wen Zhang, Jiaoyan Chen, Jeff Z. Pan, and Huajun Chen. 2019. Knowledge-Driven Stock Trend Prediction and Explanation via Temporal Convolutional Network. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 678–685. https://doi.org/10.1145/3308560.3317701

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[10] Laura Dietz, Alexander Kotov, and Edgar Meij. 2018. Utilizing Knowledge Graphs for Text-Centric Information Retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 1387–1390. https://doi.org/10.1145/3209978.3210187

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[12] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295

[13] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 5484–5495. https://aclanthology.org/2021.emnlp-main.446

[14] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access* 7 (2019), 63373–63394. https://doi.org/10.1109/ACCESS.2019.2916887

[15] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a Unified View of Parameter-Efficient Transfer Learning. *CoRR* abs/2110.04366 (2021). arXiv:2110.04366 https://arxiv.org/abs/2110.04366

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[17] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y. Chang. 2018. Improving Sequential Recommendation with Knowledge-Enhanced Memory Networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 505–514. https://doi.org/10.1145/3209978.3210017

[18] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 260–270. https://doi.org/10.18653/v1/N16-1030

[19] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). The Association for Computational Linguistics, 260–270. https://doi.org/10.18653/v1/n16-1030

[20] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 11336–11344. https://aaai.org/ojs/index.php/AAAI/article/view/6795

[21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint* abs/1908.03557 (2019). https://arxiv.org/abs/1908.03557

[22] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 4582–4597. https://doi.org/10.18653/v1/2021.acl-long.353

[23] Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. 2021. Noisy-Labeled NER with Confidence Estimation. In *Proceedings of NAACL*. Association for Computational Linguistics, 3437–3445. https://doi.org/10.18653/v1/2021.naacl-main.269

[24] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings (Lecture Notes in Computer Science, Vol. 11503)*, Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J. G. Gray, Vanessa López, Armin Haller, and Karl Hammar (Eds.). Springer, 459–474. https://doi.org/10.1007/978-3-030-21348-0_30

[25] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1990–1999. https://doi.org/10.18653/v1/P18-1185

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 13–23. https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 13–23. https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html

[28] Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. https://doi.org/10.18653/v1/p16-1101

[29] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[30] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[31] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 852–860. https://doi.org/10.18653/v1/N18-1078

[32] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91–99. https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[33] Hatem Mousselly Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, Malvina Nissim, Jonathan Berant, and Alessandro Lenci (Eds.). Association for Computational Linguistics, 225–234. https://doi.org/10.18653/v1/s18-2027

[34] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 2556–2565. https://doi.org/10.18653/v1/P18-1238

[35] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2895–2905. https://doi.org/10.18653/v1/p19-1279

[36] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SygXPaEYvH

[37] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. RpBERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13860–13868. https://ojs.aaai.org/index.php/AAAI/article/view/17633

[38] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal Knowledge Graphs for Recommender Systems. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1405–1414. https://doi.org/10.1145/3340531.3411947

[39] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*.

[40] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5100–5111. https://doi.org/10.18653/v1/D19-1514

[41] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased Scene Graph Generation From Biased Training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 3713–3722. https://doi.org/10.1109/CVPR42600.2020.00377

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10347–10357. http://proceedings.mlr.press/v139/touvron21a.html

[43] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning*. 2071–2080.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[45] Hai Wan, Manrong Zhang, Jianfeng Du, Ziling Huang, Yufei Yang, and Jeff Z. Pan. 2021. FL-MSRE: A Few-Shot Learning based Approach to Multimodal Social Relation Extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13916–13923. https://ojs.aaai.org/index.php/AAAI/article/view/17639

[46] Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. 2021. Is Visual Context Really Helpful for Knowledge Graph? A Representation Learning Perspective. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 2735–2743. https://doi.org/10.1145/3474085.3475470

[47] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. 2019. Multimodal Data Enhanced Representation Learning for Knowledge Graphs. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. IEEE, 1–8. https://doi.org/10.1109/IJCNN.2019.8852079

[48] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied Knowledge Representation Learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra (Ed.). ijcai.org, 3140–3146. https://doi.org/10.24963/ijcai.2017/438

[49] Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

[50] Zuoxi Yang. 2020. Biomedical Information Retrieval incorporating Knowledge Graph for Explainable Precision Medicine. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 2486. https://doi.org/10.1145/3397271.3401458

[51] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A Fast and Accurate One-Stage Approach to Visual Grounding.

In *ICCV*. https://doi.org/10.1109/ICCV.2019.00478

[52] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for Knowledge Graph Completion. *CoRR* abs/1909.03193 (2019). arXiv:1909.03193 http://arxiv.org/abs/1909.03193

[53] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3342–3352. https://doi.org/10.18653/v1/2020.acl-main.306

[54] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1753–1762. https://doi.org/10.18653/v1/d15-1203

[55] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 14347–14355. https://ojs.aaai.org/index.php/AAAI/article/view/17687

[56] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level Relation Extraction as Semantic Segmentation. In *Proceedings of IJCAI*, Zhi-Hua Zhou (Ed.). ijcai.org, 3999–4006. https://doi.org/10.24963/ijcai.2021/551

[57] Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. 2021. AliCG: Fine-grained and Evolvable Conceptual Graph Construction for Semantic Search at Alibaba. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 3895–3905. https://doi.org/10.1145/3447548.3467057

[58] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 5674–5681. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16432

[59] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021. Multimodal Relation Extraction with Efficient Graph Alignment. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 5298–5306. https://doi.org/10.1145/3474085.3476968

[60] Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021. MNRE: A Challenge Multimodal Dataset for Neural Relation Extraction with Visual Evidence in Social Media Posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME51207.2021.9428274