Few-Shot Multimodal Named Entity Recognition based on Mutlimodal Causal Intervention Graph

Feihong Lu[1], **Xiaocui Yang**[2], **Qian Li**[4], **Qingyun Sun**[1*], **Ke Jiang**[1], **Cheng Ji**[1], **Jianxin Li**[1,3]

[1]School of Computer Science and Engineering, BDBC, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[3]Zhongguancun Laboratory, Beijing, China
[4]Beijing University of Posts and Telecommunications, Beijing, China
{lufh,sunqy, jiangke22,jicheng,lijx}@act.buaa.edu.cn,yangxiaocui@stumail.neu.edu.cn
13240016260@163.com

Abstract

Multimodal Named Entity Recognition (MNER) models typically require a significant volume of labeled data for effective training to extract relations between entities. In real-world scenarios, we frequently encounter unseen relation types. Nevertheless, existing methods are predominantly tailored for complete datasets and are not equipped to handle these new relation types. In this paper, we introduce the Few-shot Multimodal Named Entity Recognition (FMNER) task to address these novel relation types. FMNER trains in the source domain (seen types) and tests in the target domain (unseen types) with different distributions. Due to limited available resources for sampling, each sampling instance yields different content, resulting in data bias and alignment problems of multimodal units (image patches and words). To alleviate the above challenge, we propose a novel **M**ultim**O**dal ca**US**al **IN**tervention **G**raphs (**MOUSING**) model for FMNER. Specifically, we begin by constructing a multimodal graph that incorporates fine-grained information from multiple modalities. Subsequently, we introduce the Multimodal Causal Intervention Strategy to update the multimodal graph. It aims to decrease spurious correlations and emphasize accurate correlations between multimodal units, resulting in effectively aligned multimodal representations. Extensive experiments on two multimodal named entity recognition datasets demonstrate the superior performance of our model in the few-shot setting.

Keywords: Few-shot Multimodal Named Entity Recognition, Multimodal Graph, Causal Intervention

1. Introduction

The proliferation of multimodal data on social media platforms has led to increased interest in various related tasks, including Multimodal Named Entity Recognition (MNER), which utilizes multimodal information assistance to improve the accuracy of traditional NER tasks effectively. MNER focuses on extracting and classifying named entities from the unstructured text by leveraging multimodal data, such as the text-image pair, as Figure 1 shows. This area of research has gained considerable attention in recent years, as evident from recent studies (Yang et al., 2022; Chen et al., 2023). Previous studies in MNER primarily focus on designing effective models based on extensive training data (full training datasets) to enhance performance. (Yu et al., 2020; Zhang et al., 2021; Wang et al., 2022b). However, collecting and annotating the vast amounts of multimodal data for MNER is time-consuming and labor-intensive (Zhang et al., 2018; Lu et al., 2018). Moreover, in real-world applications, a substantial portion of the data remains unlabeled, while only a limited amount of labeled data is typically available. In this paper, we concentrate on the Few-shot Multimodal Named Entity Recognition (FMNER) task to enhance the model's capability to tackle MNER tasks with limited data and identify new relation types.

Meta-Learning has demonstrated significant success in various few-shot tasks(Vinyals et al., 2016; Bao et al., 2020; Yang et al., 2023a; Ma et al., 2022b), include Few-shot Named Entity Recognition (Fritzler et al., 2019; Ma et al., 2022a), Few-shot Image Classification (Chi et al., 2022; Wang et al., 2022a), and more. Building upon these studies, we utilize Meta-Learning with the Prototypical Network (Snell et al., 2017) for FMNER. FMNER, operating with limited text-image pairs, focuses on extracting entities from unstructured text and classifying them into corresponding entity types, which are various unforeseen types during training. Different from few-shot text-based NER, we require exploring both intra-modal and inter-modal information from multiple modalities to improve the performance of FMNER by incorporating multimodal representations tailored to specific entity types. As Figure 1 shows, without considering multimodal fine-grained semantic alignment, multiple occurrences of "DELL" may receive the most attention, which may cause the model to assign the same entity type (such as "ORG") to all "DELL". By establishing the fine-grained association between image patches and

[*] Corresponding author

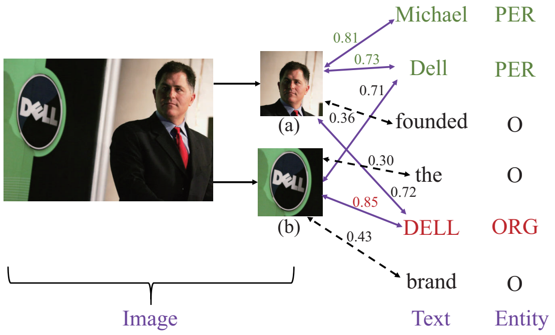


Figure 1: An example for multi-modal named entity recognition. (a) and (b) indicate multiple patches of the image. When the similarity between different modal units, *i.e.*, text words and image patches, exceeds a certain threshold, it indicates a significant correlation between them. In order to capture and represent this correlation, we establish an edge between different modal units in our approach, as solid lines show. Conversely, if the similarity falls below a certain threshold, indicating a weak correlation, we will remove the edges, as dash lines show.

text words, we can get the multimodal fusion representation to enhance the performance of MNER. For instance, "Michael Dell" can be linked to (a), while "DELL" can be connected to (b), which improves the beneficial alignment between different multimodal units.

To capture the effective aligned fine-grained representations across different modalities, we propose a novel **M**ultim**O**dal ca**US**al **IN**tervention **G**raphs (**MOUSING**) for the FMNER task, as Figure 2 shows. MOUSING consists of two modules, such as the Construction of the Multimodal Graph and the Updating of the Multimodal Graph. Since global representations of different modalities are unable to capture the effective fine-grained semantic information, as Figure 1 shows, refer to (Gao et al., 2023a), we first build the multimodal graph based on fine-grained information of different modalities, which utilizes each word and patch representation as nodes in the multimodal graph when the similarity between two nodes reaches a certain threshold, an edge is established. As mentioned by (Fan et al., 2022), FMNER faces a significant challenge referred to as data sampling bias, *i.e.*, the data distribution varies between source domains and target domains. This discrepancy can create spurious connections among the multimodal graph's nodes, exacerbating the risk of overfitting due to erroneous projections between the multimodal representation and the entity type. To alleviate the aforementioned challenge, we present a novel approach called the Multimodal Causal Intervention Strategy (MCIS) to update the multimodal graph. MCIS operates across different environments to update the original graph, with the primary objective of mitigating the effects of data bias, reducing erroneous relationship edges, and enhancing the weight of correct edges. Specifically, we simulate multiple training environments to perform causal intervention, resulting in multiple multimodal graphs. Furthermore, we propose a multi-view graph updating method that simultaneously updates multiple multimodal graphs from diverse perspectives. This approach adaptively reduces false associations between different modalities while emphasizing the correct associations among them, thereby enhancing the weight of correct patches and words for connecting edges and improving the generalization of our model in target domains. Extensive experiments on two datasets demonstrate that our approach outperforms strong baselines on the FMNRE task.

Our contributions can be summarized as follows.

- We propose a novel **M**ultim**O**dal ca**US**al **IN**tervention **G**raph (**MOUSING**) which builds deeper correlations among different modalities, to handle the Multimodal Named Entity Recognition task in a multimodal few-shot scenario. To the best knowledge, we are the first to propose the FMNER task.

- We first construct a multimodal graph to integrate fine-grained information. After that, Multimodal Causal Intervention Strategies (MCIS) are introduced to simulate multiple training environments to perform causal interventions, followed by a multi-view graph update method to improve fine-grained alignment across modalities.

- Experimental results indicate that MOUSING achieves state-of-the-art performance on the two multimodal named entity recognition datasets in the few-shot scenario.

2. Related Work

2.1. Multimodal Named Entity Recognition

With the boosting of multimodal data from social media platforms, many excellent works of MNER are constantly emerging. Wu et al. (2020a) propose a neural network that combines object-level image information and character-level text information to predict entities. Subsequently, UMT(Yu et al., 2020) extends the transformer(Vaswani et al., 2017) to a multimodal version and incorporates the auxiliary entity span detection module. Wang et al. (2022c) propose Image-text Alignments (ITA) to align image features into the textual space. Wang et al. (2023) first incorporate the transformer-based bottleneck fusion mechanism to reduce the noise propagation. Chen et al. (2022b) combines hierarchical multi-scaled visual features to generate effective and

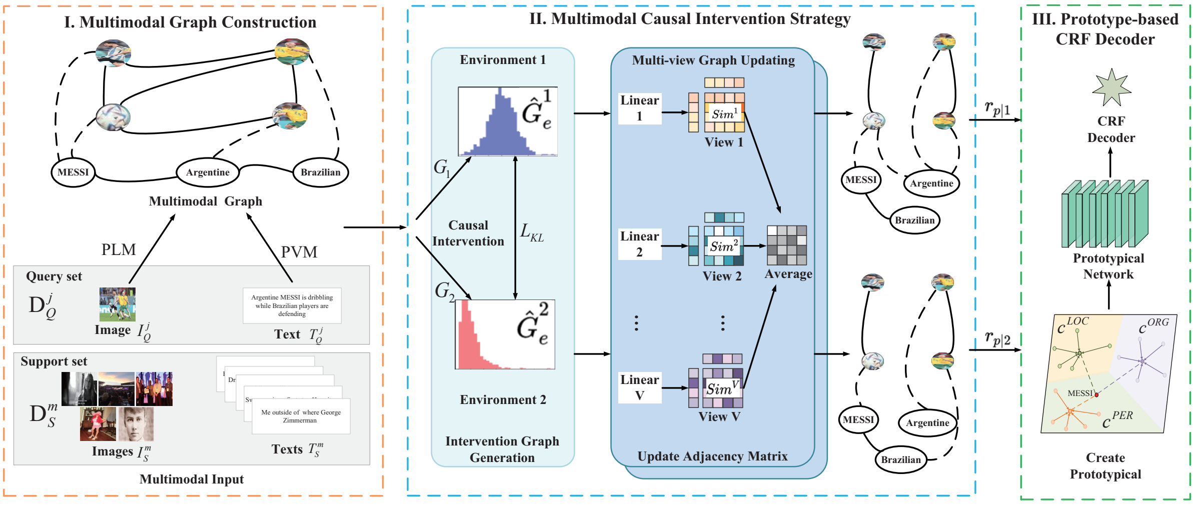


Figure 2: The overall architecture of our Multimodal Causal Intervention Graph model MOUSING."PLM" indicates the pre-trained language model while "PVM" indicates the pre-trained visual model. The multimodal graph does not undergo intervention graph generation during the testing phase. Instead, it directly updates the nodes and edges through multi-view graph updating.

robust text representations. Chen et al. (2022a) propose a hybrid transformer with multi-level fusion to enhance its adaptability in various real-world scenarios. Zhang et al. (2021) propose a unified multimodal graph fusion (UMGF) approach for MNER. However, the above MNER methods are only used in fully-supervised scenarios and can not handle unseen entity types. We are the first to propose the few-shot MNER task to detect the entity type with limited labeled data.

2.2. Few-shot Learning

Researchers propose many approaches to handle few-shot tasks, such as Matching Network (Vinyals et al., 2016), Meta-Learning (Ravi and Larochelle, 2017; Tian et al., 2020), and so on. Few-shot Named Entity Recognition aims to extract entities and classify them into the corresponding types, which are unknown in the training process, with a few support samples. As a typical Meta-Learning method, the prototypical network (Snell et al., 2017) is introduced to learn a metric space where instances of a novel specific class cluster around a single prototypical. Inspired by feature extraction and nearest neighbors, Yang and Katiyar (2020) propose NNShot and StructShot, which uses the nearest neighbor to classify entities. Yang et al. (2023b) proposes a causal intervention-based few-shot NER method, which uses context-based and prototype-based causal interventions to block the spurious correlation of different entities. However, this method is only used on unimodal data. Our method focuses on Multimodal Named Entity Detection in the few-shot scenario and achieves the alignment of fine-grained multimodal features.

3. Problem Formulation

FMNER aims to extract entities from unstructured text and classify them into the corresponding entity types which are unseen during training with limited labeled multimodal data. Like Yang and Katiyar (2020), we define the FMNER setting where the model is trained on source domains with annotations $D_S^m = (\mathcal{T}_S^m, \mathcal{I}_S^m)$ with source tag set C_S^m and then tested on target domains $D_Q^j = (\mathcal{T}_Q^j, \mathcal{I}_Q^j)$ with target tag set C_Q^j by only providing a few labeled examples per entity type, where $\mathcal{T}$ is the text modality, $\mathcal{I}$ is the image modality, m is the m-th entity type, j is j-th entity type, and $C_S \cap C_Q$= ϕ. Formally, the setting of N-way K-shot is defined as follows: given K text-image pairs for each entity type from D_Q as input, $x = (t, i)_{k=1}^{K} \in (\mathcal{T}_Q, \mathcal{I}_Q)$ and make the best tag sequence y, where $|C_Q| = N$.

4. Method

In this section, we propose a novel Multimodal causal Intervention Graphs (MOUSING) model for the FMNER task to more effectively capture the fine-grained alignment between different modalities. The overview of our framework is depicted in Figure 2. MOUSING including the construction of the multimodal graph and the updating of the multimodal graph. We depict the training and testing process of MOUSING in detail, as shown in Algorithms 1 and 2.

4.1. Multimodal Graph Construction

MNER, which involves extracting text spans from unstructured text, places a strong emphasis on leveraging the detailed information provided by multimodal data. To achieve this, refers to (Gao et al., 2021, 2023b), we construct the multimodal graph that integrates fine-grained information,*i.e.*, text words, and image patches, to facilitate a better understanding and representation of the data, which is shown in the upper left of Figure 2. Formally, the initial multimodal graph is undirected and can be formalized as $G_o = (X, A)$, where X represents the node features and A represents the adjacency matrix. Details of the multimodal graph are as follows:

Node Construction. To capture the effective representation of different modalities, we employ a pre-trained CLIP (Radford et al., 2021) model to derive the text embedding E_t and the image embedding E_i as initial node features of the multimodal graph.

$$\begin{aligned} E_t = CLIP_T(t), E_T \in \mathbb{R}^{L_t \times d_C}, \\ E_i = CLIP_I(i), E_C \in \mathbb{R}^{L_i \times d_C}, \end{aligned} \tag{1}$$

where L_t is the length of words for each t, L_i is the length of patches for each i, $d_C = 768$ is the dimension of embedding, and the number of a multimodal graph is $L = L_t + L_i$. E_t and E_i are the initial node features of G_o.

Edge Construction. We calculate the cosine similarity between the features of different nodes to capture valuable interactions between multimodal semantic units, including intra-modal nodes and inter-modal nodes. When the similarity exceeds a certain threshold, δ, that indicates a significant correlation between nodes, we establish an edge between different modal units in our approach. Conversely, if the similarity falls below a certain threshold, indicating a weak correlation, we will cancel this edge. By iteratively processing all possible pairs of nodes, we construct the initial multimodal adjacency matrix, $A_{p,q}$, as shown in Figure 2.

$$A_{p,q} = \begin{cases} 1, \cos\left(E_p, E_q\right) > \delta \\ 0, \cos\left(E_p, E_q\right) \leq \delta, \end{cases} \tag{2}$$

where p and q indicate the p-th and q-th nodes of the multimodal graph, respectively.

4.2. Multimodal Causal Intervention Strategy

Different from MNER, FMNER needs to detect entity types in the target domain, which are invisible during training, so it becomes crucial to enhance the generalization ability of the model. Therefore, multimodal graphs are inevitably biased, *i.e.*, existing edges are spuriously associated between nodes, which further leads to meaningless substructures being spuriously associated with labels.

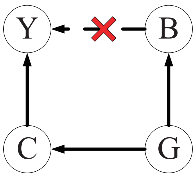


Figure 3: The causal structure of this task consists of four components. G represents the feature of multimodal graph, C represents correct correlations between different nodes, Y represents the target label, and B represents bias correlations between different nodes.

As DisC (Fan et al., 2022) said, GNNs majorly utilize bias substructure as shortcuts to make predictions, causing a large generalization performance degradation. Therefore, motivate by (Sun et al., 2022; Yuan et al., 2024), we propose a novel multimodal causal intervention strategy (MCIS) to address the issue of unstable and biased correlations caused by data selection bias in the few-shot scenario. MCIS can adaptively learn stable associations between different modalities, *i.e.*, reducing bias associations between different nodes while emphasizing the correct associations among them. MCIS contains two steps: the Intervention Graph Generation and the Multi-view Graph Updating.

4.2.1. Intervention Graph Generation

We present a multimodal causal view of the union of the model-training process and the model-detection process behind the task. Here we formalize the causal view as a multimodal causal graph by inspecting the causalities among four variables: the feature of a multimodal graph G, correct correlations between nodes of the graph C, bias correlations between nodes of the graph B, the target label Y. Figure 3 illustrates causal relationships between different variables in our task, detailed descriptions are as follows:

C ← **G** → **B**. A multimodal graph, G, consists of correct edges, C, and biased edges, B, for efficient detection of entity types.

C → **Y** . It means that the causal variable C, which is the sole endogenous parent, determines the generation of the true label Y. For instance, C consists of edges that are effective for correctly classifying entity types, which precisely explains why the label is assigned as Y, as the edge between ***Dell*** and ***the patch (a)*** shows in the Figure 1.

B → **Y**. It indicates that spurious correlations between B and Y are due to the presence of bias edges among nodes. That is edges that have high similarity but interfere with entity type identification. For example, due to the deviation in data collection under the few-shot setting, there are spurious correlations between the multimodal representation and the ground truth label, as the edge between ***Dell*** of *Miachael Dell* and ***the patch (b)*** shows in the Figure 1.

According to Figure 3, our model makes detection based on both C and B. As (Fan et al., 2022) said, GNNs majorly utilize bias substructure as shortcuts to make predictions, which can quickly achieve low loss and lead to model overfitting, *i.e.*, **B** → **Y**. Hence, we propose the multimodal causal interventions strategy to block **B** → **Y** and strengthen **C** → **Y**. In this respect, we simulate multiple training environments to perform causal intervention resulting in multiple multimodal graphs based on the same original graph.

$$P(Y = y \mid do(\mathcal{G} = G)) = \sum_{E} P(\mathcal{E} = E \mid \mathcal{G} = G)), \\ \sum_{G'} P\left(Y = y \mid \mathcal{E} = E, \mathcal{G} = G'\right) P\left(\mathcal{G} = G'\right), \tag{3}$$

where $do(.)$ represents generate different intervention graph, E indicates the specific environment E from $\mathcal{E}$, G' indicates the specific multimodal graph from $\mathcal{G}$, and y from Y indicates the ground truth label of current graph.

Specifically, to capture stable multimodal representations and reduce the influence of biased correlations, we employ distinct training environments to simultaneously train a pair of GNNs, denoted as G_1 and G_2, based on the same graph.

$$\hat{G}_1 = G_1(G_o, \theta_1), \hat{G}_2 = G_2(G_o, \theta_2), \tag{4}$$

where G_o is the initial multimodal graph from section 4.1; the θ_1 and θ_2 are parameters of G_1 and G_2, respectively; $\hat{G}_1 \in \mathbb{R}^{L\times d}$ and $\hat{G}_2 \in \mathbb{R}^{L\times d}$ are outputs of G_1 and G_2, respectively. d is the dimension of hidden representation from GNNs.

4.2.2. Multi-view Graph Updating

We further design a multi-view graph updating approach to simultaneously update multiple multimodal graphs from diverse perspectives. This approach enhances the fine-grained alignment across modalities and improves the generalization of our model in target domains. In other words, multi-view graph updating adaptively reduces bias associations between different nodes while emphasizing the correct associations among them in the few-shot setting. We update the multimodal graph from V different views to capture robust representations of multimodal graphs.

$$Sim_e^v = \cos(w_v \odot \hat{G}_e, w_v \odot \hat{G}_e), \tag{5}$$

$$Sim_e = \frac{1}{V}\sum_{v=1}^{V} Sim_e^v, \tag{6}$$

where $w_v \in \mathbb{R}^{d\times z}$ represents parameters of the v-th linear layer, z is the dimension of projection, V is the hyperparameter, and $e \in \{1, 2\}$ is the graph of different environment.

We then update the multimodal adjacency matrix.

$$(A_e)' = \gamma A_e + (1-\gamma) Sim_e, \tag{7}$$

where γ is used to balance the original graph representation and the newly learned graph knowledge.

Following (Wu et al., 2020b), we obtain $l+1$-th node features based on the l-th layer of multimodal GNNs.

$$X_{p|e}^{l+1} = \mathrm{U}(X_{p|e}^{l}, \sum_{q\in N(p|e)} \mathrm{M}(X_{p|e}^{l}, X_{q|e}^{l}, (A_{p,q|e}^{l})'), \tag{8}$$

where p indicates current node, $N(p)$ represents the set of neighbor nodes of p node, M represents the function of aggregating neighborhood information for p, and U indicates the function of updating p by aggregated information.

We capture the final node features according to features of the last layer and a multi-layer perceptron F_M.

$$r_{p|e} = F_M(X_{p|e}^{L}, \theta_M), \tag{9}$$

where L is the number of layer and θ_M represents the parameters of MLP.

4.3. Prototype-based CRF Decoder

Similar to Snell et al. (2017), we leverage Meta-Learning based on the prototypical network to handle FMNER. In each batch, we randomly sample a few instances as query set, $\mathcal{Q}$, and other K instances as support set $\mathcal{S}$, where $|\mathcal{Q}| = 1$ and $|\mathcal{S}| = 5$. $\mathcal{S}^m$ denotes the set of instances labeled with m-th entity type [1]. Each prototype is the mean vector of the embedded support points belonging to its class:

$$c^m = \frac{1}{|\mathcal{S}^m|}\sum_{(x_k,y_k)\in\mathcal{S}^m} G_e(x_k), \tag{10}$$

where G_e means operations of the multimodal graph with the e-th environment to get final node features r. x is the text-image pair, (t, i).

The prototypical networks produce a distribution over classes for each word of $x_q \in \mathcal{Q}$ based on a softmax over distances to the prototypes in the embedding space:

$$p(y_w = m|x^q, G_e) = \frac{\exp(-d(G_e(x_q), c^m))}{\sum_{m'} \exp(-d(G_e(x_q), c^{m'}))}, \tag{11}$$

where d calculates the euclidean distance and w is the word from the text of x^q.

We employ Cross-Entropy loss, CE, to calculate classification loss.

$$L_{CRF} = CE(D(P|G_1), y) + CE(D(P|G_2), y), \tag{12}$$

where P is the sequence of predicted probability for words of x_q, y is the target label, and D indicates the CRF decoder.

[1] Note that the entity type is from the C_S during training, and the entity type is from the C_Q during test.

4.4. Overall Loss Function

We calculate the KL Loss used as a constraint to increase the distribution dissimilarity of node features between the two graphs, which is shown in Eq. (13).

$$L_{KL} = -D_{KL}(F_M(\hat{G}_1, \theta'_M) \| F_M(\hat{G}_2, \theta'_M)), \tag{13}$$

where D_{KL} represents the calculation of KL divergence.

The overall loss function L_{final} is as follows,

$$L_{final} = \lambda L_{CRF} + (1 - \lambda) L_{KL}, \tag{14}$$

where λ is a hyperparameter used to balance the effects of different losses.

Algorithm 1 Training Process of MOUSING

```
Input: Source domain data $D_S^m$, Max iterations $T$ for each epoch.
Output: Learned two GNN networks ($\hat{G}_1$, $\hat{G}_2$), multiple MLPs, and CRF Decoder $D$.
Initialization: iteration $t = 0$; Initialize ($\hat{G}_1$, $\hat{G}_2$), MLPs, and $D$.
for sampled minibatch $X \in D_S^m$ and iterations $\leq T$ do
    Build the initial multimodal graph, $G_o$, including $X$, i.e., $E_t$ and $E_i$ by Eq. (1), and the adjacency matrix $A$ by Eq. (2).
    for $e \in \{1, 2\}$ do
        $G_e = G_e(G_o, \theta_e)$ by Eq. (4);
        for v in range $\{0, V\}$ do
            Calculate $Sim_e^v$ as by Eq. (5);
        end for
        Calculate $Sim_e$ by Eq. (6);
        Updated $(A_e)'$ by Eq. (7);
        Update node features by Eq. (8);
        Calculate the final node feature $r_{p|e}$ by Eq. (9);
        Random sample one instance as the query set, $\mathcal{Q}$;
        Random select five instances as the support set,$\mathcal{S}$;
        Construct the prototypical for the $m$-th entity type by Eq. (10).
    end for
    Get the output of query by Eq. (11);
    Calculate the $L_{CRF}$ by Eq. (12).
end for
Calculate the $L_{KL}$ by Eq. (13).
Calculate the $L_{final}$ by Eq. (14).
Update all parameters by backpropagation to reduce $L_{final}$.
```

5. Experiments

5.1. Datasets

Based on the Zhang et al. (2021) guidelines, we utilize two publicly available Twitter datasets, Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018), for FMNER. The statistics of the different datasets are presented in Table 1. Considering that Twitter-2015 and Twitter-2017 datasets share the same four entities, we randomly select two entity types for training and others for testing. Therefore, each dataset has 6 splits for FMNER. We conduct our main experiments on the 2-way 5-shot setup. For six different data splits, the number of each entity should be consistent with Table 1.

Algorithm 2 Testing Process of MOUSING

```
Input: Target domain data $D_Q^j$.
Output: The predicted label for each entity.
Initialization: Initialize ($\hat{G}_1$, $\hat{G}_2$), MLPs, and $D$ by trained parameters.
for sampled minibatch $X \in D_Q^j$ do
    Random sample one instance as query set,$\mathcal{Q}$;
    Random select five instances as support set,$\mathcal{S}$;
    Calculate the prototypical representation for each entity type by (10);
    Get the query data output by Eq. (11);
    Get the predicted label by the trained CRF-decoder $D$.
end for
```

Entity	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
Person	2217	552	1816	2943	626	621
Location	2091	522	1697	731	173	178
Organization	928	247	839	1674	375	395
Misc.	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351

Table 1: Statistics on two datasets.

5.2. Implementation Details

We use PyTorch as a deep learning framework to develop the few-shot MNER. The maximum length of the text is 128, the number of image patches is 9, and the batch size is 6. The learning rate, dropout rate, and trade-off parameter are set to 5e-3, 0.15, and 2, respectively. The convergence time of the model is almost 30 minutes for all six experimental settings. The weight parameters for trainable parameters are 6.4M, and the total weight parameters are 434.02M. Performance is evaluated by the Micro-F1 score on the test dataset, and the predicted entity is correct if the entity type and position match the gold-standard entity. We use the BIO mode by default to allow a fair comparison with previous studies.

Modality	Model	Twitter-2015						
		Per+Loc	Per+Org	Per+Others	Loc+Org	Loc+Others	Org+Others	Avg.
TNER	BERT	17.99	13.33	15.43	14.85	13.71	17.61	15.49
	ProtoBERT	18.52	22.26	20.71	15.89	14.80	17.83	18.34
	RoBERTa	20.11	16.73	17.71	16.25	17.56	21.16	18.25
	ProtoRoBERTa	20.51	19.41	19.52	20.19	18.57	23.53	20.29
	NNshot	18.65	27.24	28.21	20.81	28.78	25.90	24.93
	Structshot	18.66	**30.41**	28.37	24.87	31.13	**29.05**	27.08
MNER	UMT	18.57	21.43	24.24	17.14	14.22	24.26	19.98
	UMT-CLIP*	28.57	22.23	31.25	27.27	12.46	22.50	24.05
	UMGF	20.00	17.77	24.94	17.20	16.23	25.71	20.31
	UMGF-CLIP*	23.30	21.73	32.21	22.50	20.54	22.31	23.76
	ProtoUMGF	23.33	25.39	27.14	25.10	18.65	26.34	24.33
	ProtoUMGF-CLIP*	23.79	20.30	31.88	23.45	20.85	22.46	23.78
	HVPNet	24.97	23.81	26.09	14.38	21.14	19.35	21.62
	ProtoHVPNet	19.35	29.13	**37.20**	21.43	17.03	16.74	23.48
	MOUSING	**34.86**	28.43	35.32	**30.10**	**36.62**	27.99	**31.22**

Modality	Model	Twitter-2017						
		Per+Loc	Per+Org	Per+Others	Loc+Org	Loc+Others	Org+Others	Avg.
TNER	BERT	14.29	20.69	15.39	9.04	12.55	12.03	13.99
	ProtoBERT	12.60	23.26	14.03	21.95	18.67	15.87	17.73
	RoBERTa	20.59	15.39	22.73	11.76	16.44	13.14	16.68
	ProtoRoBERTa	21.21	24.10	23.14	23.36	17.51	19.67	21.50
	NNshot	24.27	**29.82**	25.29	23.23	23.82	22.78	24.86
	Structshot	25.41	29.39	22.89	24.45	24.20	**27.22**	25.59
MNER	UMT	24.00	16.76	12.50	20.59	17.54	17.39	18.13
	UMT-CLIP*	33.32	25.02	13.34	22.58	20.02	13.04	21.22
	UMGF	24.24	17.65	14.63	22.43	23.73	13.04	19.29
	UMGF-CLIP*	19.18	19.28	11.27	15.52	24.11	27.27	19.44
	ProtoUMGF	20.41	16.00	26.12	**24.62**	23.18	19.23	21.59
	ProtoUMGF-CLIP*	18.20	21.36	11.77	14.04	14.17	22.18	17.07
	HVPNet	32.50	16.28	16.67	18.23	14.07	25.41	20.53
	ProtoHVPNet	29.05	24.94	24.29	13.16	17.18	16.21	20.81
	MOUSING	**34.12**	28.10	**27.14**	24.12	**24.85**	25.79	**27.35**

Table 2: Performance of different competitive uni-modal and multimodal approaches in terms of **F1** for FMNER on Twitter-2015 and Twitter-2017. "*" indicates that the reproducible results of different models are achieved by using CLIP instead of the original image encoder.

5.3. Baselines

To ensure a comprehensive comparison, we thoroughly evaluate our model against various approaches based on unimodal and multimodal baselines. The first group is the text-based NER (TNER) approach: 1) BERT (Devlin et al., 2019) is the competitive baseline for NER. 2) ProtoBERT (Devlin et al., 2019) exploits the Prototypical Network based on BERT to solve TNER in the few-shot setting. 3) RoBERTa (Liu et al., 2019) is an improvement of BERT. 4) ProtoRoBERT (Liu et al., 2019) combines the Prototypical Network with RoBERTa. 5) NNshot (Wang et al., 2019) applies simple feature transformations on the features before nearest-neighbor classification in the few-shot TNER task. 6) Structshot (Wang et al., 2019) adds an additional Viterbi decoder based on the NNShot. The second group is competitive multimodal approaches for MNER: 1) UMT (Yu et al., 2020) extends Transformer to a multimodal version and incorporates the auxiliary entity span detection module. 2) UMGF (Zhang et al., 2021) constructs the multimodal graph and further stacks multiple multimodal fusion layers to learn node representations. 3) ProtoUMGF utilizes the Prototypical Network based on UMGF. 4) UMT-CLIP, UMGF-CLIP and ProtoUMGF-CLIP use the CLIP model as a visual and text encoder. 5) HVPNet (Chen et al., 2022b) incorporates hierarchical multi-scaled visual features to generate an effective and robust textual representation for reducing error sensitivity. 6) ProtoHVPNet exploits the Prototypical Network based on HVPNet. 7) **MOUSING** is our model, which builds deeper correlations among different modalities by Multimodal Causal Intervention Graph, to handle the FMNER task.

5.4. Main Results

To verify the effectiveness of our model, we report the results of different splits for entity types and overall average results on all entity types, as Table 2 shows, where "Per+Loc" means using "Person" and

"Location" entities to train and the rest of the entities to testing. The other five groups of experiments are the same as above. Our experimental results consistently demonstrate the superior performance of our model MOUSING, compared to both TNER and MNER methods across various entity types in both datasets. This notable improvement can be attributed to several key factors. First, we leverage a multimodal graph to explore multimodal fine-grained information that is suitable for named entity recognition. Secondly, the Multimodal Causal Intervention strategy in MOUSING facilitates the establishment of correlations and alignment of fine-grained features between different modalities. This capability allows the model to capture highly effective multimodal representations. It demonstrates the superiority of our model over other models in handling the challenges of limited labeled data and the image modality is crucial for achieving high performance in the few-shot MNER task.

5.5. Ablation Experiments

Variants	Precision	Recall	F1
w/o Image	26.59	26.96	26.72
w/o Intervention	24.94	25.09	25.00
w/o Multi-view	26.12	25.75	25.72
w/o MICS	19.36	19.11	19.06
w/ Random Intervention	29.99	30.69	30.24
w/ Gaussian Intervention	30.30	30.75	30.47
MOUSING (Ours)	**32.89**	**33.70**	**33.18**

Table 3: Average results for MOUSING ablation experiments overall splits. "w/" indicates "with" and "w/o" indicates "without".

We perform ablation experiments on the MOUSING model to assess the effectiveness of different modules and report the average results on all splits in Table 3. We first remove the image modality (w/o Image) and only leverage the text modality to accomplish MOUSING. The performance of our model drops dramatically, showing that image modality is critical for detecting named entities in the few-shot setting. When we remove the causal intervention module (w/o Intervention), spurious correlations between different nodes in the multimodal graph interfere with the training process, resulting in poor model performance. It demonstrates that constrained interventions are able to control the distribution of features in the graph to more efficiently obtain causal representations in the case of two inconsistent distributions. We also remove the Multi-view module (w/o Multi-view) to verify the utility of Multi-view updating in Multimodal Graphs. As Table 3 shows, our model performs poorly. It shows that updating multimodal graphs from multiple perspectives can preserve more correct edges and reduce spurious edges. When we remove the Multimodal Causal Intervention Strategy (MCIS) module (w/o MCIS), *i.e.*, simultaneously removing the Intervention module and Multi-view module, the model has the worst performance. It indicates that our MCIS has excellent effectiveness for FMNER. We also replaced causal interventions with random interventions (w/ Random Intervention) and Gaussian interventions (w/ Gaussian Intervention). The model underperforms MOUSING, indicating that the causal intervention is effective.

5.6. Hyperparameters Setting

We conduct experiments for different hyperparameters, including V of Eq. (6), δ of Eq. (2), γ of Eq. (7), L of Eq. (9), and λ of Eq. 14. The experimental results are shown in Figure 4.

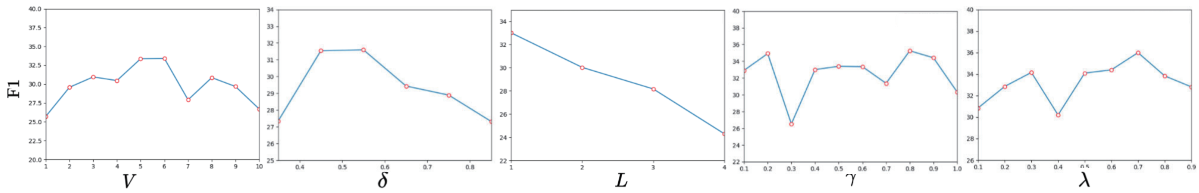
1) We investigate the effect of the number of different perspectives on the update of multimodal causal intervention graphs for multimodal graphs and text graphs, that is, setting $V \in \{1, ..., 10\}$ in Eq. (6). As Figure 4 (a) shows, MOUSING achieves the best performance when V is 6. When the value of V is smaller, the model cannot adequately capture the fine-grained alignment between different modalities, when the value of V is larger, excess perspectives will bring redundant information to the model.

2) In Figure 4 (b), the experimental results regarding δ of Eq. (2) are presented. Increasing δ from 0.35 to 0.55 consistently improves the model's performance, suggesting that a larger δ introduces redundant information that interferes with the detection of named entities. However, when δ increases from 0.55 to 0.85, the model's performance sharply declines, indicating that valuable information in the multimodal graph is lost at higher δ values.

3) L represents the representation of nodes updated with different layers of GNNs. When the number of layers of GNN is 1, the model achieves the best performance. As the number of GNN layers increases, the model tends to learn more dense node connections, resulting in more redundant information.

4) We verify the value of γ, as Figure 4 (d) shows. When the value of γ is between 0.1 and 1, the performance of the model keeps fluctuating, reaching the best performance when γ is 0.8. When γ is smaller than 0.8, the update amplitude of the new graph is too large, resulting in unstable model performance. When γ is larger than 0.8, the update amplitude of the new graph is too small, and the update is too slow.

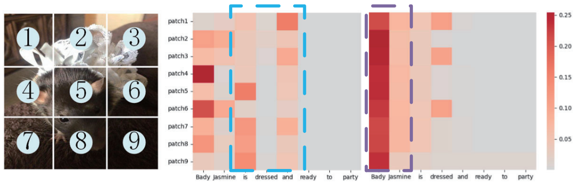
5) We test the value of λ ranges from 0.1 to 0.9, the model's performance fluctuates continuously, with the best performance achieved at λ=0.7. When the λ is smaller, the model focuses on environment generation and neglects constraints of the the CRF loss function. Therefore, the effectiveness



(a) (b) (c) (d) (e)

Figure 4: Hyperparameters experiments. **F1** comparisons of V for MOUSING.

of the model is constantly fluctuating and unstable. When the λ is larger, the model neglects the environment generation, which leads to the overfitting of the model.



(a) Image data (b) Original graph (c) Updated graph

Text input: Bady Jasmine is dressed and ready to party					
	Entity	Label	MOUSING	structshot	UMGF
	Bady	PER	√	×	×
	Jasmine	PER	√	√	√

(d) Case example

Figure 5: Case study example and the visualization of the adjacency matrix.

5.7. Cases study

To illustrate our model can effectively construct the fine-grained alignment between different modalities, we exhibit an example as Figure 5 shows. We visualize adjacency matrices of the original multimodal graph and the updated multimodal graph via MCIS, respectively, as Figure 5 (b) and Figure 5 (c) show. It can be seen that the initial multimodal graph establishes strong associations between “is, and” and all image patches, so this association cannot assist in text prediction, as Figure 5(b) shows. After updating our model, the new graph weakens invalid associations and strengthens beneficial associations between patches and words, such as the association between image patches and “Baby Jasmine”, which will more effectively assist the model in making a detection. The comparative results for the case studies are shown in Figure 5(d), where our model performs the best.

6. Conclusions

We propose a novel MultimOdal caUSal INtervention Graphs (MOUSING) model for multimodal named entity recognition in the few-shot scenario. MOUSING first builds the multimodal graph based on fine-grained information from different modalities. We then leverage the Multimodal Causal Intervention Strategy (MCIS) to strengthen collect edges and weaken spurious edges to improve the performance of FMNER. Extensive experiments conducted on the two datasets demonstrate that our approach outperforms strong baselines on the FMNRE task. We provide a new direction for related tasks of MNER in the few-shot setting. We will explore more multimodal tasks with multimodal causal intervention graph in future work. In future work, we will explore more multimodal tasks with multimodal causal intervention graph model, such as multimodal relation extraction, multimodal event extraction, and multimodal parsing.

Limitations

Our work overcomes the severe data bias impact in the few-shot MNER setting, which effectively decreases the spurious correlations and emphasizes accurate correlations between multimodal units, resulting in effectively aligned multimodal representations. Empirical experiments demonstrate that our method weakens the influence of biased data in the few-shot setting. However, there are still some limitations of our approach, which can be summarized as follows:

- Due to the limitation of the existing MNRE datasets, we only experiment on two modalities. We will study more modalities in future work.
- Our model does not consider applications on other multimodal tasks, such as multimodal relation extraction and multimodal image-text retrieval. We will study more multimodal tasks in the future.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. The authors of this paper were supported by the NSFC through grants 62225202 and 62302023.

7. Bibliographical References

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *8th International Conference on Learning Representations, ICLR*.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022a. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 904–915. ACM.

Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022b. Good visual guidance make A better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618. Association for Computational Linguistics.

Yong Chen, Xinkai Ge, Shengli Yang, Linmei Hu, Jie Li, and Jinwen Zhang. 2023. A survey on multimodal knowledge graphs: Construction, completion and applications. *Mathematics*, 11(8):1815.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. 2022. Metafscil: A meta-learning approach for few-shot class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 14146–14155. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. 2022. Debiasing graph neural networks via learning disentangled causal substructure. In *NeurIPS*.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 993–1000. ACM.

Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. 2021. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3073.

Chen Gao, Si Liu, Jinyu Chen, Luting Wang, Qi Wu, Bo Li, and Qi Tian. 2023a. Room-object entity prompting and reasoning for embodied referring expression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. 2023b. Adaptive zone-aware hierarchical planner for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14911–14920.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2020. Few-shot named entity recognition: A comprehensive study. *CoRR*, abs/2012.14978.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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, ACL 2018*, pages 1990–1999. Association for Computational Linguistics.

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022a. Decomposed meta-learning for few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1584–1596. Association for Computational Linguistics.

Yao Ma, Shilin Zhao, Weixiao Wang, Yaoman Li, and Irwin King. 2022b. Multimodality in meta-learning: A comprehensive survey. *Knowl. Based Syst.*, 250:108976.

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 Amer-*

ican Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, pages 852–860. Association for Computational Linguistics.

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

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *5th International Conference on Learning Representations, ICLR*.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.

Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and S Yu Philip. 2022. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4165–4174.

Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. 2020. Rethinking few-shot image classification: A good embedding is all you need? In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 266–282. Springer.

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*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 3630–3638.

Kai Wang, Xialei Liu, Andy Bagdanov, Luis Herranz, Shangling Jui, and Joost van de Weijer. 2022a. Incremental meta-learning via episodic replay distillation for few-shot image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,CVPR Workshops 2022*, pages 3728–3738. IEEE.

Peng Wang, Xiaohang Chen, Ziyu Shang, and Wenjun Ke. 2023. Multimodal named entity recognition with bottleneck fusion and contrastive learning. *IEICE Trans. Inf. Syst.*, 106(4):545–555.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022b. ITA: image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 3176–3189. Association for Computational Linguistics.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022c. ITA: image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 3176–3189. Association for Computational Linguistics.

Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. 2019. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *CoRR*, abs/1911.04623.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. 2022a. Handling distribution shifts on graphs: An invariance perspective. In *The Tenth International Conference on Learning Representations, ICLR 2022*.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022b. Discovering invariant rationales for graph neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020a. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020b. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Aimin Yang, Chaomeng Lu, Jie Li, Xiangdong Huang, Tianhao Ji, Xichang Li, and Yichao Sheng. 2023a. Application of meta-learning in cyberspace security: a survey. *Digit. Commun. Networks*, 9(1):67–78.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19):9691.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6365–6375. Association for Computational Linguistics.

Zhen Yang, Yongbin Liu, and Chunping Ouyang. 2023b. Causal interventions-based few-shot named entity recognition. *CoRR*, abs/2305.01914.

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*, pages 3342–3352. Association for Computational Linguistics.

Haonan Yuan, Qingyun Sun, Xingcheng Fu, Ziwei Zhang, Cheng Ji, Hao Peng, and Jianxin Li. 2024. Environment-aware dynamic graph learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36.

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*, pages 14347–14355. AAAI.

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)*, pages 5674–5681. AAAI.