applied sciences

MDPI

Article

Visual Clue Guidance and Consistency Matching Framework for Multimodal Named Entity Recognition

Li He [1,2], Qingxiang Wang [1,2,*], Jie Liu [1,2], Jianyong Duan [1,2] and Hao Wang [1,2]

[1] School of Information Science and Technology, North China University of Technology, Beijing 100144, China; heli@ncut.edu.cn (L.H.)
[2] CNONIX National Standard Application and Promotion Lab, Beijing 100144, China
[*] Correspondence: 19862128500@163.com

Abstract: The goal of multimodal named entity recognition (MNER) is to detect entity spans in given image–text pairs and classify them into corresponding entity types. Despite the success of existing works that leverage cross-modal attention mechanisms to integrate textual and visual representations, we observe three key issues. Firstly, models are prone to misguidance when fusing unrelated text and images. Secondly, most existing visual features are not enhanced or filtered. Finally, due to the independent encoding strategies employed for text and images, a noticeable semantic gap exists between them. To address these challenges, we propose a framework called visual clue guidance and consistency matching (GMF). To tackle the first issue, we introduce a visual clue guidance (VCG) module designed to hierarchically extract visual information from multiple scales. This information is utilized as an injectable visual clue guidance sequence to steer text representations for error-insensitive prediction decisions. Furthermore, by incorporating a cross-scale attention (CSA) module, we successfully mitigate interference across scales, enhancing the image's capability to capture details. To address the third issue of semantic disparity between text and images, we employ a consistency matching (CM) module based on the idea of multimodal contrastive learning, facilitating the collaborative learning of multimodal data. To validate the effectiveness of our proposed framework, we conducted comprehensive experimental studies, including extensive comparative experiments, ablation studies, and case studies, on two widely used benchmark datasets, demonstrating the efficacy of the framework.

Keywords: multimodal named entity recognition; contrastive learning; feature pyramid

check for updates

Citation: He, L.; Wang, Q.; Liu, J.; Duan, J.; Wang, H. Visual Clue Guidance and Consistency Matching Framework for Multimodal Named Entity Recognition. *Appl. Sci.* **2024**, *14*, 2333. https://doi.org/10.3390/app14062333

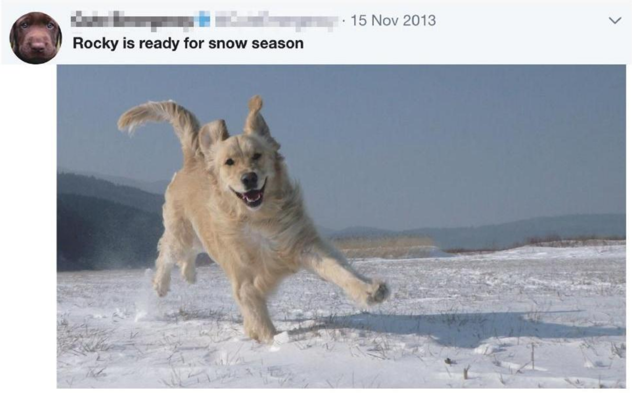
Academic Editor: Antonio Fernández-Caballero

Received: 19 January 2024
Revised: 28 February 2024
Accepted: 7 March 2024
Published: 10 March 2024

1. Introduction

Multimodal named entity recognition (MNER) has become an important research direction in named entity recognition (NER), and it can improve text-based NER by using images as additional inputs [1]. It assumes that image information can help to recognize unambiguous named entities when textual information is insufficient. Considering the given text example, "Rocky is ready for snow season", there are obvious challenges in determining named entity categories based on the textual content alone. "Rocky" could refer to a person, an animal, or some other entity type. However, when we combine this text with the corresponding image information (as shown in Figure 1), we can easily determine its type as MISC. In this paper, we investigate MNER in social media posts.

Existing work achieved good performances compared to text-based NER methods [1–14]. Refs. [2–4] are pioneering works in multimodal named entity recognition. For the first time, image information was added to assist entity recognition in text, which improved the accuracy of entity recognition. However, although word-related visual representations are generated, they are insensitive to the visual context and ignore the bias introduced by the visual context. To address these issues, ref. [5] proposes an end-to-end model for learning joint representations of text and images using a multidimensional self-attention technique that simultaneously captures text and improves the accuracy of the visual context. The authors of [4] present a novel model based on visual attention that provides deeper visual understanding in model decision-making. Ref. [6] utilizes object labels as embeddings to achieve a bridge between visual and verbal aspects and introduces an intensive co-attention mechanism for fine-grained interactions to recognize named entities more accurately using the visual context. A multimodal interaction module was designed for the first time using a Transformer to obtain information about the word representation and visual representation of an image [7], and the authors propose to utilize plain text entity span detection as an auxiliary module to mitigate the visual bias, considering both textual and visual information to improve the recognition accuracy. A unified multimodal graph fusion (UMGF) method is proposed to represent the input sentences and images as a unified multimodal graph [1], stacking multiple graph-based multimodal fusion layers, learning node representations by iteratively performing semantic interactions, and utilizing graph structures to make fine-grained semantic associations between semantic text and image units. However, there are limitations in all five approaches [1,4–7], which do not fully utilize the knowledge behind image and text pairs. Therefore, ref. [8] proposes a novel pretrained multimodal model based on Relational Inference and Visual Attention (RIVA), which employs a teacher–student semi-supervised paradigm to utilize the large unlabeled multimodal inferred corpus and labeled datasets for the classification of text–image relations. A novel MNER neural model is proposed for acquiring image attributes and image knowledge [9], and a multiple attention mechanism is designed to integrate this information. Reference [10] proposes a MoRe framework for injecting knowledge-aware information into a multimodal NER task using multimodal retrieval, which has rarely been seen in previous research. However, the above three approaches [8–10] also have drawbacks in that they cannot avoid the complexity of using external tools and datasets. To address this problem, ref. [11] proposes a new multilevel semantic alignment approach that captures coarse- to fine-grained interactions between images and language and directly utilizes learned visual features to capture the relationship between images and text more comprehensively, avoiding the complexity of using external tools and datasets. In addition, ref. [12] had a unique idea [12–14] to utilize the external matching between different (text, image) pair relationships and designed an R-GCN model to model these relationships, using the external matching relationships between different (text, image) pairs within the dataset to mitigate image noise in the MNER task while modeling both cross-modal and intra-modal relationships. Ref. [13] proposed the SD-NER model, which models the minimum distance matrix between entities and is easily transferable to other tasks. Ref. [14] proposed a multimodal Chinese named entity recognition (USAF) model using acoustic features, which unifies textual and acoustic features through a unique positional embedding and fuses the features of both features using a multi-head attention mechanism.



[redacted] · 15 Nov 2013

Rocky is ready for snow season

Figure 1. An example of multimodal tweets. In this tweet, "Rocky" is the name of the dog.

Although existing methods are a great improvement [1–14], they have three main shortcomings. Firstly, existing methods assume that each piece of text and its corresponding image are matched and that the image can help recognize named entities in the text, but not all texts are matched with their corresponding images. As shown in Figure 2, ref. [7] incorrectly predicts "Aquamarine" as ORG due to the influence of the house, indicating that the existing model has difficulties in filtering the noise introduced by the mismatched text–image pairs. According to [15], about 34.1% of the text content in Twitter-2015 has an imperfect match with images. Secondly, existing work on images usually relies only on the Residual Network (ResNet) [16] or Mask-RCNN [17] and other image encoders to extract visual features without enhancement or filtering and feed them directly into cross-modal interaction mechanisms along with text. Finally, most existing approaches fail to construct a consistent representation to bridge the semantic gap between the two modalities. As shown in Figure 3, ideally, the words "Danielle" and "Melissa" in the text should show a high degree of similarity with the region associated with the "rabbits" in the image and low similarity with other regions in the image. However, due to the inconsistent representation between the text and the image, the similarity between "Danielle" and "Melissa" in the text and the "rabbits" in the image may be lower than the similarity between "Danielle" and "Melissa" in the text and the "rabbits" in the image in the calculation of the similarity score. The similarity between "Danielle" and "Melissa" in the text and the "rabbits" in the image may be lower than the similarity with other regions in the image. To address these shortcomings, we propose a visual clue guidance and consistency matching framework (GMF). Firstly, we design a visual clue guidance module (VCG), which collects image features through the Resnet50 image encoder. On this basis, we propose a cross-scale attention (CSA) module, which combines low-level, high-resolution image information with high-level, strong semantic image information to enhance visual features. Further, feature fusion is performed by a structured mapping function $F(\theta)$, and, finally, the above vectors are processed using a gating mechanism for path decision-making. Secondly, we utilize Contrastive Language–Image Pretraining (CLIP) [18] based on the idea of multimodal contrast learning, and we design the consistency matching module (CM), defining contrast loss to bring matched image–text pairs closer together and push them away, enhancing the learning of the multimodal potential semantic space, and making the representation of two modalities more consistent. In addition, the attention mechanism of the pretrained Bert model is utilized to obtain text-aware image representations. Finally, the MNER task is performed by a Conditional Random Field (CRF) decoder.

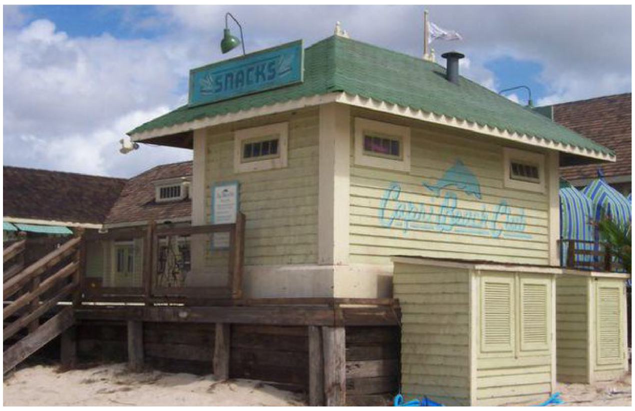
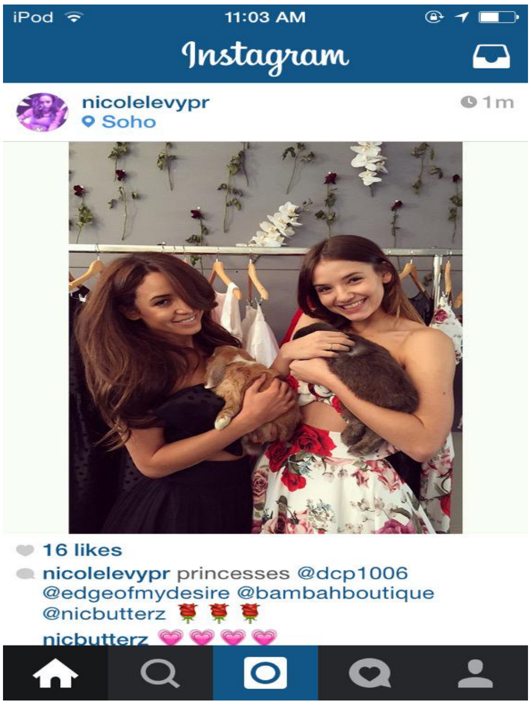


Figure 2. (Aquamarine MISC) (2006).

The main contributions of this paper can be summarized as follows.

Firstly, we propose a visual clue guidance and consistency matching framework for the MNER task, which, through the visual clue guidance (VCG) module, reduces the effect of mismatched text–image pairs using a cross-scale attention (CSA) module. This enriches the information of visual modality. In addition, a consistency matching (CM) module is used to make the representation between images and text more consistent. Secondly, the three modules we proposed (VCG, CSA, and CM), which do not require additional data annotation, can be extended to other multimodal tasks. Finally, experiments on two publicly available datasets, Twitter-2015 and Twitter-2017, show that an excellent performance is achieved, outperforming powerful existing models and achieving F1 scores of 75.81% and 87.11%, respectively. We also conducted ablation studies, case studies, and further analysis to show that the VCG module, CSA module, and CM module play an important role in our framework.



nicolelevypr princesses @dcp1006 @edgeofmydesire @bambahboutique @nicbutterz 🌹🌹🌹

nicbutterz 💗💗💗💗

Figure 3. RT @theteamof1D: (MISC Danielle) and (MISC Melissa) are friends now lmaoooo.

2. Related Work

In this section, we review and summarize the works most relevant to our study.

Starting from [2–4], multimodal named entity recognition has become an important research direction in named entity recognition (NER), which significantly extends the traditional text-based NER by using images as additional inputs [1]. The key challenge is to combine text representation with image representation. Reference [2] proposes an LSTM-CNN architecture that combines text and image information through a generic modal attention module. The authors of [3] propose an adaptive co-attention network to dynamically control the combination of text representation and image representation. The authors of [4] propose an attention-based model to extract image features from the regions in the image most related to the text and use a gate to combine text features and image features. Meanwhile, refs. [2–4] propose methods that can integrate text and image information and improve the accuracy of entity recognition. However, they have the disadvantages of ignoring the bias brought about by the visual context and insensitivity to visual context information. In order to solve the above problems, ref. [5] proposes an end-to-end model to learn joint representations of text and images using multidimensional self-attention techniques. Ref. [6] utilizes object labels as embeddings to achieve a bridge between the visual and the verbal, introducing an intensive co-attention mechanism for fine-grained interactions. The authors of [7] pioneered the use of Transformers in multimodal tasks, and they proposed a multimodal interaction module that acquires word representations and visual representations of images and utilizes plain text entity span detection as an auxiliary module to mitigate visual bias. Ref. [1] proposes a unified multimodal graph fusion (UMGF) approach, which represents input sentences and images as a unified multimodal graph, stacks multiple graph-based multimodal fusion layers, and learns node representations by iteratively performing semantic interactions. However, refs. [1,5–7] mitigated to some extent the bias caused by the visual context and the problem of insensitivity to visual contextual information. However, they suffer from two shortcomings. First, these methods assume that each text and its accompanying image are matched and that the image can be used to help named entity recognition. In addition, they cannot construct a consistent representation to bridge the semantic gap between the two modalities. Other unique ideas, e.g., the R-GCN model, were designed to model the external matching relationship between different (text, image) pairs [12], which is utilized to mitigate image noise in the MNER task by using the external matching relationship between different (text, image) pairs within the dataset and to model both cross-modal and endo-modal relationships. The authors of [10] used retrieval to inject relevant knowledge information from input text and images into a knowledge corpus into the MNER and Multimodal Relation Extraction (MRE) tasks, and they proposed the Mixed Expert (MoE) for MNER and MRE algorithms to combine the predictions of text and image models and make a final decision, which has rarely been seen in previous research. Disappointingly, refs. [10,12] also do not effectively solve the two problems mentioned above.

Therefore, we propose a visual clue guidance and consistency matching framework (GMF) that can effectively reduce the impact of mismatched text–image pairs, enrich the information of visual modalities, and make the representation between images and text more consistent.

3. Methodology

In this section, we will provide a detailed exposition of the GMF's application in multimodal named entity recognition (MNER). Prior to elucidating our proposed approach, we will commence with an overview of the MNER task.

3.1. Task Definition

The multimodal named entity recognition (MNER) task is designed to extract and categorize specific entities from a given sentence S along with its corresponding image I. The central challenge of this task is to accurately assign each entity to predefined categories. Drawing from existing research, MNER is conceptualized as a sequence-labeling task. Specifically, $S = (s_1, s_2, \dots, s_n)$ describes the composition of the sentence, where each s_n represents the nth word. Simultaneously, we use $Y = (y_1, y_2, \dots, y_n)$ to denote the entity labels associated with each word s_n in S. These labels adhere to the BIO2 annotation scheme [19].

3.2. Framework

Our visual clue guidance and consistency matching framework (GMF) is illustrated in Figure 4 and is composed of five key components: (1) input representations, (2) a visual clue guidance module, (3) a cross-scale attention module, (4) a consistency matching module, and (5) a CRF decoder. Subsequently, we will commence with an exploration of the input data transformation methods and then delve into the functionality and roles of each module. Finally, we will provide a detailed description of the training strategy for the MNER task.

3.3. Input Representations

3.3.1. Text Feature Extraction

To capture the deep semantic information in the text, we opted for BERT as the text encoder. Before presenting the sentence to BERT, we introduced specific tokens at the beginning and end positions of the sentence, namely [CLS] and [SEP]. Thus, the extended sentence representation is denoted as $S' = (s_0, s_1, s_2, \dots, s_n, s_{n+1})$, where s_0 represents the [CLS] token and s_{n+1} represents the [SEP] token. After passing S' to BERT, we obtained the encoded output sequence $C = (c_0, c_1, c_2, \dots, c_n, c_{n+1})$. Subsequently, we utilized a fully connected layer with a Tanh activation function to process c_0, obtaining the overall text representation C_g.

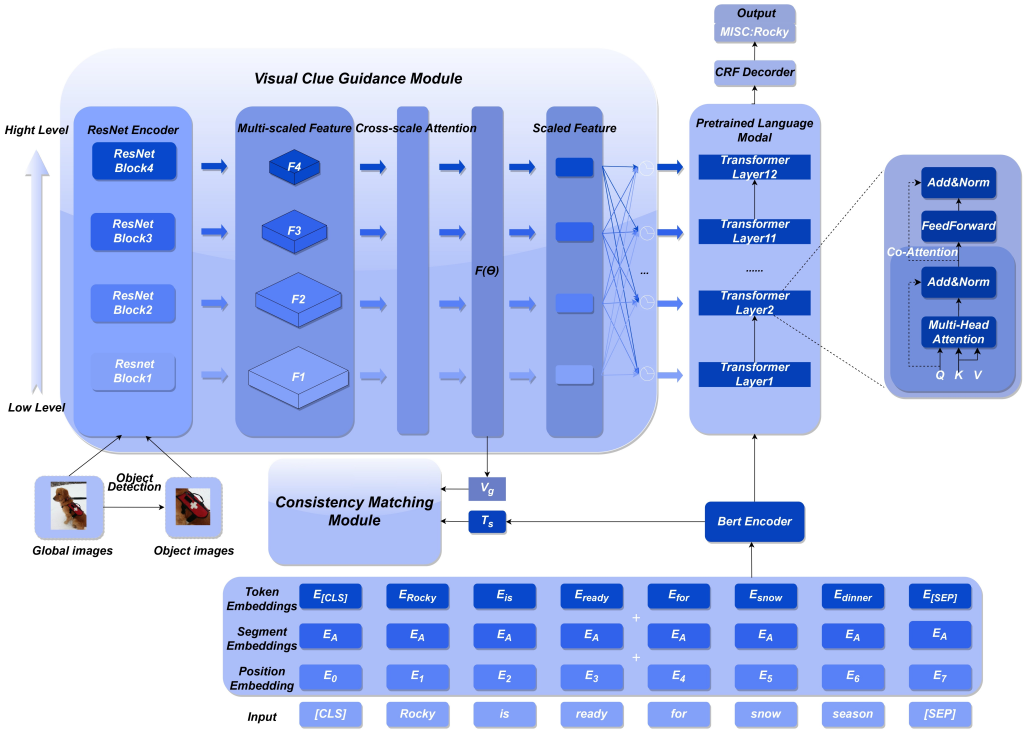


Figure 4. Overall architecture of GMF.

3.3.2. Image Feature Extraction

Given an image, we followed the approach outlined by [1,20] to first employ a visual localization toolkit to locate the top salient local objects in the image. On this basis, we adjusted these local visual objects and the global image to a consistent size of 224 × 224 pixels, serving as the global image I and visual objects $O = (o_0, o_1, o_2, \ldots, o_n)$. For image processing, we chose ResNet50 to obtain the regional and global representations of the image. The regional representation $D{=}(d_1, d_2, \ldots, d_{48}, d_{49})$ comes from the last convolutional layer of ResNet, with dimensions $2048 \times 7 \times 7$, where $7 \times 7 = 49$ denotes the number of regions in the image and 2048 is the dimension of each region's representation, with each region's size being 32×32 pixels. We utilized a 7×7 average pooling layer to obtain the global representation $V_g \in \mathbb{R}^{2048}$, representing the entire image.

3.4. Visual Clue Guidance Module

The visual clue guidance module serves two primary purposes. Firstly, it captures multiple visually relevant objects associated with the text in a mutually aligned text–image pair, thereby enhancing the semantic knowledge for information extraction. Secondly, global image features often encompass abstract conceptual information, providing the model with a weak learning signal. In light of this fact, we aggregate various visual clues to strengthen the recognition of multimodal named entities. This strategy not only incorporates local region features as crucial clues but also introduces the global image as auxiliary information. Therefore, as illustrated in Section 3.3.2, we employed the aforementioned strategy to extract image features in this study. In computer vision research, the fusion strategy of features from different blocks, implemented through pretrained models [21–23], has been proven as an effective method to enhance model performances. To fully leverage these research findings, we particularly focus on exploring the potential of applying feature pyramids in multimodal scenarios. For this purpose, we propose embedding layered image features in each Transformer layer. Initially, encoding the image based on the above strategy generates a collection of pyramid feature maps at different scales $\{F_1, F_2, \ldots, F_b\}$, as shown in Figure 5.

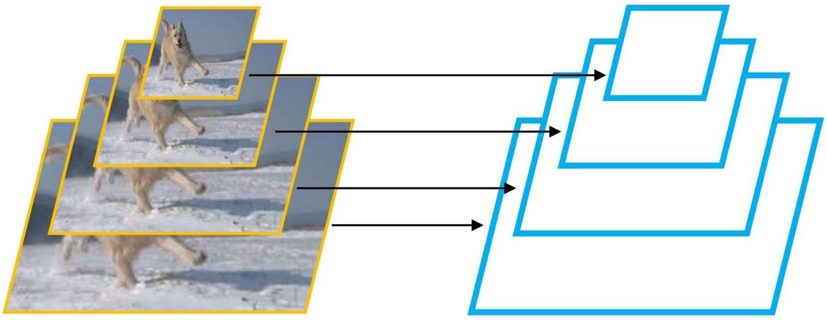


Figure 5. Pyramid feature maps.

3.4.1. Cross-Scale Attention Module

The detailed structure of the cross-scale attention (CSA) module is depicted in Figure 6. Given an input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, CSA infers a 3D attention map $\mathbf{M}(\mathbf{F}) \in \mathbb{R}^{C\times H\times W}$. The refined feature map $\mathbf{F}'$ is computed as follows:

$$\mathbf{F}' = \mathbf{F} + \mathbf{F} \otimes \mathbf{M}(\mathbf{F}) \tag{1}$$

where $\otimes$ denotes element-wise multiplication. We employed a residual learning scheme combined with attention mechanisms to facilitate gradient flow. To design an efficient yet powerful module, we initially computed channel attention $\mathbf{M_c}(\mathbf{F}) \in \mathbb{R}^{C}$ and spatial attention $\mathbf{M_s}(\mathbf{F}) \in \mathbb{R}^{H\times W}$ separately in two branches. The attention map $\mathbf{M}(\mathbf{F})$ was then calculated as follows:

$$\mathbf{M}(\mathbf{F}) = \sigma(\mathbf{M_c}(\mathbf{F}) + \mathbf{M_s}(\mathbf{F})) \tag{2}$$

where σ is the sigmoid function. The outputs from both branches were resized to $\mathbb{R}^{C\times H\times W}$ before being added.

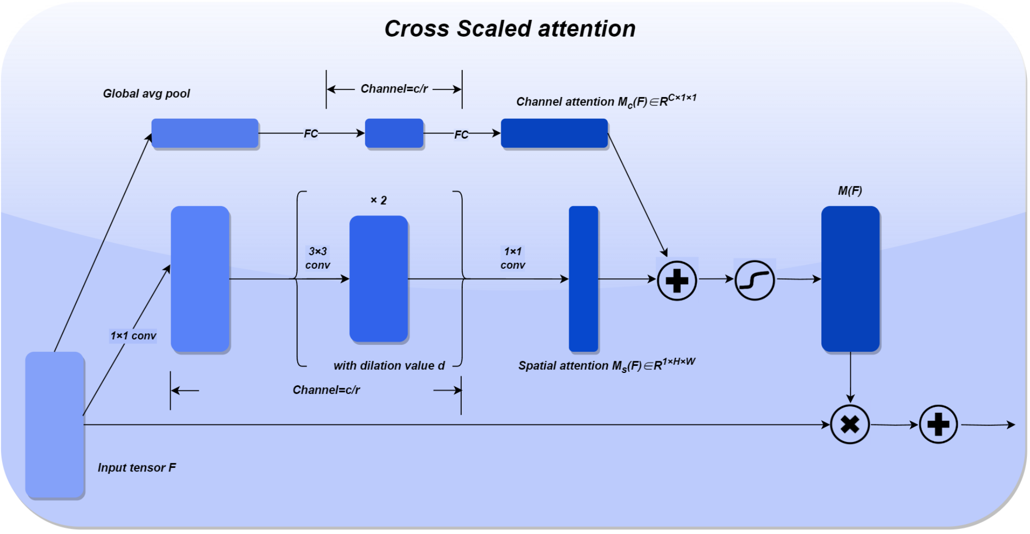


Figure 6. Structure diagram of the cross-scale attention module.

Channel attention branch. We leverage relationships between channels since each channel contains specific feature responses. To aggregate feature maps in each channel, we apply global average pooling to the feature map $\mathbf{F}$, generating a channel vector $\mathbf{F_c} \in \mathbb{R}^{C\times 1\times 1}$. This vector softly encodes global information across all channels. To estimate cross-channel attention from the channel vector $\mathbf{F_c}$, we employ a multilayer perceptron (MLP) with a single hidden layer. To save on parameter costs, the hidden activation size is set to $\mathbb{R}^{C/r\times 1\times 1}$, where r is the reduction ratio. Following the MLP, we add a batch normalization (BN) layer [24] to scale the output in proportion to the spatial branch. In summary, the channel attention computation is as follows:

$$
\begin{aligned}
\mathbf{M_c}(\mathbf{F}) &= BN(MLP(AvgPool(\mathbf{F}))) \\
&= BN(\mathbf{W_1}(\mathbf{W_0}AvgPool(\mathbf{F}) + \mathbf{b_0}) + \mathbf{b_1})
\end{aligned} \tag{3}
$$

Spatial attention branch. A spatial attention map $\mathbf{M_s}(\mathbf{F}) \in \mathbf{R}^{H\times W}$ is generated to emphasize or suppress features at different spatial positions. It is crucial to utilize contextual information to determine which spatial locations should be attended to. Effectively utilizing contextual information requires a larger receptive field. We employ dilated convolution [25] to efficiently expand the receptive field, as dilated convolution is known to construct spatial mappings more effectively than standard convolution. Our spatial branch adopts the "bottleneck structure" recommended by ResNet [16], reducing the parameter count and computational cost. Specifically, the feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ is projected to the reduced dimension $\mathbb{R}^{C/r\times H\times W}$ through a 1 × 1 convolution, integrating and compressing the feature map in the channel dimension. For simplicity, we use the same reduction ratio r as the channel branch. After reduction, two 3 × 3 dilated convolutions are applied to effectively leverage contextual information. Finally, the features are reduced to the spatial attention map $\mathbb{R}^{1\times H\times W}$ again through a 1 × 1 convolution. To adjust to the scale, a batch normalization layer is applied at the end of the spatial branch. In summary, the spatial attention computation is as follows:

$$
\mathbf{M_s}(\mathbf{F}) = BN\left(f_3^{1\times 1}\left(f_2^{3\times 3}\left(f_1^{3\times 3}\left(f_0^{1\times 1}(\mathbf{F})\right)\right)\right)\right) \tag{4}
$$

where f denotes a convolution operation, BN denotes a batch normalization operation, and the superscript denotes the size of the convolution filter. There are two 1 × 1 convolutions for channel reduction. The intermediate 3 × 3 dilated convolutions are applied to aggregate contextual information with a larger receptive field, $\mathbf{F}^{'} = \mathbf{F} + \mathbf{F} \otimes \mathbf{M}(\mathbf{F})$, which allows a model with stronger information perception for the image $\left\{F_1^{'}, F_2^{'}, \ldots, F_b^{'}\right\}$ to be obtained. Next, we adopt a structured mapping function $F(\theta)$ for feature fusion, including $Conv_{1\times 1}$, $Pool$, and Concat, represented as follows:

$$
V_b = \mathrm{Conv}_{1\times 1}(F_b^{\prime}) \tag{5}
$$

$$
V_m = \mathrm{Conv}_{1\times 1}(\mathrm{Pool}(F_m^{\prime})), m = 1, 2, b-1 \tag{6}
$$

where m represents the m-th block of the main model and b represents the number of blocks in the visual backbone model (in this case, ResNet has four blocks). $Conv_{1\times 1}$ denotes the 1 × 1 convolution operation used to adjust the number of channels, and the Pool operation ensures that all features are mapped to the same spatial dimension.

3.4.2. Dynamic Gating

Understanding that appropriate feature representations appear at corresponding scales of differently sized objects, the decision of mapping which module from ResNet50 to each layer of the Transformer is not straightforward. To address this challenge, we introduce a dense routing structure, enabling the hierarchical multiscale visual features to interact with each layer of the Transformer. The motivation behind dynamic gating is to predict a normalization vector, describing the importance of features from each visual block corresponding to a specific layer in the Transformer, for path selection. In the dynamic gating module, $g_i^{(l)} \in [0, 1]$ represents the path probability from the i-th block in ResNet50 to the l-th layer of the Transformer. The computation formula for the gate signal is given by $g^{(l)} = \mathbb{G}^{(l)}(V) \in \mathbb{R}^d$, where $\mathbb{G}^{(l)}(\cdot)$ is the gating function for the l-th layer of the Transformer and d is the number of modules in ResNet50. Now, let us discuss the logit of the gating signal.

$$
\gamma^{(l)} = \sigma\left(W_p\left(\frac{1}{d}\sum_{i=1}^{d} P(V_m)\right)\right) \tag{7}
$$

where $\sigma(\cdot)$ is the Leaky_ReLU activation function and P is the global average pooling layer. Initially, we perform average pooling on the features V_m obtained from the i-th block (with a shape of (d_i, h_i, w)). Next, we accumulate these block features to obtain an average vector. Through the MLP layer W_p, we adjust the dimensions of these features to d while considering a soft gating mechanism, generating continuous values as path probabilities. Furthermore, we obtain the probability vector $g^{(l)}$ for the l-th layer of the Transformer:

$$g^{(l)} = \text{Softmax}\left(\gamma^{(l)}\right) \tag{8}$$

3.4.3. Hierarchical Feature Fusion

Based on the dynamic gating signal $g^{(l)}$, we form the fused hierarchical visual features $V^{(l)}_{gated}$ corresponding to the l-th layer of the Transformer:

$$V^{(l)}_{\text{gated}} = g^{(l)} V^{(l)} \tag{9}$$

To more precisely represent the visual features $\tilde{V}^{(l)}_{gated}$ that correspond to the final l-th layer of the Transformer, we perform the following concatenation operation:

$$\tilde{V}^{(l)}_{\text{gated}} = \left[V^{(l,I)}_{\text{gated}}; V^{(l,o_1)}_{\text{gated}}; \ldots; V^{(l,o_n)}_{\text{gated}}\right] \tag{10}$$

This structure enhances the hierarchical representation of the text modality through the attention mechanism based on visual clue guidance.

3.4.4. Visual Clue Guidance

We describe the hierarchical multiscale visual features as visual clue guidance and insert these sequences of visual clue guidance into the text sequence in the self-attention layer of BERT. Specifically, for a given input sequence $S = (s_1, s_2, \ldots, s_n)$, the context representation $C \in \mathbb{R}^{n \times d}$ first undergoes linear projection to obtain the Q, K, and V vectors:

$$\boldsymbol{Q}^l = C\boldsymbol{W}_l^Q, \boldsymbol{K}^l = C\boldsymbol{W}_l^K, \boldsymbol{V}^l = C\boldsymbol{W}_l^V \tag{11}$$

For the integrated hierarchical visual features $\tilde{V}^{(l)}_{gated}$, we use a linear transformation $W_l^{\phi} \in \mathbb{R}^{d \times 2 \times d}$ (for the l-th layer) to project it into the embedding space matching the text representation. Furthermore, we define the operations for visual cues $\phi_k^l, \phi_v^l \in \mathbb{R}^{hw(m+1) \times d}$ as follows:

$$\left\{\phi_k^l, \phi_v^l\right\} = \tilde{V}^{(l)}_{\text{gated}} W_l^{\phi} \tag{12}$$

where $hw(m+1)$ represents the length of the visual sequence and m denotes the number of visual objects detected by the object detection algorithm. Based on the visual clue attention mechanism, the calculation is as follows:

$$\mathcal{P}refix\_Attention^l = softmax(\frac{Q^l[\phi_k^l; K^l]^T}{\sqrt{d}})[\phi_v^l; V^l] \tag{13}$$

We take the hierarchical multiscale visual features as visual cue prompts for each fusion layer and update the final hidden representation $H = (h_0, h_1, \ldots, h_{n+1})$ after text and image fusion through the multimodal attention mechanism layer by layer.

3.5. Consistency Matching Module

To address the inconsistency issue when integrating two modalities in the previous model, we designed a cross-modal consistency matching module to ensure the consistency of text and image representations. Inspired by contrastive learning [26–28], we constructed the consistency matching module. The module takes the text representation C_g and the global representation of the image P_g as inputs, and the overall process can be summarized in the following three steps. First, in the input pairs (C_g, P_g) with a batch size of N, we generate positive and negative samples. Here, C_g^m represents the text representation for the m-th pair in the batch, and P_g^n is the image representation for the n-th pair. We define positive samples as the text and image representations from the same data pair $\{(C_g^m, P_g^n)_{m=n}\}$, while negative samples are selected from different data pairs $\{(C_g^m, P_g^n)_{m \neq n}\}$. Although there might be a small number of mismatched positive samples, the literature [3] suggests that their impact can be negligible. Next, for each (C_g^m, P_g^n) example, we process them using two independent multilayer perceptrons (MLPs) to obtain text representation $C_n \in \mathbb{R}^d$ and image representation $P_n \in \mathbb{R}^d$. This MLP processing technique effectively assists the encoder in obtaining better representations, aligning with the results in the studies. Finally, we adjust the similarity of positive and negative samples by minimizing two contrastive loss functions. Specifically, the contrastive loss functions from image to text and text to image are defined as follows:

For the i-th positive sample, the loss from the image to the text is

$$\mathcal{L}_i^{(P_n \to C_n)} = -\log \frac{\exp\left(\frac{\left(P_n^i\right)^T C_n^i / \|P_n^i\| \|C_n^i\|}{\tau}\right)}{\sum_{k=1}^{N} \exp\left(\frac{\left(P_n^i\right)^T C_n^k / \|P_n^i\| \|C_n^k\|}{\tau}\right)} \tag{14}$$

where $(P_n^i)^T C_n^i / \parallel P_n^i \parallel\parallel C_n^i \parallel$ is the cosine similarity between P_n^i and $\backslash C_n^i$ and τ is the temperature coefficient. Similarly, for the i-th positive sample, the loss from the text to the image is

$$\mathcal{L}_i^{(C_n \to P_n)} = -\log \frac{\exp\left(\frac{\left(C_n^i\right)^T P_n^i / \|C_n^i\| \|P_n^i\|}{\tau}\right)}{\sum_{k=1}^{N} \exp\left(\frac{\left(C_n^i\right)^T P_n^k / \|C_n^i\| \|P_n^k\|}{\tau}\right)} \tag{15}$$

Finally, integrating all the losses for positive samples, we define the total loss as

$$\mathcal{L}_{CM\_\text{LosS}} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \mathcal{L}_i^{(P_n \to C_n)} + (1-\lambda) \mathcal{L}_i^{(C_n \to P_n)} \right) \tag{16}$$

In this equation, $\lambda \in [0, 1]$ is set as a hyperparameter. Minimizing this loss function ensures the representations of both modalities are more consistent.

3.6. CRF Decoder

After incorporating the contextual information from the image, we employ a Conditional Random Field (CRF) as the decoder for the MNER task. The CRF decoder takes the final hidden representation $\boldsymbol{H}$ described in Section 3.5 and combines it with the original text $\boldsymbol{S}$ mentioned in Section 3.1 and the corresponding image $\boldsymbol{I}$ to predict the conditional probability of the sequence $\boldsymbol{y}$. Specifically, this probability is defined as follows:

$$P(\boldsymbol{y} \mid S, I) = \frac{\exp(\text{point}(\boldsymbol{H}, y))}{Z(\boldsymbol{H})} \tag{17}$$

$$Z(H) = \sum_{y} \exp\left(\sum_{i=1}^{n} E_{h_i, y_i} + \sum_{i=0}^{n} T_{y_i, y_{i+1}} \right) \tag{18}$$

$$\text{point}(H, y) = \sum_{i=1}^{n} E_{h_i, y_i} + \sum_{i=0}^{n} T_{y_i, y_{i+1}} \tag{19}$$

$$E_{h_i y_i} = w_{MNER}^{y_i} \cdot h_i \tag{20}$$

where E_{h_i,y_i} and $T_{y_i,y_{i+1}}$, respectively, denote the emission score for the i-th label y_i and the transition score from y_i to y_{i+1}. $w_{MNER}^{y_i}$ represents the weight parameter for y_i, and the normalization factor $Z(H)$ is used to sum and normalize the emission and transition scores for all possible sequences $\boldsymbol{y}$. To train this module, we utilize the logarithmic likelihood loss function:

$$\mathcal{L}_{\text{mner}} = -\frac{1}{|D_{\text{mner}}|} \sum_{j=1}^{N} \left(\log P\left(\boldsymbol{y}^j \mid S^j, I^j \right) \right) \tag{21}$$

where $D_{\boldsymbol{mner}} = \{S^j, I^j, \boldsymbol{y}^j\}_{j=1}^{N}$ denotes the training data batch.

3.7. Model Training

In summary, our framework encompasses both a supervised learning task (MNER) and an auxiliary self-supervised learning task (consistency matching). To achieve joint training and optimization, we formulate the comprehensive loss function as follows:

$$\mathcal{L}_{Total\_Loss} = (1-\alpha)\mathcal{L}_{CM\_Loss} + \alpha L_{mner} \tag{22}$$

where $\mathcal{L}_{CM\_Loss}$ represents the loss of the consistency matching module (see Section 3.5) and $\mathcal{L}_{mner}$ corresponds to the loss of the MNER task (see Section 3.6). In this equation, α is a hyperparameter that needs to be fine-tuned.

4. Experiments

4.1. Dataset

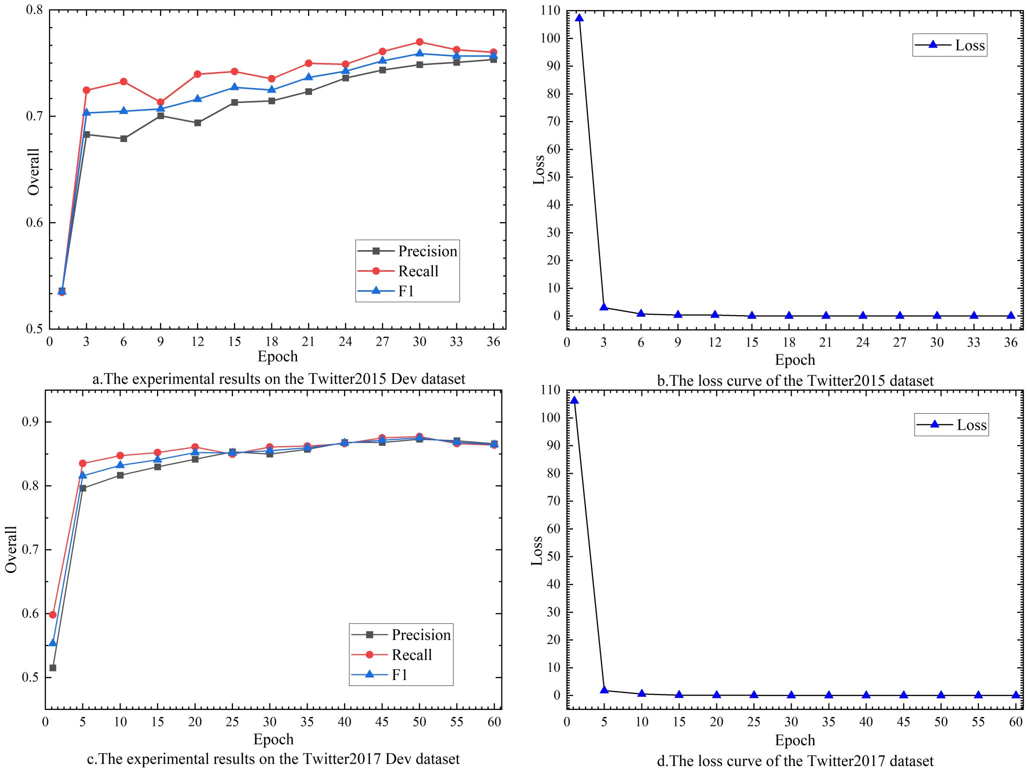
In the evaluation phase of this study, we utilized two widely adopted datasets, namely Twitter-2015 and Twitter-2017, which were provided by [3,4], respectively. Each data sample consists of a {sentence, image} pair. It is noteworthy that due to the absence of the image modality in certain samples during the collection process, we took measures to ensure the data integrity and consistency. For these instances lacking image information, we uniformly substituted them with predefined blank images. Specific details regarding data distribution and the quantities of various entity classes can be found in Table 1.

Table 1. Summary statistics of the two MNER datasets.

Type	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
PER	2217	552	1816	2943	626	621
LOC	2091	522	1697	731	173	178
ORG	928	247	839	1674	375	395
MISC	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
# Tweets	4000	1000	3257	3373	723	723

4.2. Parameter Settings

In the experimental phase, we used an NVIDIA GTX 3090 GPU and PyTorch version 1.13.1 for all relevant tasks and experiments. To ensure the accuracy and efficiency of the framework, we set specific parameters and configuration strategies for each module. For text encoding, we chose bert-base-uncased as the text encoder for our model. In the image processing phase, we used ResNet50 as the image encoder. In addition, considering two different Twitter datasets, we used the following hyperparameter settings: for the Twitter-2015 dataset, the temperature coefficient is 0.21, and for the Twitter-2017 dataset, the temperature coefficient is 0.174. In addition, our uniform hyperparameter setting is 0.7. Regarding the training strategy and details, for the Twitter- 2015 dataset, we set the training batch size to 16, the learning rate to 5×10^{-5}, and the random seed to 1234. As for the Twitter-2017 dataset, we adjusted the training batch size to 32, and, again, the learning rate was kept at 5×10^{-5}, and the random seed remained at 1234. In order to set up the most appropriate epoch, we refer to Figure 7. These parameters and configuration choices were made to ensure that our model achieves an optimal performance and accuracy on the given dataset. Finally, we set the epochs of Twitter-2015 and Twitter-2017 to 30 and 50, respectively.



a.The experimental results on the Twitter2015 Dev dataset

b.The loss curve of the Twitter2015 dataset

c.The experimental results on the Twitter2017 Dev dataset

d.The loss curve of the Twitter2017 dataset

Figure 7. GMF test results on the development set and loss on the training set.

4.3. Baselines

To comprehensively evaluate the effectiveness of our proposed visual clue guidance and consistency matching framework (GMF), we selected representative text-based and multimodal NER models for comparison. Text-based NER methods include BiLSTM-CRF [29], a classical NER model employing a bidirectional LSTM structure and an CRF layer; HBiLSTM-CRF [30], a variant of CNN-BiLSTM-CRF that replaces the CNN layer with LSTM for character-level word representations; BERT [19], serving as a multilayer bidirectional Transformer encoder followed by a softmax decoder; and BERT-CRF, similar to BERT but using a CRF decoder for label prediction; and T-NER [3,31], a NER system designed specifically for tweets, utilizing a range of effective features, including dictionaries, context, and positive word features. Multimodal NER methods include GVATT-HBiLSTM-CRF [4], which combines HBiLSTM-CRF with an attention mechanism to obtain representations merging text and visual information, and UMT-BERT-CRF [7], a leading multimodal NER model comprising a multimodal interaction module and an auxiliary module for pure text entity span detection. Additionally, we include other state-of-the-art multimodal models for comparison: MAF [32], which calculates similarity scores between text and images and uses the score to determine the proportion of visual information to retain; UMGF [1], which proposes a unified multimodal graph fusion method to capture fine-grained semantic correspondences between semantic units of different modalities; BFCL [33], which leverages a Transformer-based bottleneck fusion mechanism to reduce noise propagation in the visual modality; and DebiasCL [34], which utilizes hard sample mining and debiased contrastive loss to alleviate biases in both the quantity and entity types, enabling global learning for aligning text and image feature spaces. Lastly, GMF, the model proposed in this study, aims to enhance the MNER performance by integrating visual clue guidance and consistency matching.

4.4. Effectiveness

This thesis uses precision (P), recall (R), and the F1 score (F1) as metrics tested on two benchmark MNER datasets. For an entity, it is judged to be correctly predicted only if both its boundaries and categories are correctly predicted. Precision is the proportion of correct entities among the predicted entities, and recall is the proportion of correctly predicted entities among all entities in the sample. A high precision rate indicates that the model predicted a high percentage of correct entities but did not necessarily cover more entities. A high recall indicates that the model predicted most of the entities in the data; it covered more entities but did not necessarily predict a high proportion of correct entities. The F1 value is a weighted average of the precision and recall, taking into account both the precision of the prediction and the number of correct entities covered. Table 2 displays the performance of four text-based models and eight multimodal models on Twitter-2015, while Table 3 presents their performance on Twitter-2017. Below is a detailed analysis.

Table 2. Experimental results on the Twitter-2015 dataset.

Modality	**Methods**	**Twitter-2015**						
		Single Type (F1)				**Overall**		
		PER	**LOC**	**ORG**	**MISC**	**P**	**R**	**F1**
Text	BiLSTM-CRF	76.77	72.56	41.33	26.8	68.14	61.09	64.42
	HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17
	BERT	84.72	79.91	58.26	38.81	68.3	74.61	71.32
	BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81
Text+image	GVATT-HBiLSTM-CRF	82.66	77.21	55.06	35.25	73.96	67.9	70.8
	GVATT-BERT-CRF	84.43	80.87	59.02	38.14	69.15	74.46	71.7
	BFCL	85.6	81.77	63.81	40.3	74.02	75.07	74.54
	UMT-BERT-CRF	85.24	81.58	63.03	39.45	71.67	75.23	73.41
	MAF	84.67	81.18	63.35	41.82	71.86	75.1	73.42
	UMGF	84.26	**83.17**	62.45	42.42	74.49	75.21	74.85
	DebiasCL	85.97	81.84	**64.02**	43.38	74.45	76.13	75.28
	GMF (ours)	**87.02**	82.51	63.25	**45.2**	**75.47**	**76.15**	**75.81**

Firstly, comparing all text-based NER methods, it is evident from the data that methods based on BERT outperform others, indicating the potential advantage of optimizing pretrained models through transfer learning rather than starting from scratch. Additionally, approaches combining BERT with the CRF surpass pure BERT strategies, suggesting the crucial role of the CRF in capturing label constraints and thereby achieving more accurate label predictions.

Next, we compare MNER methods with competitive text-based models, such as GVATT-HBiLSTM-CRF and HBiLSTM-CRF. The majority of multimodal strategies notably outperform their corresponding text-only models, further indicating the contribution of image information to named entity recognition in text. Furthermore, compared to the latest MNER method, DebiasCL, our GMF method, incorporating both the proposed visual guidance module and consistency matching module, enhances image information aligned with the text.

Table 3. Experimental results on the Twitter-2017 dataset.

Modality	Methods	Twitter-2017						
		Single Type (F1)				Overall		
		PER	LOC	ORG	MISC	P	R	F1
Text	BiLSTM-CRF	85.12	72.68	72.5	52.56	79.42	73.43	76.31
	HBiLSTM-CRF	87.91	78.57	76.67	59.32	82.69	78.16	80.37
	BERT	90.88	84	79.25	64.63	82.19	83.72	82.95
	BERT-CRF	70.25	83.05	81.13	62.21	83.32	83.57	83.44
Text+image	GVATT-HBiLSTM-CRF	89.34	78.53	79.12	62.21	83.41	80.38	81.87
	GVATT-BERT-CRF	90.94	83.52	81.91	62.75	83.64	84.38	84.01
	BFCL	91.17	**86.43**	83.97	66.67	85.99	85.42	85.7
	UMT-BERT-CRF	91.56	84.73	82.24	70.1	85.28	85.34	85.31
	MAF	91.51	85.8	**85.1**	68.79	86.13	86.38	86.25
	UMGF	84.26	83.17	62.45	42.42	86.54	84.50	85.51
	DebiasCL	**93.46**	84.15	84.42	67.88	**87.59**	86.11	86.84
	GMF (ours)	93.42	84.68	84.83	**69.42**	86.95	**87.27**	**87.11**

Finally, on Twitter-2015 and Twitter-2017, our model shows overall F1 improvements of 0.53% and 0.27%, respectively. This suggests that the introduced modules can assist the model in better integrating text and image representations.

4.5. Ablation Study

In order to evaluate the contribution of the visual clue guidance module (VCG) cross-scale attention module (CSA), and consistency matching module (CM) to the overall performance of the GMF model, we conducted detailed culling experiments. The data in Table 4 show the following: (1) Overall, the visual clue guidance module, cross-scale attention module, and consistency matching module all contribute significantly to the performance of the GMF model. (2) In the absence of the visual clue guidance module (labeled "w/o VCG"), the F1 score for the Twitter-2015 dataset decreased by 0.74%, while that for the Twitter-2017 dataset decreased by 0.66%. The critical role of the visual clue guidance module in integrating and enhancing visual features is emphasized. (3) In the absence of the cross-scale attention module (labeled "w/o CSA"), the F1 scores of the Twitter-2015 dataset decreased by 0.58%, while the Twitter-2017 dataset decreased by 0.5%, and both F1 scores are smaller than w/o VCG and w/o CM, suggesting that CSA plays a role in GMF to assist other modules. (4) In the case of removing the consistency matching module (labeled "w/o CM"), the F1 scores of the Twitter-2015 dataset decreased by 0.87%, while that of the Twitter-2017 dataset decreased by 0.78%. This indicates that the consistency matching module effectively aligns the text with the relevant image regions. (5) With the removal of the visual clue guidance module, the cross-scale attention module, and the consistency matching module (labeled "w/o VCG + CSA + CM"), the F1 score for the Twitter-2015 dataset decreased by 1.18%, while that for the Twitter-2017 dataset decreased by 0.99%. It is shown that the visual clue guidance module, the cross-scale attention module, and the consistency matching module alleviate the problems caused by text–image mismatch, enrich the information of the visual modality, and ensure a high degree of consistency of modal representations between text and images.

Table 4. Ablation studies of the GMF model.

Methods	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
GMF	**0.7547**	**0.7615**	**0.7581**	**0.8695**	0.8727	**0.8711**
w/o VCG	0.7422	0.7596	0.7508	0.8557	**0.8734**	0.8645
w/o CSA	0.7504	0.7542	0.7523	0.861	0.8712	0.8661
w/o CM	0.749	0.7499	0.7494	0.8549	0.8719	0.8633
w/o VCG + CSA + CM	0.7347	0.7583	0.7463	0.8615	0.8608	0.8612

4.6. Case Study

Figure 8 presents a case study showcasing the effectiveness of our proposed GMF and DebiasCL methods. In the first scenario, DebiasCL and GMF w/o CM misclassify the PER entity "Isis" as an ORG entity, whereas our methods GMF and GMF w/o VCG accurately identify it. Similarly, in the second scenario, DebiasCL and GMF w/o CM fail to recognize the MISC entity "Oscars", while our methods GMF and GMF w/o VCG do so correctly. These examples underscore the efficacy of the CM module in associating entities in the text with relevant image regions, thereby enhancing the accuracy of entity recognition. In the third scenario, DebiasCL and GMF w/o VCG incorrectly classify the MISC entity "Mufasa" as a PER entity. This misclassification may be attributed to the metaphorical emotions depicted in the image and the significant noise introduced by the characters, posing challenges for the MNER model. Conversely, our GMF approach, along with the GMF w/o CM modules, accurately predicts the MISC entity "Mufasa". This instance illustrates the effectiveness of the VCG module in filtering image noise and leveraging deep textual semantics for accurate entity prediction. Moving forward, we intend to explore the visual clue guidance module further and explore more advanced strategies for improved entity recognition accuracy. Additionally, we aim to apply our method to other multimodal tasks such as multimodal relationship extraction and multimodal entity linking.

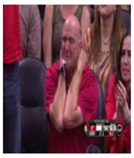
	Importance of the CM Module		**Importance of the VCG Module**
Text-image pairs	MY RELIGION PROHIBITS GAY MARRIAGE RT @samkalidi : **[Isis PER]**^{1} and **[Kim Davis PER]**^{2} have things in common.	 Actor **[Neil Patrick Harris PER]** ^{1} will host the 2015 **[Oscars MISC]**^{2}	 When **[Mufasa MISC]**^{1} dies in the **[Lion King MISC]**^{2}
DebiasCL	1-ORG × 2-PER √	1-PER √ 2-PER ×	1-PER × 2-MISC √
GMF	1-PER √ 2-PER √	1-PER √ 2-MISC √	1-MISC √ 2-MISC √
GMF w/o CM	1-ORG × 2-PER √	1-PER √ 2-PER ×	1-MISC √ 2-MISC √
GMF w/o VCG	1-PER √ 2-PER √	1-PER √ 2-MISC √	1-PER × 2-MISC √

Figure 8. Case study of GMF vs. DebiasCL methodology.

4.7. Further Analysis

In order to better understand the importance of the three main contributions (VCG, CSA, and CM modules) in our proposed GMF approach, we performed additional analysis on both test sets. In Figure 9, we show the number of entities incorrectly/correctly predicted by BERT-CRF and the number of entities correctly/incorrectly predicted by each multimodal approach. First, we can see in Figure 9a,b that our GMF method correctly identifies more entities compared to the three multimodal baselines. In addition, we can see in Figure 9c,d that the GMF method has a lower error probability compared to the three multimodal baselines, and we believe that the reason is that GMF can significantly reduce the bias caused by the visual environment by incorporating our VCG module, and at the same time, with the aid of the auxiliary CSA module, it successfully reduces the inter-scale noise and strengthens the ability of detail capturing using the CM module that narrows the semantic gap between text and images.

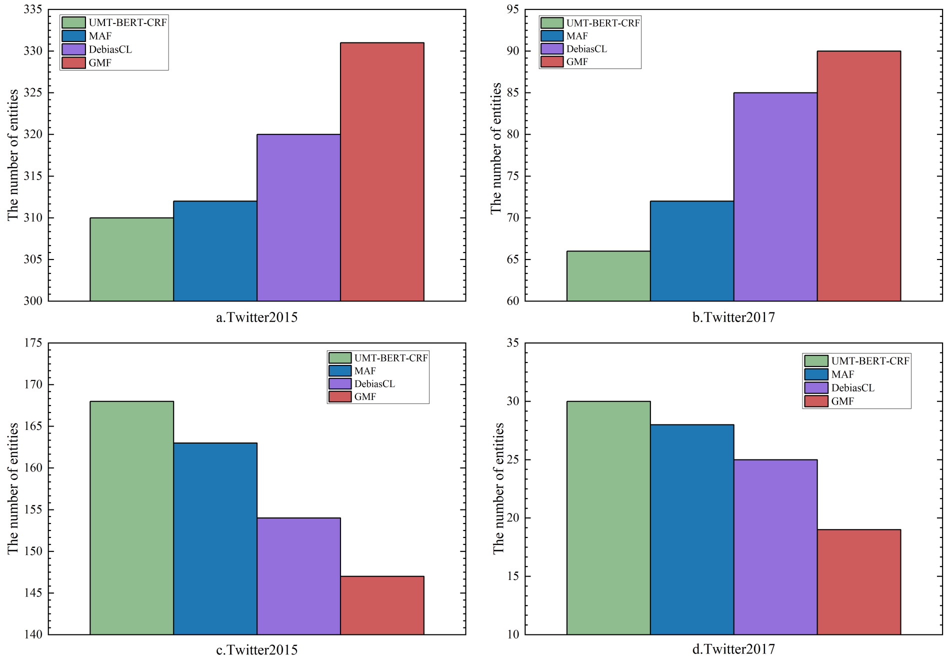


Figure 9. (**a**,**b**) both show the number of entities incorrectly predicted by BERT-CRF but corrected by each multimodal method (shown in the *y*-axis). (**c**,**d**) both show the number of entities correctly predicted by BERT-CRF but incorrectly predicted by each multimodal method (shown in the *y*-axis).

5. Conclusions

In this paper, we propose a visual clue guidance and consistency matching framework (GMF), which improves the state-of-the-art performance of multimodal named entity recognition for social media posts. Specifically, the VCG module is used to combine low-level, high-resolution image information and high-level, strong semantic image information. On this basis, the CSA module is used to enrich the information of visual modalities. In addition, the CM module is used to minimize the contrast loss and align the entities in the text with the most relevant objects in the image. We conducted a number of experiments, ablation studies, case studies, and further analyses to show that the VCG module can help the model filter out most of the image information that is irrelevant to the text and reduce the effect of image mismatch on the text. The CSA module mitigates cross-scale interference and enhances the ability of images to capture details by augmenting visual features. The CM module can help the model establish a connection between the named entities in the text and the regions in the image where the corresponding objects are located and reduce the interactions with other regions in the image.

However, the method proposed in this paper still has some shortcomings. For example, there is a lack of stress testing of the model in different environments. In the future, we plan to increase the dataset of social media posts to better reflect the diversity and uncertainty of the data and to stress test the model at different data distributions, noise levels, and so on. In addition, we would like to apply the methods in this paper to other multimodal tasks, such as multimodal relationship extraction and multimodal entity linking.

Author Contributions: Conceptualization, L.H., Q.W., J.L., J.D. and H.W.; writing—original draft preparation, L.H. and Q.W.; writing—review and editing, L.H.; data curation, Q.W.; validation, J.L.; supervision, L.H., J.L., J.D. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China (2020AAA0109700), National Natural Science Foundation of China (61972003), and Beijing Urban Governance Research Base of the North China University of Technology (2023CSZL16).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: We would like to thank the anonymous reviewers for their helpful comments. We would like to thank the referees for their comments, which helped improve this paper considerably.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, D.; Wei, S.; Li, S.; Wu, H.; Zhu, Q.; Zhou, G. Multi-modal graph fusion for named entity recognition with targeted visual guidance. *AAAI Conf. Artif. Intell.* **2021**, *35*, 14347–14355. [CrossRef]
2. Moon, S.; Neves, L.; Carvalho, V. Multimodal named entity recognition for short social media posts. *arXiv* **2018**, arXiv:1802.07862.
3. Zhang, Q.; Fu, J.; Liu, X.; Huang, X. Adaptive co-attention network for named entity recognition in tweets. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018; Volume 32.
4. Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; Ji, H. 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), Melbourne, Australia, 15–20 July 2018; pp. 1990–1999.
5. Arshad, O.; Gallo, I.; Nawaz, S.; Calefati, A. Aiding intra-text representations with visual context for multimodal named entity recognition. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 337–342.
6. Wu, Z.; Zheng, C.; Cai, Y.; Chen, J.; Leung, H.f.; Li, Q. 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, Seattle, WA, USA, 12–16 October 2020; pp. 1038–1046.
7. Yu, J.; Jiang, J.; Yang, L.; Xia, R. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. *Assoc. Comput. Linguist.* **2020**, 3342–3352.
8. Sun, L.; Wang, J.; Su, Y.; Weng, F.; Sun, Y.; Zheng, Z.; Chen, Y. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 1852–1862.
9. Chen, D.; Li, Z.; Gu, B.; Chen, Z. Multimodal named entity recognition with image attributes and image knowledge. In Proceedings of the Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, 11–14 April 2021; Proceedings, Part II 26, pp. 186–201.
10. Wang, X.; Cai, J.; Jiang, Y.; Xie, P.; Tu, K.; Lu, W. Named entity and relation extraction with multi-modal retrieval. *arXiv* **2022**, arXiv:2212.01612.
11. Liu, P.; Li, H.; Ren, Y.; Liu, J.; Si, S.; Zhu, H.; Sun, L. A Novel Framework for Multimodal Named Entity Recognition with Multi-level Alignments. *arXiv* **2023**, arXiv:2305.08372.
12. Zhao, F.; Li, C.; Wu, Z.; Xing, S.; Dai, X. Learning from Different text-image Pairs: A Relation-enhanced Graph Convolutional Network for Multimodal NER. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3983–3992.
13. Szczepanek, R. A Deep Learning Model of Spatial Distance and Named Entity Recognition (SD-NER) for Flood Mark Text Classification. *Water* **2023**, *15*, 1197. [CrossRef]
14. Liu, Y.; Huang, S.; Li, R.; Yan, N.; Du, Z. USAF: Multimodal Chinese named entity recognition using synthesized acoustic features. *Inf. Process. Manag.* **2023**, *60*, 103290. [CrossRef]
15. Vempala, A.; Preoţiuc-Pietro, D. Categorizing and inferring the relationship between the text and image of twitter posts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2830–2840.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
18. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; Luo, J. A fast and accurate one-stage approach to visual grounding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seattle, WA, USA, 12–16 October 2019; pp. 4683–4693.
21. Wang, T.; Anwer, R.M.; Cholakkal, H.; Khan, F.S.; Pang, Y.; Shao, L. Learning rich features at high-speed for single-shot object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 1971–1980.

22. Kim, S.W.; Kook, H.K.; Sun, J.Y.; Kang, M.C.; Ko, S.J. Parallel feature pyramid network for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 234–250.
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 448–456.
25. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
26. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
27. Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv* **2021**, arXiv:2104.08821.
28. Zhang, H.; Koh, J.Y.; Baldridge, J.; Lee, H.; Yang, Y. Cross-modal contrastive learning for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 833–842.
29. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
30. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. *arXiv* **2016**, arXiv:1603.01360.
31. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: an experimental study. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Scotland, UK, 27–31 July 2011; pp. 1524–1534.
32. Xu, B.; Huang, S.; Sha, C.; Wang, H. MAF: A general matching and alignment framework for multimodal named entity recognition. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event, 21–25 February 2022; pp. 1215–1223.
33. Wang, P.; Chen, X.; Shang, Z.; Ke, W. Multimodal Named Entity Recognition with Bottleneck Fusion and Contrastive Learning. *IEICE Trans. Inf. Syst.* **2023**, *106*, 545–555. [CrossRef]
34. Zhang, X.; Yuan, J.; Li, L.; Liu, J. Reducing the Bias of Visual Objects in Multimodal Named Entity Recognition. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, Singapore, 27 February–3 March 2023; pp. 958–966.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.