Journal of Computers Vol. 37 No. 5, October 2026, pp. 59-74
https://doi.org/10.63367/19911599202610370500 5

Research Progress on Multimodal Named Entity Recognition with Multi-Task Learning

Fangping Chen, Yanbo Jia*, Hairong Wang, and Zaiyuan Zhao

College of Computer Science and Engineering, North Minzu University, Yinchuan, Ningxia 750021, China

20237495@stu.nmu.edu.cn, 2001057@nmu.edu.cn, wanghr@nun.edu.cn, 20247561@stu.nmu.edu.cn

Received 8 July 2025; Revised 7 February 2026; Accepted 22 February 2026

Abstract. Multimodal Named Entity Recognition (MNER) leverages heterogeneous data sources such as text and images to improve the accuracy of named entity recognition. However, MNER performance remains limited due to the semantic gap between visual and textual representations and the bias introduced by ignoring visual context (or incorporating it inappropriately). To address these issues, multi-task learning is adopted to assist MNER from two perspectives: feature extraction and processing, and multimodal fusion, by jointly learning multiple related auxiliary tasks. This paper tracks recent advances in MNER and conducts a comparative analysis of representative mainstream approaches. According to the auxiliary tasks used in multi-task learning, existing methods are categorized into three types: visual optimization network–assisted methods, local matching–assisted methods, and text-assisted multi-task methods. The first two categories enhance model performance by improving the quality of visual features and modeling image–text similarity, respectively, whereas the third category reduces visual bias through auxiliary tasks based on textual signals. Moreover, twelve representative methods, including MFD and SMRG, are evaluated on Twitter-2015 and Twitter-2017. Experimental results demonstrate that multi-task learning effectively alleviates the semantic discrepancy between image and text modalities and mitigates visual bias, improving the accuracy and robustness of MNER systems.

Keywords: multimodal named entity recognition, multitask learning, feature extraction, multimodal fusion

1 Introduction

The widespread adoption of self-media platforms has led to an increase in multimodal data, including text and images. In social media scenarios, text-only named entity recognition (NER) methods often fail to achieve satisfactory performance. This is mainly due to the fact that social media posts are typically short and lack sufficient context, resulting in ambiguous entity boundaries and stronger semantic ambiguity. These factors make effective disambiguation challenging when relying solely on textual cues. Therefore, researchers incorporate multimodal signals such as images and videos to perform Multimodal Named Entity Recognition (MNER), which enhances the accuracy of label prediction. MNER has shown practical value in various applications, including crop pest and disease identification, local chronicles text processing, and product attribute extraction.

In the MNER task, Esteves et al. [1] first integrated visual information by extracting features from both images and text to classify named entities. Subsequently, Moon et al. [2] proposed a modality attention mechanism to adaptively attenuate or enhance word embeddings, character embeddings, and visual weights. Lu et al. [3] further introduced a visual attention mechanism to locate image regions that are relevant to the textual content. However, these methods typically rely on a single attention module to model the matching between multimodal features, which may lead to incomplete or even erroneous multimodal interactions. To address this issue, cross-modal attention mechanisms have been proposed. For example, Tian et al. [4] introduced image–text cross-modal attention and obtained the final feature representation via multi-level attention and adaptive multimodal integration. Zhong et al. [5] employed multiple cross-modal attention mechanisms to facilitate deeper semantic interactions across modalities.

Overall, the aforementioned single-task methods are trained solely with the sequence labeling loss of named entity recognition, lacking explicit supervision for image region–text span alignment and visual denoising. Such methods suffer from the following limitations: (1) Insufficient cross-modal alignment signals: single-task training

* Corresponding Author

fails to provide fine-grained constraints for word–region correspondence, making it challenging to accurately locate critical visual cues; (2) Difficulty in suppressing visual noise and contextual bias: social media images are often weakly correlated with the accompanying text, and single-task fusion may be misled by background objects, resulting in visual bias; (3) A significant semantic gap between image and text modalities: single-task learning struggles to learn stable shared representations, limiting cross-domain generalization; (4) Limited annotated data and a long-tail distribution of entity types: single-task training is prone to overfitting and leads to reduced robustness. To address these challenges, researchers have proposed multi-task learning-based MNER methods that jointly train multiple related tasks [6] to promote cross-modal fusion between text and images, thereby significantly improving the performance of MNER models.

Based on the above studies, this paper focus is placed on multi-task learning–based MNER. The overall processing pipeline is summarized, and the main ideas, advantages, and limitations of multi-task learning strategies are discussed. Existing approaches are further categorized into three groups according to the auxiliary tasks adopted in multi-task learning: visual optimization network–assisted methods, local matching–assisted methods, and text-assisted multi-task methods. To verify the effectiveness of each category, representative models from the three groups are reproduced on the Twitter-2015 and Twitter-2017 datasets, followed by a comparative performance analysis. Experimental results indicate that multi-task learning strategies can effectively alleviate visual bias and modality discrepancy, leading to substantial performance improvements.

2 MENR of Multi-task Learning

In multi-task learning–based MNER methods, the overall pipeline can be generally divided into four components: unimodal input, feature extraction, multimodal feature fusion, and label prediction. The processing workflow is illustrated in Fig. 1.

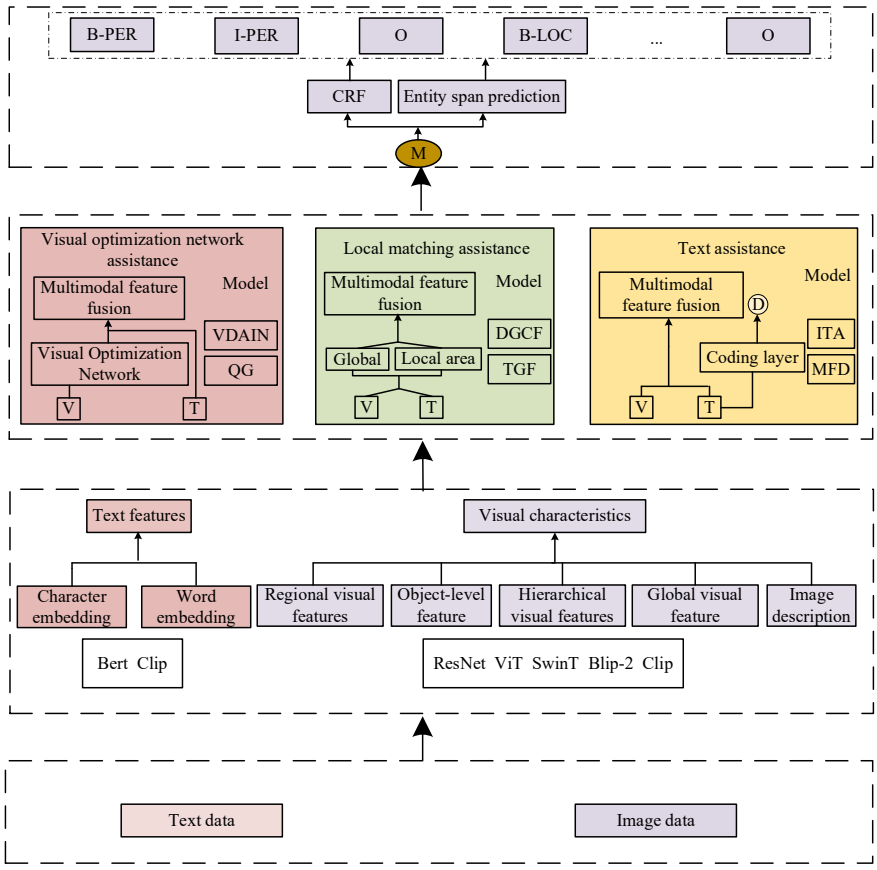


Fig. 1. The basic framework of MNER for multi-task learning

Unimodal inputs in MNER primarily consist of textual and visual data. Text representations are typically constructed using word embeddings, character embeddings, or a combination of both to capture linguistic features. Popular models for generating word embeddings include Word2Vec [7], GloVe [8], fastText [9], BERT [10], and CLIP [11]. These low-dimensional, dense real-valued vectors have been demonstrated to offer superior word representations compared to one-hot vectors or other sparse feature encodings. Character embeddings are typically obtained through BiLSTM [12] or Convolutional Neural Networks (CNNs) [13]. For joint modeling of word- and character-level information [14-15], word representations X^W and character representations X^C are concatenated to form the final textual representation $X^T = [X^W; X^C]$. In terms of visual feature extraction, existing studies generally rely on several representative models. CLIP can extract both textual and visual features within a unified framework, whereas Mask R-CNN [16] specializes in object-level feature extraction. Models like VGGNet [17], ResNet [18], and Vision Transformer (ViT) [19] are more effective at capturing region-level visual features. Additionally, BLIP-2 [20] and image captioning techniques [23] generate textual descriptions of images, which not only help bridge the semantic gap between modalities but also facilitate more accurate extraction of high-level semantic information from images. Following unimodal feature extraction, the multimodal fusion stage maps visual and textual features into a unified vector space and performs cross-modal interactions to produce a multimodal representation with rich semantics.

In the label decoding stage, multimodal representations are typically decoded using two main strategies: (1) CRF-based decoding [21]. A CRF performs sequence labeling by explicitly modeling dependencies among adjacent labels, thereby effectively utilizing contextual information and enhancing entity boundary detection accuracy. However, due to the Markov assumption, CRF-based decoding is less capable of capturing long-range dependencies or dealing with weakly correlated labels, which may restrict its effectiveness in complex semantic scenarios. (2) Span-based decoding through entity span prediction [22]. This approach identifies entities by predicting their start and end positions and then assigns corresponding category labels. It is especially effective for handling nested entities. Nonetheless, because entities of the same type can appear multiple times in a sentence, the decoding process may generate spurious entities (false positives).

Current mainstream MNER approaches mainly focus on two aspects: feature extraction and processing, and multimodal fusion. Among them, visual optimization network–assisted methods aim to extract and refine fine-grained visual features, enhancing the model's ability to understand visual semantics and thereby improving its capability to learn critical visual cues. Local matching–assisted methods compute local similarities between textual tokens and image regions to select more relevant visual content, effectively reducing the interference of irrelevant images on entity prediction and improving the efficiency of multimodal information fusion. In contrast, text-assisted methods introduce auxiliary modules based solely on textual signals, guiding the model to capture semantic correlations more accurately during fusion, mitigating the negative effects of visual bias, and enabling more robust cross-modal semantic alignment. Table 1 summarizes the key characteristics, advantages, and limitations of these three categories of methods.

Table 1. Comparison of three categories of multi-task learning–based MNER methods

Method category	Core characteristics/Key mechanisms	Advantages	Limitations
Visual optimization network–assisted	Introduces visual optimization/ enhancement modules to extract multiple types of visual features, perform selection and guidance, and then fuse them with textual representations	Improves the quality of fine-grained visual cues; suppresses noise from irrelevant objects; yields more stable multimodal fusion	Highly dependent on detection/ visual feature quality; incurs high computational cost and architectural complexity; may still introduce bias when images are weakly related to the text
Local matching–assisted	Incorporates a local matching mechanism on top of global image–text alignment, aggregating visual regions that are more relevant to the text	Alleviates the semantic gap; better exploits region-level information for precise image–text alignment	Local matching is sensitive to noise; may filter out important visual information when image–text discrepancy is large
Text-assisted	Constructs text-only auxiliary modules to provide priors during fusion or decoding, calibrating multimodal representations and mitigating visual bias	Strengthens semantic alignment and interactions across modalities; reduces visual bias; improves textual semantic understanding via text-only supervision	May weaken the contribution of visual signals; overall model design can be complex; still limited when text is ambiguous or lacks sufficient context

3 Visual Optimization Network-assisted Multitasking Methods

Visual optimization network–assisted multi-task methods significantly improve the quality of visual features by introducing a dedicated visual optimization network. Such approaches effectively enhance the fusion of visual and textual representations while suppressing irrelevant modal noise, thereby improving the overall performance of multimodal tasks. The architecture of this category of methods is illustrated in Fig. 2.

This category of methods typically obtains textual representations (X^t) and visual representations (X^v) from pretrained models, and then refines the visual features (X^v) through a visual optimization network to produce optimized visual features, which helps filter out noise from irrelevant information. After fusing the optimized visual features with textual features, the model yields a multimodal representation *M*, which is then input into a CRF layer for decoding. Following the main theme of improving visual feature quality, this line of work explores different technical routes—from early approaches that filter noisy objects via adversarial learning [24], to more fine-grained optimization strategies such as introducing visual prefixes [25], dynamic gating [27], and hierarchical features [28]. Specifically, Zheng et al. [24] proposed an adversarial bilinear attention network to capture the correlation between visual objects and textual entities. In their framework, adversarial learning is used to construct a shared subspace for better image–text feature fusion, while the bilinear attention network facilitates the alignment between entities and relevant visual objects, thereby filtering out irrelevant visual objects. Chen et al. [25] introduced a Hierarchical Visual Prefix Fusion Network (HVPNeT), which treats object-level visual representations in BERT as prefixes for each self-attention layer, enabling visual prefix–guided fusion. Moreover, a dynamic gating mechanism is employed to fully exploit hierarchical visual features and reduce sensitivity to errors caused by irrelevant objects in images.

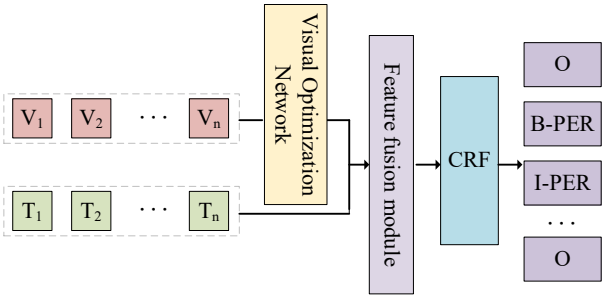


Fig. 2. Structural diagram of the MNER method assisted by the visual optimization network

Although the above two methods alleviate multimodal noise to some extent, a substantial modality gap still exists between image and text representations. To address this issue, Zuo et al. [26] extended the approach of Chen et al. [25] by introducing a visual description–enhanced fusion network, which extracts semantic information from images and effectively reduces the semantic gap during multimodal fusion. In addition, ViT is incorporated to extract higher-quality visual features, thereby improving cross-modal integration capability to a certain degree. Different from Zuo et al. [26], Liu et al. [27] construct multi-granularity visual context as a supplement to visual semantics and further design a dynamic filtering network to remove visual noise introduced by irrelevant images. Dai et al. [28] propose an Alignment and Matching Network with hierarchical visual feature fusion (AMNeT) to reduce the adverse impact of mismatched images. Specifically, ViT is utilized to combine and concatenate hierarchical visual features from global regions and visual objects, which are used as an improved visual prefix for multimodal fusion, effectively suppressing modality noise. Jia et al. [29] design a unified query with prior information as a navigator to guide joint training, while extracting multi-scale visual features. Two interaction mechanisms are further introduced, including a multi-scale cross-spatial interaction module and a presence-aware unimodal interaction module, which enrich textual and visual information and enhance image–text fusion. Wang et al. [30] propose a multi-scale visual semantic enhanced MNER method (MSVSE), which generates and fuses multi-scale visual semantic features by exploring semantic interactions between text and multiple types of visual features. Zheng et al. [31] adopt a hierarchical fusion strategy to dynamically integrate image and text representations at different granularities, and address the difficulties of diverse multimodal fusion via a “decompose, prioritize, and eliminate” strategy. Zhang et al. [32] align modality features at different

granularities to capture both global and local information, and achieve deep fusion by combining a cross-modal attention mechanism, thereby significantly improving representational capacity and generalization in multimodal tasks. Moreover, Zhang et al. [33] propose a visual object density–guided hard example mining strategy by selecting text–image pairs with high object density as hard samples during training. To mitigate bias caused by imbalanced entity-type distributions, a debiased contrastive loss is further introduced to replace conventional contrastive learning objectives, enabling more effective balancing of learning weights across different categories during optimization and improving overall performance and robustness in MNER. The debiased contrastive loss is formulated as follows:

$$L(X^+, X^-) = -log \frac{\sum_{x \in X^+} f(x)}{\sum_{x \in X^+} f(x) + \frac{Q}{N} \sum_{x \in X^-} f(x)} \tag{1}$$

where X^+ and X^- denote the positive and negative samples, respectively. $f(x)$ is the index of $exp(d(x)/\tau)$, $d(\bullet)$ denotes the cosine similarity function, τ is the temperature parameter, W is the weighting parameter, and N represents the number of negative samples. The weighting term dynamically adjusts the impact of negative samples. By maximizing the similarity of positive pairs while minimizing that of negative pairs, the objective encourages the model to learn a more discriminative cross-modal representation space.

Visual optimization network–assisted multi-task methods improve MNER performance by refining visual features; however, such methods are often structurally complex during visual feature extraction and optimization. In contrast, local matching–assisted multi-task methods primarily focus on modeling image–text similarity, thereby effectively filtering out interference from image content that is irrelevant to the text.

4 Multi-Task Method with Local Matching Assistance

Local matching–assisted multi-task methods enhance the effectiveness of global matching through a dedicated local matching mechanism. Global matching measures the alignment degree between an image and a text by computing their similarity in a shared semantic space, which encourages consistent image–text representation learning and alleviates semantic mismatch across modalities. In contrast, local matching operates on region-level fine-grained similarities and adopts a dual-mechanism cross-modal alignment strategy to explore fine-grained correspondences between image regions and text spans, thereby producing more accurate image–text matching scores. Such methods facilitate precise alignment between visual cues and textual entities in named entity recognition, effectively filtering out visual distractions that are irrelevant to target entities and improving the model's discriminative capability and robustness. The architecture of local matching–assisted multi-task methods is illustrated in Fig. 3. This paper further provides a systematic analysis of how existing approaches achieve alignment from global to local, such as constructing dynamic graphs [36], introducing graph neural networks (GNNs) [37], or applying multi-level contrastive learning [40]. We also discuss the advantages and limitations of these strategies in bridging the semantic gap and enabling precise cross-modal alignment.

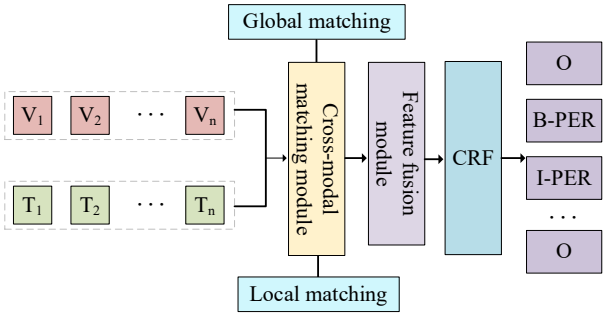


Fig. 3. Architecture of the MNER method with local matching assistance

To construct consistent image–text representations and bridge the semantic gap between the two modalities, Xu et al. [34] uniformly partition the feature map of the whole image into multiple patches and model the correlations between the textual sequence and each visual region via a gating mechanism, thereby learning more semantically informative visually grounded token representations. Cheng et al. [35] measure image–text similarity using high-level semantic features from both modalities and capture the relationships of objects across images and texts.

The aforementioned methods primarily focus on global image-text matching while neglecting local matching, thereby limiting their ability to enhance text-image alignment and reduce cross-modal semantic discrepancies. Specifically, existing approaches inadequately address both global and local image-text matching in a self-supervised manner. Moreover, static graphs constructed in many MNER models struggle to bridge the semantic gap across modalities and fail to effectively exploit fine-grained visual information from images. To overcome these limitations, local matching-assisted multi-task methods have been proposed. For example, Mai et al. [36] introduced a novel Dynamic Graph Construction Framework (DGCF) tailored for MNER. DGCF designs three image-text matching mechanisms—region-level similarity computation, global similarity computation, and cross-modal matching measurement—to comprehensively capture semantic associations between text and images at multiple levels. The region-level similarity computation is formulated as follows:

$$s_i^t = \frac{W_s^t \mid x_i - a_i^t \mid^2}{\left\| W_s^t \mid x_i - a_i^t \mid^2 \right\|_2} \tag{2}$$

$$\mathrm{s}_j^v = \frac{W_s^v \mid o_j - a_j^v \mid^2}{\left\| W_s^v \mid o_j - a_j^v \mid^2 \right\|_2} \tag{3}$$

The global similarity computation is formulated as follows:

$$s^{glo} = \frac{W_s^g \mid o^{glo} - x^{glo} \mid^2}{\left\| W_s^g \mid o^{glo} - x^{glo} \mid^2 \right\|_2} \tag{4}$$

where $a_i^t = \sum_{j=1}^{49} \alpha_{ij} o_j$, $\alpha_{ij} = soft\ max(c_{i1}, c_{i2}, ..., c_{i49})$, c_{ij} denote the attention weights between the *i*-th word and the *j*-th image region obtained via a multi-head mechanism, $s_i^t \in R^{d_{sim}}$ denotes the similarity vector of the *i*-th word with respect to the image, $x_i \in R^d$ represents the textual representation, $o_j \in R^d$ is the global image representation obtained via an average pooling layer, W_s^t, W_s^v, W_s^g are learnable weight matrices, $s_j^v \in R^{d_{sim}}$ is the similarity vector of the *j*-th image region with respect to the text, and x^{glo} denotes the overall text representation. Region-level similarity modeling produces fine-grained similarity vectors, enabling more precise cross-modal alignment. By combining global matching with region-level similarity, the model forms a multi-granularity matching mechanism, which enhances its capability to assess image–text relevance. Liu et al. [37] employ a Transformer and R-GCN to learn textual representations and local/global visual representations, respectively. The relevance between text and images is evaluated, and visual information at different granularities is integrated based on the estimated relevance. Semantic representations are further refined through iterative cross-modal interactions and a co-attention mechanism. Zhang et al. [38] construct a multimodal graph using textual tokens from BERT and visual tokens from Swin Transformer [39], and feed it into a multi-layer Transformer to enable interactive fusion both within and across modalities. In addition, multiple contrastive learning objectives are introduced to perform global and local cross-modal alignment, further enhancing semantic consistency between text and images. Yan et al. [40] proposed a more fine-grained rank-based contrastive learning framework (FRCL) for MNER. This framework introduces a global contrastive learning strategy to effectively align the semantic features of text and images. At the same time, it employs a Top-K rank-based masking approach to construct positive and negative sample pairs, thereby helping the model learn more deeply about the fine-grained interactions between modalities.

Local matching–assisted multi-task methods primarily emphasize fine-grained processing of multimodal features, aiming to reduce the negative impact of visual bias on model performance. By computing local

similarities between images and texts, visually relevant regions with higher textual correlation are selected, enabling more precise cross-modal alignment. However, when substantial semantic discrepancies exist between image and text, local matching may filter out critical visual cues, resulting in the loss of important image content. In contrast, text-assisted multi-task methods focus on enhancing and exploiting textual features. By constructing text-only auxiliary modules, additional linguistic context is provided for multimodal named entity recognition. Such methods not only improve textual semantic understanding, but also guide visual features to align more effectively with text during fusion, thereby alleviating the cross-modal semantic gap and enriching the final multimodal representation. In summary, the two categories optimize multimodal feature fusion from complementary perspectives—visual and textual—jointly improving robustness and recognition accuracy in complex cross-modal scenarios.

5 Multi-Task Method with Text-Based Assistance

Text-assisted multi-task methods mainly aim to address the bias introduced by visual context by incorporating a text-only auxiliary module during either the feature fusion stage or the label decoding stage. Specifically, textual features C and visual features V are first extracted using pretrained models such as BERT and ResNet, respectively. A feature fusion module then integrates multimodal inputs to obtain a multimodal representation *M*, which is subsequently decoded. Meanwhile, the text is further encoded, and its representation is fed together with *M* into a multimodal fusion layer to produce the final label predictions. In this framework, the multimodal fusion layer improves recognition performance by leveraging fusion strategies and optimizing the model through appropriate loss functions. The architecture of text-assisted multi-task methods is illustrated in Fig. 4.

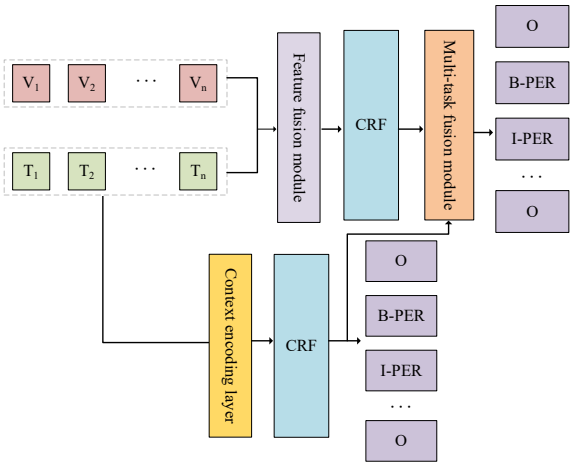


Fig. 4. Architecture of the MNER method with text-based assistance

This article divides it into two subcategories based on the stage of auxiliary task intervention: “label decoding stage” and “feature fusion stage”. This subdivision helps readers understand how the former (such as UAMNER [42], M3S [44]) uses text to generate candidate labels to correct visual biases; and how the latter (such as HVPNeT [25], P-MNER [49]) guides at the feature level through cross-modal consensus or text correction modules.

5.1 Label Decoding Stage

For approaches that introduce auxiliary tasks at the label decoding stage, the primary goal is to preserve the integrity of visual semantics and leverage fine-grained visual features to reduce the visual bias caused by image content. A text-only auxiliary module is then combined with the multimodal representation to generate the final predicted label sequence.

During the label decoding stage, Yu et al. [41] proposed a text-only Entity Span Detection (ESD) module as an auxiliary task to mitigate the bias introduced by visual context and guide the final predictions. The formulation is given as follows:

$$P(\mathrm{z} \mid S) = \frac{\exp(\sum_{i=0}^{n} T_{z_i, z_{i+1}} + \sum_{i=1}^{n} w_{ESD}^{z_i} \cdot t_i)}{\sum_{z'} \exp(\sum_{i=0}^{n} T_{z'_i, z'_{i+1}} + \sum_{i=1}^{n} w_{ESD}^{z'_i} \cdot t_i)} \tag{5}$$

where $P(z \mid S)$ denotes the probability of predicting the label sequence Z given the input text S, while $z_i \in Z$, $Z = \{B, I, O\}$, and $\mathrm{W}_{ESD}^{z_i} \in R^d$ are task-specific parameters, and t_i represents the textual features. By introducing a text-only auxiliary task at the decoding stage, the model can reduce the bias brought by visual context. Moreover, a CRF explicitly models dependencies among adjacent labels by incorporating transition terms (e.g., a B-ORG tag is more likely to be followed by I-ORG), thereby improving entity boundary detection and label consistency. Liu et al. [42] employ a Bayesian neural network to generate candidate labels from textual input, and further introduce a multimodal Transformer to fuse textual and visual information for refining and optimizing these uncertainty-aware candidate labels. The overall computation is summarized by the following equations:

$$\mathrm{p} \approx \frac{1}{T} \sum_{\tau=1}^{T} \frac{exp(W_l h_i \mid \varpi_\tau^*)}{\sum_{j=1}^{C} exp(W_l h_j \mid \varpi_\tau^*)} \tag{6}$$

$$u_i = h(p_i) = -\sum_{c=1}^{C} p_c \, log(p_c) \tag{7}$$

where p_i denotes the distribution of candidate labels, W_l is a $2d^*d_h$ matrix, d_h equals the number of output classes, represents the sampling weights, c is the number of label categories, u_i is the uncertainty value for each token, and T is the text representation. When u_i is large, it indicates that the entity label is difficult to determine from text alone; thus, the model relies more heavily on the multimodal branch for correction. In contrast, when u_i is small, the textual prediction is more reliable, and the fusion module tends to preserve the text prior to avoid interference from visual noise. Therefore, the candidate distribution provides a set of "plausible answers," whereas the uncertainty score indicates the "degree of multimodal intervention" required; together, these two components improve robustness and interpretability at the decoding stage. Wang et al. [43] proposed a confidence-threshold–guided label fusion approach, which leverages the vision–language pretrained model BLIP-2 to generate image captions. A confidence threshold is further applied to select high-precision labels from both the text predictions and the candidate label set, thereby improving overall performance. Wang et al. [44] introduced a scene graph–driven multimodal multi-granularity multi-task learning framework (M3S). It employs a cascaded task design that combines Named Entity Segmentation (NES) with Named Entity Classification (NEC), enabling deep modeling of both textual and visual information. In addition, M3S incorporates a Multi-Granularity Gated Aggregation (MGA) module to fuse textual and visual inputs and capture cross-modal semantic interactions at different granularities. Wang et al. [45] proposed an Image–Text Alignment (ITA) method, which aligns the output distributions of the cross-modal view (I+T) and the text-only view (T) via a Cross-View Alignment (CVA) module. The corresponding formulation is given as follows:

$$M = Transformer([W^T; W^G; W^L]) \tag{8}$$

$$L_{ITA} = L_{CVA} + L_T + L_{I+T} \tag{9}$$

Where W^T denotes the text, W^G represents the image caption, and W^L is the visual object label.

In addition, to explore fine-grained semantic interactions between text and images, Wang et al. [46] introduced

a Graph Convolutional Network (GCN) mechanism to model semantic feature interactions across modalities. By leveraging the feature propagation capability of GCNs, their method captures rich semantic relations between image and text representations, and further employs a dynamic gated fusion network to derive the probability of matching the fused multimodal features with the original textual input. Xu et al. [47] proposed a span-based method, namely SMNER, which incorporates a cross-modal information bottleneck module to make visual semantics closer to textual semantics, thereby filtering out visual noise. In addition, an attention mechanism is introduced to achieve more accurate alignment between text and images.

5.2 Feature Fusion Stage

Auxiliary tasks introduced at the feature fusion stage aim to mitigate visual bias during multimodal fusion by using the original textual features to calibrate the fused representations. After obtaining unimodal representations, the model repeatedly integrates the original text features with the fused multimodal features, iteratively refining and correcting the fused representation, which effectively reduces visual bias.

To preserve the unique information contained in the original textual features, Huang et al. [48] adopt a cross-modal consensus fusion strategy to obtain shared cross-modal features (AC), which helps reduce distribution discrepancy between modalities. The text information T is used as a supporting modality after fusion (ST), preventing the loss of textual uniqueness and yielding the final multimodal representation M. The computation formula is as follows:

$$g_c = R(W_C[A_C; S_T] + b_c) \in R^{d \times l} \tag{10}$$

$$g_I = R(W_I[A_I; S_T] + b_I) \in R^{d \times l} \tag{11}$$

$$H = g_c(W_C^{'} A_C) + g_I(W_I^{'} A_I) + b \in R^{d \times l} \tag{12}$$

$$M = S_T + min(\frac{\left\| S_T \right\|_2}{\left\| H \right\|_2} \beta, 1) H \tag{13}$$

Where, g_c and g_I denote the gating weights. W_C, $W_I \in R^{d \times 2d}$ and $W_C^{'}$, $W_I^{'} \in R^{l \times d}$ are learnable weight matrices, $R(x)$ denotes the nonlinear activation function, A_I represents the visual features, β is a hyperparameter, b_c, b_I and b are scalar biases, and $||ST||_2$, $||H||_2$ represent the L_2-norms of S_T and H, respectively. The ratio reflects the similarity between the two. M denotes the final multimodal representation. By introducing the gating and normalization mechanisms, the method ensures semantic fusion while effectively suppressing visual bias. Wang et al. [49] proposed a prompt-based multimodal named entity recognition network, namely P-MNER, which leverages fine-grained visual features to enhance semantic understanding of entities in text, thereby providing more effective support for MNER. In addition, to address potential visual bias introduced during multimodal fusion, a Text Calibration Fusion Module (TCFM) is incorporated to mitigate fusion-induced bias. Xu et al. [50] exploited semantic complementarity across features using a Transformer and a cross-modal attention mechanism. Specifically, two types of textual features (word-level textual features and character-level features) and three types of visual features (global visual features, visual descriptions, and region-level visual features) are fused to obtain a semantically completed textual representation and a semantically enhanced multimodal representation. Cai et al. [51] introduced two auxiliary tasks to predict boundary labels and category labels, vectorized the outputs, and performed MNER using label embeddings that integrate multimodal features (text and image) together with auxiliary-task predictions. Zhao et al. [52] proposed a reinforcement learning–based discriminator guided by human preference, which distinguishes images according to relevance and models the image–text relationship as a heterogeneous graph. A graph transformer architecture is employed to enable information interaction across modalities. Ding et al. [53] proposed an Entity Label–Guided Graph Fusion approach (ELGF) for multimodal named entity recognition. Entity label information is utilized as a bridge between text and images to construct a multimodal interaction graph, and attention and gating mechanisms are further applied to interactively fuse multimodal information.

Text-assisted multi-task methods enhance image–text fusion by performing deep encoding of textual features and constructing text-only auxiliary modules, thereby strengthening semantic alignment and cross-modal interactions between text and images, and improving the accuracy and robustness of entity recognition. However, such methods often involve complex model architectures and multi-level interaction mechanisms, which increases the overall model complexity.

6 Method Analysis

To validate the effectiveness of multi-task learning–based MNER methods, an experimental setup is established and evaluations are conducted on two multimodal benchmark datasets, Twitter-2015 and Twitter-2017. The datasets involve four entity types, namely PER, LOC, ORG, and MISC. Following the dataset split protocol of Zhang et al. [54], each dataset is divided into a training set (Train), a validation set (Dev), and a test set (Test), as summarized in Table 2, which reports the number of instances in each split as well as the counts of each entity type.

Table 2. Data statistics of the multimodal datasets

Entity type	Twitter-2015			Twitter-2017		
	Train	Dev	Test	Train	Dev	Test
PER	2217	552	1816	2943	626	621
LOC	2091	552	1697	731	173	178
ORG	928	247	839	1647	375	395
MISC	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
Tweets	4000	1000	3257	3373	723	723

The F1 score is adopted as the evaluation metric. Representative models from the three categories of multi-task MNER methods, including MVPN [27], QG [29], and DGCF [36], are reproduced, and the experimental results are reported in Table 3 and Table 4.

Table 3. Comparison results of the MNER method for multi-task learning on Twitter-2015

Method category		Model	Twitter-2015				
			PER	LOC	ORG	Misc	F1
Vision optimization network assisted		MVPN* [27]	86.69	83.26	62.45	43.63	75.72
		QG [29]	85.68	81.42	63.62	41.53	74.94
		MSVSE [30]	86.72	81.63	64.08	38.91	75.11
Local matching assisted		DGCF [36]	84.89	83.32	64.08	38.07	75.13
		TGF [38]	85.18	83.06	63.42	42.48	74.91
		HamLearning [37]	85.28	82.84	64.46	42.52	76.49
Text-based assisted	Label decoding stage	CLGLF [43]	86.30	81.92	63.59	38.52	75.09
		M3S [44]	86.05	81.32	62.97	41.36	75.03
		SMRG* [46]	86.96	82.75	64.66	41.95	76.01
	Feature fusion stage	P-MNER [49]	-	-	-	-	79.43
		MFD* [51]	-	-	-	-	77.05
		ELGF* [53]	-	-	-	-	89.30

Note: The data in the table was obtained by running the source code provided in the original paper. "*" indicates that the experimental results in the original paper are cited, and "-" means that the corresponding results are not displayed.

Table 4. Comparison results of the MNER method for multi-task learning on Twitter-2017

Method category		Model	Twitter-2017				
			PER	LOC	ORG	Misc	F1
Vision optimization network assisted		MVPN* [27]	93.26	85.47	86.10	72.03	87.80
		QG [29]	93.17	86.02	84.64	71.83	87.25
		MSVSE [30]	93.24	85.96	85.22	70.00	87.34
Local matching assisted		DGCF [36]	93.80	87.86	85.71	70.90	88.07
		TGF [38]	93.84	86.86	84.89	70.70	87.70
		HamLearning [37]	91.43	86.26	86.66	69.17	87.13
Text-based assisted	Label decoding stage	CLGLF [43]	93.59	85.87	84.87	70.06	87.40
		M3S [44]	92.73	84.81	82.49	69.53	86.06
		SMRG* [46]	93.28	85.31	85.82	71.86	87.66
	Feature fusion stage	P-MNER [49]	-	-	-	-	91.31
		MFD* [51]	-	-	-	-	87.31
		ELGF* [53]	-	-	-	-	91.43

Note: The data in the table was obtained by running the source code provided in the original paper. "*" indicates that the experimental results in the original paper are cited, and "-" means that the corresponding results are not displayed.

According to the experimental results in Table 3 and Table 4, among the text-assisted multi-task learning methods, P-MNER, MFD, and ELGF achieve F1-score improvements of 3.42%, 1.04%, and 13.29% over SMRG, respectively, on the Twitter-2015 dataset. Notably, on Twitter-2017, P-MNER and ELGF show more competitive performance, reaching F1 scores of 91.31% and 91.43%, which are 3.65% and 3.77% higher than that of SMRG, respectively. These results demonstrate that introducing text-assisted information at the feature fusion stage generally yields larger performance gains than introducing auxiliary tasks at the label decoding stage. Feature-fusion–stage methods utilize the original textual features as guidance to prompt and calibrate multimodal fusion representations. In contrast, label-decoding–stage methods encode the original text and combine it with the multimodal representation to generate final label predictions. Such approaches emphasize capturing fine-grained visual features during visual feature extraction and filtering out visual content that is irrelevant to the text. However, compared with decoding-stage assistance, feature-fusion–stage assistance exhibits clear advantages: it preserves the distinctive characteristics of the original textual features while more effectively reducing interference from irrelevant visual information, thereby substantially improving the performance of Multimodal Named Entity Recognition (MNER).

The key advantage of P-MNER lies in the incorporation of prior knowledge from pretrained language models via prompt learning, together with the introduction of a Text Calibration Fusion Module (TCFM) to explicitly alleviate visual bias during multimodal fusion. Consequently, P-MNER often exhibits superior performance in the following scenarios: (1) Short texts with strong entity ambiguity, such as frequent use of aliases/abbreviations and blurred entity boundaries, where prompt learning can leverage the knowledge and contextual modeling capacity of pretrained language models to improve prediction stability; (2) Weakly related images or images containing substantial background noise, where TCFM functions as a "text-dominant calibration" mechanism over fused representations, reducing misleading effects from irrelevant visual content and preventing the model from being distracted by salient yet entity-irrelevant regions. The core idea of ELGF is to use entity label information as a semantic "bridge" between text and images. Specifically, a multimodal interaction graph is constructed based on text representations, entity label representations, and visual representations, followed by interactive fusion via attention and gating mechanisms. This design is particularly effective for social media MNER. On the one hand, compared with alignment strategies that rely solely on word–region similarity, entity labels provide more stable and structured semantic anchors, alleviating alignment instability caused by weak image–text relevance. On the other hand, the gating mechanism suppresses entity-irrelevant visual information and retains object cues that are most relevant to entity labels, making ELGF more advantageous for samples where images contain many objects but only a few are related to the target entities.

Among visual optimization network–assisted and local matching–assisted methods, MVPN achieves an F1 score that is 0.59% and 0.81% higher than DGCF and TGF, respectively, on Twitter-2015, whereas DGCF outperforms MVPN by 0.27% in terms of F1 on Twitter-2017. Overall, the two categories exhibit comparable performance, while adopting different strategies for visual feature processing. Visual optimization network–assisted methods (e.g., QG, MVPN, and MSVSE) typically adopt a hierarchical visual feature mapping mechanism, constructing multi-granularity visual prompts as prefixes to guide feature fusion. This strategy

provides several advantages: 1) visual guidance is supplied to each layer of the self-attention network; 2) noise interference in visual features is effectively suppressed; the semantic gap between modalities is substantially reduced. In contrast, local matching–assisted methods (e.g., DGCF, TGF, and HamLearning) emphasize region-level visual feature extraction, characterized by: 1) establishing fine-grained correspondences between textual tokens and image regions; 2) eliminating interference from irrelevant regions via local matching; 3) enhancing local cross-modal alignment capability. Although the two categories differ in visual feature processing—global hierarchical guidance versus local region matching—the underlying objective remains consistent: optimizing visual representations and filtering image noise to improve MNER performance. In terms of scenario adaptability, when images contain substantial noise, the dynamic filtering mechanism in visual optimization–based methods is more likely to preserve critical features; when image–text relevance is stronger and region-level cues are explicit, the fine-grained alignment mechanism of local matching–based methods tends to be more advantageous.

7 Conclusion and Future Work

This paper elaborates the overall pipeline of multi-task learning–based MNER methods and categorizes existing approaches into three major groups according to the types of auxiliary tasks employed: visual optimization network–assisted methods, local matching–assisted methods, and text-assisted multi-task methods. The main principles, representative methods, as well as the strengths and limitations of each category are systematically summarized. Experimental validation on the Twitter-2015 and Twitter-2017 datasets indicates that current MNER systems still face several challenges, including weak image–text relevance that provides insufficient grounding for entity recognition, inadequate multimodal fusion, ambiguous textual semantic expressions, and missing semantic information. Future research should emphasize improving the discrimination of easily confused entities to enhance overall performance, and further exploiting textual semantics while preserving key visual cues, so as to achieve more complementary and synergistic understanding of image–text information.

In recent years, with the development of large language models (LLMs), significant achievements have been made in various natural language processing (NLP) tasks and multimodal tasks. Specifically, for the MNER task, models like RoBERTa [55] are used as task encoders to extract textual features. For instance, CAT [56] employs both BERT-large and RoBERTa-large during testing, as these models provide more accurate textual semantic representations, leading to improved performance. Guo et al. [57] utilize XLM-RoBERTa and VinVL as task encoders to extract textual and visual features, respectively, leveraging their strong capabilities in contextual and image modeling. ChatGPT-assisted multimodal named entity recognition, such as that proposed by Li [58-59], enhances prediction performance through the design of a task-specific prompt template, enabling ChatGPT to generate refined auxiliary knowledge. In addition, domestic open-source models such as Qwen, DeepSeek, and ChatGLM have demonstrated unique advantages in Chinese understanding and multimodal fusion. Therefore, future MNER methods can improve performance from the following two aspects:

(1) LLMs-assisted MNER represents a promising research direction. Fully exploiting the strong text understanding capability of large language models enables deeper analysis of complex linguistic structures and semantics, thereby effectively improving the accuracy and coverage of named entity recognition. In addition, large language models exhibit the potential to deeply integrate textual descriptions with visual information, which facilitates more precise cross-modal alignment and significantly enhances model robustness and generalization. Furthermore, the superiority of large language models in contextual modeling leads to stronger comprehension and adaptability when recognizing entities that rely heavily on contextual cues, providing solid technical support for entity recognition in multimodal scenarios.

LLMs exhibit stronger capabilities in instruction understanding, contextual reasoning, and knowledge transfer, making them applicable to generating entity candidates, providing type priors, explaining modality conflicts, and performing cross-modal disambiguation. However, integrating LLMs into MNER still faces several key technical challenges: 1) Computational cost and inference latency: LLMs have large parameter scales and high end-to-end fine-tuning costs, which motivates the exploration of parameter-efficient fine-tuning (e.g., LoRA/Adapters), model distillation, and two-stage lightweight frameworks. 2) Mismatch between sequence labeling and generative outputs: MNER typically requires BIO-style token-aligned labels, whereas LLMs-generated outputs must be strictly aligned with spans in the original sentence. Consistency can be improved through structured prompting, constrained decoding, and posterior validation. 3) Hallucination and unreliability: LLMs may produce entities that do not appear in the text or assign incorrect entity types, and cross-modal hallucinations are more likely when images are weakly related or contain substantial noise. Controllable inference can be supported

by combining retrieval augmentation, confidence calibration, and text-anchored constraints. 4) Data and evaluation: Existing MNER annotations are limited in scale, and LLMs-driven data augmentation may introduce noise, requiring high-quality filtering strategies and more fine-grained error analysis metrics. Addressing these challenges is essential for converting the potential of LLMs into stable and reproducible performance improvements in MNER.

(2) Beyond textual and image data, constructing datasets that incorporate additional modalities such as video and audio can enhance perceptual capacity and deepen semantic understanding, promoting the development of MNER toward settings with more modalities, multi-scale representations, and multi-task learning, thereby improving the accuracy and robustness of MNER systems.

Acknowledgement

This work has been supported by Open Research Project of the Key Laboratory of Arid Region Characteristic Agricultural Meteorological Disaster Monitoring, Early Warning, and Risk Management under the China Meteorological Administration (CAMF-202502) and the Graduate Education Quality Improvement Project of North Minzu University (YJZT202424).

References

[1] D. Esteves, R. Peres, J. Lehmann, G. Napolitano, Named entity recognition in twitter using images and text, in: Proc. of the International Conference on Web Engineering, 2017.
https://doi.org/10.1007/978-3-319-74433-9_17

[2] S. Moon, L. Neves, V. Carvalho, Multimodal named entity recognition for short social media posts, in: Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.
https://doi.org/10.18653/v1/N18-1078

[3] D. Lu, L. Neves, V. Carvalho, N. Zhang, H. Ji, Visual Attention Model for Name Tagging in Multimodal Social Media, in: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, 2018.
https://doi.org/10.18653/v1/P18-1185

[4] Y. Tian, X. Sun, H.F. Yu, Y. Li, K. Fu, Hierarchical self-adaptation network for multimodal named entity recognition in social media, Neurocomputing 439(2021) 12-21.
https://doi.org/10.1016/j.neucom.2021.01.060

[5] W.X. Zhong, H.R. Wang, D. Wang, M. Che, Image-Text Joint Named Entity Recognition Method Based on Multi-modal Semantic Interaction, Guangxi Science 29(4)(2022) 681-690.
https://doi.org/10.13656/j.cnki.gxkx.20220919.008

[6] Z. Chen, V. Badrinarayanan, C.Y. Lee, A. Rabinovich, Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, International conference on machine learning. PMLR 80(2018) 794-803.
https://proceedings.mlr.press/v80/chen18a.html

[7] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. <https://doi.org/10.48550/arXiv.1301.3781>, 2013 (accessed 12.03.2025).

[8] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.
https://doi.org/10.3115/v1/D14-1162

[9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5(2017) 135-146.
https://doi.org/10.1162/tacl_a_00051

[10] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language under-standing, in: Proc. of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies 1, 2019.
https://doi.org/10.18653/v1/N19-1423

[11] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Proc. of the 33rd International Conference on Neural Information Processing Systems, 2019.
https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

https://doi.org/10.18653/v1/N16-1030
[13] Z.W. Wu, C.M. Zheng, Y. Cai, J.Y. Chen, H.F. Leung, Q. Li, Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts, in: Proc. of the 28th ACM International conference on multimedia, 2020.
https://doi.org/10.1145/3394171.3413650
[14] I.J. Unanue, E.Z. Borzeshi, M. Piccardi, Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition, Journal of Biomedical Informatics 76(2017) 102-109.
https://doi.org/10.1016/j.jbi.2017.11.007
[15] G.Z. Tang, Named entity recognition in chinese electronic medical records based on ALBERT-IDCNN-CRF, in: Proc. 2022 IEEE 8th International Conference on Computer and Communications (ICCC), 2022.
https://doi.org/10.1109/ICCC56324.2022.10066018
[16] K.M. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proc. of the IEEE international conference on computer vision, 2017.
https://doi.org/10.1109/ICCV.2017.322
[17] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. <https://doi.org/10.48550/arXiv.1409.1556>, 2014 (accessed 15.03.2025).
[18] K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2016.
https://doi.org/10.1109/CVPR.2016.90
[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale. <https://doi.org/10.48550/arXiv.2010.11929>, 2020 (accessed 15.03.2025).
[20] J.N. Li, D.X. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, Proc. of the 40th International Conference on Machine Learning, PMLR 202(2023) 19730-19742.
https://proceedings.mlr.press/v202/li23q.html
[21] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proc. of the Eighteenth International Conference on Machine Learning Francisco, 2001.
[22] J.L. Fu, X.J. Huang, P.F. Liu, SpanNER: named entity re-/recognition as span prediction, in: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.
https://doi.org/10.18653/v1/2021.acl-long.558
[23] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proc. of the IEEE conference on computer vision and pattern recognition, 2015.
https://doi.org/10.1109/CVPR.2015.7298935
[24] C.M. Zheng, Z.W. Wu, T. Wang, Y. Cai, Q. Li, Object-aware multimodal named entity recognition in social media posts with adversarial learning, IEEE Transactions on Multimedia 23(2020) 2520-2532.
https://doi.org/10.1109/TMM.2020.3013398
[25] X. Chen, N.Y. Zhang, L. Li, Y.Z. Yao, S.M. Deng, C.Q. Tan, F. Huang, L. Si, H.J. Chen, Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction, in: Proc. Findings of the Association for Computational Linguistics: NAACL 2022, 2022.
https://doi.org/10.18653/v1/2022.findings-naacl.121
[26] M. Zuo, Y.J. Wang, W. Dong, Q.C. Zhang, Y.Y. Cai, J.L. Kong, Visual description augmented integration network for multimodal entity and relation extraction, Applied Sciences 13(10)(2023) 6178.
https://doi.org/10.3390/app13106178
[27] W. Liu, A.Q. Ren, C. Wang, Y. Peng, S.R. Xie, W.M. Li, MVPN: Multi-granularity visual prompt-guided fusion network for multimodal named entity recognition, Multimedia Tools and Applications 83(28)(2024) 71639-71663.
https://doi.org/10.1007/s11042-024-18472-w
[28] Y.L. Dai, F. Gao, D.J. Zeng, An alignment and matching network with hierarchical visual features for multimodal named entity and relation extraction, in: Proc. International Conference on Neural Information Processing, 2023.
https://doi.org/10.1007/978-981-99-8148-9_24
[29] M.H.Z. Jia, J. Shen, X. Shen, L.J. Liao, M. Chen, X.D. He, Z.D. Chen, J.Q. Li, MNER-QG: an end-to-end MRC framework for multimodal named entity recognition with query grounding, Proceedings of the AAAI Conference on Artificial Intelligence 37(7) 8032-8040.
https://doi.org/10.1609/aaai.v37i7.25971
[30] H.R. Wang, X. Xu, T. Wang, F.P. Chen, Multi-scale Visual Semantic Enhancement for Multimodal Named Entity Recognition Method, Acta Automatica Sinica 50(6)(2024) 1234-1245.
https://doi.org/10.16383/j.aas.c230573
[31] Z.H. Zheng, Z.H. Zhang, Z.X. Wang, R.J. Fu, M. Liu, Z.Y. Wang, B. Qin, Decompose, prioritize, and eliminate: dynamically integrating diverse representations for multimodal named entity recognition, in: Proc. of the 2024 Joint Inter-national Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024.

https://aclanthology.org/2024.lrec-main.403/
[32] Y. Zhang, G.X. Li, H. Gao, D.P. Dang, Multi-scale interaction network for multimodal entity and relation extraction, Information Sciences 699(2025) 121787.
https://doi.org/10.1016/j.ins.2024.121787
[33] X. Zhang, J.L. Yuan, L. Li, J.Q. Liu, Reducing the bias of visual objects in multimodal named entity recognition, in: Proc. of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023.
https://doi.org/10.1145/3539597.3570485
[34] B. Xu, S.Z. Huang, C.F. Sha, H.Y. Wang, MAF: A general matching and alignment framework for multimodal named entity recognition, in: Proc. of the fifteenth ACM international conference on web search and data mining, 2022.
https://doi.org/10.1145/3488560.3498475
[35] J. Cheng, K.F. Long, S. Zhang, T. Zhang, L.B. Ma, S. Cheng, Text-image scene graph fusion for multimodal named entity recognition, IEEE Transactions on Artificial Intelligence 5(6)(2024) 2828-2839.
https://doi.org/10.1109/TAI.2023.3326416
[36] W.X. Mai, Z.X. Zhang, K.T. Li, Y. Xue, F.H. Li, Dynamic graph construction framework for multimodal named entity recognition in social media, IEEE Transactions on Computational Social Systems 11(2)(2024) 2513-2522.
https://doi.org/10.1109/TCSS.2023.3303027
[37] P.P. Liu, H. Li, Y.M. Ren, J. Liu, S.Z. Si, H.S. Zhu, L.M. Sun, A novel framework for multimodal named entity recognition with multi-level alignments. <https://arxiv.org/abs/2305.08372v1>, 2023 (accessed 23.03.2025).
[38] Z.X. Zhang, W.X. Mai, H.L. Xiong, C.H. Wu, Y. Xue, A token-wise graph-based framework for multimodal named entity recognition, in: Proc. of the IEEE International Conference on Multimedia and Expo (ICME), 2023.
https://doi.org/10.1109/ICME55011.2023.00368
[39] Z. Liu, Y.T. Lin, Y. Cao, H. Hu, Y.X. Wei, Z. Zhang, S. Lin, B.N. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proc. of the IEEE/CVF international conference on computer vision, 2021.
https://doi.org/10.1109/ICCV48922.2021.00986
[40] T.W. Yan, S. Zhao, W.T. Ma, S.Z. Song, C.Y. Wang, Z.B. Rao, S.Z. Chen, Z.G. Luo, X.W, Liu, FRCL-MNER: A finer grained rank-based contrastive learning framework for multimodal NER, IEEE Transactions on Neural Networks and Learning Systems 36(6)(2025) 10779-10793.
https://doi.org/10.1109/TNNLS.2025.3528567
[41] J.F. Yu, J. Jiang, L. Yang, R. Xia, Improving multimodal named entity recognition via entity span detection with unified multimodal transformer, in: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
https://doi.org/10.18653/v1/2020.acl-main.306
[42] L.P. Liu, M.L. Wang, M.Z. Zhang, L.B. Qing, X.H. He, Uamner: uncertainty-aware multimodal named entity recognition in social media posts, Applied Intelligence 52(4)(2022) 4109-4125.
https://doi.org/10.1007/s10489-021-02546-5
[43] H.R. Wang, T. Wang, X. Xu, B.X. Jing, F.P. Chen, CLGLF: Confidence Learning Guides Label Fusion for Multimodal Named Entity Recognition Method, Acta Electronica Sinica 52(7)(2024) 2429-2437.
https://doi.org/10.12263/DZXB.20231160
[44] J. Wang, Y. Yang, K.Y. Liu, Z.P. Zhu, X.R. Liu, M3S: Scene graph driven multi-granularity multi-task learning for multi-modal NER, IEEE/ACM Transactions on Audio, Speech, and Language Processing 31(2022) 111-120.
https://doi.org/10.1109/TASLP.2022.3221017
[45] X.Y. Wang, M. Gui, Y. Jiang, Z.X. Jia, N., Bach, T. Wang, Z.Q. Huang, F. Huang, K.W. Tu, ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition, in: Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022.
https://doi.org/10.18653/v1/2022.naacl-main.232
[46] Y.P. Wang, C.M. Jiang, Fine-grained multimodal named entity recognition with heterogeneous image-text similarity graphs, International Journal of Machine Learning and Cybernetics 16(4)(2025) 2401-2415.
https://doi.org/10.1007/s13042-024-02398-8
[47] Y.X. Xu, H. Xu, H.Y. Huang, S.Y. Cui, M.H. Tang, L.Z. Wang, H.B. Xu, An effective span-based multimodal named entity recognition with consistent cross-modal alignment, in: Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024.
https://aclanthology.org/2024.lrec-main.95/
[48] G.H. Huang, Q. He, Z.H. Dai, G. Zhong, X.C. Yuan, C.M. Pun, GDN-CMCF: a gated disentangled network with cross-modality consensus fusion for multimodal named entity recognition, IEEE Transactions on Computational Social Systems 11(3)(2024) 3944-3954.
https://doi.org/10.1109/TCSS.2023.3323402
[49] Z. Wang, Y.J. Zhang, K. An, X.Y. Zhou, M.Y. Lu, H.F. Lin, P-MNER: Cross modal correction fusion network with prompt learning for multimodal named entity recognition, in: Proc. of the 22nd China National Conference on Chinese Computational Linguistics, 2023.
https://doi.org/10.1007/978-981-99-6207-5_13
[50] X. Xu, H.R. Wang, T. Wang, H. Ma, Textual-visual semantics-enhanced multimodal named entity recognition method, Application Research of Computers 41(6)(2024) 1679-1685.
https://doi.org/10.19734/j.issn.1001-3695.2023.09.0439

[51] C.R. Cai, Q.L. Wang, B. Qin, R.F. Xu, A multi-task framework based on decomposition for multimodal named entity recognition, Neurocomputing 604(2024) 128388.
https://doi.org/10.1016/j.neucom.2024.128388
[52] J.C. Zhao, S.Z. Huang, X. Lin, A graph interaction framework on relevance for multimodal named entity recognition with multiple images, in: Proc. of the 31st International Conference on Computational Linguistics, 2025.
https://aclanthology.org/2025.coling-main.82/
[53] G.H. Ding, W.J. Tang, Z.Y. Yuan, Entity label-guided graph fusion multi-modal named entity recognition, in: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2024.
https://doi.org/10.1109/SMC54092.2024.10831708
[54] Q. Zhang, J.L. Fu, X.Y. Liu, X.J. Huang, Adaptive co-attention network for named entity recognition in tweets, Proceedings of the AAAI conference on artificial intelligence 32(1)(2018) 5674-5681.
https://doi.org/10.1609/aaai.v32i1.11962
[55] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 2019.
https://doi.org/10.18653/v1/2020.acl-main.747
[56] X.W. Wang, J.B. Ye, Z.X. Li, J.F. Tian, Y. Jiang, M. Yan, J. Zhang, Y.H. Xiao, CAT-MNER: multimodal named entity recognition with knowledge-refined cross-modal attention, in: Proc. of the IEEE International Conference on Multimedia and Ex-po, 2022.
https://doi.org/10.1109/ICME52920.2022.9859972
[57] A.B. Guo, X. Zhao, Z. Tan, W.D. Xiao, MGICL: multi-grained interaction contrastive learning for multimodal named entity recognition, in: Proc. of the 32nd ACM International Conference on Information and Knowledge Management, 2023.
https://doi.org/10.1145/3583780.3614967
[58] J.Y. Li, H. Li, Z. Pan, G. Pan, Prompt ChatGPT In MNER: Improved multimodal named entity recognition method based on auxiliary refining knowledge from ChatGPT. <https://arxiv.org/abs/2305.12212v1>, 2023 (accessed 16.04.2025).
[59] J.Y. Li, H. Li, D. Sun, J.H. Wang, W.K. Zhang, Z. Wang, G. Pan, LLMs as Bridges: Reformulating grounded multimodal named entity recognition, <https://doi.org/10.48550/arXiv.2402.09989>, 2024 (accessed 16.04.2025).