

Enhancing Multimodal Named Entity Recognition through Adaptive Mixup Image Augmentation

Bo Xu¹, Haiqi Jiang¹, Jie Wei¹, Hongyu Jing¹, Ming Du^{1,*}, Hui Song¹,
Hongya Wang¹ and Yanghua Xiao²

¹School of Computer Science and Technology, Donghua University,

²Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University,
xubo@dhu.edu.cn, {2222674, weijie, 2232833}@mail.dhu.edu.cn,
{duming, songhui, hywang}@dhu.edu.cn, shawyh@fudan.edu.cn

Abstract

Multimodal named entity recognition (MNER) extends traditional named entity recognition (NER) by integrating visual and textual information. However, current methods still face significant challenges due to the text-image mismatch problem. Recent advancements in text-to-image synthesis provide promising solutions, as synthesized images can introduce additional visual context to enhance MNER model performance. To fully leverage the benefits of both original and synthesized images, we propose an adaptive mixup image augmentation method. This method generates augmented images by determining the mixing ratio based on the matching score between the text and image, utilizing a triplet loss-based Gaussian Mixture Model (TL-GMM). Our approach is highly adaptable and can be seamlessly integrated into existing MNER models. Extensive experiments demonstrate consistent performance improvements, and detailed ablation studies and case studies confirm the effectiveness of our method.

1 Introduction

Multimodal named entity recognition (MNER) extends traditional named entity recognition (NER) by integrating visual and textual information (Zhang et al., 2021). Unlike conventional NER, which relies solely on text, MNER incorporates images to enhance contextual understanding. This proves particularly beneficial in scenarios such as multimedia news extraction and product information retrieval on online platforms (Zheng et al., 2021). Current research in MNER typically focuses on optimizing modality representations (Zhang et al., 2018b; Chen et al., 2021), achieving modality alignment and fusion (Lu et al., 2018; Bao et al., 2023; Guo et al., 2023; Zeng et al., 2024; Zhou

Corresponding Author: Ming Du. The work reported in this paper is partially supported by the Fundamental Research Funds for the Central Universities 2232023D-19 and the NSF of Shanghai under grant number 22ZR1402000.

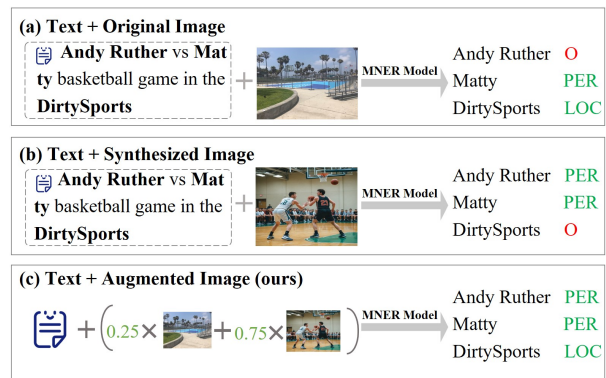


Figure 1: An example of using different image augmentation methods for the MNER task.

et al., 2024), and mitigating image noise interference (Sun et al., 2021; Xu et al., 2022; Zhang et al., 2023; He et al., 2024).

Despite significant advancements, current MNER methods struggle with the text-image mismatch problem. When the visual content of an image does not align with the corresponding textual information, it forces models to rely solely on text or, worse, make incorrect predictions due to image noise. For example, in Figure 1(a), the image lacks entities mentioned in the text, highlighting the inefficiency of existing methods in handling modality mismatches, thereby limiting model performance and accuracy.

Recent advancements in text-to-image synthesis offer promising solutions to these limitations (Ramesh et al., 2022; Nichol et al., 2021). Technologies such as Stable Diffusion (Rombach et al., 2021) and DALL-E (Ramesh et al., 2022) can generate visually relevant content based on text inputs, enhancing MNER model performance. While synthesized images align better with textual information, reducing mismatches, they often lack the rich semantic detail in real images. Conversely, though semantically richer, real images tend to suffer from text-image mismatches.

To leverage the benefits of both real and synthe-

sized images, we propose using the mixup image augmentation method (Zhang et al., 2018a). Traditional mixup techniques usually apply random ratios for data and label mixing. Since labels cannot be altered in the MNER context, the mixing ratio becomes particularly crucial. We introduce an adaptive mixup image augmentation model that determines the ratio of original to synthesized images based on a matching score derived from a triplet loss-based Gaussian mixture model (TL-GMM). This adaptable method can be seamlessly integrated into existing MNER models, replacing original images with augmented ones to significantly enhance performance.

Our main contributions can be summarized as follows:

- We are the first to propose the use of synthesized images to address the text-image mismatch problem in MNER tasks, enhancing performance by integrating additional visual context.
- We introduce a novel adaptive mixup image augmentation method, employing a triplet loss-based Gaussian mixture model to determine the mixing ratio of original and synthesized images. This method acts as a plugin that seamlessly integrates into existing multimodal named entity recognition models.
- We conduct extensive experiments on multiple MNER models, demonstrating consistent performance improvements with our augmented images. Detailed ablation studies and case analyses confirm the effectiveness and advantages of our adaptive mixing ratio setting.

2 Overview

2.1 Problem Formulation

The task of multimodal named entity recognition (MNER) aims to extract named entities from a given text $T = \{t_1, t_2, \dots, t_n\}$ and its associated original image I , classifying these entities into predefined categories to produce an output set $Y = \{y_1, y_2, \dots, y_n\}$, where each y_i is a label selected from a predefined label set according to the BIOES-style annotation scheme.

This paper focuses on an enhanced version of the MNER task that incorporates image augmentation. Initially, an augmented image V_{mix} is generated using the text T and the original image I . The

objective is to perform multimodal named entity recognition based on the text T and the augmented image V_{mix} .

2.2 Framework

As shown in Figure 2, the left diagram outlines the overall architecture of the MNER task, which includes four components: input, adaptive mixup image augmentation model (AMIA), MNER model, and output. Our proposed AMIA model, which is the core of our approach, aims to generate augmented images that better match the text, replacing the original images as the input for the MNER model.

The right diagram in Figure 2 details the structure of our proposed AMIA model. It consists of four main modules: the input representation module, the text-image matching module, the text-to-image generation module, and the mixup image augmentation module. Firstly, the input representation module generates representations for both the text and the original image. Secondly, the text-image matching module calculates the text-image matching score using a triplet loss-based Gaussian mixture model (TL-GMM) to obtain an adaptive matching score. Thirdly, the text-to-image generation module generates a synthesized image that matches the text. Lastly, the mixup image augmentation module blends the original and synthesized images based on the adaptive matching score to produce the augmented image.

These augmented images, along with the text, serve as inputs to the MNER model. The MNER model, which can be any existing model, processes these inputs to enhance overall performance.

3 Method

In this section, we provide a detailed explanation of how augmented images are obtained using our proposed adaptive mixup image augmentation model, which consists of four main modules: input representation, text-image matching, text-to-image generation, and mixup image augmentation.

3.1 Input Representations Module

The input representations module is responsible for extracting representations of the text and the original image, which are crucial for the subsequent text-image matching calculation. As illustrated in Figure 2, we employ the multimodal vision and language pre-trained model CLIP (Radford et al.,

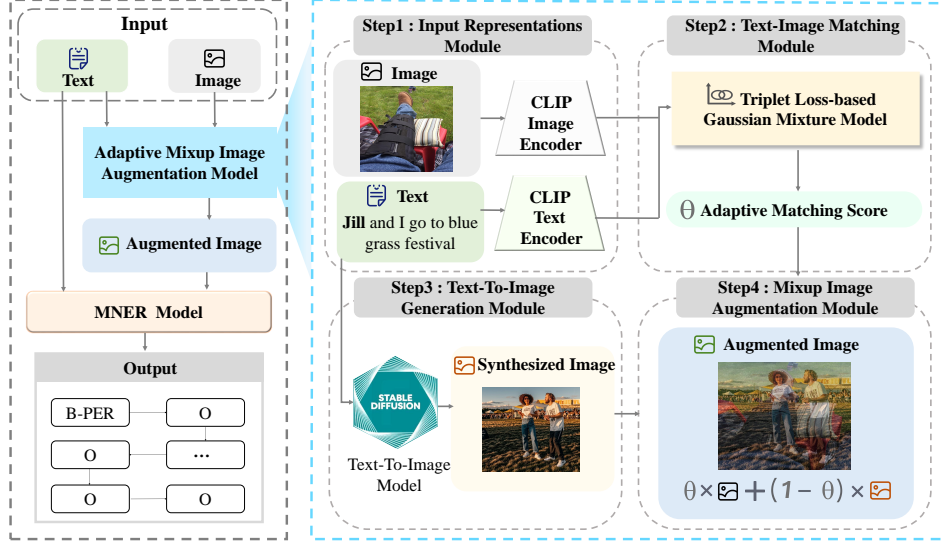


Figure 2: The workflow of our proposed adaptive mixup augmentation framework.

2021) as our modality-specific encoder to obtain these representations. CLIP’s text and image encoders project the visual and textual modalities into a shared embedding space, thus facilitating effective alignment between the two.

For the text input, the CLIP text encoder is used to encode it. Given an input text T , we first tokenize it using byte pair encoding (BPE), resulting in a token sequence (t_1, t_2, \dots, t_n) , where n denotes the sequence length. Special tokens $[SOS]$ and $[EOS]$ are added at the beginning and end of the sequence, respectively, forming $([SOS], t_1, t_2, \dots, t_n, [EOS])$. The representation of the entire text, denoted as $T_s \in \mathbb{R}^d$, is then obtained from the activation of the $[SOS]$ token in the last layer of the CLIP text encoder.

For the image input, the CLIP image encoder is employed to encode the image. An input original image I is first resized to 224×224 pixels. The image is then divided into non-overlapping 16×16 patches, which are linearly embedded to produce their representations $(i_1, i_2, \dots, i_{196})$. A $[CLS]$ token, with the same dimension as the patch embeddings, is prepended to the sequence, resulting in $([CLS], i_1, i_2, \dots, i_{196})$. The representation of the image, denoted as $I_s \in \mathbb{R}^d$, is derived from the activation of the $[CLS]$ token in the last layer of the CLIP image encoder.

3.2 Text-Image Matching Module

The text-image matching module is designed to evaluate the compatibility between text and original images by calculating a matching score. Inspired

by prior research on noisy label learning (Han et al., 2023; Huang et al., 2021), which revealed that clean data tends to have a lower loss than noisy data during early stages of training due to the memory effect of deep neural networks (Arpit et al., 2017), we utilize this loss difference to distinguish between matched and mismatched text-image pairs. Building on this insight, we employ a triplet loss-based Gaussian mixture model (TL-GMM) to assess text-image matching and generate adaptive matching scores θ . The specific process is as follows.

Firstly, we compute the triplet loss between text-image pairs. Given an input text-image pair (T_s, I_s) in a batch of size N , where T_s and I_s are obtained from the input representation module, we employ a bidirectional triplet loss to comprehensively measure text-image matching. This loss includes both image-to-text and text-to-image triplet losses. The image-to-text triplet loss is defined as follows:

$$L_s^{(I \rightarrow T)} = \max(0, m - pos_s^{(I \rightarrow T)} + neg_s^{(I \rightarrow T)}) \quad (1)$$

$$pos_s^{(I \rightarrow T)} = \frac{I_s}{\|I_s\|} \cdot \frac{T_s}{\|T_s\|} \quad (2)$$

$$neg_s^{(I \rightarrow T)} = \max_{i \in N} \left(\frac{I_s}{\|I_s\|} \cdot \frac{T_i}{\|T_i\|} \right), \quad i \neq s. \quad (3)$$

Here, m is a positive margin coefficient. The text-to-image triplet loss is defined as follows:

$$L_s^{(T \rightarrow I)} = \max(0, m - pos_s^{(T \rightarrow I)} + neg_s^{(T \rightarrow I)}) \quad (4)$$

$$pos_s^{(T \rightarrow I)} = \frac{T_s}{\|T_s\|} \cdot \frac{I_s}{\|I_s\|} \quad (5)$$

$$neg_s^{(T \rightarrow I)} = \max_{i \in N} \left(\frac{T_s}{\|T_s\|} \cdot \frac{I_i}{\|I_i\|} \right), \quad i \neq s. \quad (6)$$

We sum up the two losses to obtain the final triplet loss L_s :

$$L_s = L_s^{(I \rightarrow T)} + L_s^{(T \rightarrow I)} \quad (7)$$

Secondly, we fit these triplet losses using a Gaussian mixture model (GMM). The triplet loss L_s is modeled with a two-component GMM to effectively distinguish between matched and mismatched text-image pairs. Representing the triplet losses with two Gaussian components helps separate the lower losses (typically corresponding to matched pairs) from the higher losses (typically corresponding to mismatched pairs). We optimize the GMM using the expectation-maximization algorithm and calculate the posterior probability of the two components, which serves as a measure of text-image matching quality. The posterior probability is defined as follows:

$$p(k|L_s) = \frac{p(k)p(L_s|k)}{p(L_s)}. \quad (8)$$

Here, $k \in \{0, 1\}$ indicates whether it is a matched or mismatched component.

Finally, we choose the posterior probability of $k = 0$ as the adaptive matching score θ for the text-image pair (T_s, I_s) . The final adaptive matching score is $\theta = p(k = 0|L_s)$.

3.3 Text-To-Image Generation Module

The text-to-image generation module is designed to generate synthesized images based on the input text. Models utilizing generative adversarial networks and variational autoencoders (Nichol et al., 2021; Ramesh et al., 2022; Saharia et al., 2022) are capable of capturing subtle semantic information in textual descriptions and converting it into visual features. Leveraging this capability, we use these models to generate images that match the text, thereby mitigating noise introduced by text-image mismatch.

Specifically, we employ the *stable diffusion* (SD) model (Rombach et al., 2021) for image generation. The stable diffusion model is renowned for its ability to generate high-quality images that accurately correspond to the input text. By inputting the text T into the stable diffusion model, we obtain the synthesized images $V = SD(T)$. For instance, as illustrated in Figure 2, using the text "Jill and I go to a bluegrass festival" as input, the Stable

diffusion model generates a synthesized image that accurately reflects the textual description.

3.4 Mixup Image Augmentation Module

The mixup image augmentation module is designed to combine the original image I with the synthesized image V using an adaptive matching score θ to generate augmented images.

Specifically, we employ the mixup method (Zhang et al., 2018a) to linearly interpolate between the original and synthesized images, thus producing an augmented image. Unlike traditional methods, which mix both data and labels, we perform mixups solely on the images and do not alter the labels. To balance the contributions of the original and synthesized images, we introduce the adaptive matching score θ . The process is detailed as follows:

$$V_{mix} = \theta \times I + (1 - \theta) \times V. \quad (9)$$

Here, θ represents the adaptive matching score derived from the text-image matching module. This dynamic score adjusts the weight between the original and synthesized images, ensuring that the augmented image V_{mix} provides a balanced and contextually relevant visual representation. This augmented image, along with the text, serves as input to the MNER model, thereby enhancing overall recognition performance.

4 Experiment

4.1 Dataset

We evaluate our proposed method on two widely used datasets in MNER tasks: Twitter 2015 and Twitter 2017 (Xu et al., 2022; Lu et al., 2018). Each dataset consists of text-image pairs where the textual content may or may not correspond to the content in the image. Additionally, the text may contain zero or more named entities. The entities are categorized into four types: Person (PER), Organization (ORG), Location (LOC), and Miscellaneous (MISC).

4.2 Evaluation Metrics

For the MNER task, an entity is deemed accurately identified if both its span and entity type match the gold standard. We evaluate the performance of our proposed method using overall precision (P), recall (R), and F1 score (F1), which are standard metrics in MNER tasks (Chen et al., 2022; Zhou et al., 2022).

Data	Method	Twitter 2015			Twitter 2017		
		P	R	F1	P	R	F1
Text	BiLSTM-CRF (Huang et al., 2015)	68.14	61.09	64.42	79.42	73.43	76.31
	CNN-BiLSTM-CRF (Ma and Hovy, 2016)	66.24	68.09	67.15	80.00	78.76	79.37
	HBiLSTM-CRF (Lample et al., 2016)	70.32	68.05	69.17	82.69	78.16	80.17
	BERT (Devlin, 2018)	68.30	74.61	71.32	82.19	83.72	82.95
	BERT-CRF (Devlin, 2018)	69.22	74.59	71.81	83.32	83.57	83.44
Text+Original Image	GVATT-HBiLSTM-CRF (Lu et al., 2018)	69.15	74.46	71.70	83.64	84.38	84.01
	AdaCAN-CNN-BiLSTM-CRF (Zhang et al., 2018b)	69.87	74.59	72.15	85.13	83.20	84.10
	RpBERT (Sun et al., 2021)	71.15	74.30	72.69	82.85	84.38	83.61
	ViLBERT (Wei et al., 2024)	73.00	74.37	73.68	83.63	85.86	84.73
	UMGF (Zhang et al., 2021)	74.49	75.21	74.85	86.54	84.50	85.51
	MAFN (Zhou et al., 2024)	71.99	75.19	73.56	85.66	85.79	85.72
	SMVAE (Zhou et al., 2022)	74.40	75.76	75.07	85.77	86.97	86.37
	DebiasCL (Zhang et al., 2023)	74.45	76.13	75.28	87.59	86.11	86.84
	MRC-MNER (Jia et al., 2022)	78.10	71.45	74.63	88.78	85.00	86.85
	HVPNeT (Chen et al., 2022)	73.87	76.82	75.32	85.84	87.93	86.87
	R-GCN (Zhao et al., 2022)	73.95	76.18	75.00	86.72	87.53	87.11
	UMT* (Yu et al., 2020)	70.24	75.671	72.86	84.04	85.34	84.69
	MAF* (Xu et al., 2022)	70.98	75.05	72.96	85.08	84.83	84.95
	BERT-ResNet-CRF* (Wang et al., 2022)	73.17	75.98	74.55	86.11	86.75	86.43
Text+Augmented Image	UMT* + AMIA	73.48	73.68	73.58	85.16	85.79	85.47
	MAF* + AMIA	73.33	73.95	73.64	85.09	86.16	85.62
	BERT-ResNet-CRF* + AMIA	75.08	76.21	75.62	87.23	88.00	87.62

Table 1: Comparison results on two MNER datasets. Methods marked with an * are reproduced by ours.

4.3 Parameter Settings

All experiments are conducted on an NVIDIA RTX 3090 GPU using PyTorch 1.7.1. The parameter settings for our framework are as follows: For the input representation module, we use CLIP to obtain the representations of text and original images. For the text-image matching module, we set the batch size to 64 and the positive margin to 0.1. For the text-to-image generation module, we utilize stable diffusion turbo to generate synthesized images from textual descriptions.

4.4 Baselines

To evaluate the effectiveness of our proposed method, we apply it to several existing MNER models and compare their performance under identical settings. Specifically, we compare the performance of these models when using text and original images versus using text and augmented images generated by our adaptive mixup image augmentation method. Additionally, we compare our approach against the state-of-the-art (SOTA) methods to provide a comprehensive evaluation. The baselines we consider are as follows.

For text-based models, we select five methods: BiLSTM-CRF (Huang et al., 2015), CNN-BiLSTM-CRF (Ma and Hovy, 2016), HBiLSTM-CRF (Lample et al., 2016), BERT (Devlin, 2018),

<https://huggingface.co/openai/clip-vit-base-patch16>
<https://huggingface.co/stabilityai/sdxl-turbo>

and BERT-CRF (Devlin, 2018).

For multimodal models, we select fourteen methods: GVATT-HBiLSTM-CRF (Lu et al., 2018), AdaCAN-CNN-BiLSTM-CRF (Zhang et al., 2018b), RpBERT (Sun et al., 2021), ViLBERT (Wei et al., 2024), UMT (Yu et al., 2020), UMGF (Zhang et al., 2021), MAFN (Zhou et al., 2024), MAF (Xu et al., 2022), SMVAE (Zhou et al., 2022), DebiasCL (Zhang et al., 2023), HVPNeT (Chen et al., 2022), MRC-MNER (Jia et al., 2022), R-GCN (Zhao et al., 2022). Specifically, to validate the effectiveness of our method, we reproduce the UMT, MAF, and BERT-ResNet-CRF models and compare the effects of using original images versus augmented images.

4.5 Overall Performance

We conducted experiments on two multimodal datasets, Twitter 2015 and Twitter 2017. As shown in Table 1, we report the overall Precision (P), Recall (R), and F1 score (F1) for both datasets.

Firstly, we compared multimodal models with text-based unimodal methods and observed that all multimodal models outperform text-based methods. This finding demonstrates that incorporating images enhances model performance in multimodal named entity recognition tasks. The images provide additional context and details to the text, especially for ambiguous entities, helping the model to better understand and distinguish them. Moreover, multimodal models capture semantic correlations and consistency more effectively, thereby

improving the robustness and accuracy of the overall model.

Secondly, we compared the performance of three classic MNER models (i.e., UMT, MAF, and BERT-ResNet-CRF) using both original and augmented images. The results indicate that integrating augmented images significantly outperforms using original images. This validates the generalizability and effectiveness of the augmented images in enhancing MNER model performance. One reason for this improvement is that augmented images reduce noise from mismatched images and provide additional image semantics that complement the text. For the BERT-ResNet-CRF model, using augmented images resulted in the best performance, highlighting the benefits of noise reduction and additional semantics provided by augmented images. For the UMT and MAF models, although our reproduced results are slightly lower than those reported in the original papers, using augmented images still improved performance. This suggests that our augmented images are robust across different models and settings.

4.6 Ablation Study

In this section, we conduct ablation studies to verify the effectiveness of our proposed model. Specifically, we compare the results of using different mixing strategies and examine the effects of using matched and mismatched text-image pairs.

4.6.1 Comparison of Different Mixing Strategies

We compare the results of various mixing strategies: 1) no mixing, meaning using either only synthesized images ($\theta = 0$) or only original images ($\theta = 1$); 2) fixed mixing ratios ($\theta = 0.3, 0.5, 0.7$); and 3) dynamic mixing ratios, including cosine similarity between text and images, and the text-image matching module proposed in this paper. The results are presented in Table 2.

Firstly, we compare the effectiveness of dynamic mixing ratios against no mixing at all. The non-mixing methods involve using only original images ($\theta = 1$) or only synthesized images ($\theta = 0$). The experimental results across three different multimodal named entity recognition models show that dynamic mixing ratios outperform non-mixing methods. This indicates that original and synthesized images are complementary, and the best performance is achieved when both are utilized together.

Method	Mixing Ratios	Twitter 2015			Twitter 2017		
		P	R	F1	P	R	F1
UMT+AMIA	Ours	73.48	73.68	73.58	85.16	85.79	85.47
	cosine	71.91	74.84	73.37	85.11	85.05	85.08
	$\theta = 1.0$	70.24	75.67	72.86	84.07	85.21	84.64
	$\theta = 0.0$	71.44	75.67	73.49	85.08	84.31	85.06
	$\theta = 0.7$	70.01	75.16	72.50	82.82	86.69	84.71
	$\theta = 0.5$	72.01	74.02	73.00	83.06	86.67	84.83
	$\theta = 0.3$	71.51	74.42	72.94	84.94	84.81	84.87
MAF+AMIA	Ours	73.33	73.95	73.64	85.09	86.16	85.62
	cosine	73.31	75.25	73.28	85.14	85.64	85.39
	$\theta = 1.0$	70.98	75.05	72.96	85.08	84.83	84.95
	$\theta = 0.0$	70.68	74.62	72.65	83.35	86.23	84.74
	$\theta = 0.7$	70.11	75.09	72.58	83.81	85.89	84.84
	$\theta = 0.5$	70.39	75.17	72.70	82.79	87.29	84.98
	$\theta = 0.3$	71.42	74.99	73.16	84.66	85.58	85.12
BERT-ResNet-CRF+AMIA	Ours	75.08	76.21	75.62	87.23	88.00	87.62
	cosine	74.59	75.89	75.24	87.04	87.50	87.26
	$\theta = 1.0$	73.17	75.98	74.55	86.11	86.75	86.43
	$\theta = 0.0$	73.71	76.27	74.97	85.86	88.49	87.16
	$\theta = 0.7$	74.37	74.85	74.61	86.73	86.60	86.67
	$\theta = 0.5$	74.03	75.00	74.51	86.87	86.68	86.77
	$\theta = 0.3$	74.91	74.97	74.94	86.10	87.56	86.83

Table 2: Comparison results of different MNER models using different mixing strategies.

Secondly, we compare the performance of dynamic mixing ratios with fixed mixing ratios. The experimental results across three different multimodal named entity recognition models indicate that dynamic mixing ratios outperform fixed mixing ratios, underscoring the necessity of the proposed dynamic mixing approach.

Lastly, we compare the performance of different dynamic mixing methods. The experimental results across three different MNER models show that the dynamic mixing methods based on the triplet loss-based Gaussian Mixture Model proposed in this paper outperform those based on cosine similarity. This demonstrates the effectiveness of the triplet loss-based Gaussian mixture model methods proposed in this paper.

4.6.2 Effects of Mismatched and Matched Text-Image Pairs

We compare the performance of different MNER models on matched and mismatched text-image pairs. The determination of matched and mismatched pairs is based on the matching score proposed in this paper. If $\theta > 0.5$, it indicates that the text and image are matched; otherwise, they are mismatched. We split the test sets of Twitter2015 and Twitter2017 into two parts and compared the performance of different models on these distinct subsets. The results are shown in Table 3.

Firstly, we compare the performance of all methods on matched and mismatched text-image pairs. The experimental results consistently show that using matched text-image pairs outperforms using mismatched pairs. This aligns with common understanding and further validates that our pro-

Method	Image	Twitter 2015			Twitter 2017		
		P	R	F1	P	R	F1
Mismatched Text-Image Pairs							
UMT	Original	70.73	73.75	72.21	82.32	85.89	84.07
	Synthesized	70.87	74.48	72.63	84.59	84.53	84.56
	Augmented	70.06	75.49	72.68	83.38	86.20	84.76
MAF	Original	70.52	74.12	72.28	83.61	84.65	84.13
	Synthesized	71.10	73.84	72.76	83.43	85.89	84.64
	Augmented	70.77	75.16	72.86	83.56	85.89	84.71
BERT-ResNet-CRF	Original	73.21	75.49	74.33	85.85	86.51	86.18
	Synthesized	72.80	76.36	74.54	86.51	86.90	86.71
	Augmented	74.00	76.66	74.82	86.98	86.81	86.90
Matched Text-Image Pairs							
UMT	Original	72.87	76.04	74.42	85.53	86.16	85.84
	Synthesized	73.13	76.04	74.54	85.25	86.16	85.70
	Augmented	73.07	76.32	74.66	86.05	85.94	85.98
MAF	Original	74.16	75.08	74.55	84.83	86.53	85.67
	Synthesized	73.36	75.77	74.51	84.95	86.53	85.73
	Augmented	73.59	75.86	74.71	85.88	85.71	85.78
BERT-ResNet-CRF	Original	73.21	75.49	74.33	85.78	88.82	87.23
	Synthesized	74.00	76.66	74.48	87.84	87.19	87.51
	Augmented	74.19	74.81	74.50	86.84	88.45	87.64

Table 3: Comparison results of different MNER models on mismatched and matched text-image pairs.

posed method can effectively distinguish between matched and mismatched text-image pairs.

Secondly, we compare the performance of different image strategies on matched text-image pairs. The experimental results indicate that when text and images are matched, using either original images or synthesized images achieves similar performance, both of which are lower than using augmented images. This suggests that synthesized images can indeed complement original images to enhance the performance of multimodal named entity recognition tasks.

Lastly, we compare the performance of different image strategies on mismatched text-image pairs. The experimental results show that when text and images are mismatched, synthesized images outperform original images, and both are less effective than using augmented images. This also demonstrates the effectiveness of our proposed mixing strategy.

4.7 Case Study

We select two representative samples from the test set to verify the effectiveness of our proposed adaptive mixup image augmentation method. The details are as follows and are illustrated in Table ??:

In the first case, the augmented image supplements the missing semantic information of the original image through a synthesized image. The text contains two entities, "Manchester" and "Ariana Grande", but the original image only depicts a scene related to "Manchester". The augmented image adds synthesized content related to "Ariana Grande". The augmented image incorporates in-

formation from both entities by utilizing adaptive matching scores, correcting the model's initial error of recognizing only "Manchester".

In the second case, the augmented image effectively filters out noise that does not match the original image context. The original image contains only a segment of text, causing the model to incorrectly identify the entity "Hemingway" as miscellaneous. The synthesized image includes a portrait of Hemingway, and the augmented image, optimized through adaptive matching scores, emphasizes the relevant synthesized content while reducing the noise impact from the original image. This allows the model to correctly recognize "Hemingway" as the correct entity type.

5 Related Work

5.1 Multimodal Named Entity Recognition

Multimodal named entity recognition (MNER) has garnered significant attention in recent years. It aims to enhance the accuracy and robustness of entity recognition by integrating information from both text and images (Zhang et al., 2023; Liu et al., 2024; Zhou et al., 2024). Existing research primarily concentrates on modality fusion and alignment and mitigating image noise interference. ITA (Wang et al., 2022) aligns images with regional object tags, image-level captions, and optical characters as visual contexts. These are concatenated with input texts to form a new cross-modal input, which is then fed into a pre-trained textual embedding model. HamLearning (Liu et al., 2023) proposes dynamically aligning image and text sequences to achieve multi-level cross-modal learning, thereby enhancing text word representation. The cross-modal matching module of MAF (Xu et al., 2022) and the fine-grained visual feature extraction method of P-MNER (Zhuang et al., 2023) aims to reduce noise by selectively filtering out irrelevant regions of the image. Additionally, (Lu et al., 2018) proposes an attention mechanism-based model to extract visual features from image areas most relevant to the text while ignoring irrelevant visual information. However, these models do not truly address the issue of mismatched images and texts. When the images are mismatched, these models can only rely on text information. This paper proposes using a text-to-image model to generate images related to the text, addressing the limitations of the above multimodal named entity recognition models.






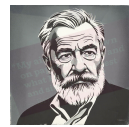
	Case 1			Case 2		
Image						
Text	Ariana Grande [PER] arrives in Manchester [LOC]			The strategy of Hemingway [PER]		
Adaptive Matching Score	0.4739			0.0103		
BERT-ResNet-CRF	O, LOC X			MISC X		
BERT-ResNet-CRF+AMIA	PER, LOC ✓			PER ✓		

Table 4: Two cases demonstrating the importance of the AMIA model.

5.2 Cross-modal Matching

Cross-modal matching aims to establish a correspondence between two modalities (Huang et al., 2021; Zha et al., 2024). It has extensively been researched within the multimodal field and serves as the basis for tasks such as cross-modal retrieval (Huang et al., 2023), visual question answering (Guo et al., 2019), and text-image matching. Existing matching methods primarily learn modality-specific feature representations and then align the two modalities in a common embedding space (Chen et al., 2021; He et al., 2021). VSE++ (Faghri et al., 2017) proposed a hard negative sample mining strategy applied to the ranking loss to improve discriminative embeddings of each specific modality. CHAN (Pan et al., 2023) uses hard assignment codes to mine informative region-word pairs and filters out mismatched alignments. These approaches are based on the ideal assumption of perfect cross-modal matching. However, most data are not perfectly matched, introducing noise that reduces model performance. Consequently, noisy correspondence learning has emerged as an important research direction in this field (Han et al., 2023; Huang et al., 2024). Noisy correspondence rectify (NCR) (Huang et al., 2021) involves processing image data with label noise and then dividing the data into clean and noisy datasets based on a Gaussian mixture model fitted with each sample loss. In this paper, we propose a new noisy label paradigm by replacing traditional label noise with images that do not match the text, thereby measuring the degree of text-image matching.

5.3 Image Augmentation

Image augmentation research primarily seeks to enhance image quality and visual effects, expand existing datasets, and improve model generaliza-

tion capabilities (Wang et al., 2024). Traditional methods often rely on geometric transformations such as rotation, translation, cropping, resizing, and flipping (Karen, 2014; Zhong et al., 2020). With the advent of deep learning, techniques like AutoAugment (Cubuk et al., 2019) and RandAugment (Cubuk et al., 2020) automatically select augmentation operations based on search strategies. Additionally, mixup augmentation (Zhang et al., 2018a) generates more diverse image features through linear interpolation between samples. However, these methods are limited to transformations of the original image and fail to fully utilize advanced techniques, such as text-to-image models, to enhance the alignment between images and task objectives, thus restricting their performance in more complex tasks. In this paper, the mixup method is employed to combine original and synthetic images, generating augmented images to enhance the performance of the MNER model.

6 Conclusion

In this paper, we propose an adaptive mixup image augmentation model to address the text-image mismatch issue in multimodal named entity recognition (MNER) tasks. Our method employs a triplet loss-based Gaussian mixture model to determine the matching score between original text-image pairs, generates synthesized images using a text-to-image model, and then mixes the original and synthesized images based on matching scores to create an augmented image. Extensive experiments demonstrate consistent performance improvements across various MNER models. Detailed ablation studies and case analyses confirm the effectiveness of our approach, which can be seamlessly integrated into existing MNER frameworks to enhance their robustness and accuracy.

Limitations

While our proposed adaptive mixup image augmentation encoder can act as an image input plugin for other MNER models, it has two main limitations. First, existing methods might employ multi-granularity image information, such as object labels and scene graphs. Using our augmented images in such contexts may lead to information confusion, limiting the applicability of our plugin to MNER models that require multi-granularity image processing. Second, the quality of our synthesized images is constrained by the performance of current text-to-image models, which can affect the overall enhancement effect.

References

- Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR.
- Xigang Bao, Mengyuan Tian, Zhiyuan Zha, and Biao Qin. 2023. Mpmrc-mner: A unified mrc framework for multimodal named entity recognition based multimodal prompt. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 47–56.
- Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798.
- Xiang Chen, Ningyu Zhang, Lei Li, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1607–1618.
- Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123.
- Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.
- Aibo Guo, Xiang Zhao, Zhen Tan, and Weidong Xiao. 2023. Mgiel: multi-grained interaction contrastive learning for multimodal named entity recognition. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 639–648.
- Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10434–10443.
- Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. 2023. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7517–7526.
- Li He, Qingxiang Wang, Jie Liu, Jianyong Duan, and Hao Wang. 2024. Visual clue guidance and consistency matching framework for multimodal named entity recognition. *Applied Sciences*, 14(6):2333.
- Yi He, Xin Liu, Yiu-Ming Cheung, Shu-Juan Peng, Jinhan Yi, and Wentao Fan. 2021. Cross-graph attention enhanced multi-modal correlation learning for fine-grained image-text retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1865–1869.
- Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. 2023. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6565–6574.
- Zhenyu Huang, Peng Hu, Guocheng Niu, Xinyan Xiao, Jiancheng Lv, and Xi Peng. 2024. Learning with noisy correspondence. *International Journal of Computer Vision*, pages 1–22.
- Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Meihuizi Jia, Xin Shen, Lei Shen, Jinhui Pang, Lejian Liao, Yang Song, Meng Chen, and Xiaodong He. 2022. Query prior matters: A mrc framework for multimodal named entity recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3549–3558.

- Simonyan Karen. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.
- Peipei Liu, Hong Li, Yimo Ren, Jie Liu, Shuaizong Si, Hongsong Zhu, and Limin Sun. 2023. A novel framework for multimodal named entity recognition with multi-level alignments. *arXiv preprint arXiv:2305.08372*.
- Wei Liu, Aiqun Ren, Chao Wang, Yan Peng, Shaorong Xie, and Weimin Li. 2024. Mvnp: Multi-granularity visual prompt-guided fusion network for multimodal named entity recognition. *Multimedia Tools and Applications*, pages 1–25.
- Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.
- Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.
- Zhengxin Pan, Fangyu Wu, and Bailing Zhang. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19275–19284.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv e-prints*, pages arXiv–2204.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. [High-resolution image synthesis with latent diffusion models](#). *Preprint*, arXiv:2112.10752.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.
- Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.
- Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022. Ita: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189.
- Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C Aggarwal, Jian Pei, and Yuanchun Zhou. 2024. A comprehensive survey on data augmentation. *arXiv preprint arXiv:2405.09591*.
- Pengfei Wei, Hongjun Ouyang, Qintai Hu, Bi Zeng, Guang Feng, and Qingpeng Wen. 2024. Vecmner: Hybrid transformer with visual-enhanced cross-modal multi-level interaction for multimodal ner. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 469–477.
- Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. 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*, pages 1215–1223.
- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.
- Qingyang Zeng, Minghui Yuan, Jing Wan, Kunfeng Wang, Nannan Shi, Qianzi Che, and Bin Liu. 2024. Icka: An instruction construction and knowledge alignment framework for multimodal named entity recognition. *Expert Systems with Applications*, 255:124867.
- Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. 2024. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 852–861.
- Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multimodal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the*

- AAAI conference on artificial intelligence*, volume 35, pages 14347–14355.
- Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018a. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.
- Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018b. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Xin Zhang, Jingling Yuan, Lin Li, and Jianquan Liu. 2023. 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*, pages 958–966.
- Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. 2022. 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*, pages 3983–3992.
- Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*.
- Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008.
- Baohang Zhou, Ying Zhang, Kehui Song, Wenya Guo, Guoqing Zhao, Hongbin Wang, and Xiaojie Yuan. 2022. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6293–6302.
- Xiaoying Zhou, Yijia Zhang, Zhuang Wang, Mingyu Lu, and Xiaoxia Liu. 2024. Mafn: multi-level attention fusion network for multimodal named entity recognition. *Multimedia Tools and Applications*, 83(15):45047–45058.
- Wang Zhuang, Zhang Yijia, An Kang, Zhou Xiaoying, Lu Mingyu, and Lin Hongfei. 2023. P-mner: Cross modal correction fusion network with prompt learning for multimodal named entity recognition. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 689–700.