Consistency-Heterogenity Balanced Fake News Detection via Cross-modal Matching

Ying Guo, Bingxin Li, Kexin Zhen, Jie Liu, Gaolei Li, Qi Wang, Yong-Jin Liu

***Abstract*—Generating synthetic content through Generative AI (GAI) presents considerable hurdles for current fake news detection methodologies. Many existing detection approaches concentrate on feature-based multi-modal fusion, neglecting semantic relationships such as correlations and diversities. In this study, we introduce an innovative cross-modal matching-driven approach to reconcile semantic relevance (text-image consistency) and semantic gap (text-image heterogeneity) in multi-modal fake news detection. Unlike the conventional paradigm of multi-modal fusion followed by detection, our approach integrates textual modality, visual modality (images), and text embedded within images (auxiliary modality) to construct an end-to-end framework. This framework considers the relevance of contents across different modalities while simultaneously addressing the gap in structures, achieving a delicate balance between consistency and heterogeneity. Consistency is fostered by evaluating inter-modality correlation via pairwise-similarity scores, while heterogeneity is addressed by employing cross-attention mechanisms to account for inter-modality diversity. To achieve equilibrium between consistency and heterogeneity, we employ attention-guided enhanced modality interaction and similarity-based dynamic weight assignment to establish robust frameworks. Comparative experiments conducted on the Chinese Weibo dataset and the English Twitter dataset demonstrate the effectiveness of our approach, surpassing the state-of-the-art by 7% to 13%.**

***Impact Statement*—Fake news detection is a popular technology in social media's security. It helps to clean social contents and create harmonious environments to a large degree by means of Artificial Intelligence. However, recent research has demonstrated that two paradoxical concepts, i.e. text-image consistency and text-image heterogeneity contribute to the authenticity of a news together. The consistency-heterogeneity balanced algorithm we introduce in this paper overcame these limitations. With a significant increase in detection accuracy by 7% to 13% on Chinese Weibo dataset and the English Twitter dataset, the technology is ready to support users in a wide variety of applications including governmental decisions, industrial designs, and individual overcomes.**

***Index Terms*—multimodal fake news detection, cross-modal matching, text-image consistency, text-image heterogentiy.**

I. Introduction

Ying Guo, Bingxin Li , Kexin Zhen and Jie Liu are with the School of Information Science, North China University of Technology, Beijing, 100144, China (email: guoying@ncut.edu.cn; liujxxxy@126.com; libingxin1126@163.com; kexinzhenzhen@163.com). Gaolei Li is with the institute of Cyber Science and Technologyemail: gaoleili@sjtu.edu.cn). Qi Wang is with the State Key Laboratory of Public Big Data, Guizhou University, Guiyang, 550025, China (email: qiwang@gzu.edu.cn). Yong-Jin Liu is with the Department of Computer Science and Technology, Tsinghua University, Beijing, China(email: liuyongjin@tsinghua.edu.cn) (Corresponding author: Yong-Jin Liu)

Ying Guo is also with the departmentment of computer science, Tsinghua University, Beijing, 100084, China

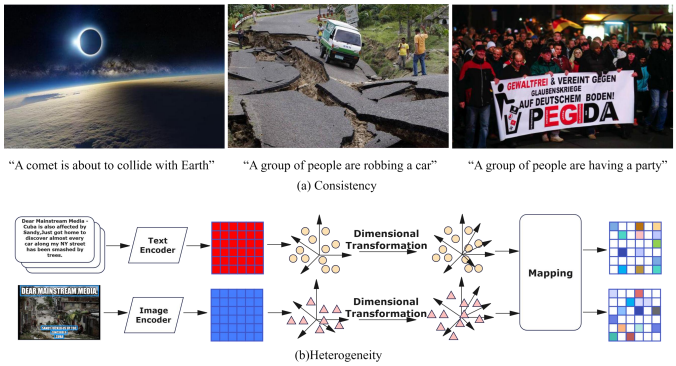


Fig. 1. Concerns Regarding Text-Image Consistency and Heterogeneity: As illustrated in Figure (a), an incongruity between the accompanying text and the image is readily apparent, giving rise to suspicions about the news being characterized as misinformation. In Figure (b), a notable semantic disparity is present between the text and the image, potentially affecting the overall comprehensibility.

GENERATIVE Artificial Intelligence (GAI), including diffusion models, deepfakes, and ChatGPT, has emerged as a foundational element within social networks. The rapid progression of GAI has led to a significant surge in the proliferation of fake news. This deliberately fabricated information has inflicted considerable harm upon the public, underscoring the importance of implementing measures to combat this issue and ensure the dissemination of reliable and accurate information. Fake news has garnered widespread attention in recent years due to its profound negative impact on significant public events, such as the U.S. presidential election, the Russia-Ukraine conflict, the COVID-19 pandemic, and others. It is worth emphasizing that rumors spread on social media may impact the opinions of billions of people [1]. Consequently, the detection of fake news assumes paramount importance and serves as the primary focus of this paper.

Current techniques for detecting fake news can be broadly categorized into two groups: single-modal based [2][3][4] and multi-modal based [5][6][7]. Given that instances of fake news often encompass a blend of diverse data types—such as user comments, manipulated images, fabricated videos, and other variables—detecting multi-modal fake news remains an intricate challenge. To date, diverse multi-modal approaches have concentrated on distinct modality fusion techniques to detect fake news [8][9][10]. For instance, Y. Wang et al. [8] utilized a straightforward splicing approach to amalgamate features from various modalities, whereas others had enhanced fusion capabilities by employing strategies such as Attention [9] and Bilinear Pooling [10].

Nowadays, semantics are proving to provide evident clues for identifying fake news. In the incorporation of semantic features, the SAFE framework [11] was the first to introduce image-text consistency, while the MCNN framework [12] introduced image tampering to propose a refined model based on it. Despite these breakthroughs, the text-image consistency is limited to the main modalities and neglects the textual parts embedded in visual elements, which may also be useful [13]. On the other hand, there exit complex mapping relationships in terms of "whether text and image are consistent or not" and "whether the news is true or not". Evidently, the discord between the semantics of text and image stands as a pivotal indicator of fabricated news, as depicted in Figure 1(a). But as illustrated in Figure 1(b), even when dealing with accurate news, a semantic gap can emerge due to the distinct modalities. Therefore, the text-image consistency generally reclines to the content-related consistency and neglects the structure-related inconsistency. That is, the text-image heterogeneity, which is an intrinsic but losing aspect, goes hand in hand with the text-image consistency. It is desired to solve the issues below since consistency and heterogeneity are contradictory concepts: 1) How to cooperate between the contents of texts and images, concerning the semantic relevance; 2) How to debate between the structures of texts and images, concerning the semantic gap; 3) How to reconcile these two conflicting concepts to attain equilibrium.

In this paper, our objective is to strike a delicate balance between text-image consistency and text-image heterogeneity. To achieve this, we propose an end-to-end **M**ulti-modal **F**ake News **D**etection framework based on **C**ross-**M**odal **M**atching (MFND-CMM), comprising four key components: *Modality Integration*: We employ a novel multi-modal framework to extract textual clues from attached images, enriching the content with more diverse and comprehensive information. *Feature Extraction*: We utilize an improved Bert encoder, an enhanced ResNet encoder, and a refined OCR encoder for extracting textual, visual, and auxiliary features, respectively. *Feature Interaction*: To facilitate interaction among different features, we employ cross-attention mechanisms, addressing semantic heterogeneity arising from inter-modality diversity across different modalities. *Feature Matching*: In this component, we assess the semantic consistency among different modalities by evaluating inter-modality correlation through pairwise-similarity scores.

Building upon this foundation, this end-to-end network adeptly achieves a balance between consistency and heterogeneity, by employing attention-guided enhanced modality interaction and similarity-based dynamic weight assignment. These techniques contribute to the development of robust frameworks designed to effectively predict the news labels. The primary contributions of this paper are outlined as follows:

- We propose a novel multi-modal fake news detection framework driven by cross-modal matching. The model effectively extracts fine-grained salient features of news, promoting semantic consistency among different modalities while alleviating semantic heterogeneity.
- We consider the text embedded within the image as one of the crucial determinants to integrate a multi-modal encoder. On the basis of it, we emphasize inter-modality correlation by pairwise-similarity scores, retain inter-modality diversity by cross-attention mechanism, and achieve a trade-off between consistency and heterogeneity via attention-guided enhanced modality interaction and similarity-based dynamic weight assignment.
- We empirically show that the proposed model can effectively identify fake news and outperform the state-of-the-art multi-modal fake news detection models on two large-scale real-world datasets, especially on English Twitter.

II. Related Work

A. Fake News Detection

Both language-based and vision-based fake news detection have received widespread attention. For the language-based methods, Castilo C et al. [2] utilized some word-level metrics such as the number of words and the number of links to infer fake news. Qazvinian et al.[4] used Bayesian network as a classifier to detect fake news and mainly studied the topic features of the text. This method of manually extracting features has major limitations. It is not only time-consuming and labor-intensive, but also cannot capture complex feature relationships. Ma et al. [3] used recurrent neural networks to extract implicit text information from texts and pioneered a rumor text feature extraction method based on deep learning. Qian et al. [14] developed a text-based method to extract semantic information from article text and proposed a user-response generation model to assist in fake news detection. Bhattarai et al. [15] used the vocabulary and semantic attributes of text to detect fake news. And for the vision-based methods, Gupta A et al. [16] conducted preliminary work based on images, and for the first time established a classification model for false image recognition on Twitter by artificially extracting relevant information such as the time when false images appeared on Twitter. However, this method cannot obtain the semantic features of the image. Thanks to the development of deep learning, many studies use convolutional neural networks for image feature extraction. P. Qi et al. [17] proposed to identify fake news by extracting image spatial and frequency domain fe atures to determine image tampering. Cao et al. [18] argued that typical image manipulation detection methods are useful in revealing traces of news tampering. Zhang H et al. [19] used pre-trained convolutional neural networks to model fake news images and extracted deep visual features.

However, for multimodal news, these methods fail to detect cross-modal correlations (although these unimodal features can be explored, and they do play a key role in distinguishing fake news, correlations and consistency are ignored and other multi-modal features, which may harm the overall performance of these single-modal schemes on multi-modal news). The key issue in multimodal fake news detection is adapting linguistic and visual representations.

B. Multi-modal Fusion

In recent years, some new fusion techniques have been introduced into fake news detection, replacing the simple

concatenation. Wang[8] et al. proposed EANN, which uses an auxiliary task of event classification to improve versatility. Singhal[20] et al. proposed Spotfake, which uses VGG and BERT to extract image and text features respectively, and connect them for classification. However, since these two methods only rely on fusion features obtained directly using connection or attention mechanisms, but the text and image features extracted respectively are not in the same semantic space, and the relevant information of text and image is not noticed during the fusion process, Therefore these fused features cannot provide sufficient identification power to classify fake news. Dhruv[21] et al. proposed MVAE, which trains a decoder to reconstruct original text and low-level image features from the fused features. Although the ability of reconstruction means that the fused features can contain more information, the necessity of these auxiliary tasks in fake news detection remains unknown and computationally expensive. The MKEMN method proposed by Zhang et al. [7] combined aligned embeddings of text, image, and knowledge to learn multi-modal representations for each post, applied in multi-modal fake news detection. Wu et al. [6] proposed MCAN, which stacked multiple co-attention layers to fuse multi-modal features, allowing the learning of inter-dependencies between multiple modalities. Wei et al. [22] introduced CMC, where the network trained two single-modality networks through contrastive learning to learn cross-modal relevance. The aforementioned works mainly focus on feature-based fusion, thereby weakening the semantic relationships between modalities.

C. Multi-modal Matching

Nowadays, semantics have been introduced into fake news detection, by some modal matching techniques, showing a novel fake news detection method. Zhou et al. [11] proposed SAFE, which fed the correlation between news textual and visual information into a classifier to detect fake news. However, the semantic gap of translating images to texts might have limited the ability to effectively utilize the cross-modal consistency. Xue et al. [12] introduced MCNN, which integrated textual semantic features, visual tampering features, and the similarity between textual and visual information for fake news detection. However, the aggregation process simply concatenated all features and did not consider the issue of cross-modal heterogeneity. Chen Y et al. [23] adopted a cross-modal alignment method to train the encoder model, mapping text and visual data into a shared semantic space, and then utilize fused features for classification. Due to the limited number of training datasets and the adoption of suboptimal labeling methods, this may limit the effectiveness of the encoder, resulting in a semantic gap remaining between text and image features. This method uses cross-modal ambiguity scores to reweight multi-modal features, but processing single-modal features through variational autoencoders with non-shared weights may affect the handling of ambiguity and thus model performance.

Based on the above related work, this paper proposes a model that not only considers the consistency between multi-modalities but also solves their heterogeneity, making it form a relationship of cooperation and competition. We also use attention-guided enhanced modal interaction and dynamic weight assignment based on similarity to create robust frameworks designed to enhance the detection of AI-generated fake news. Unlike traditional post-fusion detection paradigms, this approach integrates text modality, visual modality (images), and embedded text within images (auxiliary modality) into a unified end-to-end framework. This framework encompasses four components: modality integration, feature extraction, feature interaction, and feature matching, for comprehensive news content analysis. By meticulously addressing the complex relationships between texts and images, this method improves detection accuracy and efficiency for fake news created by generative AI technologies.

III. Methodology

The objective of this study is to ascertain the authenticity of news sourced from generative AI-enabled social networks or other platforms, based on an analysis of its textual and visual components. Given a news article $A = (T, I, E)$ containing text (T), image (I), and image-embedded content (E). Our objective is to discern the authenticity of the news, classifying it as either fake $(y = 1)$ or real $(y = 0)$. This determination relies on evaluating the correlations represented by $S(T^I, I^T)$, which means the correlations between the textual and visual components, where $S(T, E)$ refers to the correlations between the original text and embedded content within the image, and $S(E^I, I^E)$ refers to the correlations between the image-embedded content and the image itself. The model under consideration is denoted as MFND-CMM, and its architectural representation is illustrated in Figure 2.

A. Multi-modal Feature Extraction

1) Textual Feature Extraction Module: For the extraction of textual features, this paper employs a pre-trained Bert[24] model comprising 12 encoder layers. This model is utilized to characterize news text, as represented by the following:

$$F_t^i = \text{Bert}(T^i), T = [T^0, T^1, ..., T^{L-1}] \tag{1}$$

where T signifies the input sentence, L denotes the sentence length, and i signifies the i-th word within the sentence. The term "Bert" encapsulates pre-trained models designed for both Chinese and English languages.

This paper incorporates a self-attention mechanism [25] into the top layer of the Bert model. This augmentation enables the model to precisely capture the significance of each position within the input sequence, moving beyond a mere averaging of outputs from individual positions[26]. The calculation process is outlined as follows:

$$Q_s = (W_q F_t^i), K_s = (W_k F_t^i), V_s = (W_v F_t^i) \tag{2}$$

$$F_T = \text{SoftMax}\left(\frac{Q_s K_s^T}{\sqrt{d}}\right) V_s \tag{3}$$

where Q_s, K_s, and V_s are derived from textual features, W_q, W_k, and W_v represent weight matrices, d is the dimension of the input features, and F_T represents the textual features with attention. This feature extraction module is defined as Bert_Att.

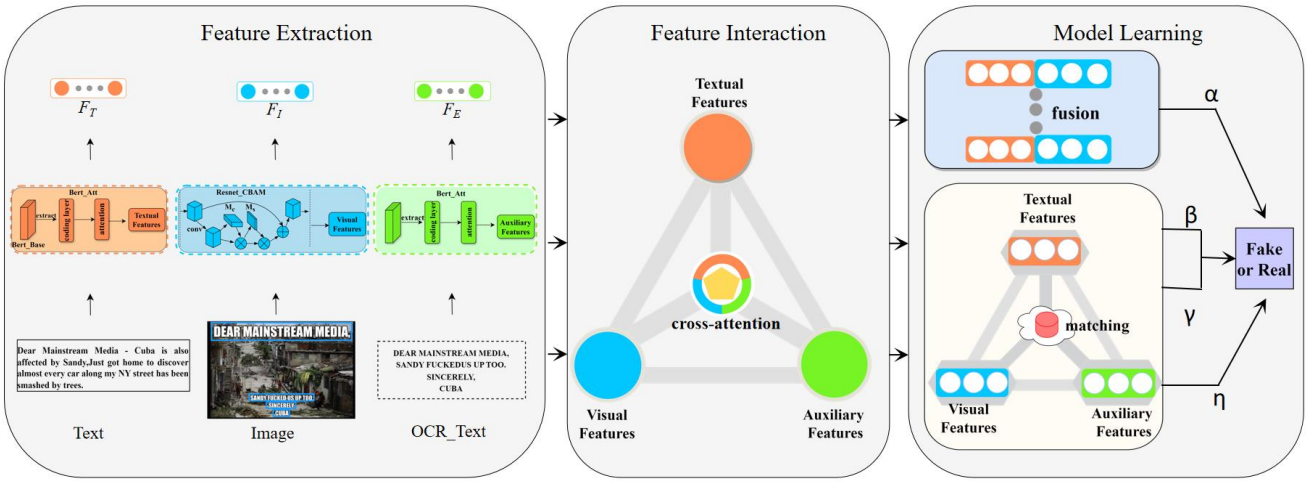


Fig. 2. The MFND-CMM framework involves the utilization of three distinct encoders: Textual Encoder for extracting textual features, Visual Encoder for extracting visual features, and Auxiliary Encoder for extracting image-embedded content, referred to as OCR_text, serving as auxiliary features. The extracted textual and visual features are amalgamated, and similarities among textual, visual, and auxiliary features are computed individually. Subsequently, various weights are assigned to facilitate a comprehensive and integrated decision-making process. The auxiliary modality is delineated as OCR_text in the diagram.

2) Visual Feature Extraction Module: In this study, an enhanced ResNet-50[27] model is chosen for visual feature extraction. When presented with an image I, the model processes it through ResNet to obtain an intermediate feature map F_i:

$$F_i = \text{ResNet}(I) \tag{4}$$

The CBAM framework [28] integrates channel attention and spatial attention, enhancing ResNet's capability to extract more prominent features. This paper incorporates CBAM into ResNet-50, denoting the modified model as ResNet_CBAM. Channel attention explicitly captures the interdependencies among feature channels, enabling the network to autonomously learn the significance of each channel. Consequently, the intermediate feature map F_i undergoes processing through channel attention, resulting in the refined feature map F_i':

$$M_c(F_i) = \sigma\{\,\text{MLP}[\text{AvgPool}(F_i)] + \text{MLP}[\text{MaxPool}(F_i)]\} \tag{5}$$

$$F_i' = M_c(F_i) \otimes F_i \tag{6}$$

where σ represents the activation function, MLP is shared for both inputs, $M_c(F_i)$ is a one-dimensional channel attention map, and $\otimes$ means element-wise multiplication.

Spatial attention functions as an adaptive mechanism for selecting spatial regions, directing the model to focus more on key feature areas. Subsequently, the refined feature map F_i' is input into the spatial attention module:

$$M_s(F_i') = \sigma\{f^{7\times7}([\text{AvgPool}(F_i'); \text{MaxPool}(F_i')])\} \tag{7}$$

$$F_i'' = M_s(F_i') \otimes F_i' \tag{8}$$

$$F_I = F_i'' \oplus F_i \tag{9}$$

where σ represents the activation function, $f^{7\times7}$ denotes a convolution operation with a filter size of 7×7, $M_s(F_i')$ is a two-dimensional spatial attention map, and F_I represents the visual features with attention.

3) Image-Embedded Content Extraction Module: Upon observation, it has been noted that numerous news article images encompass substantial embedded information. This embedded information commonly constitutes text and encapsulates the core theme of the news[29]. This study deems image-embedded information as an auxiliary modality. The open-source model, PP-OCRv3 [30], is employed for text extraction. Specifically, the pre-trained PP-OCRv3 model is utilized in this paper to extract the image-embedded content (E):

$$E = \text{OCR}(I) \tag{10}$$

where I refers to the image.

Likewise, Bert_Att is employed for feature extraction yielding vectorized representations and thereby producing the auxiliary feature F_E:

$$F_E = \text{Bert\_Att}(E^i), E = [E^0, E^1, ..., E^{L-1}] \tag{11}$$

B. Multi-modal Feature Interaction

In addressing the interplay between textual and visual components in news articles, this section strives to establish alignment and interaction of information between the textual content and fine-grained images. To accomplish this, cross-attention[31] is employed to facilitate the integration of cross-modal information between texts and images.

In the textual branch, the visual feature F_I undergoes mapping through three separate fully connected layers to serve as Q_{c1}, while the textual feature F_T is simultaneously mapped as both K_{c1} and V_{c1}. Following this, the cross-modal attention process is employed to generate the textual feature F_{T^I}, establishing correlation with the images:

$$Q_{c1} = (W_q F_I), K_{c1} = (W_k F_T), V_{c1} = (W_v F_T) \tag{12}$$

$$F_{T^I} = \text{SoftMax}\left(\frac{Q_{c1}K_{c1}^T}{\sqrt{d}}\right)V_{c1} \tag{13}$$

Similarly, within the image branch, the textual feature F_T is mapped through three separate fully connected layers to function as Q_{c2}, whereas the visual feature F_I is simultaneously mapped as both K_{c2} and V_{c2}. Following this, the cross-modal attention process is applied, resulting in the visual feature F_{I^T} that correlates with the texts:

$$Q_{c2} = (W_q F_T), K_{c2} = (W_k F_I), V_{c2} = (W_v F_I) \tag{14}$$

$$F_{I^T} = \text{SoftMax}\left(\frac{Q_{c2}K_{c2}^T}{\sqrt{d}}\right) V_{c2} \tag{15}$$

Via the cross-modal cross-attention process, the model has extracted crucial and pertinent information from the original textual feature F_T and visual feature F_I, denoted as F_{T^I} and F_{I^T}, respectively.

The cross-modal attention mechanism between the auxiliary modality and image is analogous to the cross-modal attention between text and image, so it will not be reiterated here. Consequently, the model has extracted significant relevant information denoted as F_{E^I} and F_{I^E} from the original cross-modal feature F_E and visual feature F_I, respectively.

C. Multi-modal Feature Fusion

Up to this point, textual and visual features have been acquired. Subsequently, the final dimension is obtained by concatenating their last dimensions without explicitly considering their interrelationship, aiming to accurately predict the likelihood that the news is false. This can be mathematically defined as:

$$C = \text{SoftMax}(W_C(F_{T^I} \oplus F_{I^T}) + B_C) \tag{16}$$

where $\oplus$ is the symbol of splicing operation, C represents the probability of class prediction, W_C and B_C are weight parameters and offset items respectively.

To make the calculated possibility of falsification of a news article close to its true label y, a loss function based on cross-entropy is defined as:

$$\begin{aligned}\mathcal{L}_c\left(\theta_{T^I}, \theta_{I^T}, \theta_c\right) = -\mathbb{E}_{(a,y)\sim(A,Y)} \\ [y \cdot \log C + (1-y) \cdot \log(1-C)]\end{aligned} \tag{17}$$

where Y is the set of news true and fake label categories, A is the news set, and y refers to the actual label of news a. θ_{T^I}, θ_{I^T} and θ_c respectively represent the corresponding parameter matrices.

D. Multi-modal Feature Matching

The correlation between the textual and visual elements of news articles can serve as a means to assess the authenticity of news stories. Creators of fake news may intentionally select unrelated images as a basis for deceptive claims, aiming to capture readers' attention.

This paper assesses the concordance between the two modalities utilizing a slightly modified cosine similarity for simplicity in calculation and precise measurement. Initially, examine the coherence of the textual feature F_{T^I} and visual feature F_{I^T}, defined mathematically as follows:

$$S\left(F_{T^I}, F_{I^T}\right) = \frac{F_{T^I} \cdot F_{I^T} + \parallel F_{T^I} \parallel \parallel F_{I^T} \parallel}{2 \parallel F_{T^I} \parallel \parallel F_{I^T} \parallel} \tag{18}$$

Similarly, when examining the coherence between the auxiliary feature F_{E^I} and the visual feature F_{I^E}, their mathematical definition aligns with the calculation of consistency between textual and visual features. Consequently, their similarity can be expressed as $S\left(F_{E^I}, F_{I^E}\right)$. In evaluating the coherence between the original textual feature F_T and the original auxiliary feature F_E, as they belong to the same modality, cross-modal attention interaction is unnecessary. This is mathematically defined as $S(F_T, F_E)$.

Then, define the following three cross-entropy loss functions to represent that from the perspective of similarity:

$$\begin{aligned}\mathcal{L}_s\left(\theta_{T^I}, \theta_{I^T}\right) = -\mathbb{E}_{(a,y)\sim(A,Y)} \{y \cdot \log[1 - S(F_{T^I}, F_{I^T})] \\ +(1-y) \cdot \log S(F_{T^I}, F_{I^T})\}\end{aligned} \tag{19}$$

The mathematical definitions of $\mathcal{L}_s\left(\theta_{E^I}, \theta_{I^E}\right)$ and $\mathcal{L}_s\left(\theta_T, \theta_E\right)$ are the same as $\mathcal{L}_s\left(\theta_{T^I}, \theta_{I^T}\right)$ and will not be repeated here.

E. Multi-modal Model Learning

Ultimately, the correct identification of fake news is achieved through the fusion and assessment of similarities among multi-modal features of news. Here, α, β, γ, and η represent the loss function weights for the four branch networks. To incorporate the significance of all four modalities, the final loss function formulations are as follows:

$$\begin{aligned}\mathcal{L}\left(\theta_T, \theta_E, \theta_c, \theta_{T^I}, \theta_{I^T}, \theta_{E^I}, \theta_{I^E}\right) = \alpha\mathcal{L}_c\left(\theta_{T^I}, \theta_{I^T}, \theta_c\right) \\ +\beta\mathcal{L}_s\left(\theta_{T^I}, \theta_{I^T}\right) + \gamma\mathcal{L}_s\left(\theta_T, \theta_E\right) + \eta\mathcal{L}_s\left(\theta_{E^I}, \theta_{I^E}\right)\end{aligned} \tag{20}$$

$$\begin{aligned}\left(\hat{\theta}_T, \hat{\theta}_E, \hat{\theta}_c, \hat{\theta}_{T^I}, \hat{\theta}_{I^T}, \hat{\theta}_{E^I}, \hat{\theta}_{I^E}\right) = \\ \text{argmin}\,\mathcal{L}\left(\theta_T, \theta_E, \theta_c, \theta_{T^I}, \theta_{I^T}, \theta_{E^I}, \theta_{I^E}\right)\end{aligned} \tag{21}$$

IV. Experiments

A. Datasets

To comprehensively assess the performance of MFND-CMM, experiments are conducted on two publicly available datasets. Initial statistics reveal that over 70% of the images in both datasets incorporate embedded text. The dataset statistics are presented in Table I. The following provides a description of the datasets:

TABLE I
DATASET STATISTIC

Dataset	Label	Number	All
Weibo	fake real	4749 4779	9528
Twitter	fake real	7021 5974	12995

1)The Weibo dataset [32] is sourced from the Weibo social platform. This Weibo dataset of true information is collected from authoritative Chinese sources, and fake information is obtained through the official Weibo rumor suppression system. The dataset contains 4,749 pieces of fake news data and 4,779 pieces of real news data, along with 9,528 news images. Each data instance includes both a news text and a corresponding news image.

2)The Twitter dataset [33] is sourced from the Twitter social platform. The content of each tweet have a brief text message and extra images or videos.There are around 6,000 rumors and 5,000 non-rumor tweets in the development set from 11 rumor-related events. It includes 7,021 pieces of fake news data, and 5,974 pieces of real news data, and 517 news images.

They are split into 7:1:2 ratios for training, validation, and test sets.

B. Experimental Settings

The dimensionality of textual features obtained from Bert_Att is 768. Images are first resized to 224x224x3 and then fed to ResNet_CBAM. The dimensionality of visual features obtained from ResNet_CBAM is 2,048. The ultimately obtained cross-modal features are all of dimensions 32x32.

In experiments, Adam is selected as the optimizer. The batch size is 32, and the epoch is 100. Weibo's learning rate is set to 0.001, whereas Twitter's learning rate is set to 0.0001, and 0.5 is the dropout rate. After comparing several groups of experiments and assigning weights according to the importance of branches and manual experience, the weights of the four network branches in the model are finally determined as α=0.6, β=0.2, γ= 0.1 and η=0.1.

In the setup of metrics for experimental results, precision refers to the proportion of samples that are actually positive among those predicted as positive by the model. It measures the accuracy of the model's predictions. Recall refers to the proportion of samples that are actually positive and are correctly predicted as positive by the model. It measures the model's ability to identify positive samples. The F1 Score is the harmonic mean of precision and recall, taking into account the importance of both precision and recall. It is a comprehensive metric that combines the two.

C. Baselines

- **att-RNN** [5]: The att-RNN model incorporates unidirectional LSTM and VGG-19 for the extraction of image features, context features, and textual features. Subsequently, it employs an RNN with attention mechanisms to discern and identify fake news.
- **EANN** [8]: EANN employs Text-CNN for textual feature extraction and VGG for visual feature extraction. It then obtains news features by concatenating textual and visual information, utilizing an event discriminator to detect fake news.
- **MVAE** [21]: MVAE utilizes bidirectional LSTM for textual feature extraction and VGG for visual feature extraction. These features are then fused into an autoencoder to reconstruct the correlations between different modes of feature learning.
- **Spotfake** [20]:Spotfake utilizes Bert for textual feature extraction and VGG-19 for visual feature extraction. These features are subsequently concatenated to facilitate the detection of fake news.
- **MKEMN** [7]:MKEMN considers texts, images, and retrieved knowledge embeddings as stacked channels and performs fusion through a convolutional operation.
- **BDANN** [34]: BDANN uses BERT to extract text features and VGG-19 to extract image features. It also applies domain classifiers to eliminate dependencies between features, which is ultimately used for news detection.
- **SAFE** [11]: SAFE converts the image into text and extracts textual features using Text-CNN. Finally, cosine similarity is employed to detect the similarity between textual representations.
- **MCNN** [12]: MCNN integrates textual semantic features, visual tampering features, and the similarity of textual and visual information computed through cosine similarity for use in fake news detection.
- **MCAN** [6]: MCAN stacks multiple co-attention layers to fuse the multimodal features.
- **CAFE** [23]: CAFE quantifies cross-modal ambiguity to adaptively aggregate unimodal features and cross-modal correlations.
- **LIIMR** [35]: LIIMR identifies and suppresses information from weaker modalities and extracts relevant information from the strong modality on a per-sample basis.
- **MRHFR** [36]: MRHFR integrates multimodal news features by simulating three reading habits of humans, and utilizes a consistency-constrained reasoning layer to detect inconsistencies between different modalities, thereby achieving multimodal fake news detection.

D. Performance Comparison

Table II juxtaposes the performance of MFND-CMM against baseline models. Notably, MFND-CMM attains the pinnacle of accuracy on both real-world datasets, registering impressive values of 90.3% and 93.2%, respectively. Particularly noteworthy is the F1 score in this study, which stands out as the highest among all metrics. Given that F1 embodies a delicate equilibrium between Precision and Recall, the overall evaluation underscores that this paper attains the most favorable metrics.

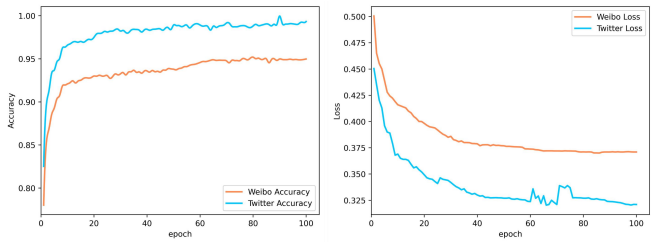


Fig. 3. Visualization results of the experimental training process on two real-world datasets

From the perspective of feature extraction, Spotfake and MKEMN models adopt pre-trained feature extractors without considering the relative importance of each feature. Experimental results on the Weibo dataset reveal that MFND-CMM outperforms both Spotfake and MKEMN, achieving an accuracy that is 2.1% higher than Spotfake and 7.9% higher than MKEMN. This substantiates the assertion that the MFND-CMM model discerns the relative importance of textual and image features through attention levels, thereby enhancing the

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS FOR METHODS. THE BEST INDEXES ARE IN BOLD AND THE SECOND BEST ONES ARE IN RED.

Dataset	Method	Accuracy	Real Information			Fake Information		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	att-RNN[5]	0.788	0.738	0.890	0.807	0.862	0.686	0.764
	EANN [8]	0.816	0.810	0.810	0.810	0.820	0.820	0.820
	MVAE[21]	0.824	0.802	0.875	0.837	0.854	0.769	0.809
	Spotfake [20]	0.892	0.847	0.656	0.739	0.902	**0.964**	**0.932**
	MKEMN [7]	0.824	0.723	0.819	0.798	0.823	0.799	0.812
	BDANN [34]	0.842	0.850	0.820	0.830	0.830	0.870	0.850
	SAFE[11]	0.816	0.816	0.818	0.817	0.818	0.815	0.817
	MCNN [12]	0.823	0.787	0.848	0.816	0.858	0.801	0.828
	MCAN [6]	0.899	0.884	0.909	0.897	0.913	0.889	0.901
	CAFE [23]	0.840	0.825	0.851	0.837	0.855	0.830	0.842
	LIIMR [35]	0.900	0.882	0.823	0.847	0.908	0.941	0.925
	MRHFR [36]	**0.907**	0.939	0.869	**0.903**	0.879	0.931	0.904
	MFND-CMM	0.903	**0.885**	**0.912**	0.898	**0.914**	0.923	0.918
Twitter	att-RNN [5]	0.682	0.603	0.770	0.676	0.780	0.615	0.689
	EANN[8]	0.719	0.771	0.870	0.817	0.642	0.474	0.545
	MVAE[21]	0.745	0.689	0.777	0.730	0.801	0.719	0.758
	Spotfake[20]	0.777	0.832	0.606	0.701	0.751	0.900	0.820
	MKEMN [7]	0.714	0.634	0.814	0.831	0.814	0.756	0.708
	BDANN [34]	0.830	0.830	0.930	0.880	0.810	0.630	0.710
	SAFE [11]	0.762	0.695	0.811	0.748	0.831	0.724	0.774
	MCNN[12]	0.784	0.790	0.787	0.788	0.778	0.781	0.779
	MCAN [6]	0.809	0.732	0.871	0.795	0.889	0.765	0.822
	CAFE [23]	0.806	0.805	0.813	0.809	0.807	0.799	0.803
	LIIMR [35]	0.831	0.836	0.832	0.830	0.825	0.830	0.827
	MRHFR [36]	0.921	0.976	0.828	0.896	0.876	**0.981**	**0.926**
	MFND-CMM	**0.932**	**0.922**	**0.941**	**0.931**	**0.922**	0.921	0.922

efficacy of feature representation and, consequently, improving the effectiveness of fake news detection. From the perspective of feature fusion, the MFND-CMM model exhibited a notable enhancement in accuracy on both the Weibo dataset (7.9% increase) and the Twitter dataset (18.7% increase) when compared to the MVAE model without a fusion mechanism. This underscores the efficacy of the feature fusion mechanism within the MFND-CMM model, aiding both modalities in uncovering more crucial information and effectively elevating the expressiveness of modal features.

From the perspective of feature matching, the absence of cross-modal interaction between the SAFE model and MCNN model results in a challenge of modal heterogeneity. Experimental outcomes on the Weibo dataset reveal that MFND-CMM surpasses both SAFE and MCNN, showcasing an accuracy rate that is 8.7% higher than SAFE and 8% higher than MCNN. This substantiates that the MFND-CMM model maximizes the utilization of multi-modal interaction data, enhancing the association features between text and image, and effectively addressing the issue of modal heterogeneity. Furthermore, MCAN does not pay enough attention to the consistency between features, and CAFE has limitations on single-modal features, which will affect the robustness of the model.

In order to further prove the validity and correctness of the model proposed in the text, the training process during the experiment is retained and visualized in this paper, as shown in Figure 3.

E. Architecture Ablation Analysis

The article conducts a series of comparative experiments employing the original Bert and ResNet to elucidate the efficacy of the proposed feature extractor. Furthermore, component ablation experiments are executed to underscore the effectiveness of each constituent of MFND-CMM. The resultant experimental findings encompass evaluation metrics for the Fake Information class, encompassing Accuracy, Precision, Recall, and F1 score.

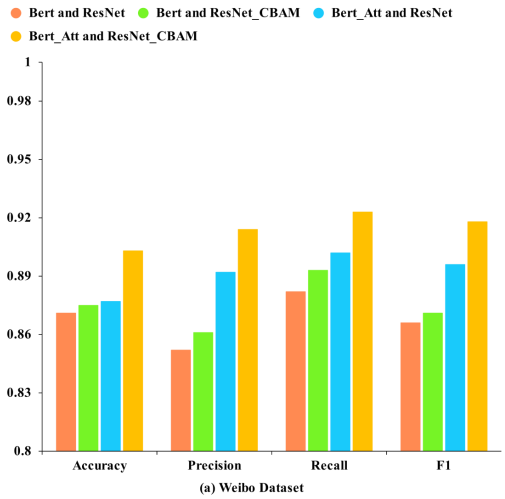


Fig. 4. Architecture ablation analysis of Feature Extractor.

1) Ablation Experiments of Feature Extractor: The pertinent experimental outcomes are depicted in Figure 4 and Figure 5. Notably, the original pre-trained feature extractor yields suboptimal results. Conversely, the enhanced model exhibits the most favorable experimental outcomes, showcasing the efficacy of the proposed feature extractor. The discernible salient features extracted aptly capture the distinctive characteristics

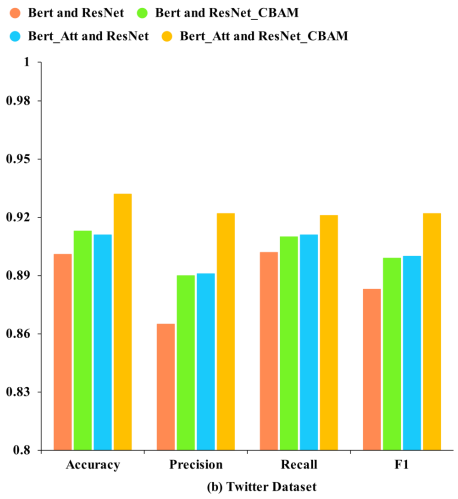


Fig. 5. Architecture ablation analysis of Feature Extractor.

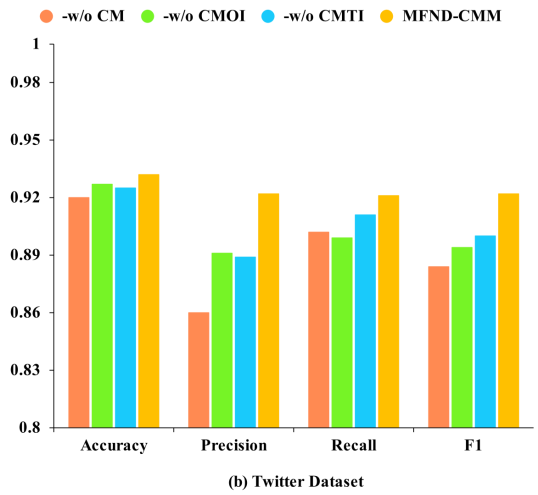


Fig. 7. Architecture ablation analysis of Feature Interaction

of news, further validating the effectiveness of the feature extractor delineated in this paper.

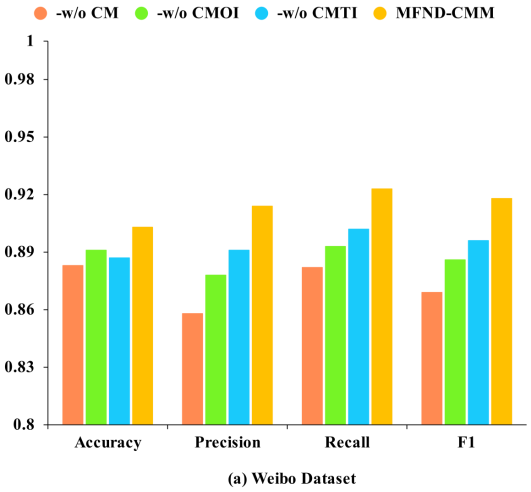


Fig. 6. Architecture ablation analysis of Feature Interaction

2) Ablation Experiments of Feature Interaction:

- -w/o CM: Experiments with no cross-modal;
- -w/o CMOI: Remove cross-modal updates between OCR_text and image. Conduct experiments with cross-modal interactions between text and image;
- -w/o CMTI: Remove cross-modal updates between text and image. Conduct experiments with cross-modal interactions between OCR_text and image;
- MFND-CMM: Experiments with cross-modal.

The pertinent experimental outcomes are illustrated in Figure 6 and Figure 7. The cross-attention mechanism facilitates dynamic integration between modalities, wherein visual updates retain pivotal textual information, and textual updates preserve essential visual details. This concurrent process effectively mitigates heterogeneity across modalities. Consequently, it can be inferred that the feature interaction module constructed in this paper proves to be effective.

3) Ablation Experiments of Individual Components of MFND-CMM:

- -w/o $\alpha\beta\gamma$: Experiments with image and image-embedded content matching;
- -w/o $\alpha\beta\eta$: Experiments with text and image-embedded content matching;
- -w/o $\alpha\gamma\eta$: Experiments with text-image matching;
- -w/o $\beta\gamma\eta$: Experiments with text-image fusion;
- -w/o $\gamma\eta$: Text-image fusion, text-image matching, experiments with weights of 0.6 and 0.4 respectively;
- -w/o η: Text-image fusion, text and image content matching, image and image-embedded content matching, experiments with weights of 0.6, 0.2 and 0.2 respectively;

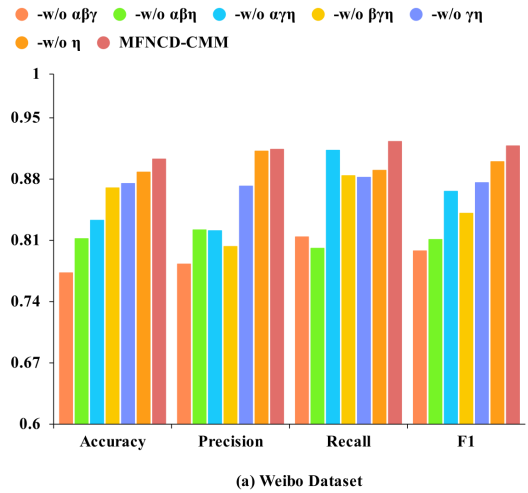
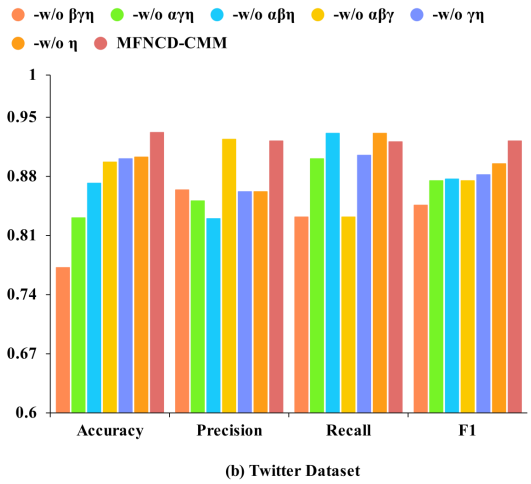


Fig. 8. Architecture ablation analysis of MFND-CMM

Figure 8 and Figure 9 reveals that text-image fusion exerts the most substantial influence on fake news detection. The semantic consistency achieved through similarity-based dynamic weight assignment emerges as a pivotal factor in enhancing the model's performance. It is evident that mere text-image fusion and text-image matching fall short of achieving the



(b) Twitter Dataset

Fig. 9. Architecture ablation analysis of MFND-CMM

desired impact. However, with the incorporation of image-embedded content, a notable improvement in the indicator is observed. This underscores the effectiveness of the text-image consistency detection method and underscores the significance of modeling image-embedded content.

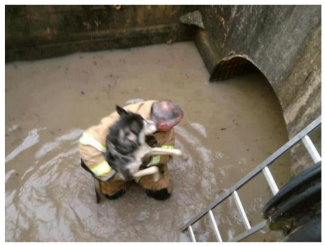
F. Case Study

"There's a new Olympic game called Turd Dropping in Sochi."

"Qianxi bus exploded! My friend witnessed this terrible explosion! How many innocent people have died like this!"

(a) Fake News

"This is a photo I took in Hejiang Province in 2007 about two Vietnamese Hmong children snuggling together"



"Dog rescued from hurricane Sandy! Firemen are not only saving the people!"

(b) Real News

Fig. 10. Cases of fabricated news and real news in real-world datasets successfully identified by the MFND-CMM model.

In our case study, we seek to substantiate the following inquiries: Do instances of fake news exist in the real world where the texts and images exhibit inconsistency? Can our model adeptly discern and accurately identify this incongruence, thus flagging the news as falsified? In pursuit of these objectives, we scrutinized news articles from two datasets, aligning their ground truth labels with the predicted labels from our model. Our findings indicate that our model adeptly identifies the consistency relationship between texts and images, serving as a pivotal criterion for discerning authenticity. Figure 10 show cases several examples, revealing a discernible correspondence between the textual and visual components of genuine news, while the text of fake news struggles to find support from unaltered images.

Future research could consider improvements in the following areas:

1) Cross-modal information integration. Current multi-modal research primarily focuses on the integration of news text and images. Future directions could be dedicated to exploring more efficient ways to merge various modalities such as text, images, and videos.

2) System upgrades. In light of the rapid dissemination and dynamic nature of fake news, future studies should focus on developing a fake news detection system capable of real-time monitoring and dynamic response.

V. Conclusion

This paper suggests a novel multi-modal fake news detection framework through cross-modal matching, which holds the potential for application in future social networks. The proposed framework comprises a feature extraction module, a feature interaction module, a fusion module, and three matching modules. The feature extraction module captures essential features from news articles. In the fusion module, textual and visual features are merged. The matching modules evaluate inter-modality correlations using pairwise-similarity scores. Simultaneously, the feature interaction module utilizes the cross-attention mechanism to introduce dynamic integration across modalities. To strike a balance between consistency and heterogeneity, attention-guided modality-enhanced interaction and similarity-based dynamic weight assignment are used, contributing to the creation of a robust framework. Experiments show that our method outperforms existing baseline methods. In the future work, a robust framework aimed at improving the detection of AI-generated fake news is still open.

VI. Acknowledgement

This research was funded by the MOE (Ministry of Education in China) Liberal arts and Social Sciences Foundation (24YJC86007), the Key Supported Program of Joint Fund of the National Natural Science Foundation of China (U23B2029), National Natural Science Foundation of China (62076167) and the Yuxiu Innovation Project of NCUT (2024NCUTYXCX102).

References

[1] L. Tan, G. Wang, F. Jia, and X. Lian, "Research status of deep learning methods for rumor detection," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 2941–2982, 2023.

[2] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *The Web Conference*, 2011.

[3] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," 2016.

[4] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Empirical Methods in Natural Language Processing*, 2011.

[5] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *the 2017 ACM*, 2017.

[6] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of ACL 2021*, 2021.

[7] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *the 27th ACM International Conference*, 2019.

[8] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *ACM SIGKDD*, 2018, pp. 849–857.

[9] Y. J. Lu and C. T. Li, "Gcan: Graph-aware co-attention networks for explainable fake news detection on social media," 2020.

[10] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[11] X. Zhou, J. Wu, and R. Zafarani, "Safe: similarity-aware multi-modal fake news detection (2020)," *Preprint. arXiv*, vol. 200304981, p. 2, 2020.

[12] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Information Processing & Management*, vol. 58, no. 5, p. 102610, 2021.

[13] J. Chen, C. Jia, H. Zheng, R. Chen, and C. Fu, "Is multi-modal necessarily better? robustness evaluation of multi-modal fake news detection," *IEEE Transactions on Network Science and Engineering*, 2023.

[14] F. Qian, C. Gong, K. Sharma, and Y. Liu, "Neural user response generator: Fake news detection with collective user intelligence," 2018.

[15] B. Bhattarai, O. C. Granmo, and L. Jiao, "Explainable tsetlin machine framework for fake news detection with credibility score assessment," 2021.

[16] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 729–736.

[17] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE international conference on data mining (ICDM)*. IEEE, 2019, pp. 518–527.

[18] J. Cao, P. Qi, Q. Sheng, T. Yang, and J. Li, "Exploring the role of visual content in fake news detection," 2020.

[19] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *ACM international conference on multimedia*, 2019, pp. 1942–1951.

[20] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 2019, pp. 39–47.

[21] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world wide web conference*, 2019, pp. 2915–2921.

[22] Z. Wei, H. Pan, L. Qiao, X. Niu, P. Dong, and D. Li, "Cross-modal knowledge distillation in multi-modal fake news detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4733–4737.

[23] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2897–2905.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 2017.

[26] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[29] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, and Y. Yu, "Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1212–1220.

[30] C. Li, W. Liu, R. Guo, X. Yin, K. Jiang, Y. Du, Y. Du, L. Zhu, B. Lai, X. Hu *et al.*, "Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system," *arXiv preprint arXiv:2206.03001*, 2022.

[31] M. Gheini, X. Ren, and J. May, "Cross-attention is all you need: Adapting pretrained transformers for machine translation," *arXiv preprint arXiv:2104.08771*, 2021.

[32] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[33] Z. Jin, J. Cao, Y. Zhang, and Y. Zhang, "Mcg-ict at mediaeval 2015: Verifying multimedia use with a two-level classification model." in *MediaEval*, 2015.

[34] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[35] S. Singhal, T. Pandey, S. Mrig, R. R. Shah, and P. Kumaraguru, "Leveraging intra and inter modality relationship for multimodal fake news detection," in *Companion Proceedings of the Web Conference 2022*, 2022, pp. 726–734.

[36] L. Wu, P. Liu, and Y. Zhang, "See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 736–13 744.