IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 April 2024, accepted 14 April 2024, date of publication 22 April 2024, date of current version 9 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3392409

RESEARCH ARTICLE

BC-FND: An Approach Based on Hierarchical Bilinear Fusion and Multimodal Consistency for Fake News Detection

YAHUI LIU, (Member, IEEE), WANLONG BING, SHUAI REN, AND HONGLIANG MA
School of Information Science and Technology, Shihezi University, Shihezi 832003, China

Corresponding author: Yahui Liu (lyh@shzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62062060, in part by the Bingtuan Science and Technology Program under Grant 2023CB005 and Grant 2022CB002-08, and in part by Shihezi University High-Level Talent Research Start-Up Project under Grant RCZK2018C11 and Grant RCZK2018C38.

ABSTRACT Fake news with multimedia on social media is deceptive, widely spread, and has serious negative effects. Therefore, multimodal fake news detection has become a popular and extensively studied topic. However, the existing methods have two shortcomings. 1) Different types of extractors are used for text and images, making it difficult to align the extracted features to the same embedding space. 2) The complex fusion approach leads to an increase in the number of features and parameters that generate redundancy and noise easily. To address these problems, we propose a simple yet powerful multimodal fake news detection model (BC-FND). It utilizes contrastive learning of CLIP to align textual and visual features to the same embedding space while using a consistency loss function to learn consistency between real news text and images as well as inconsistency between fake news text and images. Additionally, BERT is employed for extracting semantic and contextual information from text while a hierarchical bilinear fusion network is designed to achieve full complementarity between textual and visual features. Cross-entropy and consistency loss functions jointly optimize BC-FND for improved accuracy in detecting fake news. We also introduce the Weibo23 dataset which is more challenging since it's closer to the real social media environment. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods on two public datasets and the Weibo23 dataset.

INDEX TERMS Fake news detection, social media, multimodal learning.

I. INTRODUCTION

The popularity of social media has made it more convenient for people to access information. However, it has also provided a broader network space for the breeding and propagation of fake news. Fake news is defined as a news article that is intentionally fabricated and verified as fake [1]. Fake news may cause misunderstandings, hatred, and other negative emotions, resulting in significant adverse effects and chaos in society [2]. As shown in Fig. 1, a piece of fake news replete with official corruption that has elicited anger, sarcasm, and abuse among the general public. Subsequently, the official report refuted the dissemination of fake news. The allocation of 3.4 billion RMB towards river management was elucidated, revealing that a mere 0.2 percent was dedicated to installing dry-hanging stones along the riverbank. This contradicts the belief held by netizens that the entire 3.4 billion RMB was used for installing dry-hanging stone materials. Although these clarifications come from authoritative sources, some conspiracy theorists still do not believe in official response. The spread of fake news has greatly damaged government credibility. Despite the emergence of numerous fake news verification platforms in the industry, most rely on public reporting and artificial verification methods to confirm the authenticity of the news, leading to a time lag that allows fake news to have a

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

FIGURE 1. Zhengzhou City spent 3.4 billion RMB to install dry-hanging stones on the riverbank of the Jinshui River.

wide-ranging impact on society before being identified. Therefore, it is imperative to develop automated detection systems on social media platforms to detect fake news early and mitigate the harm they cause.

As the saying goes, *"A picture is worth a thousand words"*. News with pictures offers a more comprehensive information experience, and fake news with images is more confusing for readers and is more likely to attract their attention while spreading widely [3], [4]. Consequently, purveyors of fake news are prone to employing a combination of images and text when disseminating misleading news [5], as depicted in Fig. 1. Notably, images also serve as valuable indicators for identifying fake news. For instance, fake news often exhibits characteristics such as image blurring, image tampering, and inconsistency between text content and accompanying images. Therefore, we were motivated to develop a multimodal fusion approach utilizing text, images and their relationships for fake news detection.

In recent years, an increasing number of scholars have focused on multimodal fake news detection (MFND). Existing research has primarily treated fake news detection as a binary classification task. Some studies have employed adversarial networks [6] and variational autoencoders [7] to design auxiliary tasks to extract potentially integrated features from text and images. However, these auxiliary tasks necessitate additional information, such as social context and event labels, which undoubtedly increases the labor costs required for data collection and classification. Moreover, training auxiliary tasks, such as text and image reconstruction using a variational autoencoder prolongs the model training time, posing challenges for the practical implementation of fake news detection applications.

With the emergence of various pre-trained models, such as BERT (Bidirectional Encoder Representation from Transformers) [8], XLNet [9], and ViT (Vision Transformer) [10], they have demonstrated significant advantages in the feature extraction of text and images through fine-tuning alone, without requiring additional information. Therefore, many researchers have utilized pre-trained models to extract textual and visual features and subsequently integrate them by straightforward concatenation [11], [12], [13]. In comparison to traditional networks like CNN (Convolutional Neural Network) [14] and RNN (Recurrent Neural Network) [15], exploiting pre-trained models for feature extraction substantially enhances the performance of fake news detection. However, straightforward concatenation fails to fuse textual and visual features adequately. Consequently, some scholars have explored the consistent relationship between text and images for fake news detection [16], [17]. Qi et al. [18] extracted textual information from images using OCR techniques and combined it with original text for fake news detection. Other researchers have adopted multimodal enhancement techniques such as co-attention [19] mechanisms and self-attention [20] mechanisms to construct complex networks that facilitate deep fusion between modalities [21], [22] or hierarchical modal fusion [2] for fake news detection. Although these multimodal enhancement methods imitate human reading habits while enhancing integration between modalities, they also introduce redundancy in fusion [23].

Overall, the current approaches to MFND generally have two shortcomings. First, it is difficult to align the extracted textual and visual features to the same embedding space using completely different types of feature extractors, such as BERT and Swin-T (Hierarchical Vision Transformer using Shifted Windows) [24], which interfere with the subsequent fusion operation and prevent the model from achieving an optimal fusion effect [25]. Second, the unsuitability multimodal fusion method substantially increases the number of features and parameters, leading to potential redundancy and noise points that can adversely impact model performance [26]. Therefore, the primary focus in MFND is on achieving alignment between modalities' features in the same embedding space while designing a more effective fusion method to enhance information complementarity across modalities and ultimately improve overall performance.

To achieve these goals, we propose a simple and powerful multimodal fake news detection model (BC-FND) that leverages the CLIP (Contrastive Language-Image Pretraining) [27] to align textual and visual features in a unified embedding space. Within the same embedding space, consistency loss is designed to learn the dissimilarity between fake news text and image features as well as the similarity between real news text and image features. BERT is employed to extract semantic and contextual information from the text content. A hierarchical bilinear fusion network is devised to achieve full complementarity between textual and visual features, with the fusion results being utilized for fake news detection and the temporary result being fed into the consistency loss function. The accuracy of fake news detection was improved by combining consistency loss and cross-entropy loss to optimize the model. The contributions of this study can be summarized as follows:

- To the best of our knowledge, this study represents the first proposed hierarchical bilinear approach for integrating BERT and CLIP features while capturing

relationship features between text and images through the consistency loss function.
- We investigated the impact of various feature fusion methods on the accuracy of fake news detection. The findings suggest that the hierarchical bilinear approach exhibits superior capability in effectively integrating textual and visual features.
- We present a more challenging multimodal fake news dataset that has made fake and real news samples nearly congruent with the distribution of textual content. This enables the model to focus on learning the crucial discrepancies between fake and real news rather than disparities in content.
- The BC-FND model was validated on two publicly available datasets as well as our proposed dataset, showing superior performance compared to the current state-of-the-art models.

II. RELATED WORKS

Existing fake news detection models can be classified into two categories: single-modal-based and multimodal-based.

A. SINGLE-MODAL FAKE NEWS DETECTION

We review related work on fake news detection using textual and visual features separately.

1) TEXTUAL FEATURES

Initially, researchers exploit only statistical [30], [31] and semantic features such as the number of emotive words, emoticons, punctuation marks, URLs [32], language styles [33], and writing styles [34] contained in the text. These hand-crafted features are not only inefficient but also neglect the connections between text contexts, resulting in poor performance in fake news detection. In recent years, deep learning has shown great strength in extracting text representations. RNN [15] is employed to capture the potential feature representation of text content. Yu et al. [14] leverage CNN to learn the relationships between words and capture the representation of textual features. Vaibhav et al. [28] utilize GCN (Graph Convolutional Neural Network) based on news sentences to extract a complete text representation. Jwa et al. [29] collect additional news data to pre-train BERT and adopt the model to analyze the relationship between news headlines and body text for fake news detection. These results show that textual features extracted by the deep learning model significantly improve the effectiveness of fake news detection.

2) VISUAL FEATURES

Features such as the number of images contained in the news [35], type of image, and sharpness score of the image are used as evidence for fake news detection. Qi et al. [36] utilize CNN to extract pixel domain and frequency domain features from images to detect fake news. However, CNN requires large-scale training samples to ensure the validity of the model, which is unsuitable for fake news detection tasks. Most existing studies have leveraged vision models pre-trained on large-scale samples, such as VGG-19 (Visual Geometry Group 19) [37], ResNet [38], Swin-T, and ViT, to extract visual features [39]. However, fake news detection that relies only on visual features has achieved limited success. Therefore, an increasing number of scholars are focusing on multimodal fake news detection methods that integrate textual and visual features.

B. MULTIMODAL FAKE NEWS DETECTION

In this section, multimodal fake news detection is categorized into multimodal complementation, consistency, enhancement, and vision-language model-based approaches.

1) MULTIMODAL COMPLEMENTATION

Some researchers believe that different modalities represent different aspects of news information and that multimodal complementarity will help detect fake news. Therefore, they incorporate an image encoder to extract visual features and a text encoder to extract textual features and concatenate their features to form a complete feature representation of news. Singhal et al. [11] exploit pre-trained BERT and XLNet to extract textual features, and VGG-19 to extract visual features. Wang et al. [6] design an event classifier as a subtask for fake news detection and maximized the event classification loss to extract event-invariant joint representations of text and images to improve the generalization performance of the model. Similarly, Khattar et al. [7] reconstruct text and images based on concatenated features to obtain potentially shared representations of text and images for detecting fake news.

2) MULTIMODAL CONSISTENCY

Fake news is more likely to contain inconsistencies between the text and image content. Based on this property, some researchers have utilized inconsistent relationships between modalities when fusing multimodal features. SAFE [16] is proposed to measure the correlation between textual and visual features by invoking deformed cosine similarity and defining a multimodal consistency loss function based on cross-entropy to assist fake news detection. Xue et al. [17] develop weight-sharing networks to extract cross-modal features and calculate the similarity between modalities. However, the semantic gap between text and visual features is so large that it is difficult to accurately compute the consistency between them, which affects the fake news detection performance. Recently, multimodal models such as CLIP have been able to align textual and visual features in the same embedding space, making it possible to capture the semantic relationship between news text and images.

3) MULTIMODAL ENHANCEMENT

Some researchers have argued that news text and image fusion via coarse-grained concatenation is insufficient to mine multimodal high-level semantic information [2], [21].

TABLE 1. Comparison of fake news detection models.

Models	Text Encoder	Visual Encoder	Datasets	Fusion Methods
GCN [28]	GCN	-	LUN; SLN; RPN	-
exBAKE [29]	BERT	-	FNC-1	-
EANN [6]	Text-CNN	VGG-19	Weibo16; Twitter	concatenation; event discriminator
MVAE [7]	Bi-LSTM	VGG-19	Weibo16; Twitter	concatenation; reconstructed decoder; MLP
SAFE [16]	Text-CNN	VGG-19	PolitiFact; GossipCop	concatenation; similarity-aware
AMFB [25]	Bi-LSTM	CNN-GRU	Weibo16; Twitter	single layer bilinear; MLP
SpotFake [11]	BERT	VGG-19	Weibo16; Twitter	concatenation; MLP
MCAN [21]	BERT	CNNs;VGG-19	Weibo16; Twitter	concatenation; co-attention
HMCAN [2]	BERT	ResNet	Weibo16; Twitter	hierarchical attention
MMFN [22]	BERT; CLIP-Text	CLIP-Image; Swin-T	Weibo16; Twitter; GossipCop	concatenation; co-attention

Therefore, they propose the mutual enhancement method for the textual and visual features. MCAN [21] is an enhancement model that fuses spatial and frequency domains and textual features through co-attention repeatedly to detect fake news. Kumari and Ekbal [25] employ single-layer bilinear and MLP (Multilayer Perceptron) to fuse textual and visual features, which are extracted by BiLSTM (Bi-directional Long Short-Term Memory) and CNN-GRU. However, these methods cannot align text and image features in the same embedding space, which limits the effectiveness of model fusion.

4) VISION-LANGUAGE MODEL

The vision-language model provides a new approach for multimodal fake news detection because it can align text and image features in the same embedding space. Luo et al. [40] construct the NewsCLIPpings dataset generated automatically to provide misinformation samples of mismatches between text and images, and benchmark it by invoking vision-language models such as CLIP. However, this study has not been applied to real-world datasets in an open environment, necessitating further verification using collected datasets of both real and fake news. Huang et al. [41] only aggregate CLIP and VinVL to detect inconsistencies between text and images of news to identify misinformation in social media but ignore the complementarity between text and image models that would also contribute to fake news detection. Zhou et al. [22] utilize visual-language models to extract multimodal features, integrate the BERT text feature extractor and Swin-T visual feature extractor, and combine them through multi-granularity fusion. However, multi-granularity fusion methods and multiple pre-trained models increase the model complexity, fusion redundancy, and deployment cost. Therefore, it is necessary to explore simpler and more effective methods to detect fake news.

Table 1 presents a comparison of fake news detection models. Single-modal fake news detection relies solely on single-modal information, such as textual modality, which may limit the effectiveness of model detection. In contrast, multimodal fake news detection integrates text, images, and their relationships to provide a more comprehensive feature representation, thus enhancing the overall efficacy of fake news detection. However, previous studies have employed diverse feature extractors to capture textual and visual features, posing challenges in aligning different modality features in the same embedding space. Despite attempts to exploit consistent relationships between modalities, detection performance remains unsatisfactory owing to significant disparities between text and image features across distinct modal embedding spaces. Moreover, multimodal enhancement often increases the number of features and parameters, which could generate redundant information and noise points when utilizing inputs from different embedding spaces, thereby hindering optimal fusion effects [23].

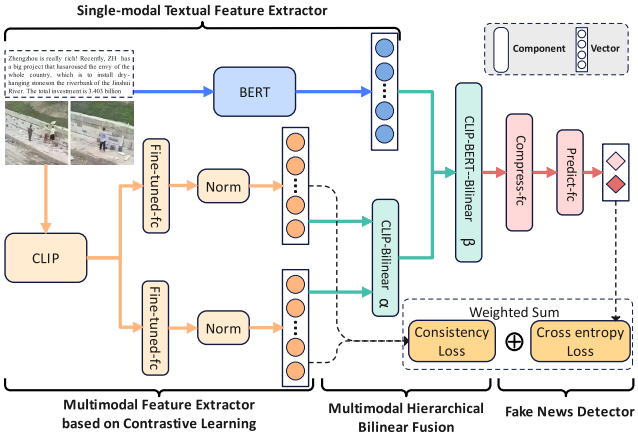

FIGURE 2. The architecture of BC-FND.

To address the limitations of existing studies, the BC-FND leverages the vision-language model to embed multimodal features into the same semantic space. It introduces a hierarchical bilinear fusion approach that effectively combines textual and visual features for fake news detection. The consistency loss function is proposed to capture the dissimilarity between fake news text and image features and the similarity between real news text and image features. This not only enables a comprehensive fusion of textual and visual information but also captures consistent relationships across multimodal. Detailed descriptions of the model are provided in subsequent subsections.

III. MODEL

A. MODEL OVERVIEW

The proposed BC-FND model aims to extract multimodal feature representations in a unified embedding space and

effectively fuse these features to enhance the accuracy of fake news detection. It consists of four key components: a contrastive learning-based multimodal feature extractor, a single-modal textual feature extractor, a multimodal hierarchical bilinear fusion module, and a fake news detector (as depicted in Fig. 2).

B. MULTIMODAL FEATURE EXTRACTOR BASED ON CONTRASTIVE LEARNING

The CLIP model demonstrated robust transfer learning capabilities and achieved success across diverse domains, including image classification, object detection, and image-text retrieval tasks. It excels at extracting multimodal features from a unified embedding space. Specifically, the model inputs the text into the Text Encoder and the image into the Image Encoder. The resulting textual and visual features are combined to form an $N \times N$ feature matrix. Matched image-text pairs correspond to the N elements on the diagonal of this feature matrix, serving as positive samples; while the remaining $N^2 - N$ elements represent unmatched negative samples. Through contrastive learning, CLIP trains its model in an unsupervised manner by optimizing similarity among positive sample pairs and dissimilarity among negative sample pairs.

Through this comparative learning process, the embedding representations of positive sample pairs gradually converge, while those of negative sample pairs diverge further. Moreover, the proximity between text and image in the feature space accurately reflects their semantic closeness. Conversely, a wider semantic gap results in a greater distance between text and image embeddings. This alignment ensures the coexistence of multimodal features within the shared embedding space. Therefore, we employe CLIP to extract textual and visual features from news articles, effectively addressing the issue of disparate embedding spaces encountered in previous fake news detection tasks. Specifically, we utilize a pre-trained Text Encoder for extracting textual features and an Image Encoder for extracting visual features as depicted below.

$$\mathbf{t} = \{t_1, t_2, \ldots, t_n\} = \mathit{CLIP\text{-}TextEncoder}(\mathbf{w}) \tag{1}$$

$$\mathbf{v} = \{v_1, v_2, \ldots, v_n\} = \mathit{CLIP\text{-}ImageEncoder}(\mathbf{i}) \tag{2}$$

where given the text sequence $\mathbf{w} = \{w_1, w_2, \ldots, w_m\}$, m represents the number of words in the sequence, and the image content $\mathbf{i} = \{i_1, i_2, \ldots, i_k\}$, k denotes the number of pixels in the preprocessed image. The feature vectors $\mathbf{t}$ and $\mathbf{v}$ correspond to the outputs of the Text Encoder and Image Encoder, respectively, with both having a dimension n equal to 768.

To accelerate the training, the PEFT (Parameter-Efficient Fine-Tuning) [42] strategy is utilized. This entails storing the multimodal features extracted by the CLIP on the local disk, which is equivalent to effectively freezing the CLIP model parameters. Subsequently, the multimodal features efficient fine-tuning by fully connected and normalized layers for fake news detection, as described below:

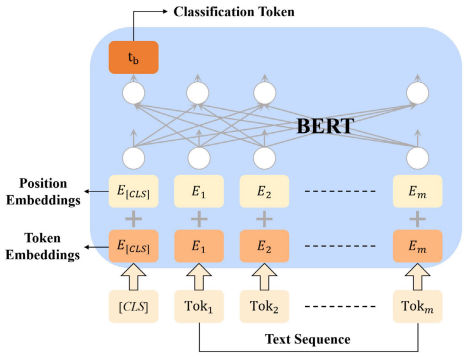

FIGURE 3. Architecture of the BERT model.

$$\mathbf{t_c} = \frac{W_{pt}\mathbf{t} + b_{pt}}{\|W_{pt}\mathbf{t} + b_{pt}\|}, \mathbf{v_c} = \frac{W_{pv}\mathbf{v} + b_{pv}}{\|W_{pv}\mathbf{v} + b_{pv}\|} \tag{3}$$

where $\mathbf{t_c}$ and $\mathbf{v_c}$ are the feature vectors, transformed and normalized from $\mathbf{t}$ and $\mathbf{v}$. W_{pt}, W_{pv} represent the weights and b_{pt}, b_{pv} represent the bias.

It is found that the textual and visual features of real news have semantic closeness, while the semantic closeness of fake news is relatively poor. Therefore, the consistency loss function is proposed to further fine-tune the relationship between textual and visual features to capture the difference between real and fake news. It is deformed from the cosine similarity [16] and cross-entropy functions, as shown in (4) and (5):

$$\mathcal{S}(\mathbf{t_c}, \mathbf{v_c}) = \frac{\mathbf{t_c} \cdot \mathbf{v_c} + \|\mathbf{t_c}\| \|\mathbf{v_c}\|}{2\|\mathbf{t_c}\| \|\mathbf{v_c}\|} \tag{4}$$

$$\mathcal{L}(\theta_c) = -[y log(1 - \mathcal{S}(\mathbf{t_c}, \mathbf{v_c})) + (1 - y) log \mathcal{S}(\mathbf{t_c}, \mathbf{v_c})] \tag{5}$$

where the similarity $\mathcal{S}(\mathbf{t_c}, \mathbf{v_c}) \in [0, 1]$; y is news label, 0 represents real news, 1 represents fake news, θ_c represents the set of learnable parameters in the fine-tuned fully connected layers. The similarity between the textual and visual features of real news increases and the similarity between the textual and visual features of fake news decreases by minimizing $\mathcal{L}(\theta_c)$. It helps the model learn the consistency of real news text and images as well as the inconsistency of fake news to capture the difference between them in terms of the multimodal consistency relationship.

C. SINGLE-MODAL TEXTUAL FEATURE EXTRACTOR

Research indicates that textual features are crucial for fake news detection [6]. However, CLIP alone can not perfectly capture the textual features of fake news. Therefore, BERT is employed to extract textual features to fully leverage the semantic and contextual information present in news text. BERT has been proven effective in translation, text classification, question and answer, and many other areas of NLP (Natural Language Processing). It utilizes the encoder component of the Transformer, specifically the Bert-Base-Chinese, which consists of 12 hidden layers

(encoders). The overall architecture for BERT is depicted in Figure 3.

The input text is processed for word segmentation using the tokenizer, resulting in a total of m tokens obtained. The token sequence is encoded in whole and fed into the 12 hidden layers. Each hidden layer applies multi-head self-attention containing K, Q, and V matrices and transmits the results through a feed-forward neural network. Therefore, regardless of the sequence length, each token learns the mutual relationship between it and other tokens in the sequence. Furthermore, the input of BERT incorporates position embedding of tokens to differentiate the positions of tokens. This enables BERT to effectively capture textual semantic and contextual information in its extracted features.

To enhance the effectiveness of downstream tasks, BERT uses $[CLS]$ token to represent the classification token, which is fed into the hidden layers for training. Therefore, we employ the output corresponding to the $[CLS]$ token as the textual feature of the news article, as shown in (6),

$$\mathbf{t_b} = \{b_1, b_2, \ldots, b_s\} = BERT_{[CLS]}(\mathbf{w}) \tag{6}$$

where $\mathbf{w}$ represents a given text sequence, $\mathbf{t_b}$ denotes the feature vector of the $[CLS]$, and s corresponds to the dimensionality of the feature vector, which is set at 768.

D. MULTIMODAL HIERARCHICAL BILINEAR FUSION

To achieve comprehensive integration of textual and visual features, BC-FND employs a hierarchical bilinear fusion network, as shown in Fig. 2. The fusion process involves two key points, namely α and β. Initially, we utilize bilinear to combine $\mathbf{t_c}$ and $\mathbf{v_c}$ at α point, resulting in the fused representation $\mathbf{c_{vt}}$. Subsequently, we apply the same approach to fuse $\mathbf{t_b}$ with $\mathbf{c_{vt}}$ at β point, yielding the hierarchical fused result $\mathbf{f_{cb}}$. The detailed calculations are shown in (7) and (8),

$$\mathbf{c_{vt}} = \mathbf{t_c^T A v_c} + \mathbf{b}_\alpha \tag{7}$$

$$\mathbf{f_{cb}} = \mathbf{t_b^T B c_{vt}} + \mathbf{b}_\beta \tag{8}$$

where $\mathbf{A}$ represents the parameter matrix for α point fusion, $\mathbf{B}$ represents the parameter matrix of β point fusion, the dimensions of $\mathbf{A}$ are $768 \times 512 \times 512$, while $\mathbf{B}$ has dimensions of $512 \times 768 \times 768$, $\mathbf{b}_\alpha$ and $\mathbf{b}_\beta$ denote the biases associated with bilinear fusion at α and β points respectively.

E. FAKE NEWS DETECTOR

The fused multimodal feature $\mathbf{f_{cb}}$ serves as the input for the fake news detector. Owing to the high dimensionality of $\mathbf{f_{cb}}$ features, we compress them using a compressed fully connected layer (Compress-fc) to achieve better fake news detection results. It is then fed into the predicted fully connected layer (Predict-fc) to obtain the probability of fake news using softmax. The complete equation is as follows:

$$p = softmax(W_p(\sigma(W_c\mathbf{f_{cb}} + b_c)) + b_p) \tag{9}$$

where the parameter W_c is to be learned by the compressed fully connected layer, while W_p is to be learned by the predicted fully connected layer. The activation function $\sigma(\cdot)$ is applied between these two fully connected layers, and b_c and b_p represent their respective biases. The model is trained through minimizing the cross-entropy loss,

$$\mathcal{L}(\theta_p) = -[ylog(p) + (1-y)log(1-p)] \tag{10}$$

where θ_p is denoted as the set of learnable parameters in the model. To formulate the overall loss function, we integrated both cross-entropy and consistency losses:

$$\mathcal{L}(\theta_c, \theta_p) = \gamma\mathcal{L}(\theta_c) + (1-\gamma)\mathcal{L}(\theta_p) \tag{11}$$

where γ represents the experimentally optimized weight hyperparameter.

IV. EXPERIMENTS

A. DATASET

To evaluate the efficacy of BC-FND, we conduct extensive experiments on two publicly available real-world datasets as well as our proposed dataset.

1) WEIBO16

The Weibo16 dataset [43] consists of two categories of news articles: those classified as fake by the official Weibo system and verified as real by the Xinhua News Agency. It encompasses textual content, images, user profiles, and social context data including follower counts, likes, and retweets. Our primary focus lies in the multimodal detection of fake news utilizing both text and image features. To maintain consistency with previous researchers, we filtered out news articles that contained only text or images. In this experiment, we employed the original training set for training purposes while using the original testing set for validation and testing to facilitate a clear contrast effect.

2) WECHAT18

Wang et al. [44] compiled a dataset from WeChat's official accounts, covering the period from March 2018 to October 2018, which included both labeled and unlabeled samples containing information such as titles, images, and user reports. In this study, we focus on labeled data comprising text and images. The statistical details of the refined datasets are listed in Table 2. The data collected from March 2018 to September 2018 is utilized as our training set, while the data from October 2018 is for validation and testing purposes, which is similar to previous studies.

3) WEIBO23

To enhance the evaluation of our model and facilitate research on fake news, we propose a more challenging dataset known as Weibo23. By amalgamating all available fake news from the Weibo Management Community until March 2023 with existing samples from public datasets [45], we formed a comprehensive collection of fake news for Weibo23. Fabricated news articles were thoroughly examined and authenticated by certified experts. To facilitate the

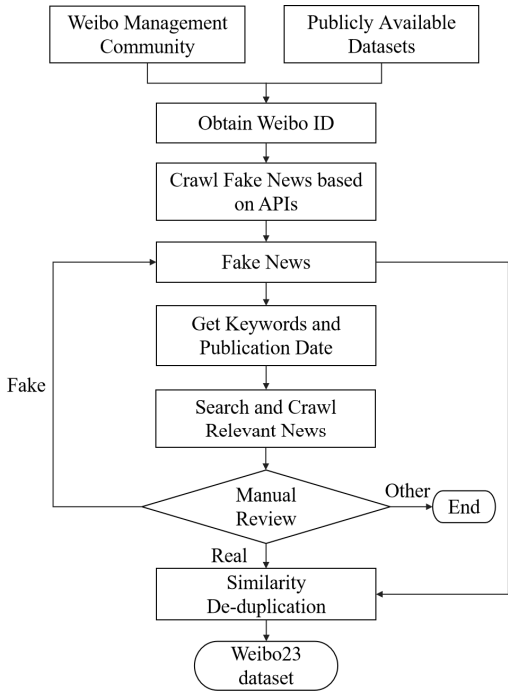

FIGURE 4. Flowchart for Weibo23 dataset collection.

accurate differentiation between fake and real news, it is imperative to minimize content-related disparities between them. Otherwise, the model tends to excessively rely on specific content cues to identify fake news. Therefore, for each instance of fake news, we employed the Baidu API to extract keywords and retrieved relevant news on Weibo based on the publication date and keywords. Subsequently, through manual scrutiny, all collected pertinent news was categorized into fake news, real news, and others (tweets about personal life, emotions, entertainment, etc.). The HANLP API was utilized to compute the similarity between samples within the same class and eliminate duplicates based on their similarity scores. A flowchart of this process is shown in Fig. 4. Furthermore, samples collected before December 31, 2021 were partitioned into training and validation sets, while those obtained from January 1, 2022 onwards constituted the testing set. Table 2 presents relevant statistics.

B. EXPERIMENTAL SETTINGS

1) IMPLEMENTATION DETAILS

We utilized BERT (bert-base-chinese) to extract textual features from WeChat18, Weibo16, and Weibo23 datasets. The maximum text lengths were set to 150, 200, and 300, respectively. This decision was based on the fact that the text in WeChat18 corresponds to typically short news headlines. Prior to 2020, posts on the Weibo platform consisted of short texts while long texts became supported thereafter. Additionally, the CLIP version was clip-vit-large-patch14_336_zh [46]. Original images were cropped and resized to achieve dimensions of $3 \times 336 \times 336$. The input and output dimensions of the "Fine-tuned-fc" layer were

TABLE 2. Statistics of datasets.

	Weibo16		WeChat18		Weibo23	
	Fake	Real	Fake	Real	Fake	Real
Train Samples	3347	2807	2743	2743	1831	1831
Val Samples	432	418	741	741	458	458
Test Samples	432	417	741	741	500	500

768 and 512, respectively. As the "CLIP-Bilinear" layer, the output dimension was 768, while the "CLIP-BERT-Bilinear" layer had an output dimension of 512. The hidden layer size of the "Compress-fc" layer was 512, and for the "Predict-fc" layer was 32. The Adam optimizer was employed to train BC-FND model with a learning rate of $2e-5$ and batch size of 32. The weights γ for the consistency and cross-entropy losses were determined through experimental investigation and were set to 0.5. The maximum number of epochs was 50, and the model was trained until either the maximum epoch size was reached, or the loss converged.

2) EVALUATION METRICS

In fake news detection tasks, accuracy is commonly used as an evaluation metric for models. In this experiment, we aimed to comprehensively assess the effectiveness of the model by incorporating Precision, Recall, and F1 scores as complementary evaluation metrics. The top five results were averaged based on their accuracy rankings.

C. BASELINES

To validate the efficacy of BC-FND, we conducted a comprehensive comparative analysis with state-of-the-art single-modal and multimodal models for fake news detection.

1) SINGLE-MODAL MODELS

CNN-Text [14] utilizes a convolutional neural network to extract textual features of news, followed by fully connected layers for accurate classification. **BERT-Text** employs a pre-trained BERT model to extract informative textual features from news content and subsequently applies an MLP network for fake news detection. **Swin-T** [24] leverages the Swin-T model to extract visual features from images, which are transformed through a fully connected layer for precise classification. **CLIP-I** relies solely on the visual modality features of the CLIP model, while **CLIP-T** exclusively utilizes textual features.

2) MULTIMODAL MODELS

EANN [6] exploits LSTM and CNN to extract textual and visual features, respectively, which are subsequently concatenated. This model incorporates an event discriminator that learns event-invariant features through a gradient inversion layer. **EANN-** is a model without an event discriminator, used on the WeChat18 and Weibo23 datasets, where event labels are unavailable. **MVAE** [7] learns complementary shared representations of text and images by decoding and feeding them into an MLP network for classification. **SpotFake** [11]

TABLE 3. Results of baselines and BC-FND on datasets.

	Methods	Accuracy	Fake News			Real News		
			Precision	Recall	F1 score	Precision	Recall	F1 score
Weibo16	CNN-Text	0.772	0.825	0.702	0.758	0.733	0.845	0.785
	Swin-T	0.770	0.774	0.773	0.774	0.766	0.767	0.767
	BERT-Text	0.878	0.903	0.852	0.876	0.857	0.904	0.879
	EANN	0.806	0.808	0.806	0.807	0.804	0.806	0.805
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SAFE	0.785	0.823	0.736	0.777	0.753	0.836	0.793
	SpotFake	0.903	0.911	0.900	0.905	0.898	0.907	0.902
	CLIP	0.907	0.917	0.898	0.908	0.897	0.916	0.907
	MCAN	0.897	0.898	0.900	0.899	0.897	0.894	0.896
	MMFN	0.893	0.891	0.901	0.895	0.897	0.884	0.890
	BC-FND	**0.942**	**0.946**	**0.940**	**0.943**	**0.939**	**0.944**	**0.942**
WeChat18	CNN-Text	0.740	0.828	0.606	0.700	0.689	0.874	0.771
	Swin-T	0.730	0.795	0.621	0.697	0.689	0.839	0.757
	BERT-Text	0.850	0.903	0.785	0.839	0.811	0.914	0.859
	EANN-	0.743	0.816	0.640	0.717	0.694	0.850	0.764
	SpotFake	0.853	**0.910**	0.784	0.842	0.810	0.922	0.862
	CLIP	0.814	0.902	0.705	0.791	0.758	**0.923**	0.832
	MCAN	0.862	0.904	0.811	0.855	0.829	0.914	0.869
	BC-FND	**0.881**	0.901	**0.857**	**0.878**	**0.864**	0.906	**0.884**
Weibo23	CNN-Text	0.779	0.773	0.790	0.781	0.785	0.768	0.776
	Swin-T	0.688	0.728	0.574	0.642	0.649	0.786	0.711
	CLIP-I	0.778	0.761	0.811	0.785	0.799	0.744	0.770
	CLIP-T	0.777	0.762	0.806	0.783	0.795	0.748	0.770
	BERT-Text	0.816	0.797	0.850	0.823	0.839	0.784	0.811
	EANN-	0.775	0.761	0.790	0.775	0.788	0.760	0.774
	SpotFake	0.813	0.800	0.836	0.817	0.829	0.791	0.809
	CLIP	0.812	0.792	0.849	0.819	0.838	0.776	0.805
	MCAN	0.825	0.793	0.880	0.834	0.865	0.770	0.815
	MMFN	0.816	**0.830**	0.796	0.812	0.805	**0.836**	0.820
	BC-FND	**0.845**	0.800	**0.920**	**0.856**	**0.906**	0.770	**0.832**

combines textual features extracted by BERT with visual features and feeds them into a fully connected layer for fake news detection. **CLIP** employs a vision-language model to extract both textual and visual features, performing fake news prediction through bilinear fusion. **SAFE** [16] leverages both image and textual features and incorporates consistent information to detect fake news. **MCAN** [21] adopts a co-attention mechanism to fuse textual, visual, and frequency-domain features of images for fake news detection. **MMFN** [22] integrates BERT, CLIP, and Swin-T models to extract textual and visual features while utilizing a multilevel co-attention mechanism for detecting fake news.

D. PERFORMANCE COMPARISON

The experimental results of all baseline methods and BC-FND on two publicly available real datasets, as well as the Weibo23 dataset, are presented in Table 3. Based on these findings, the following conclusions were drawn:

(1) By comparing the single-modal models of CNN-Text, BERT-Text, Swin-T, and CLIP-I, it can be concluded that textual features outperform visual features. Especially, the textual features extracted using BERT demonstrate good performance across all three datasets. However, it is worth noting that the textual features extracted by CLIP-T showed significantly lower results than those extracted by BERT for fake news detection on the Weibo23 dataset. This discrepancy may arise from CLIP being primarily designed for multimodal tasks rather than refined processing of single textual features. Therefore, this observation motivates us to integrate BERT into CLIP to extract more informative textual features and enhance the accuracy of fake news detection.

(2) The multimodal EANN and SAFE models employ CNN for textual feature extraction, exhibiting superior performance compared to the single-modal CNN-Text on Weibo16 dataset. This is attributed to the utilization of VGG-19 for visual feature extraction, which effectively enhances model performance. These findings suggest that integrating text and vision information along with their relationships plays a supportive role in fake news detection.

However, the opposite scenario occurred on Weibo23, where the multimodal models EANN and SpotFake underperformed their respective single-modal models CNN-Text and BERT-Text. This is because the Weibo23 test set contains a considerable amount of COVID-19-related fake news, mostly in the form of document images such as notices and chat screenshots. Ordinary vision models encounter challenges in accurately capturing semantic features from document images. It can be observed from the performance of Swin-T, which is considerably lower on Weibo23 than on Weibo16 and WeChat18. The visual features extracted by Swin-T contain substantial text noise, which affects multimodal fusion. Most notably, CLIP-I outperformed Swin-T significantly, indicating that vision language models have an advantage in detecting fake news that includes document images. This advantage is attributed to their ability

TABLE 4. Results of models with different fusion methods.

Weibo16 Methods	Accuracy	Fake News		
		Precision	Recall	F1 score
3Cat	0.884	0.908	0.858	0.882
Cat-Cat	0.908	0.925	0.892	0.908
Cat-Cat-cos	0.909	0.933	0.886	0.909
BiL-Cat	0.921	0.947	0.894	0.920
BiL-Cat-cos	0.922	**0.953**	0.892	0.921
Cat-BiL	0.927	0.950	0.904	0.926
Cat-BiL-cos	0.929	0.927	0.935	0.931
BiL-Bil	0.941	0.946	0.938	0.942
BiL-Bil-cos	**0.942**	0.946	**0.940**	**0.943**

to align textual and visual features within the same semantic embedding space.

(3) The model's performance on the WeChat18 dataset is comparatively lower than that on the Weibo16 dataset, potentially because of the limited semantic information provided by short texts consisting solely of news headlines on the Weibo18 dataset. Additionally, the forward-looking prediction further diminishes the accuracy of the model. The WeChat18 dataset utilizes samples from March 2018 to September 2018 for predicting fake news in October 2018, presenting a more challenging task for the model. The model's accuracy is also notably diminished in the Weibo23 dataset due to its utilization of historical data spanning over a decade for predicting samples from January 2022 to March 2023, resulting in a longer span of forward-looking prediction. On the other hand, the Weibo23 dataset selected real news that was close to fake news in terms of publication time and keywords. The text similarity between samples must not exceed 0.6 within the same class; otherwise, the deduplication operation will be executed. Therefore, this approach enhances the challenge of detecting fake news within the Weibo23 dataset while aligning more closely with the real scene in social media.

(4) The proposed BC-FND outperformed all baseline models across multiple datasets, which can be attributed to the integration of CLIP. CLIP is a dedicated multimodal model that aligns textual and visual features in the same semantic embedding space using contrastive learning. This alignment of multimodal features effectively enhances the accuracy of fake news detection, particularly when news contains document images such as chat screenshots. Moreover, BC-FND benefits from the hierarchical bilinear fusion method and consistency loss function. The hierarchical bilinear fusion enables fine-grained fusion of textual and visual features, thereby avoiding excessive redundancy in fusion. Therefore, BC-FND offers a more streamlined and efficient approach than MMFN, which employs BERT, CLIP, and Swin-T for feature extraction. The consistency loss function facilitates the model in capturing both the inconsistency between fake news text and image, as well as the consistency between real news text and image, thereby enhancing its performance for fake news detection. Further experiments will validate the improvement of the consistency loss function and hierarchical bilinear fusion methods on the model's performance.

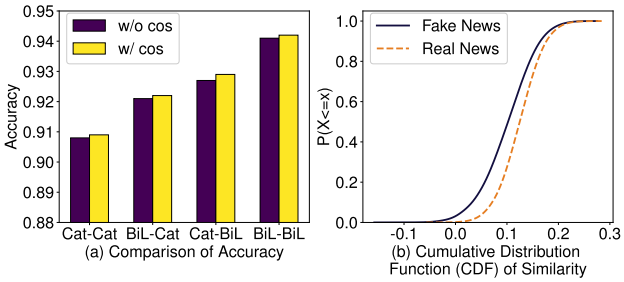

FIGURE 5. Accuracy and cumulative distribution functions.

E. STUDY OF FUSION METHODS

The fusion methods, including straightforward concatenation and bilinear transformation, were experimented with at points α and β in Fig. 2. Furthermore, the impact of multimodal consistency on the fusion method was investigated by excluding the consistency loss function.

For convenience, we adopt the following abbreviations: "BiL" for bilinear fusion, "Cat" for concatenation fusion, and "cos" for consistency loss function. For instance, "Cat-BiL-cos" denotes concatenation fusion for α point and bilinear fusion for β point while utilizing a consistency loss function. To validate the merits of the hierarchical design, we attempted to integrate textual and multimodal features in a single layer referred to as "3Cat". The experimental results on the Weibo16 dataset are presented in Table 4, and the following conclusions can be drawn.

(1) The comparison between "Cat-Cat" and "3Cat" reveals a noticeable increase in the accuracy rate by 2.4% and an improvement in the F1 score by 2.6%. Consequently, it can be inferred that the hierarchical fusion method outperforms the approach of integrating BERT and CLIP features within a single layer.

(2) Based on Fig. 5 (a), a significant enhancement in performance is observed with the utilization of the hierarchical bilinear fusion approach (BiL-BiL). Compared to simple concatenation fusion (Cat-Cat), there is an improvement of 3.3% and 3.4% in accuracy and F1 score, respectively. The increased adoption of bilinear fusion (BiL) in α and β points led to improved model performance, indicating that bilinear fusion achieves fine-grained complementarity between textual and visual features, and enhances fake news detection.

(3) As depicted in Fig. 5 (b), the cumulative distribution functions (CDF) of multimodal similarity for fake news and real news exhibit distinct patterns, indicating disparities in multimodal consistency between the two types. By analyzing the results in Table 4 and Fig. 5 (a), it observed slight enhancements in the accuracy and F1 scores for models utilizing different fusion methods after applying the consistency loss function. This finding further confirms that the consistency loss function could capture both the inconsistency between fake news text and image, as well as the consistency between real news text and image. Therefore, exploiting the consistent relationship between multimodal features allows

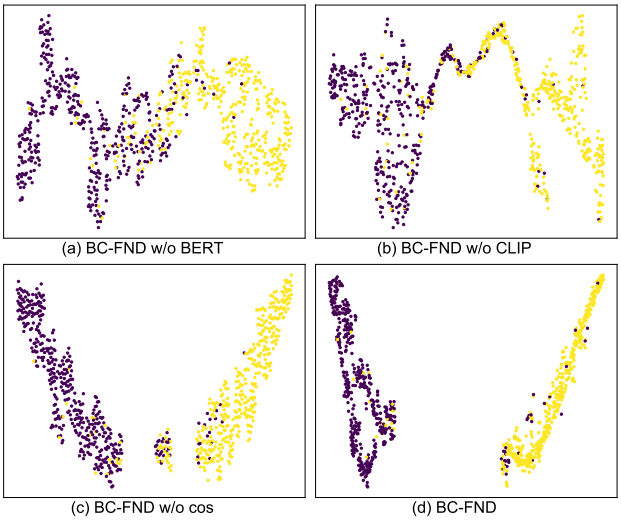

FIGURE 6. Visualization of potential feature representations of different components.

us to identify disparities between fake and real news, leading to an improved model performance.

In summary, the hierarchical fusion method outperforms single-layer fusion, while the bilinear method exhibits a significant advantage over straightforward concatenation. In addition, CLIP aligns textual and visual features in the same embedding space, making it possible to detect fake news by analyzing inconsistencies between images and text.

F. ABLATION ANALYSIS

In Tables 3 and 4, we conduct a quantitative analysis of the performance exhibited by each component of the BC-FND. To further evaluate the effectiveness of each component, this section qualitatively assesses the separability between fake and real news using the Weibo16 dataset. Specifically, we extract the outputs from the penultimate layer of our models and employ t-SNE [47] for dimensionality reduction and visualization purposes. The findings are presented in Fig. 6.

(1) **BC-FND w/o BERT**, when the BERT component is removed, fake news can be distinguished from real news to some extent. However, there is a significant overlap in the decision boundaries between them. Approximately one-third of the feature space contains real and fake samples with similar densities, posing challenges for distinguishing between them.

(2) **BC-FND w/o CLIP**, with the CLIP component removed, exhibits strong discrimination against fake news. However, a "line" connecting fake and real news was evident in the feature space. Consequently, there is no clear decision boundary, and the model struggles to discriminate samples lying along the "line". This suggests that it is difficult to discriminate between these samples based solely on textual features.

(3) As shown in Table 4, **BC-FND w/o cos** (BiL-BiL), which removes the consistency loss function, experiences a slight decrease in accuracy and an F1 score of 0.1%. Nevertheless, comparing Fig. 6 (c) and (d), it is evident that removing the consistency loss function results in a distribution of positive and negative sample features in the feature space, and a small number of samples cannot be accurately distinguished.

(4) **BC-FND** (BiL-BiL-cos), with all components intact, yields a clearer decision boundary and a greater distance between the positive and negative samples in the feature space. The learned features exhibited a greater discriminative power for the two classes sample. In summary, the BC-FND leverages the benefits of various components, each playing a vital role in enhancing performance.

FIGURE 7. Word clouds for the textual content of Weibo23.

V. LIMITATIONS

The performance of our model on the time-divided Weibo23 dataset was unsatisfactory, prompting us to conduct further analysis. By comparing Fig. 7 (a) and (b), the content of the testing set for fake news word clouds shifted considerably over time. This shifting trend may significantly affect model performance. To verify this hypothesis, we reshuffled the Weibo23 dataset and divided it into an 8:1:1 ratio. The reshuffled dataset yielded 90.0% accuracy for BC-FND, which was 5.5% higher than that achieved using the initial time-divided dataset. It is apparent that our model has limitations in detecting future fake news. Notably, despite employing the same shuffling division approach for the Weibo23 and Weibo16 datasets, the model's accuracy on the Weibo23 dataset is lower (90.0% < 94.2%). The distribution of text content between fake news and real news in the testing set, as depicted in Fig. 7 (b) and (c), appears to be relatively similar due to the inclusion of real news based on keywords associated with fake news. This similarity poses a challenge for detecting fake news, which might have contributed to the diminished accuracy of our model.

VI. CONCLUSION

In this paper, we propose BC-FND, a powerful multimodal fake news detection model that offers several advantages: (1) It leverages contrastive learning of CLIP to extract textual and visual features from the united semantic space. (2) We design a hierarchical bilinear fusion method to enhance fine-grained feature complementarity and achieve superior multimodal feature representation. (3) BC-FND is optimized by a jointly of multimodal consistency and cross-entropy losses to learn differences in text and image consistency between fake and real news. Quantitative and qualitative analyses demonstrate significant contributions

of each component towards improving model performance. Experimental results on two public datasets and our dataset confirm the effectiveness and robustness of BC-FND compared to state-of-the-art baselines.

In future work, we will focus on developing models that capture essential differences between fake news and real news more accurately, particularly for cases where their content (keyword) closely resembles each other. For instance, detecting conflicts or evidence from the external environment (hot topics or events) rather than solely relying on its own textual and visual features can aid in identifying fake news effectively. Furthermore, we believe that future research should aim at mining time-invariant features of fake news to adapt to complex and changing social media environments.

ACKNOWLEDGMENT

(Yahui Liu, Wanlong Bing, Shuai Ren, and Hongliang Ma are co-first authors.)

REFERENCES

[1] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. ACM Conf. Inf. Knowl. Manage.*, NY, USA: ACM, Nov. 2017, pp. 797–806.

[2] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA: ACM, Jul. 2021, pp. 153–162.

[3] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal fake news detection with textual, visual and semantic information," in *Proc. 23rd Int. Conf.*, 2020, pp. 30–38.

[4] S. Hangloo and B. Arora, "Combating multimodal fake news on social media: Methods, datasets, and future perspective," *Multimedia Syst.*, vol. 28, no. 6, pp. 2391–2422, Dec. 2022.

[5] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21503–21517, Dec. 2022.

[6] 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 *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, Jul. 2018, pp. 849–857.

[7] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf.*, NY, USA: ACM, May 2019, pp. 2915–2921.

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

[9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. 33rd Conf. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 5753–5763.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[11] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*. Piscataway, NJ, USA: IEEE, Sep. 2019, pp. 39–47.

[12] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 10, pp. 13915–13916.

[13] B. Palani, S. Elango, and V. Viswanathan K, "CB-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and BERT," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5587–5620, Feb. 2022.

[14] F. Yu, Q. Liu, S. Wu, L. Wang, and T. Tan, "A convolutional approach for misinformation identification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3901–3907.

[15] 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," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 3818–3824.

[16] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," 2020, *arXiv:2003.04981*.

[17] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Inf. Process. Manage.*, vol. 58, no. 5, Sep. 2021, Art. no. 102610.

[18] 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 *Proc. 29th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2021, pp. 1212–1220.

[19] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.

[21] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Proc. IJCNLP*, 2021, pp. 2560–2569.

[22] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multi-modal fake news detection on social media via multi-grained information fusion," in *Proc. ACM Int. Conf. Multimedia Retr.*, New York, NY, USA: ACM, Jun. 2023, pp. 343–352.

[23] W. Zhuang and S. Jie, "Multimodal rumor detection model based on multilevel fusion," *Comput. Eng. Des.*, vol. 43, no. 6, pp. 13915–13916, Jun. 2022.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. NJ, USA: IEEE, Oct. 2021, pp. 9992–10002.

[25] R. Kumari and A. Ekbal, "AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115412.

[26] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[28] V. Vaibhav, R. Mandyam, and E. Hovy, "Do sentence interactions matter? Leveraging sentence level representations for fake news classification," in *Proc. 13th Workshop Graph-Based Methods Natural Lang. Process.*, 2019, pp. 134–139.

[29] H. Jwa, D. Oh, K. Park, J. Kang, and H. Lim, "ExBAKE: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT)," *Appl. Sci.*, vol. 9, no. 19, p. 4062, Sep. 2019.

[30] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 647–653.

[31] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," 2017, *arXiv:1702.05638*.

[32] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. Conf. World Wide Web*. New York, NY, USA: ACM, Mar. 2011, pp. 675–684.

[33] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, Jul. 2012, pp. 171–175.

[34] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as 'False news,'" in *Proc. ACM Workshop Multimodal Deception Detection*. New York, NY, USA: ACM, Nov. 2015, pp. 15–19.

[35] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina Weibo by propagation structures," in *Proc. IEEE 31st Int. Conf. Data Eng.*, NJ, USA: IEEE, Apr. 2015, pp. 651–662.

[36] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*. Piscataway, NJ, USA: IEEE, Nov. 2019, pp. 518–527.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. NJ, USA: IEEE, Jun. 2016, pp. 770–778.
[39] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," *AI Open*, vol. 3, pp. 133–155, Sep. 2022.
[40] G. Luo, T. Darrell, and A. Rohrbach, "NewsCLIPpings: Automatic generation of out-of-context multimodal media," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, PA, USA: ACL, 2021, pp. 6801–6817.
[41] M. Huang, S. Jia, M.-C. Chang, and S. Lyu, "Text-image de-contextualization detection using vision-language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. IEEE, May 2022, pp. 8967–8971.
[42] V. Lialin, V. Deshpande, and A. Rumshisky, "Scaling down to scale up: A guide to parameter-efficient fine-tuning," 2023, *arXiv:2303.15647*.
[43] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. 25th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2017, pp. 795–816.
[44] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, "Weak supervision for fake news detection via reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 1, pp. 516–523.
[45] Q. Nan, J. Cao, Y. Zhu, Y. Wang, and J. Li, "MDFEND: Multi-domain fake news detection," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.* NY, USA: ACM, Oct. 2021, pp. 3343–3347.
[46] A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, and C. Zhou, "Chinese CLIP: Contrastive vision-language pretraining in Chinese," 2022, *arXiv:2211.01335*.
[47] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

YAHUI LIU (Member, IEEE) received the Ph.D. degree in computer software and theory from the Institute of Computing Technology, Chinese Academy of Sciences. She is currently an Associate Professor with Shihezi University, China. Her research interests include data mining, databases, social computing, and information security.

WANLONG BING received the B.E. degree in software engineering from Shihezi University, China. He is currently pursuing the master's degree (research) in electronic information with Shihezi University. His research interests include fake news detection and multimodal learning.

SHUAI REN received the B.E. degree in information and computational science from North China Electric Power University, China. He is currently pursuing the master's degree (research) in electronic information with Shihezi University, China. His research interests include rumor detection and machine learning.

HONGLIANG MA received the Ph.D. degree in information security from Beijing Jiaotong University, China. He is currently an Associate Professor with Shihezi University, China. His research interests include network security, data security, artificial intelligence security, and anomaly detection.