Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060025

ARTICLE

Fake News Detection Based on Cross-Modal Ambiguity Computation and Multi-Scale Feature Fusion

Jianxiang Cao[1], Jinyang Wu[1], Wenqian Shang[1,*], Chunhua Wang[1], Kang Song[1], Tong Yi[2,*], Jiajun Cai[1] and Haibin Zhu[3]

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, 100024, China
[2]School of Computer Science and Engineering, Guangxi Normal University, Guilin, 541004, China
[3]Department of Computer Science, Nipissing University, North Bay, ON P1B 8L7, Canada

*Corresponding Authors: Wenqian Shang. Email: shangwenqian@163.com; Tong Yi. Email: yitong@mailbox.gxnu.edu.cn

Received: 22 October 2024; Accepted: 21 February 2025; Published: 16 April 2025

ABSTRACT: With the rapid growth of social media, the spread of fake news has become a growing problem, misleading the public and causing significant harm. As social media content is often composed of both images and text, the use of multimodal approaches for fake news detection has gained significant attention. To solve the problems existing in previous multi-modal fake news detection algorithms, such as insufficient feature extraction and insufficient use of semantic relations between modes, this paper proposes the MFFFND-Co (Multimodal Feature Fusion Fake News Detection with Co-Attention Block) model. First, the model deeply explores the textual content, image content, and frequency domain features. Then, it employs a Co-Attention mechanism for cross-modal fusion. Additionally, a semantic consistency detection module is designed to quantify semantic deviations, thereby enhancing the performance of fake news detection. Experimentally verified on two commonly used datasets, Twitter and Weibo, the model achieved F1 scores of 90.0% and 94.0%, respectively, significantly outperforming the pre-modified MFFFND (Multimodal Feature Fusion Fake News Detection with Attention Block) model and surpassing other baseline models. This improves the accuracy of detecting fake information in artificial intelligence detection and engineering software detection.

KEYWORDS: Fake news detection; multimodal; cross-modal ambiguity computation; multi-scale feature fusion

1 Introduction

With the rise of the Internet and mobile devices, social media has become the primary platform for sharing and accessing information. Weibo, China's leading social media platform, has 550 million monthly active users, while Twitter and Facebook have approximately 3 billion active users worldwide [1–3]. Many fake news is deliberately created to attract attention, serve economic interests, or pursue political agendas. If not controlled, they could lead to economic losses and social unrest [4–6]. Information on social platforms has quickly evolved from plain text to a mix of text and images. In recent years, short video content has also grown. This shows that social media news is shifting towards a multimodal trend [7,8]. Therefore, research on multimodal fake news detection on social platforms has become increasingly urgent and important, attracting the attention of many leading researchers.

Jin et al. [9] and Singhal et al. [10] proposed attRNN and MKEMN, respectively. However, both methods require extensive background information, making them unsuitable for verifying authenticity before news

publication. Wang et al. [11] proposed the EANN framework, which uses image-text feature linking for basic cross-modal feature complementation. MCAN was proposed by Wu et al. [12], which stacks multiple attention layers to mine the relationship between text and vision, but it ignores the semantic relationship between modalities. The CAFE model introduced by Chen et al. [13] adaptively fuses single-modal and cross-modal features using cross-modal ambiguity, but it does not account for the frequency domain information of the modalities. In the area of multimodal fake news detection, the research mentioned above has advanced by implementing cutting-edge feature extraction, fusion methods, and model optimization strategies. These advancements have enhanced the accuracy and robustness of fake news detection, but they still have the following limitations:

(1) The problem of fewer modal features is considered, such as only focusing on content level features and ignoring fake features at the physical level.

(2) During the feature extraction and fusion of modalities like text and image, the semantic relationship between different modalities has not been deeply explored.

Based on this, this paper proposes the MFFFND-Co model, which integrates three multimodal features, namely text content, image content, and image frequency domain features. This paper investigates the semantic alignment between text and image content, optimizing multimodal feature integration to boost fake news detection effectiveness. The key contributions of this study are as follows:

(1) To address the issue of limited features in multimodal fake news detection, this paper incorporates frequency domain features of images to enhance detection accuracy.

(2) On the basis of fully integrating the information of each modality, the ambiguity calculation between images and texts is carried out, the semantic relationship between modalities is mined, and the efficiency of fake news detection is optimized through semantic deviation.

(3) In this paper, we propose the MFFFND-Co model, which comprehensively considers multi-scale features and calculates cross-modal ambiguity to verify the authenticity of news.

2 Related Work

Multimodal fake news detection integrates information from various modalities to identify fake news. Jin et al. [9] proposed the att-RNN, which employs a recurrent neural network with an attention mechanism to simultaneously process image, text, and social context information, then combines these features for authenticity classification. However, the social context features in their model face challenges related to manual annotation and extraction. Khattar et al. [14] proposed the MVAE method, which uses a variable autoencoder to reconstruct the text and image representation, and it quantifies the relationship of text and image between modalities. MVAE has good performance, but it has a high computational cost. The EANN method proposed by Wang et al. [11] connects image and text features, and then distinguishes the event type of the currently analyzed news through a multi-task learning framework, and removes some irrelevant event information to assist in the final judgment. Singhal et al.'s [10] method was the first to introduce pre-trained language models into multimodal fake news detection. It uses BERT to extract text features, VGG19 to extract image features, and performs detection by concatenating them. These methods employed the pre-trained VGG19 network to extract visual features. However, the fusion of modalities was carried out through simple concatenation, lacking a more interactive integration of modal information.

Relatively speaking, the multimodal fusion method can achieve better results. The MKEMN method introduced by Zhang et al. [15] employs an attention mechanism to fuse text and image representations. Additionally, the model incorporates external knowledge to learn the representation of each multimodal news item. In Duc Tuan et al.'s [16] method, fine-tuned BERT is used to extract text embeddings, and

VGG19 extracts image information. Attention mechanisms are employed to fuse multimodal information, and multiple One-dimensional Convolutional Neural Networks (CNNs) are used to further compress the text content. The MCAN method proposed by Wu et al. [12] uses the Co-Attention network to repeatedly fuse text and image information. At the same time, this method also introduces frequency domain information, and it further uses the image tampering period information embodied in the frequency domain to assist detection, but this method ignores the auxiliary effect brought by the semantic consistency between modalities [17]. Qian et al. [18] introduced the Hierarchical Multimodal Contextual Attention Network (HMCAN), which integrates multimodal context with hierarchical semantic information from text and fuses text and image features to enhance news prediction. Zhang et al. [19] proposed the MVAE method, a BERT-driven multimodal framework designed for detecting unreliable COVID-19 news. This approach utilizes a contrastive learning strategy to extract textual and visual features from unreliable articles for fake news identification. Fung et al. [20] constructed a new knowledge element-level benchmark dataset, and proposed a knowledge element-level fake news detection method, to combat the fake news generated by neural networks, but it isn't convenient to detect artificial fake news. Chen et al. [13] calculated the similarity between modes by KL divergence, and adaptively fused the features of different modes for detection, but this method was relatively deficient in feature extraction. Lao et al. [21] pioneered the FSRU method, incorporating frequency-domain information for fake news detection. However, it fails to fully utilize the abundant information present in the spatial domain [22–24]. Shang et al. [25] introduced the DGExplain method, utilizing an object-aware multimodal feature encoder to capture essential information from news content and comments. It also combines text information generated by image guidance, image information generated by text guidance, and content generated by a comment interpreter to determine the authenticity of news through concatenation. However, these methods have some problems, such as ignoring the semantic consistency relationship between texts and images, and less modal feature extraction.

Therefore, this paper proposes the MFFFND-Co model, which not only fuses the spatial domain features of different modalities, but also introduces the frequency domain features of the image. By calculating the cross-modal consistency information, the Co-Attention mechanism is combined for feature fusion, to effectively detect fake news.

3 Method

3.1 Model Structure

Multimodal fake news detection generally includes three main steps: extracting multimodal features, fusing these features, and classifying fake news. Unlike single-modal detection, multimodal methods can leverage features from different modalities, improving detection effectiveness. However, current approaches often fail to fully utilize modality features, and the integration of multimodal data remains limited. In many studies, image modality usually only uses spatial domain information for feature extraction, but image tampering and image compression also exist in the frequency domain. To fully fuse the information of various modalities and further utilize the features of various modalities, this paper proposes the MFFFND-Co model, with the structure shown in Fig. 1.

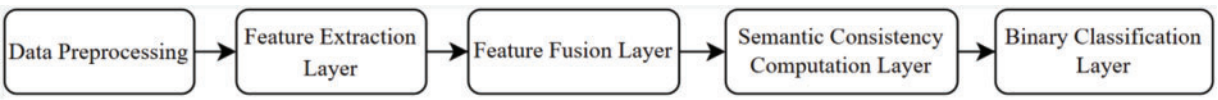


Figure 1: MFFFND-Co model flow chart

The model is composed of four key modules. A simplified flowchart is provided to give an overview of the MFFFND-Co architecture, as shown in Fig. 2.

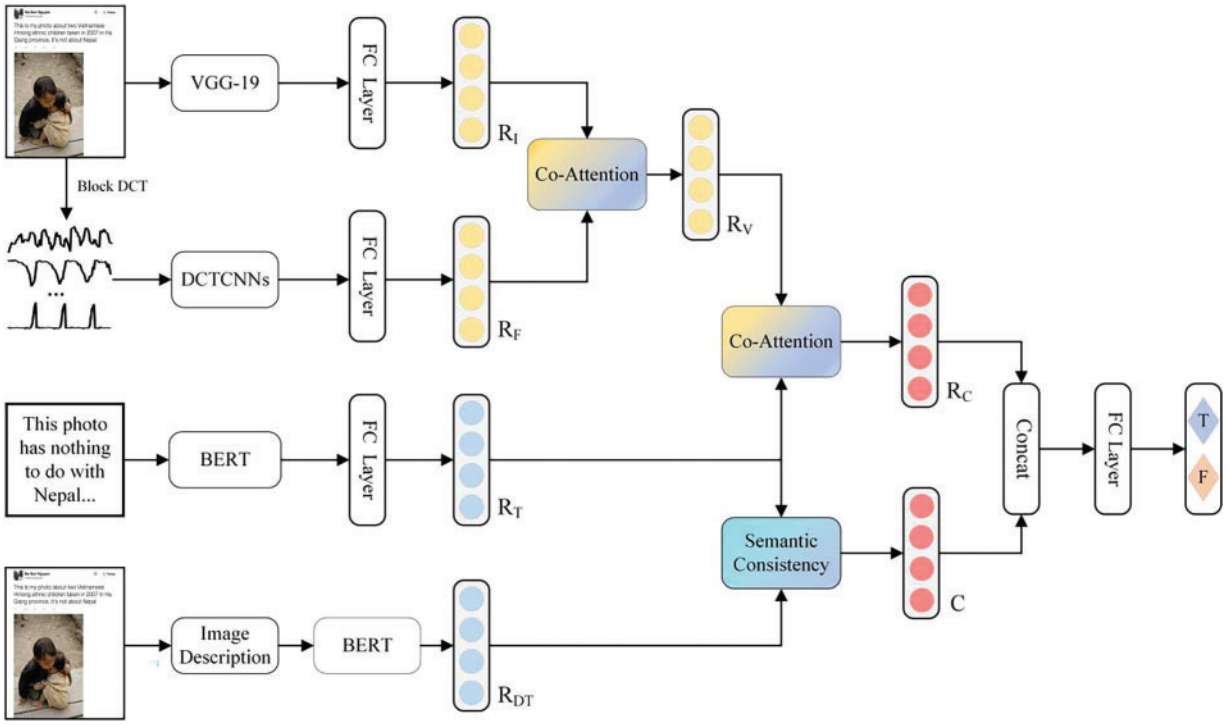


Figure 2: MFFFND-Co model structure diagram

(1) Feature extraction layer: extract the embedding vector information of the input text and image, including the text in the image and the generated image caption text; (2) Feature fusion layer: fusing features extracted from text and image; (3) Semantic consistency calculation layer: calculate the semantic consistency of different modal descriptions to obtain consistency metrics; (4) Binary classification layer: the information in (2) and (3) is combined to generate multimodal news representation vectors and perform binary classification. The following sections describe each layer in detail.

3.2 Feature Extraction Layer

3.2.1 Text Feature Extraction

In this paper, BERT is employed as the text feature extractor, as it outperforms other models in capturing semantic relationships between words and their contexts [26–30]. The input news text undergoes encoding, BERT processing, and fully connected activation to generate the vector representation of the text.

Before being input into the Transformer encoder of BERT, the text undergoes segmentation and encoding. The text is represented as a sequence of words $T = [T_1, T_2, \ldots, T_n]$, where n denotes the total number of words. In the encoding stage, each word is converted into its corresponding word vector using word embedding techniques. Afterward, the word vectors are passed through BERT, where the multi-layer Transformer encoder processes both semantic and positional encoding information for each word, resulting in the hidden state sequence output $O = [o_1, o_2, \ldots, o_n]$, and the process is shown in Eqs. (1) and (2).

$$I_{text} = WordEmbedding(T) \tag{1}$$

$$O = BERT(I_{text}) \tag{2}$$

O can be fine-tuned as a result of $BERT$ output to be used for downstream fake news detection tasks, because the vector O_j of each dimension in O has global context information. Through the fully connected layer of $ReLU$ activation function, the hidden state sequence output O is transformed into $R_T = [r_0, r_1, \ldots, r_{63}]$, as the final representation sequence of the text, R_T is a vector of dimension d × 1, and the process is shown in Eq. (3).

$$R_T = ReLU(FC(O)) \tag{3}$$

where FC represents the fully connected layer and $ReLU$ represents the $ReLU$ activation function.

3.2.2 Image Feature Extraction

To effectively extract information from images in news, this section captures both spatial and frequency domain representations, serving as a shared representation of the two modalities of images.

Spatial domain features: Spatial information represents the semantics conveyed by an image, such as the event occurring, the main visual subjects, etc. [31,32]. To obtain the semantic representation of a given news image, the VGG19 network is used to extract spatial domain features [33]. After image compression, VGG19 encoding, and activation through fully connected layers, a vector representation of the image's spatial domain is obtained. The VGG19 network is used o fine-tune the given news data network, and the low-level semantic features of the image can be extracted to help the detection, as shown in Eq. (4).

$$R_I = ReLU(VGG19(I)) \tag{4}$$

where R_I represents the spatial representation of the image and I represents the input image.

Frequency domain features: Frequency domain information reflects compression and modifications in images. Compressed and spliced images, common in fake news, exhibit distinct periodic patterns in the frequency domain, making them easily identifiable by CNNs [34]. To more effectively extract frequency domain features, the DCT-CNN (discrete cosine transform convolutional neural network) is employed, inspired by the design of VGG19 and Inception networks. Through deep convolution and pooling operations, the network deeply digs the compression and tampering features in the image frequency domain, providing important evidence for the identification of fake news. Its structure is shown in Fig. 3.

In the first stage, continuous single convolution and pooling operations are performed, like VGG19, using multiple small convolution kernels for the convolution process. In the second stage, a multi-branch network like Inception V3 is used for convolution, and the outputs between the branch networks are concatenated. In the third stage, through pooling and convolution once, the final output is obtained by combining a layer of a fully connected layer with the ReLU activation function. The frequency domain is represented as R_F, the process is shown in Eq. (5).

$$R_F = DCTCNNs(Fourier(I)) \tag{5}$$

where $Fourier$ refers to the Fourier transform, which shifts an image from the spatial domain to the frequency domain.

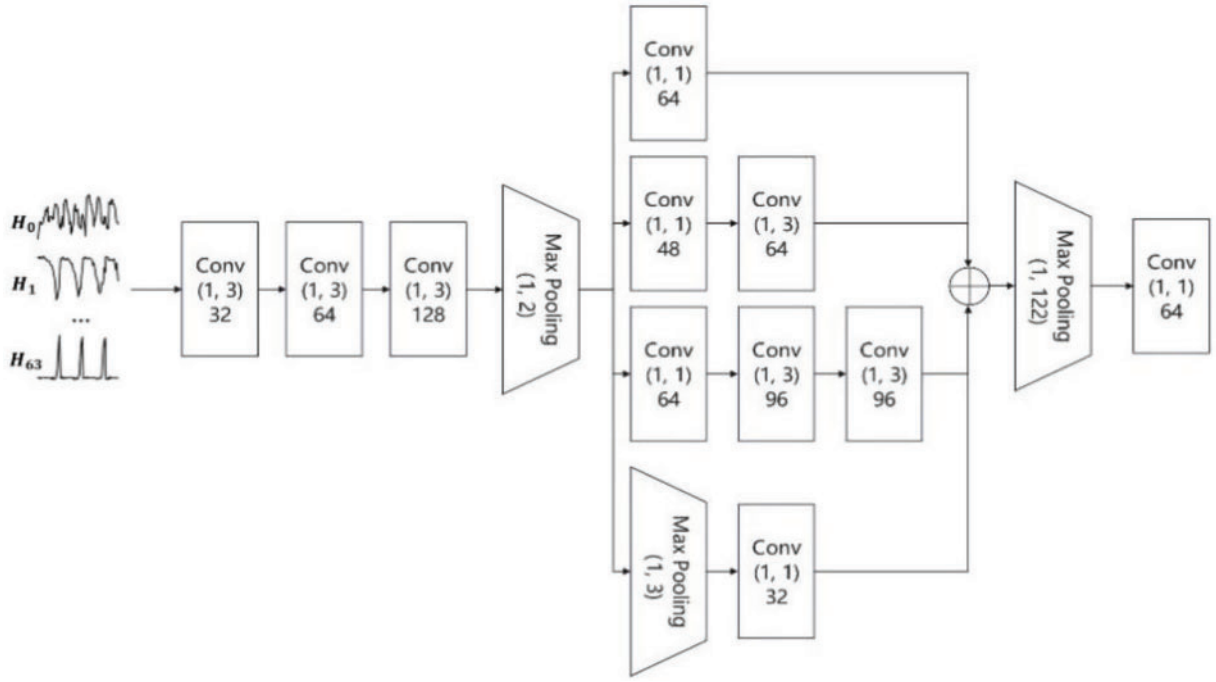


Figure 3: Structure of DCT-CNN

3.3 Feature Fusion Layer

The Co-Attention mechanism calculates the information of the two modalities as Q, K, and V each other, and has achieved excellent performance in cross-modal tasks such as VQA and Text2Image. Similar to the VQA problem, when reading multimodal news, readers usually repeatedly observe images and text, which is very similar to the idea of Co-Attention, so we introduce this mechanism to fuse the multimodal information of fake news [35].

Co-Attention, a modification of the multi-head attention mechanism, captures global dependencies by treating one modality as the Query and the other as the Key and Value. The rest of the process is consistent with the standard multi-head attention mechanism.

For a certain modality, Co-Attention generates a representation which is based on another modality. For example, if Q comes from the frequency domain, K and V come from the spatial domain, then the computed result is the frequency domain representation based on the spatial guidance, as shown in Eqs. (6)–(8).

$$Q_i = Q_I W_i^Q, K_i = K_F W_i^K, V_i = V_F W_i^V \tag{6}$$

$$head_i = Attention(Q_i, K_i, V_i) \tag{7}$$

$$MultiHead(Q, K, V) = Concact(head_1, head_2, \ldots, head_n) W^o \tag{8}$$

where Q_i comes from spatial domain of the image, K_F and V_F come from frequency domain of the image. The attention calculation process of $head_i$ is shown in Eq. (9).

$$head_i = Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_h}}\right) V_i \tag{9}$$

In MFFFND-Co, first through the airspace guided generated based on frequency domain airspace guided, said a joint representation of images. Then, image guidance is used to generate an image-guided text representation as a "text-image" joint representation. Both undergo the residual fully connected layer operation during the joint process, with the detailed procedure outlined in Eqs. (10)–(13).

$$H_v = CA(R_I, R_F) \tag{10}$$

$$R_V = H_v + FC(H_v) \tag{11}$$

$$H = CA(R_V, R_T) \tag{12}$$

$$R_C = H + FC(H) \tag{13}$$

where CA represents the Co-Attention mechanism layer, R_V is the "time domain-frequency domain" joint representation of the image, R_C is the "text-image" joint representation after the residual. H_v is the visual joint representation calculated by the Co-Attention mechanism layer, and H is the news content representation by the Co-Attention mechanism layer.

3.4 Semantic Consistency Calculation Layer

The difficulty in detecting fake news stems from the intentional combination of images and text to create misleading content. However, due to the natural connection between cross-modal semantic information, it is difficult to fully disguise it. The conflict between the semantics of the image and text becomes a key clue for identifying fake news.

The semantic consistency calculation method is implemented by cosine similarity calculation. The news text provides an overview of the story, while the news image offers a spatial and visual interpretation of the content. Specifically, this paper first uses the image description API to convert news images into text descriptions, to provide a unified representation for cross-modal comparison. Then, through the pre-trained BERT model, the representation vector R_{DT} of the image description text is extracted, and combined with the news text representation vector R_T, the cosine similarity between R_{DT} and R_T is calculated as the semantic consistency measure between the two modalities. The process is shown in Eqs. (14)–(16).

$$D_T = Image\_Caption(I) \tag{14}$$

$$R_{DT} = BERT(D_T) \tag{15}$$

$$Consistency(T, V) = \frac{R_{DT} * R_T}{||R_{DT}||||R_T||} \tag{16}$$

where D_T represents the text description of the news image and $Image\_Caption$ represents the image caption API. The computed consistency metric is concatenated with the "text-visual" joint representation. Then, the concatenated result is fed into a binary classification layer for fake news prediction, as shown in Eq. (17).

$$R = concatenate(R_C, Consistency\,(T, V)) \tag{17}$$

where R is the final representation of the news and $concatenate$ represents vector concatenation.

The MFFFND-Co model, with its semantic consistency detection module, effectively leverages both the spatial and visual information of images and the semantic information of text. This addresses the shortcomings of traditional single-modality models, enhancing the accuracy of fake news detection.

3.5 Binary Classification Layer

Following the image and text representation extraction layers, along with the feature fusion layer, a "text-visual" joint representation of the news article is obtained, encompassing image visual features, frequency domain features, and text features. To predict fake news, a fully connected layer with a Softmax activation function computes the probabilities of the joint representation R corresponding to true or fake news. The process is shown in Eq. (18).

$$p = softmax(WR + b) \tag{18}$$

where $p = [p_0, p_1]$ predicts the possibility of the news being real or fake. Where W is the matrix weight and$b$is the bias term. Here, the cross-entropy is chosen as the loss function, and the objective of the model is to minimize the function value, whose function expression is shown in Eq. (19).

$$L = -[y\ log p_0 + (1 - y) log p_1] \tag{19}$$

where y ∈ {0, 1} represents fact labels, which influence all learnable parameters through model training and backpropagation of parameters.

4 Experimental Results and Analysis

4.1 Dateset

To evaluate the performance of the MFFFND-Co model presented in this paper, experiments were conducted using two datasets: the Weibo dataset and the English news dataset [9,31]. These two datasets were selected because one is an all-Chinese dataset and the other is an all-English dataset, both of which are multimodal news datasets. The dataset description is given in Table 1.

Table 1: Dataset information

Dataset	Language	Tag category	Total quantity	Number of positive samples	Number of negative samples
Weibo	Chinese	2	7850	4211	3639
English news	English	2	5669	2825	2844

The Weibo dataset comes from the largest social media platform in China, in which the true news is verified by Xinhua, which is the official authoritative news agency of China, and the fake news is provided by Weibo and the national rumor-refuting platform of the Ministry of Public Security. The data source of Weibo is true and reliable, and it is the most commonly used dataset in multimodal fake news detection research. Since multimodal fake news detection was proposed, a large number of studies have used it for performance analysis. The fake news of the English news dataset comes from Kaggle, and the fake news is composed of the classic fake news dataset BS Detector, which contains the content collected from 244 fake news websites. The authentic news data is sourced from reputable international English news outlets, such as the New York Times.

By cleaning the txt file and picture information of the dataset, the data items with both images and text were filtered. For news with multiple pictures, the first one was selected as its image data. Finally, the data was extracted, labeled, and organized into csv files for better experimental performance.

4.2 Experiment Setting

The experimental environment involved in the model experiments in this paper is the same machine, and the configuration of the experimental environment is shown in Table 2.

Table 2: Experimental environment configuration table

Experimental environment	Configuration
OS	Windows11 (64-bit)
CPU	11th Gen Intel(R) Core(TM) i7-11800H
GPU	GeForce 3070 Laptop
Memory	16 GB
Disk	1 T
Languages	Python3.8
Deep learning framework	Tensorflow2.10.0

In this paper, a pre-trained BERT model is utilized to generate word vector representations, with the maximum sequence length set to 256, matching the vector dimension derived from the images. The sentence vector is obtained by extracting the content corresponding to the [CLS] tag. For models that don't include BERT, Word2Vec is used to obtain vector representations of their words. For text entities, this paper extracts them through API and open-source libraries. For the Co-Attention module, eight headers are used, along with a 256-dimensional fully connected layer output. During training, the size of a Batch is set to 64. The model is trained for 120 epochs, and early stopping is used during training to prevent overfitting. For the training optimization method, this paper uses the Adam method for optimization, using ReLU as the activation function, and the dropout rate is set to 0.3.

4.3 Evaluation Metrics

This paper employs accuracy, precision, recall, and the F1 score as evaluation metrics to quantify the effectiveness of fake news detection. The formulas are given in Eqs. (20)–(23).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

$$Recall = \frac{TP}{TP + FN} \tag{22}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{23}$$

Here, TP denotes the count of samples where true news is correctly identified as true, FP refers to the samples where fake news is mistakenly classified as true, FN represents the cases where true news is incorrectly predicted as fake, and TN indicates the number of samples where fake news is accurately recognized as fake.

4.4 Performance Study

To assess the performance of the proposed model in differentiating between real and fake news, it is compared with several recent representative multimodal fake news detection models as baseline models.

Additionally, various unimodal models are included to assess the performance improvement achieved by the multimodal approach.

Unimodal model:

Bi-LSTM: Using a bidirectional LSTM-based network to distinguish the authenticity of a piece of news text.

BERT: A pre-trained BERT model is employed to generate the text representation, followed by a fully connected layer to assess the authenticity of the news.

VGG19: Using the fine-tuned VGG19 model to distinguish the authenticity of news from news images.

Multimodal model:

att-RNN [9]: The model combines RNN and Attention mechanism to fuse text, social context and image features for fake news detection. LSTM processes text and social context features, merging them with image features, while Neural Attention from LSTM is applied during visual feature fusion. Social context data is removed here for a reasonable comparison.

MVAE [14]: This model uses a bimodal variational autoencoder along with a binary classifier for fake news detection. It consists of a text and image encoder, a decoder, and a detection module for fake news.

MKEMN [15]: This model improves the semantic understanding of short news texts by extracting entities from a knowledge base, using conceptual knowledge to enhance rumor detection accuracy. It also introduces a multi-channel CNN for aligning text and visual knowledge to fuse multimodal information.

CAFE [13]: The model introduces an auxiliary task to align text and image features, evaluates their similarity using KL divergence, and further weight the unimodal information and cross-modal fusion information.

MCAN [12]: This model divides the image into two modules: frequency domain and spatial domain, and it uses the Co-Attention mechanism for the first time to repeatedly fuse the information between each modality. The news authenticity is judged by the vector representation output to the fully connected layer after four fusions.

MFFFND: The comparison model is modified by MFFFND-Co, the Co-Attention Block is modified to the Attention Block, and the semantic consistency layer is deleted to compare the effects of Co-Attention and semantic consistency calculations.

The experimental results are shown in Table 3.

As can be seen from Table 3: (1) Most multimodal models outperform single-modal models, suggesting that incorporating additional features generally enhances model performance, though the relationship is not strictly linear. (2) Using pre-trained deep learning models to process text features has better performance, even outperforming some multimodal methods using traditional deep learning models. (3) The accuracy of BERT and Bi-LSTM on the two datasets is at least 5.5% higher than that of VGG19, which shows that text features are very important in the field of fake news detection. (4) The MFFFND model performs worse than MCAN, mainly because of the lack of fusion of Co-Attention mechanism. MFFFND-Co outperforms MCAN in most cases, benefiting from the auxiliary role of semantic consistency measure, which helps to detect more semantically inconsistent samples. However, although MFFFND-Co outperforms MCAN in most metrics, the improvement is small, which is mainly due to the insufficient accuracy of cosine similarity calculation caused by the anisotropy of BERT vector space. (5) The F1 score of the MFFFND-Co model on the Weibo dataset is 90.0%, and the F1 score on the Twitter dataset is 94.0%, which exceeds other multimodal fusion models, indicating that the multimodal fusion model combined with semantic consistency measurement has a certain performance improvement for fake news detection.

Table 3: Experimental results of model comparison. The best results are in boldface

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Weibo	Bi-LSTM	0.785	0.851	0.692	0.763
	BERT	0.830	0.977	0.675	0.798
	VGG19	0.730	0.789	0.626	0.698
	att-RNN [9]	0.808	0.882	0.711	0.787
	MVAE [14]	0.797	0.827	0.751	0.787
	MKEMN [15]	0.805	0.865	0.722	0.787
	CAFE [13]	0.840	0.855	0.830	0.842
	MCAN [12]	0.899	**0.898**	0.899	0.898
	MFFFND	0.873	0.881	0.879	0.880
	MFFFND-Co	**0.901**	**0.898**	**0.902**	**0.900**
English news	Bi-LSTM	0.864	0.877	0.843	0.859
	BERT	0.873	0.869	0.875	0.872
	VGG19	0.773	0.783	0.744	0.764
	att-RNN [9]	0.872	0.861	0.882	0.871
	MVAE [14]	0.879	0.902	0.848	0.874
	MKEMN [15]	0.889	0.846	0.929	0.886
	CAFE [13]	0.806	0.807	0.799	0.803
	MCAN [12]	**0.942**	0.931	**0.947**	0.939
	MFFFND	0.904	0.920	0.913	0.916
	MFFFND-Co	0.939	**0.933**	**0.947**	**0.940**

4.5 Ablation Study

To verify the efficiency of each component of MFFFND-Co, this section compares the performance of each part of the overall model architecture after removing the function. The overall model architecture is MFFFND-Co with all modules, including Co-Attention fusion layer, frequency domain information extraction layer, spatial domain information extraction layer, text information extraction layer, and semantic consistency calculation layer. The removal of the feature extraction part means that the relevant Co-Attention fusion layer is also removed, and the removed model group is as follows:

w/o CA: MFFFND-Co removes the Co-Attention mechanism fusion and instead uses splicing to send to the fully connected layer for fusion.

w/o Text: MFFFND-Co removes the text information extraction as well as the fusion layer that combines text and image joint information, and it directly uses image information and semantic consistency information for the final discrimination.

w/o Frequency: MFFFND-Co removes the extraction of image frequency domain information, as well as the fusion layer that combines frequency domain information and image spatial domain information, allowing image spatial features to be directly fused with text features.

w/o Space: MFFFND-Co removes the image spatial domain information as well as the fusion layer that combines spatial domain information and image frequency domain information, allowing for direct fusion of image frequency domain features with text features.

w/o Semantic: MFFFND-Co removes the semantic consistency calculation layer and only remains the final "text-visual" joint representation for fake news detection.

The results of the ablation experiments are shown in Table 4.

Table 4: Results of ablation experiments. The best results are in boldface

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Weibo	MFFFND-Co	**0.901**	**0.898**	**0.902**	**0.900**
	MFFFND	0.873	0.881	0.879	0.880
	w/o CA	0.854	0.863	0.865	0.864
	w/o Text	0.855	0.864	0.875	0.869
	w/o Frequency	0.893	0.884	0.893	0.888
	w/o Space	0.885	0.883	0.893	0.888
	w/o Semantic	0.899	0.898	0.899	0.898
English news	MFFFND-Co	**0.939**	**0.933**	0.947	**0.940**
	MFFFND	0.904	0.920	0.913	0.916
	w/o CA	0.875	0.884	0.847	0.865
	w/o Text	0.864	0.891	0.868	0.879
	w/o Frequency	0.928	0.929	**0.948**	0.938
	w/o Space	0.901	0.896	0.904	0.900
	w/o Semantic	0.937	0.931	0.947	0.939

Table 4 shows that: (1) All the components contribute to the performance of MFFFND-Co. (2) Each modal information improves the final accuracy detection performance, and the text information has the greatest contribution. On the Weibo dataset, the accuracy of text is improved by 4.6%, and on the Twitter data set, the accuracy is improved by 7.5%. (3) Co-Attention to various modal information fusion also have significant effect, compared with the performance of the MFFFND, after the removal of Co-Attention on the accuracy of the two dataset with F1 scores dropped by at least 2.8%. (4) Both semantic consistency detection and frequency domain information are helpful to improve the overall performance, and news with significant semantic deviation and serious image tampering can be detected.

4.6 Qualitative Analysis

To further show the fusion effect of Co-Attention (-ca), this section also tests the performance of multimodal information fusion layer of MFFFND-Co. The fusion methods in the frequency domain and time domain fusion, as well as the image-text fusion stages, were modified using concatenation (-c), attention mechanism (-a), and multi-head self-attention mechanism (-ma) for performance evaluation. The results are shown in Fig. 4.

Fig. 4 shows that the accuracy of information fusion using the concatenation method is the lowest, which is due to the lack of attention to key information compared with the attention mechanism method. Multi-head self-attention mechanisms perform better. This is because they can simultaneously explore internal relationships among different parts of the concatenated sequence, providing more information than a single attention mechanism. Since the Co-Attention method uses the information of different modalities to query during fusion, it can better simulate the process of human reading multimodal news, so it has achieved

better performance. In summary, using the Co-Attention method can more fully learn the representation difference between real news and fake news, and get better performance.

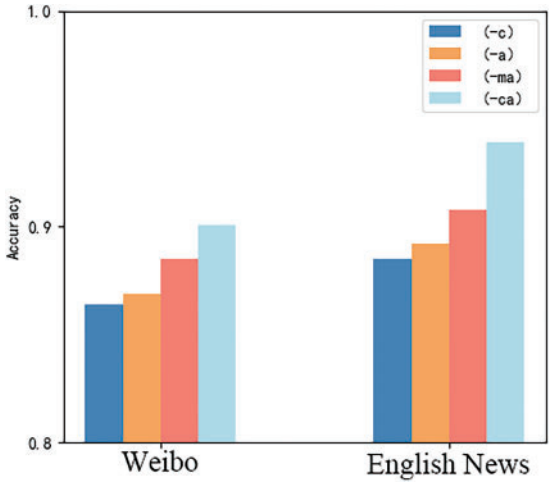


Figure 4: Experimental bar chart of fusion mode

For the frequency domain extraction model, a similar comparison scheme is adopted. The frequency domain feature extraction module is modified to include a frequency domain coefficient fully connected layer (-f) and DCT-CNN. Performance is compared with DCT-CNNs, and the results are shown in Fig. 5.

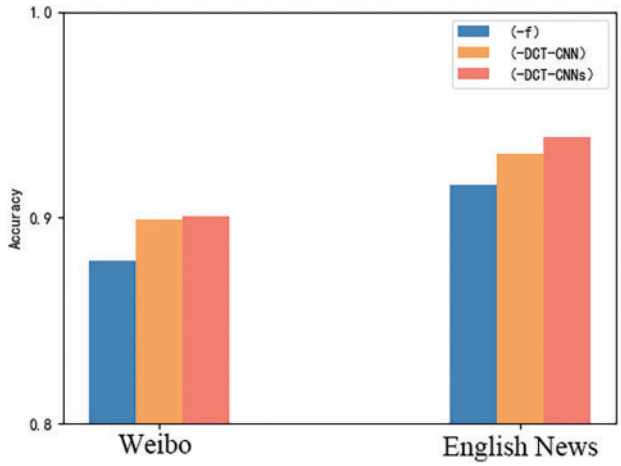


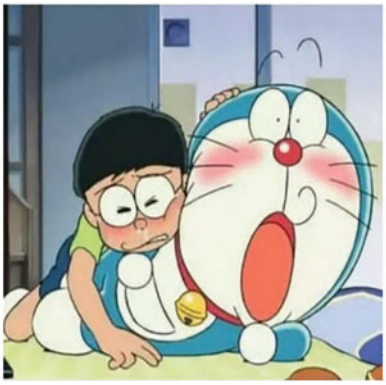
Figure 5: Histogram of frequency domain extraction experiments

From Fig. 5, it is evident that using frequency domain information with a fully connected layer directly results in poorer performance. This is because the coefficient information cannot be adequately extracted and utilized by the fully connected network. In contrast, DCT-CNN performs better. The continuous convolution and pooling in the CNN network help summarize the frequency domain periodic information. The DCT-CNNs method benefits from deeper convolutional layers and network branches, inspired by Inception and VGG networks. This allows it to extract features from different segments of the frequency domain, leading to the best performance.

4.7 Case Analysis

To further demonstrate the role of the semantic consistency calculation module and the frequency domain extraction network in MFFFND-Co, this paper analyzes several fake news cases that could only be detected after using these modules.

As shown in (a) in Fig. 6, this is a case that is successfully detected by the MFFFND-Co model, but missed by the w/o Semantic model. In this case, the text describes crime related content, while the captions are cartoon characters, showing a clear semantic inconsistency. (b) in Fig. 6 illustrates the news cases detected by MFFFND-Co, which failed to be detected by the w/o Frequency model. Although the semantic information of the news image is relatively normal, the image has been compressed many times and has a large number of watermarks, which obviously has frequency domain characteristics.



(a) Chinese prostitute arrested in Thailand, police statement shines!

(b) [Zhang Haidi has denied that she holds German citizenship] However, according to internal public security records, her household registration was transferred from Qingdao, Shandong, to Osaka, Japan.

Figure 6: News cases-semantic consistency and frequency domain

5 Conclusion

In this paper, we proposed a multimodal feature fusion fake news detection algorithm which combined with semantic consistency, called the MFFFND-Co algorithm. Based on the fake news detection algorithm that uses multi-head self-attention to fuse image frequency domain, spatial domain, and news text content, this algorithm improves the method of extracting frequency domain information. The improvement addresses issues such as insufficient feature extraction and fusion, as well as the underutilization of semantic relationships between modalities. Meanwhile, the Co-Attention mechanism is incorporated to enhance modality fusion, while a semantic consistency detection module is implemented to further boost the effectiveness of fake news detection. The experimental results show that the MFFFND-Co algorithm through multi-scale feature extraction, feature fusion and more fully semantic consistency detection module, improve the detection accuracy of fake news. However, although the semantic consistency detection module uses BERT's embedding vectors, which solve the cross-modal semantic measurement problem and utilize relationships between modalities, the anisotropy in the BERT vector space reduces the accuracy of cosine similarity calculations. To address this, improving the semantic consistency calculation method will further enhance fake news detection performance. Additionally, as fake news increasingly appears in various forms such as audio and video, we plan to extend the multi-scale feature extraction method and cross-modal ambiguity calculation from MFFFND-Co to support detection across more modalities.

Acknowledgement: The authors sincerely appreciate the anonymous reviewers for their insightful feedback and constructive suggestions.

Funding Statement: This work was supported by Communication University of China (HG23035) and partly supported by the Fundamental Research Funds for the Central Universities (CUC230A013).

Author Contributions: The authors' contributions to the paper are as follows: study conception and design: Jianxiang Cao, Jinyang Wu, Wenqian Shang, Tong Yi; data collection: Tong Yi, Chunhua Wang, Kang Song; analysis and interpretation of results: Tong Yi, Jiajun Cai, Haibin Zhu; draft manuscript preparation: Jianxiang Cao, Jinyang Wu, Tong Yi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The Twitter data used to support the findings of this study is available at the following website: https://github.com/MKLab-ITI/image-verification-corpus (accessed on 20 February 2025). The Weibo data used in this study can be accessed through the following link: https://docs.google.com/forms/d/e/1FAIpQLSdI0H90LyFs3ZZdH0hIP45FUl_kgsIqRfsfING1WJgYJmbeKw/viewform (accessed on 20 February 2025).

Ethics Approval: No applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Mitra T, Wright GP, Gilbert E. A parsimonious language model of social media credibility across disparate events. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing; 2017; Portland, OR, USA: ACM. p. 126–45. doi:10.1145/2998181.2998351.
2. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media. SIGKDD Explor Newsl. 2017;19(1):22–36. doi:10.1145/3137597.3137600.
3. Jin Z, Cao J, Zhang Y, Zhou J, Tian Q. Novel visual and statistical image features for microblogs news verification. IEEE Trans Multimed. 2017;19(3):598–608. doi:10.1109/TMM.2016.2617078.
4. Fang X, Wu H, Jing J, Meng Y, Yu B, Yu H, et al. NSEP: early fake news detection via news semantic environment perception. Inf Process Manag. 2024;61(2):103594. doi:10.1016/j.ipm.2023.103594.
5. Laato S, Islam AKMN, Islam MN, Whelan E. What drives unverified information sharing and cyberchondria during the COVID-19 pandemic?. Eur J Inf Syst. 2020;29(3):288–305. doi:10.1080/0960085x.2020.1770632.
6. Osmundsen M, Bor A, Vahlstrup PB, Bechmann A, Petersen MB. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. Am Polit Sci Rev. 2021;115(3):999–1015. doi:10.1017/S0003055421000290.
7. Farhangian F, Cruz RMO, Cavalcanti GDC. Fake news detection: taxonomy and comparative study. Inf Fusion. 2024;103(1):102140. doi:10.1016/j.inffus.2023.102140.
8. Lao A, Shi C, Yang Y. Rumor detection with field of linear and non-linear propagation. In: Proceedings of the Web Conference 2021; 2021; Ljubljana, Slovenia: ACM. p. 3178–87. doi:10.1145/3442381.3450016.
9. Jin Z, Cao J, Guo H, Zhang Y, Luo J. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia; 2017; Mountain View, CA, USA: ACM. p. 795–816. doi:10.1145/3123266.3123454.
10. Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh S. SpotFake: a multi-modal framework for fake news detection. In: IEEE Fifth International Conference on Multimedia Big Data (BigMM); 2019 Sep 11–13; Singapore, Singapore: IEEE; 2019. p. 39–47. doi:10.1109/bigmm.2019.00-44.
11. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, et al. EANN: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018; London, UK: ACM. p. 849–57. doi:10.1145/3219819.3219903.

12. Wu Y, Zhan P, Zhang Y, Wang L, Xu Z. Multimodal fusion with co-attention networks for fake news detection. In: Findings of the association for computational linguistics. Association for Computational Linguistics; 2021. p. 2560–9. doi:10.18653/v1/2021.findings-acl.226.
13. Chen Y, Li D, Zhang P, Sui J, Lv Q, Tun L, et al. Cross-modal ambiguity learning for multimodal fake news detection. In: Proceedings of the ACM Web Conference 2022; 2022. p. 2897–2905. doi:10.1145/3485447.3511968.
14. Khattar D, Goud JS, Gupta M, Varma V. MVAE: multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference; 2019; San Francisco, CA, USA: ACM. p. 2915–21. doi:10.1145/3308558.3313552.
15. Zhang H, Fang Q, Qian S, Xu C. Multi-modal knowledge-aware event memory network for social media rumor detection. In: Proceedings of the 27th ACM International Conference on Multimedia; 2019; Nice, France: ACM. p. 1942–51. doi:10.1145/3343031.3350850.
16. Duc Tuan NM, Quang Nhat Minh P. Multimodal fusion with BERT and attention mechanism for fake news detection. In: 2021 RIVF International Conference on Computing and Communication Technologies (RIVF); 2021 Aug 19–21; Hanoi, Vietnam: IEEE; 2021. p. 1–6. doi:10.1109/rivf51545.2021.9642125.
17. Chaudhari S, Mithal V, Polatkan G, Ramanath R. An attentive survey of attention models. ACM Trans Intell Syst Technol. 2021;12(5):1–32. doi:10.1145/3465055.
18. Qian S, Wang J, Hu J, Fang Q, Xu C. Hierarchical multi-modal contextual attention network for fake news detection. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2021; Canada: ACM. p. 153–62. doi:10.1145/3404835.3462871.
19. Zhang W, Gui L, He Y. Supervised contrastive learning for multimodal unreliable news detection in COVID-19 pandemic. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management; 2021; Queensland, Australia: ACM. p. 3637–41. doi:10.1145/3459637.3482196.
20. Fung Y, Thomas C, Reddy RG, Polisetty S, Ji H, Chang S-F, et al. Infosurgeon: cross-media fine-grained information consistency checking for fake news detection. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; 2021. p. 1683–98.
21. Lao A, Zhang Q, Shi C, Cao L, Yi K, Hu L, et al. Frequency spectrum is more effective for multimodal representation and fusion: a multimodal spectrum rumor detector. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024; Vancouver, BC, Canada. p. 18426–34. doi:10.1609/aaai.v38i16.29803.
22. Fenza G, Loia V, Stanzione C, Di Gisi M. Robustness of models addressing information disorder: a comprehensive review and benchmarking study. Neurocomputing. 2024;596(3):127951. doi:10.1016/j.neucom.2024.127951.
23. Biamby G, Luo G, Darrell T, Rohrbach A. Twitter-COMMs: detecting climate, COVID, and military multimodal misinformation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2022; Seattle, WA, USA. p. 1530–49.
24. Alam F, Cresci S, Chakraborty T, Silvestri F, Dimitrov D, Martino GDS, et al. A survey on multimodal disinformation detection; Gyeongju, Republic of Korea; 2022. p. 6625–43.
25. Shang L, Kou Z, Zhang Y, Wang D. A duo-generative approach to explainable multimodal COVID-19 misinformation detection. In: Proceedings of the ACM Web Conference 2022; 2022; Lyon, France: ACM. p. 3623–31. doi:10.1145/3485447.3512257.
26. Devlin J, Chang MW, Lee K, Toutanova K, Hulburd E, Liu D, et al. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018.
27. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
28. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014; Doha, Qatar. p. 1746–51.
29. Ying Q, Hu X, Zhou Y, Qian Z, Zeng D, Ge S. Bootstrapping multi-view representations for fake news detection. Proc AAAI Conf Artif Intell. 2023;37(4):5384–92. doi:10.1609/aaai.v37i4.25670.
30. Elman J. Finding structure in time. Cogn Sci. 1990;14(2):179–211. doi:10.1207/s15516709cog1402_1.
31. Yang Y, Zheng L, Zhang J, Cui Q, Li Z, Yu PS. TI-CNN: convolutional neural networks for fake news detection. arXiv:1806.00749. 2018.

32. Qi P, Cao J, Yang T, Guo J, Li J. Exploiting multi-domain visual information for fake news detection. In: IEEE International Conference on Data Mining (ICDM); 2019 Nov 8–11; Beijing, China: IEEE; 2019. p. 518–27. doi:10.1109/icdm.2019.00062.
33. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
34. Xu K, Qin M, Sun F, Wang Y, Chen YK, Ren F. Learning in the frequency domain. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 1740–49. doi:10.1109/cvpr42600.2020.00181.
35. Lu J, Batra D, Parikh D, Lee S. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems; 2019; Vancouver, BC, Canada.