MViR: Multi-View Visual-Semantic Representation for Fake News Detection

Anonymous ACL submission

Abstract

With the rise of online social networks, detecting fake news accurately is essential for a healthy online environment. While existing methods have advanced multimodal fake news detection, they often neglect the multi-view visual-semantic aspects of news, such as different text perspectives of the same image. To address this, we propose a Multi-View Visual-Semantic Representation (MViR) framework. Our approach includes a Multi-View Representation module using pyramid dilated convolution to capture multi-view visual-semantic features, a Multi-View Feature Fusion module to integrate these features with text, and multiple aggregators to extract multi-view semantic cues for detection. Experiments on benchmark datasets demonstrate the superiority of MViR. The codes will be released.

1 Introduction

Fake news refers to deliberately spreading false or misleading information with the aim of deceiving the public, creating confusion, manipulating public opinion, or achieving specific political, economic, or social objectives. Online social networks (OSNs) have increased the convenience of real-time information dissemination, but they also lead to the rapid and widespread dissemination of fake news, causing detrimental effects on the online environment (Aïmeur et al., 2023). Detecting fake news has thus become a current research hotspot.

Early works primarily focused on manually extracting features from text content (Choudhary and Arora, 2021), such as the proportion of negation words, writing style, and language styles. However, traditional methods are inefficient and unable to handle large amounts of data. Therefore, researchers began to focus on deep learning-based automatic fake news detection. Bhattarai et al. (Bhattarai et al., 2021) captured the lexical and semantic properties of news text. Jin et al. (Jin et al., 2016) detected fake news by leveraging significant disparities in image distributions.

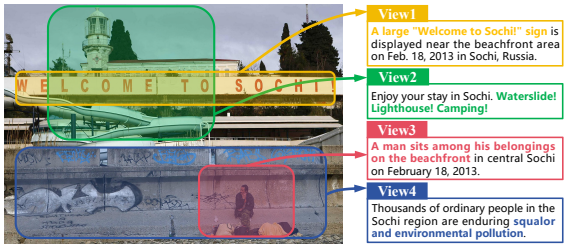


Figure 1: Motivation of our proposed MViR. We can see that different news texts describe the same image from various perspectives. For instance, some focus on the background building, others on the sign, and some on the person.

With the development of OSNs (Aïmeur et al., 2023), multimodal fake news (Zhou and Zafarani, 2020; Singh et al., 2021), which includes text, images, and videos, has emerged. These forms are often more attractive and have a broader reach than traditional unimodal fake news. EANN (Wang et al., 2018) introduces an event discriminator to detect fake news. MVAE (Khattar et al., 2019) incorporates a multimodal variational autoencoder for multimodal fake news detection. MCAN (Wu et al., 2021) designs a co-attention network to better fuse multimodal features.

However, existing multimodal fake news detection methods struggle to effectively capture the multi-view visual-semantic relationships present in news content. News images often contain information from various perspectives. For instance, as shown in Figure 1, the left side illustrates an image associated with fake news, while the right side presents the accompanying text. It is evident that the text reports multiple aspects of the image from different viewpoints, a common phenomenon in real-world news articles. This poses a challenge for previous detection methods (Tufchi et al., 2023), as they often fail to consider the multi-view semantics of the content for trustworthy fake news detection.

We propose a multi-view visual-semantic rep-

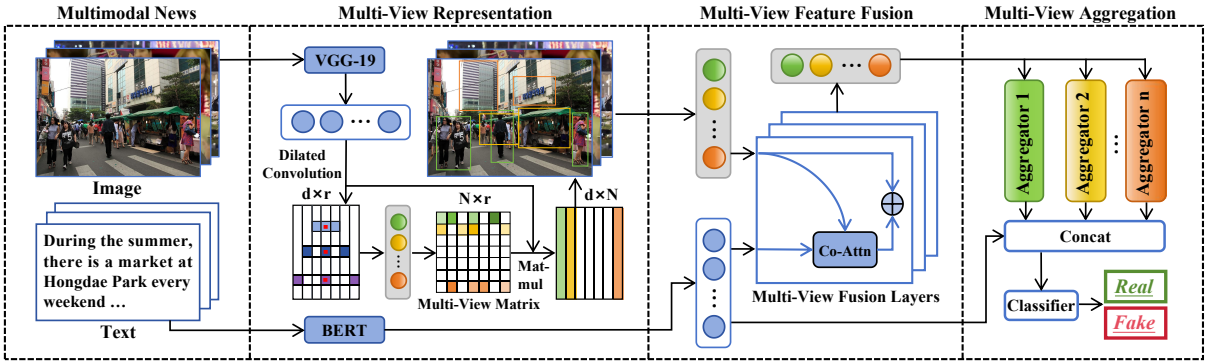


Figure 2: The MViR framework consists of three modules: Multi-View Representation (MVR), Multi-View Feature Fusion (MVFF), and Multi-View Aggregation (MVA). It extracts image and text features, learns multi-view visual-semantic representations via MVR, fuses features with MVFF, and uses MVA to generate embeddings and predict fake news probabilities.

resentation for fake news detection (MViR) to address the above issues. Specifically, we propose a multi-view representation module to extract multi-view fine-grained features from images, thereby providing the model with comprehensive multi-view visual-semantic information. Afterward, we use a multi-view feature fusion module to fuse image and text information, further enhancing the representation capability of multi-view features. Finally, we use a Multi-View Aggregation module to process the fused features and extract multi-view semantic cues to enhance fake news detection. Our main contributions include:

(1) We design a multi-view representation module, which can explicitly model the multi-view semantics in news images and capture the multi-view features within the images.

(2) We design a multi-view aggregation module, which can explicitly learn multi-view embeddings, extract multi-view semantic cues from news and utilize them to enhance the model's ability to identify fake news.

(3) Experiments conducted on the widely used datasets show that MViR significantly outperforms previous approaches.

2 Methodology

As shown in Figure 2, MViR consists of three main parts: Multi-View Representation (MVR), Multi-View Feature Fusion (MVFF), and Multi-View Aggregation (MVA).

2.1 Feature Extraction

For a image I from multimodal news, we utilize VGG-19 (Simonyan and Zisserman, 2014) to extract visual features. These features are subsequently projected into a d-dimensional space via a fully connected (FC) layer, yielding image features represented as $V = [v_1, v_2, \ldots, v_r] \in \mathbb{R}^{r \times d}$, where r denotes the number of extracted regions. Similarly, for a text T containing m words, word embeddings are extracted using a pre-trained BERT (Devlin, 2018), which are mapped to the d-dimensional space using a FC layer. Text features are expressed as $T = [t_1, t_2, \ldots, t_m] \in \mathbb{R}^{m \times d}$.

2.2 Multi-View Representation

We propose a Multi-View Representation (MVR) module to capture the multi-view semantics of images. Understanding an image from different perspectives means that each region requires different attention from different perspectives. To achieve this, we employ a pyramid dilated convolution (Yu, 2015; Qu et al., 2020) layer with K parallel kernels to aggregate multi-scale contextual information from the image, which serves as the basis for calculating view-specific importance scores.

$$s_i^k = Convd(V, w^k, d^k), \quad k = 1, 2, \ldots, K \tag{1}$$

where w^k and d^k denote its kernel size and dilation rate, s_i^k denotes the output of the k-th kernel. We then concatenate these outputs:

$$s_i = Concat(s_i^1, ..., s_i^K), \tag{2}$$

where $Concat(\cdot)$ denotes the concatenation of vectors. Afterward, a FC followed by a softmax activation is applied to compute the multi-view matrix $\hat{S} = [\hat{s}_1; ...; \hat{s}_r] \in \mathbb{R}^{r \times N}$, where N is the number of views, and $\hat{s}_i$ is the i-th row vector. The above process can be summarized as follows:

$$\hat{s}_{ij} = \frac{\exp\left((W_s s_i + b_s)_j\right)}{\sum_{j=1}^{r} \exp\left((W_s s_i + b_s)_j\right)}, \tag{3}$$

where $(W_s s_i + b_s) \in \mathbb{R}^{r \times N}$, $\hat{s}_i \in \mathbb{R}^N$ represents the importance scores of the i-th region over N views, $W_s \in \mathbb{R}^{N \times d}$ and $b_s \in \mathbb{R}^{1 \times N}$ are the learnable weights and bias, respectively. Finally, the image features can be summarized into a multi-view representation $\mathbf{V}^* \in \mathbb{R}^{N \times d}$ as follows:

$$\mathbf{V}^* = \hat{\mathbf{S}}^T \mathbf{V}. \tag{4}$$

Table 1: Results on two datasets. The best performance is in bold, while underlining highlights the follow-up.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1 score	Precision	Recall	F1 score
Weibo	EANN (Wang et al., 2018)	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	SAFE (Zhou et al., 2020)	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCAN (Wu et al., 2021)	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	CAFE (Chen et al., 2022)	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	FND-CLIP (Zhou et al., 2023)	0.907	0.914	0.901	0.907	0.917	0.901	0.908
	MSACA (Wang et al., 2024)	0.903	0.935	0.873	0.903	0.872	0.935	0.902
	EVENT-RADAR (Ma et al., 2024)	0.919	0.924	0.905	0.914	**0.932**	0.915	0.924
	MViR (Ours)	**0.924**	**0.944**	**0.906**	**0.920**	0.906	**0.941**	**0.928**
GossipCop	EANN (Wang et al., 2018)	0.864	0.702	0.518	0.594	0.887	0.956	0.920
	SAFE (Zhou et al., 2020)	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	SPOTFAKE (Singhal et al., 2020)	0.858	0.732	0.372	0.494	0.866	0.962	0.914
	CAFE (Chen et al., 2022)	0.867	0.732	0.409	0.587	0.887	0.957	0.921
	FND-CLIP (Zhou et al., 2023)	0.880	0.761	0.549	0.638	0.899	0.959	0.928
	MSACA (Wang et al., 2024)	0.887	**0.816**	0.538	0.648	0.897	**0.971**	0.933
	RaCMC (Yu et al., 2024)	0.879	0.745	0.563	0.641	0.902	0.954	0.927
	MViR (Ours)	**0.895**	0.784	**0.619**	**0.692**	**0.914**	0.963	**0.937**

2.3 Multi-View Feature Fusion

We propose a Multi-View Feature Fusion (MVFF) module that combines multi-view image features with the corresponding text features, further enhancing their representational capacity. MVFF consists of l multi-view fusion layers, each containing a co-attention mechanism (Lu et al., 2019) and a feed-forward network (FFN). Both components are enclosed by a residual connection and followed by layer normalization. The co-attention extends the standard multi-head attention by using queries (Q) from one modality and keys (K) and values (V) from another. In our approach, Q comes from the multi-view features of the image or fused features, while K and V come from the text.

$$Q_i = \mathbf{V}^* W_i^Q, \quad K_i = TW_i^K, \quad V_i = TW_i^V, \tag{5}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{1\times d_h}$ are the projection matrices for the i-th head, H denotes the number of heads, $d_h = d/H$ is the dimension of the output feature of each head. The calculation process of the co-attention can be presented as follows:

$$MultiHead(Q, K, V) = Concat(h_1, h_2, ..., h_H)W^O + V \tag{6}$$

where $W^O \in \mathbb{R}^{d\times d}$ is learnable weights, and $h_i = Att(Q_i, K_i, V_i)$. Att denotes the scaled-dot product attention, defined as follows:

$$Att(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^\top}{\sqrt{d_h}}\right) V_i. \tag{7}$$

To enhance the representational capacity of the fused features, the output of the co-attention is processed through an FFN. It is implemented as a two-layer multi-layer perceptron (MLP) with a ReLU activation function applied between the layers. The above process can be summarized as follows:

$$\mathbf{X} = FFN(MultiHead(Q, K, V)) \oplus Q, \tag{8}$$

where $\mathbf{X} \in \mathbb{R}^{N\times d}$ denotes the features after fusion, $\oplus$ represents the fusion operation, e.g., vector concatenation or elementwise add.

2.4 Multi-View Aggregation

The uniqueness of the Multi-View Aggregation (MVA) module lies in its use of multiple aggregators to generate a set of embeddings from the fused features, explicitly modeling multi-view features. This allows for the evaluation of news authenticity through multi-view semantic cues. Specifically, after obtaining a fused feature set $\{\mathbf{x}_n\}_{n=1}^N$, a series of feature aggregators $\{f_n^v\}_{n=1}^N$ are used to aggregate $\{\mathbf{x}_n\}_{n=1}^N$ into a set of semantic cues embeddings $\{\hat{\mathbf{x}}_n\}_{n=1}^N$:

$$\hat{\mathbf{x}}_n = f_n^v\left(\{\mathbf{x}_n\}_{n=1}^N\right), \tag{9}$$

where $\hat{\mathbf{x}}_n \in \mathbb{R}^d$. Each $\hat{\mathbf{x}}_n$ represents the semantic cues of a set of views. N is the number of views.

Next, we combine the semantic cues from each view with the text features and use a decision network to assess whether the news is true or fake. We employ only a single aggregator to obtain the text embedding:

$$z_n = \max\left(0, W_f \text{Concat}(\hat{\mathbf{x}}_n, f^t(T))\right), \tag{10}$$

$$\hat{y} = \max_{n=1}^{N}\left(\text{softmax}\left(z_n W_l\right)\right), \tag{11}$$

where $\hat{y}$ represents the probability of the news being fake, W_f represents the parameters of the fully connected layers, and W_l represents the parameters of the linear layer within the softmax function. For multiple semantic cues from different views of news, if any of these cues is detected as fake, the news is classified as fake.

2.5 Objective Function

We leverage cross-entropy to measure the classification loss and train our model:

$$\mathcal{L} = \sum_{\mathrm{i}=1}^{\mathcal{N}} -\left[y_{\mathrm{i}} * \log\left(\hat{y}_{\mathrm{i}}\right) + \left(1 - y_{\mathrm{i}}\right) * \log\left(1 - \hat{y}_{\mathrm{i}}\right)\right], \tag{12}$$

where $\mathcal{N}$ denotes the number of news reports, y_{i} represents the ground-truth label of the i-th news. Labels 0 and 1 refer to real news and fake news, respectively.

Table 2: Ablation study on two datasets.

	Method	Accuracy	F1 score	
			Fake News	Real News
Weibo	MViR (Ours)	**0.924**	**0.920**	**0.928**
	w/o MVR	0.901	0.893	0.909
	w/o MVFF	0.894	0.884	0.902
	w/o MVA	0.907	0.904	0.909
	Max Probability(Real)	0.912	0.908	0.915
	Average Probability	0.918	0.915	0.921
GossipCop	MViR (Ours)	**0.895**	**0.692**	**0.937**
	w/o MVR	0.883	0.648	0.930
	w/o MVFF	0.881	0.667	0.927
	w/o MVA	0.886	0.639	0.932
	Max Probability(Real)	0.874	0.655	0.915
	Average Probability	0.884	0.653	0.929

Table 3: Analysis for different numbers of MVFF layer.

	Layers	Accuracy	F1 score	
			Fake News	Real News
Weibo	2	0.912	0.908	0.915
	3	**0.924**	**0.920**	**0.928**
	4	0.917	0.915	0.921
GossipCop	2	0.884	0.648	0.930
	3	**0.895**	**0.692**	**0.937**
	4	0.891	0.674	0.935

3 Experiments

3.1 Datasets and Experimental Settings

We performed experiments using two real-world datasets: Weibo (Jin et al., 2017) and GossipCop (Shu et al., 2020). The Weibo dataset contains a total of 9,528 news, with 7,532 used for training and 1,996 used for testing. The GossipCop dataset contains a total of 12,840 news, with 10,010 used for training and 2,830 used for testing. For fairness in comparison, we adhered to the data-splitting protocol and processing steps used in prior works (Wu et al., 2021; Ying et al., 2023).

Experiments were carried out on an NVIDIA Tesla A100 GPU. For the Weibo and GossipCop datasets, we adopted bert-base-chinese and bert-base-uncased respectively to extract text features. More details of the implementation can be found in the appendix A.

3.2 Performance Comparison

Table 1 shows the performance comparison between MViR and the baseline methods. MViR demonstrated excellent performance across all metrics, including accuracy, precision, recall, and F1 score. MViR achieved an average accuracy of 92.4% on the Weibo dataset and 89.5% on the GossipCop dataset, outperforming the best existing models by 1.9 and 1.7 percentage points.

While many methods, such as MCAN, detect fake news by fusing multimodal features, they do not consider the multi-view characterization of fake news. In contrast, MViR effectively captures the multi-view features of images, thus improving the performance of multimodal fake news detection.

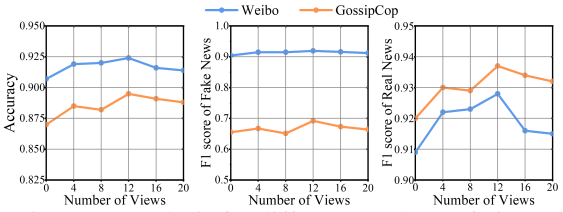


Figure 3: Analysis for different numbers of views.

3.3 Ablation Study

To evaluate the contribution of each component in MViR, we removed the MVR, MVFF, and MVA modules individually. Table 2 shows that removing any module leads to a significant performance drop, confirming that the components complement each other to improve fake news detection. Additionally, we compared decision networks using maximum real news probability, average probability, and maximum fake news probability. The results indicate that the network using maximum fake news probability performs best, as it can capture more reliable key features of fake news.

3.4 Parameter Sensitivity Analysis

Impact of the Number of Feature Fusion Layers: As shown in Table 3. The experimental results demonstrate that appropriately increasing the number of layers can promote interactions between different modality features, thereby enhancing model performance. However, when the number of layers becomes larger, further increases result in performance degradation, likely due to excessive model complexity leading to overfitting.

Impact of the Number of views: We also analyzed the impact of the number of viewpoints on model performance, as shown in Figure 3. It can be observed that appropriately increasing the number of viewpoints helps capture richer details and more diverse features, thereby improving performance. However, when the number of viewpoints increases further, the model's performance may decline due to the introduction of redundant information, which could affect its generalization ability. In our experiments, MViR performed best with 12 viewpoints.

4 Conclusion

In this work, we propose MViR, a novel framework for fake news detection using multi-view visual-semantic representations. Our approach includes a multi-view representation module to extract visual-semantic features from images, a feature fusion module to combine image and text features, and a Multi-View Aggregation module to learn multi-view embeddings. Experiments on two benchmark datasets show that MViR outperforms existing state-of-the-art methods.

Limitations

In this work, we propose MViR for detecting multimodal fake news. Although our approach demonstrates excellent performance, it still has the following limitations:

- Although MVIR performs best on most metrics, it underperforms EVENT-RADAR and MSACA on a few individual metrics (e.g., real news precision on the Weibo dataset). This may stem from dataset-specific biases or the model's sensitivity to certain semantic cues, which warrants further investigation.
- Experiments were conducted on the Weibo and GossipCop datasets, which are relatively small and domain-specific. The model's generalization capability on larger, cross-domain datasets (e.g., Twitter, news articles) requires further validation.
- The multimodal processing pipeline introduces additional complexity, but the computational overhead (e.g., training time, memory usage) has not been quantified. Efficiency analysis is critical for practical deployment, especially on social platforms requiring real-time detection. In future work, we will further evaluate MViR's efficiency and conduct more comprehensive experiments on additional datasets.

Ethical Statement

Our proposed multimodal detection method aims to reduce false news on the internet and maintain a healthy online environment. This paper adheres to the ACM Code of Ethics and Professional Conduct. All experimental data used in our research originates from publicly accessible resources, and before utilization, we confirm that it complies with relevant usage regulations and does not contain sensitive private information. Additionally, appropriate citations are given to the sources of related papers and pre-trained models utilized in our work. Lastly, our code will be released under the license applicable to any artifacts used.

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Liesbeth Allein, Marie-Francine Moens, and Domenico Perrotta. 2021. Like article, like audience: Enforcing multimodal correlations for disinformation detection. *arXiv preprint arXiv:2108.13892*.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2021. Explainable tsetlin machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114*.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *WWW*.

Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACM MM*.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE TMM*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *WWW*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-aware multi-view summarization network for image-text matching. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1047–1055.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Vivek K Singh, Isha Ghosh, and Darshan Sonagara. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 72(1):3–17.

Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *AAAI*.

Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, 12(2):28.

Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024. Fake news detection via multiscale semantic alignment and cross-modal attention. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2406–2410.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *KDD*.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with coattention networks for fake news detection. In *ACL-IJCNLP*.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *AAAI*.

F Yu. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Xinquan Yu, Ziqi Sheng, Wei Lu, Xiangyang Luo, and Jiantao Zhou. 2024. Racmc: Residual-aware compensation network with multi-granularity constraints for fake news detection. *arXiv preprint arXiv:2412.18254*.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *ICME*.

A Implementation

We implemented MViR using PyTorch 2.3.1 and conducted all experiments on a single NVIDIA Tesla A100 GPU. For text feature extraction, we used bert-base-chinese for the Weibo dataset with a maximum sequence length of 160, and bert-base-uncased for the GossipCop dataset with a maximum sequence length of 394. Images were resized to 224×224 to match the input dimensions of the pre-trained VGG-19 model. The dimensions of image and text features d were set to 256, with the number of heads set to 4 and a dropout rate of 0.5. We trained the model using AdaBelief for 50 epochs with a batch size of 32 and an initial learning rate of 1e-4. Additional implementation details can be found in the code.

The specific Pyramid Dilated Convolutional Layers are shown in Table 4, where k denotes the kernel number, w^k represents the kernel size, d^k represents the dilation rate, and s^k represents output channel of k-th convolution kernel, respectively. As the dilation rate increases, the receptive field of the kernel is enlarged without reducing the regional resolution.

We also provide a statistical overview of the detailed parameters for the two datasets (Weibo and GossipCop) used in our study, as shown in Table 5.

Table 4: Configurations of pyramid dilated convolution

k	1	2	3	4	5	6	7
w^k	1	3	3	3	5	5	5
d^k	1	1	2	3	1	2	3
s^k	256	128	128	128	128	128	128

Table 5: Statistics of datasets.

	Weibo	GossipCop
Total news	9528	12840
Images	13272	15488
Fake news	3783(Train)	2036(Train)
	1000(Test)	545(Test)
Real news	3749(Train)	7974(Train)
	996(Test)	2285(Test)

B Baselines

To evaluate the performance of MViR, we compared it with several baseline approaches in our experiments, all of which are classic schemes in the field of fake news detection. We provide a brief introduction to each of them:

EANN (Wang et al., 2018) enhances the detection capability of fake news in new events by building an end-to-end framework that includes multi-modal feature extraction, fake news detection, and event discrimination to learn event-invariant feature representations.

SAFE (Zhou et al., 2020) effectively identifies fake news by jointly learning the textual and visual features of news articles and their cross-modal relationships, utilizing a similarity-aware approach to detect content mismatches.

MCAN (Wu et al., 2021) better fuses textual and visual features for fake news detection by learning the inter-dependencies among multimodal features.

CAFE (Chen et al., 2022) is a method for detecting fake news that improves accuracy by assessing ambiguities between different media types and capturing how they relate to each other.

FND-CLIP (Zhou et al., 2023) is a framework that uses CLIP technology to combine text and image information for better fake news detection.

MSACA (Wang et al., 2024) is a network that improves fake news detection by aligning text and images at multiple scales and using attention to select the best features.

EVENT-RADAR (Ma et al., 2024) is a framework that detects fake news by analyzing multi-modal information in events and calculating the credibility of each view.

SPOTFAKE (Singhal et al., 2020) is a method that uses transfer learning to combine text and image information for more accurate fake news detection.

RaCMC (Allein et al., 2021) enhances the differences between real and fake news by utilizing multi-scale feature interaction and fusion, along with multi-granularity constraints, to improve the accuracy of fake news detection.

C Case Study

To further illustrate the importance of multi-perspective analysis for fake news detection, we compared the detection results of MViR with baseline approache MVAE and showcased some fake news instances (translated into English) that were correctly identified by MViR but overlooked by MVAE, as shown in the Figure 4. A common feature of these fake news items is their rich image perspective information; previous works, which ig-

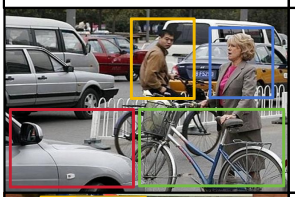
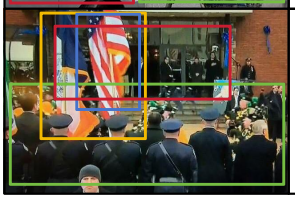
Figure	Post
	In Beijing, a car entered a non-motorized zone and was noticed by a foreigner who politely asked the driver to leave. The driver refused, honked loudly, and threatened to damage her bicycle. No bystanders intervened to condemn the driver's actions.
	A Chinese-American firefighter lost his life while performing a firefighting mission. To honor his bravery and selfless dedication, the U.S. government held a state funeral, lowered flags nationwide, and provided significant compensation to his family.

Figure 4: Case study.

nored the multi-perspective characteristics of fake news, resulted in inaccurate identification. In contrast, MViR successfully recognized these fake news instances by capturing multi-perspective features from both images and text. These examples demonstrate that a single image representation struggles to comprehensively describe image information, leading to misclassification of certain fake news items that contain multi-perspective details.