Intra and Inter-modality Incongruity Modeling and Adversarial Contrastive Learning for Multimodal Fake News Detection

Siqi Wei
Beijing University of Posts and Telecommunications
Beijing, China
Weisiqi@bupt.edu.cn

Bin Wu*
Beijing University of Posts and Telecommunications
Beijing, China
Wubin@bupt.edu.cn

ABSTRACT

Multimodal fake news detection (FND) is significant in safeguarding network security and societal safety. Most existing studies only focus on common semantic features between different modalities and utilize simple cross-entropy loss for model training. However, these studies overlook the incongruent semantic features in multimodal news data, which can arise within or between modalities. Moreover, the utilization of simple cross-entropy loss may not provide the model with robustness against well-designed forged fake news. To address the above issues, we propose a novel approach named Signed Attention-based Graph Transformer with Adversarial Contrastive Learning (SAGT-ACL) for the detection of multimodal fake news. SAGT-ACL models fine-grained semantic associations in multimodal news articles by constructing a fully connected multimodal graph and reframes the fake news classification task as a graph classification problem. Additionally, SAGT-ACL incorporates a signed attention-based graph transformer module to identify both common and incongruent semantics within and across modalities. Finally, SAGT-ACL proposes an adversarial data augmentation mechanism to simulate malicious forgeries by fake news creators and designs an auxiliary adversarial contrastive learning task to help the model learn more discriminative news representations from the adversarial samples for robust and effective detection. Extensive experiments demonstrate that SAGT-ACL outperforms existing methods, with detection accuracy improvements of 4.95%, 6.01%, and 5.68% on Weibo, Twitter, and Gossipcop datasets, respectively.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Multimedia information systems**; *Web mining*.

KEYWORDS

Multimodal Fake News Detection; Intra and Inter-modality Incongruity Modeling; Adversarial Data argumentation; Contrastive Learning

ACM Reference Format:
Siqi Wei and Bin Wu. 2024. Intra and Inter-modality Incongruity Modeling and Adversarial Contrastive Learning for Multimodal Fake News Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652583.3658118

1 INTRODUCTION

The swift advancement of multimodal social media enables news to transition from its conventional single-text form to a multimodal form that incorporates both images and text [32]. Multimodal news possesses a heightened visual impact, resulting in a broad-reaching dissemination effect [1]. Nevertheless, the absence of effective network management results in the widespread propagation of fake news, posing a significant threat to both information security and societal stability [15]. Therefore, how to automatically and effectively detect multimodal fake news is an important problem to be solved urgently.

The key to multimodal fake news detection lies in the extraction of distinctive features from various modalities. The advancement of deep learning has led to the emergence of numerous fake news detection approaches that extract multimodal features automatically [5]. These approaches can be broadly classified into three categories. Firstly, **joint representation methods** [8, 17, 18, 23] concatenate features from image and text modalities to create representations of news, yet they overlook the semantic interactions between different modalities. Secondly, **similarity representation methods** [28, 31, 33] assess the similarities between different modalities to model their information interactions at a coarse-grained level, but they struggle to capture semantic interactions fine-grainedly. Moreover, they tend to focus on common semantics while neglecting inconsistent ones. To model the fine-grained semantic interactions between modalities, **alignment representation methods** [6, 13, 19, 22, 24, 30] align the common semantics between modalities through a well-designed attention mechanism at a fine-grained level. However, traditional attention mechanisms based on the softmax function tend to assign positive attention weights, preserving only consistent features between modalities and overlooking irrelevant, inconsistent, or even conflicting features commonly present in fake news.

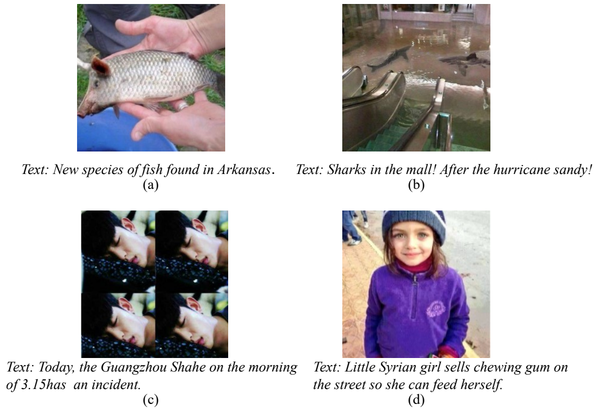
While the approaches above can enhance the efficacy of multimodal fake news detection (FND), they face the following challenges:

(1) *Neglect of intra and inter-modality incongruity features.* Prior studies primarily concentrate on common semantics across various modalities, neglecting intra and inter-modality incongruent semantic features. Incongruent semantics encompass irrelevant, inconsistent, or conflicting semantic features that serve as crucial

[*] Corresponding author.

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06…$15.00
https://doi.org/10.1145/3652583.3658118



Text: New species of fish found in Arkansas.
(a)

Text: Sharks in the mall! After the hurricane sandy!
(b)

Text: Today, the Guangzhou Shahe on the morning of 3.15has an incident.
(c)

Text: Little Syrian girl sells chewing gum on the street so she can feed herself.
(d)

Figure 1: Some examples of multimodal fake news. (a) The image contains intra-modality incongruity features. (b) The image and text both contain intra-modality incongruity features. (c) The image and the text are irrelevant. (d) The image and the text are conflicting.

clues for identifying misinformation. Fake news is often generated by tampering or forging by the creators. Rough forgery techniques can easily cause semantic incongruity. To illustrate, several examples in Figure 1 demonstrate the prevalence of semantic incongruity in multimodal fake news: (a) the body of a fish and the head of a pig are semantically contradictory; (b) a shark and a shopping mall are semantically conflicting; (c) the event described in the text is unrelated to the Korean star in the image; and (d) the Syrian war depicted in the text conflicts with the smiling girl in the image. Consequently, effective mining of intra-modality and inter-modality incongruent semantic features is the major challenge to multimodal fake news detection.

(2) *Lack of robustness to adversarial samples.* With the improvement of fake news forgery techniques [12], some creators use well-designed forgery methods to avoid detection by the model. The simple classification task makes the model lack robustness in the face of the elaborately forged samples. In some cases, just introducing a simple adversarial perturbation may lead to misclassification of labels, which presents a significant risk to the fake news detection system [3]. Consequently, enhancing the robustness of FND models against adversarial samples is another critical challenge that needs to be addressed.

To address the above challenges, we propose a novel approach for detecting multimodal fake news, termed Signed Attention-based Graph Transformer with Adversarial Contrastive Learning (SAGT-ACL), as shown in Figure 2. Specifically, we first employ pre-trained unimodal encoders to extract fine-grained image and text features and map them into a unified embedding space. Subsequently, we construct a fully connected multimodal graph for each news item using the representations in the unified space and transform the fake news detection task into a graph classification task. Based on the multimodal graph, we propose a novel signed attention-based graph transformer (SAGT) to explore the consistent and incongruent features within and between different modalities. Unlike traditional attention mechanisms, SAGT separates the sign and weight of the attention, allowing for the adaptive preservation of both positive and negative associations for each node pair. To enhance the robustness of the model, we utilize an adversarial training mechanism to simulate the well-designed forgery of fake news creators. Moreover, we propose an auxiliary adversarial contrastive learning task to pull the distance between the original and adversarial samples of the same class in the hidden space and to push the distance between the samples of different classes,

The main contributions of this paper are summarized as follows:

- To capture complex, fine-grained semantic associations in multimodal news, we construct a fully connected graph for each news item and design a signed attention-based graph transformer module (SAGT) for learning this graph. The SAGT module separates the sign and the weight of attention to preserve consistent and incongruent semantic features within and across modalities.
- To identify carefully forged fake news, we introduce an adversarial contrastive learning (ACL) auxiliary task. The ACL task employs adversarial data augmentation to generate challenging samples and supervised contrastive learning to help the model learn more discriminative features from these challenging samples, leading to robust and effective detection.
- We perform comprehensive experiments using three representative datasets to validate the effectiveness of our proposed SAGT-ACL model in detecting multimodal fake news.

2 RELATED WORK

Multimodal fake news detection (FND) seeks to detect the truthfulness of news based on its multimodal content features. The existing multimodal FND approaches can be categorized into three categories: joint representation methods, similarity representation methods, and alignment representation methods.

The joint representation methods utilize unimodal encoders to learn visual and textual representations, respectively, followed by simple vector concatenation to obtain the multimodal representation of the news. Based on this, Wang et al. [23] add an event classification auxiliary task to help the model better understand multimodal content. Similarly, Khattar et al. [8] add a news reconstruction auxiliary task. Benefiting from the strong representation capability of the pre-trained models [10], Singhal et al. [17, 18] utilize the pre-trained language model BERT [7] or XLNet [29] and the visual model VGG19 [16] to extract text and image features, respectively, and then concatenate them to obtain the multimodal news representation for classification. However, simple concatenation fails to capture the complex semantic interactions between different modalities, leading to limitations.

Similarity representation methods coarse-grainedly model semantic interactions between different modalities by comparing the similarity between them. Zhou et al. [31] transform the image into caption text by utilizing an image caption model and measure the Similarity between the news text and the caption text. Xue et al. [28] map the image and the text into the same semantic space and calculate their cosine similarity. Zhou et al. [33] use CLIP scores to measure the similarity between modalities. Despite achieving some results, similarity representation methods can only capture information interactions at the coarse-grained level and cannot describe semantic interactions at the fine-grained level.

To address this issue, alignment representation methods model fine-grained semantic interactions by aligning the consistent details of the image and text through well-designed attention mechanisms. Jin et al. [6] use an attention-based RNN to highlight the text token associated with the image to enhance the model's understanding of news. Inspired by the transformer model's self-attention mechanism [21], some studies utilize modified co-attention mechanisms [13, 19, 24, 30] to align and fuse the image and the text features. Nevertheless, the aforementioned approaches fail to consider the inconsistency between different modalities. Blindly aligning mismatched images and text can lead to the introduction of unpredictable noise. Therefore, Wang et al. [22] propose a masked attention mechanism to mask out irrelevant information between modalities and only align modality-common information, thus avoiding the introduction of noise. However, they ignore modality-specific information, leading to performance loss.

Although the above methods achieve good results, they only focus on the common features while ignoring the intra and inter-modality incongruity features. In addition, they lack robustness in the face of elaborate fake news.

3 PROBLEM FORMULATION

The multimodal fake news detection task can be considered as a binary classification problem. Specifically, we let $\mathcal{D} = \left\{n_1, n_2, \ldots, n_{|\mathcal{D}|}\right\}$ denote the set of news, where $|\mathcal{D}|$ denotes the total number of news. Each news item in the dataset can be represented as $n_i = \{(T_i, V_i), y_i\}$, where T_i is the text, V_i is the image, and $y_i \in \{0, 1\}$ is the ground-truth label of the news (i.e., real or fake). Formally, the objective of fake news detection is to learn a projection $F(T, V) \rightarrow \{0, 1\}$.

4 METHODOLOGY

4.1 Feature Extraction

Given a piece of multimodal news, the feature extraction module aims to extract the features from text and image modality and map them into a unified embedding space.

4.1.1 Text Feature Extraction. To effectually capture word semantics and linguistic contexts, we utilize the Bidirectional Encoder Representations from Transformers (BERT) [7] to extract textual features of news. Specifically, the text T is tokenized into a sequence list of words $T = \{t_1, t_2, \ldots, t_k\}$, where k denotes the number of words in the text. Then, we input the word sequence into a pre-trained BERT model to get the transformed embeddings as:

$$\mathbf{E}^T = \left\{\mathbf{e}_1^T, \mathbf{e}_2^T, \ldots, \mathbf{e}_k^T\right\} = \mathrm{BERT}(T), \tag{1}$$

where $\mathbf{e}_i^T \in \mathbb{R}^{d_t}$ is the output word embedding in BERT and d_t is the dimension of the word embedding.

4.1.2 Visual Feature Extraction. Benefiting from the effectiveness of the Vision Transformer (ViT) [4] in visual understanding tasks, we use ViT to extract visual features. Specifically, we reshape the image $V \in \mathbb{R}^{C \times H \times W}$ into a sequence of flattened 2D patches $\hat{V} \in \mathbb{R}^{n \times (P^2 \cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $n = HW/P^2$ is the resulting number of patches. Then, we input the patch sequences into a pre-trained ViT to get patch embeddings as:

$$\mathbf{E}^V = \left\{\mathbf{e}_1^V, \mathbf{e}_2^V, \ldots, \mathbf{e}_l^V\right\} = \mathrm{ViT}(V), \tag{2}$$

where $\mathbf{e}_i^V \in \mathbb{R}^{d_v}$ is the output patch embedding in ViT and d_v is the dimension of the patch embedding.

After capturing visual and textual features, we utilize a single feed-forward layer followed by a non-linearity exponential linear unit (ELU) [20] to map them into a uniform embedding space as:

$$\begin{cases} \mathbf{m}_j^T = \mathrm{ELU}\left(\mathbf{W}_m \mathbf{e}_i^T + \mathbf{b}_k\right), \forall \mathbf{e}_i^T \subset \left\{\mathbf{e}_1^T, \mathbf{e}_2^T, \ldots, \mathbf{e}_k^T\right\}, \\ \mathbf{m}_i^V = \mathrm{ELU}\left(\mathbf{W}_m \mathbf{e}_j^V + \mathbf{b}_l\right), \forall \mathbf{e}_j^V \subset \left\{\mathbf{e}_1^V, \mathbf{e}_2^V, \ldots, \mathbf{e}_l^V\right\}, \end{cases} \tag{3}$$

where $\mathbf{W}_m \in \mathbb{R}^{d \times d_t}$ and $\mathbf{b}_m \in \mathbb{R}^d$ are learnable parameters of the multimodal projection layer. $\mathbf{m}_i^T$ and $\mathbf{m}_j^V$ are representations of each token and patch in the unified multimodal space.

4.2 Multimodal Graph Construction

Benefiting from the fact that the graph structure can be used to mine semantic associations between nodes [25], we construct a multimodal graph $\mathcal{G}_M = (\mathcal{V}_M, \mathcal{E}_M)$ to simultaneously model the intra and inter-modality semantic relationships. The node set $\mathcal{V}_M$ consists of all image patches and text tokens and we let the representations within the unified multimodal space as the initialized embeddings of the nodes $\mathbf{H}^{\mathrm{in}} = \left\{\mathbf{m}_1^V, \ldots, \mathbf{m}_l^V, \mathbf{m}_1^T, \ldots, \mathbf{m}_k^T\right\}$. The edge set $\mathcal{E}_M = \{\mathcal{E}_T, \mathcal{E}_V, \mathcal{E}_I\}$ consists of three heterogeneous types of edges: image intra-modality edges $\mathcal{E}_V$, text intra-modality edges $\mathcal{E}_T$, and inter-modality edges $\mathcal{E}_I$.

For text intra-modality edges, we assume that there exists a relationship between each text token. Therefore, the text nodes are fully connected with unweighted and bi-directional edges. Similarly, any two image nodes are connected. For inter-modality edges, we connect each text node to every image node and vice-versa. These intra and inter-modality connecting edges help the model learn from dependencies arising from both within and across modalities concurrently at a more granular level.

4.3 Signed Attention-based Graph Transformer

In order to model the positive and negative relationships between nodes in the multimodal graph, we propose a signed attention-based graph transformer, namely SAGT, as shown in Figure 2 (b). Similar to the original transformer, SAGT includes multi-head signed attention blocks (MH-SA), feed-forward blocks (FFN), and layer normalization blocks (LN).

4.3.1 Signed Multi-head Attention Mechanism. The conventional self-attention mechanism, constrained by the property of the softmax function, can only capture consistent (positive) associations between nodes but ignores incongruent (negative) associations between nodes. Fake news frequently includes irrelevant, inconsistent, or even conflicting features within or between different modalities, making it challenging for self-attention mechanisms to identify and process. To fully encompass the consistent and incongruent

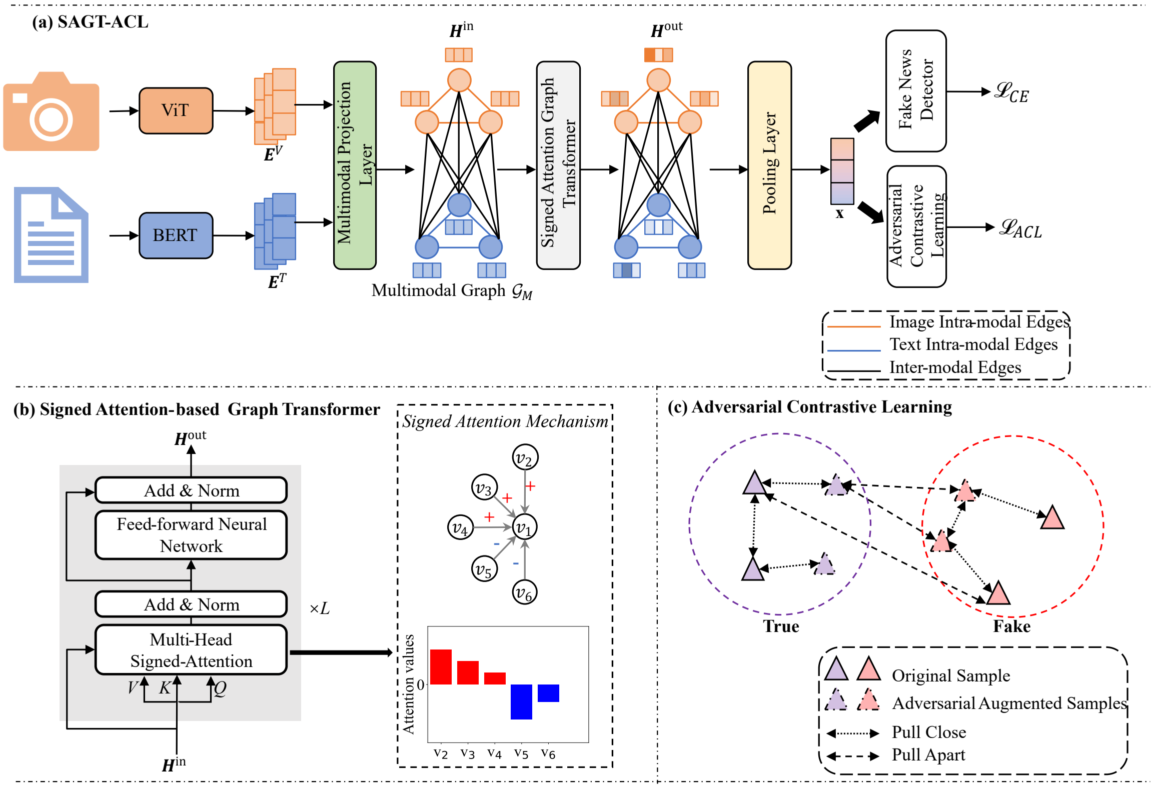

(a) SAGT-ACL
ViT
BERT
E^V
E^T
Multimodal Projection Layer
H^{in}
H^{out}
Multimodal Graph $\mathcal{G}_M$
Signed Attention Graph Transformer
Pooling Layer
x
Fake News Detector
Adversarial Contrastive Learning
$\mathcal{L}_{CE}$
$\mathcal{L}_{ACL}$
Image Intra-modal Edges
Text Intra-modal Edges
Inter-modal Edges
(b) Signed Attention-based Graph Transformer
H^{out}
Add & Norm
Feed-forward Neural Network
Add & Norm
Multi-Head Signed-Attention
V K Q
H^{in}
×L
Signed Attention Mechanism
v_1 v_2 v_3 v_4 v_5 v_6
+ + + - -
Attention values
0
v2 v3 v4 v5 v6
(c) Adversarial Contrastive Learning
True
Fake
Original Sample
Adversarial Augmented Samples
Pull Close
Pull Apart

Figure 2: (a) The overall architecture of SAGT-ACL. It consists of a feature extraction module, multimodal graph construction module, signed attention-based graph transformer module, fake news detector module, and adversarial contrastive learning module. (b) The signed attention-based graph transformer module explores intra and inter-modality consistent and incongruent semantic features in news data. (c) The adversarial contrastive learning module pulls the distance between samples with the same label close and pushes apart the distance between samples with different labels.

semantic features in multimodal news, a viable approach is to produce signed attention values based on the input node pairs. Positive values can retain consistent relationships, while negative values can capture incongruent relationships. Inspired by this, we design a new multi-head signed attention mechanism (MH-SA) based on the original multi-head self-attention calculation.

Specifically, given the input feature matrix $\mathbf{H}^{\text{in}} \in \mathbb{R}^{(k+l)\times d}$, we first define a set of weight matrices $\mathbf{W}_i^K, \mathbf{W}_i^Q, \mathbf{W}_i^V \in \mathbb{R}^{d\times d_h}$ to obtain query, key, and value vectors as:

$$\mathbf{Q}_i = \mathbf{H}^{\text{in}}\mathbf{W}_i^Q, \mathbf{K}_i = \mathbf{H}^{\text{in}}\mathbf{W}_i^K, \mathbf{V}_i = \mathbf{H}^{\text{in}}\mathbf{W}_i^V, \tag{4}$$

Then we calculate the signed attention as:

$$\text{Signed-att}\left(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\right) = \text{sgn}\left(\mathbf{Q}_i\mathbf{K}_i^\top\right) \cdot \text{softmax}\left(\frac{\left|\mathbf{Q}_i\mathbf{K}_i^\top\right|}{\sqrt{d_h}}\right)\mathbf{V}_i, \tag{5}$$

where $\text{sgn}(\cdot)$ denotes the sign function, $|\cdot|$ denotes the absolute value, i denotes the i–th attention head, d_h denotes the output dimension of single-headed signed attention. The principle of Equation 5 is to decompose the sign and weight of attention. The sign information indicates the semantic correlation polarity between nodes. Specifically, when two nodes are semantically incongruent, $\text{sgn}(\cdot)$ will produce a negative signal, capturing incongruity features. When two nodes are semantically correlated, $\text{sgn}(\cdot)$ will produce a positive signal, capturing the common features. The absolute value of the dot product result $\left|\mathbf{Q}_i\mathbf{K}_i^\top\right|$ is taken and subsequently normalized, ensuring that the softmax function places greater emphasis on attention weights.

Like the original transformer, we also boost the model's representational power by extending the signed attention to multi-head signed attention (MH-SA). This is denoted as:

$$\begin{aligned} \text{MH-SA}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) &= \left(head_1 \parallel head_2 \parallel \cdots \parallel head_h\right)\mathbf{W}_o \\ head_i &= \text{Signed-att}\left(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i\right), i \in \{1, 2, \cdots, h\} \end{aligned} \tag{6}$$

where $(\cdot \parallel \cdot)$ denotes vector concatenation operation, $\mathbf{W}_o \in \mathbb{R}^{h\cdot d_h \times d}$ represents the learnable parameter matrix, h denotes the total number of attention heads and $d_h = d/h$.

4.3.2 Graph Transformer. Similar to the original transformer model [21], we also employ feed-forward layers (FFN) with residual connections and layer normalization techniques (LN) to prevent the loss of initial information. These procedures are represented as:

$$\begin{aligned}\mathbf{H}'^{(L)} &= \mathrm{LN}\left(\text{MH-SA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)+\mathbf{H}^{L-1}\right)\\ \mathbf{H}^{L} &= \mathrm{LN}\left(\mathrm{FFN}\left(\mathbf{H}'^{(L)}\right)+\mathbf{H}'^{(L)}\right)\end{aligned} \tag{7}$$

where the superscript L denotes the number of layers of the signed attention-based graph transformer, $\mathbf{H}^L$ denotes the node embedding of the $L-$th layer output.

4.4 Fake News Detector

After $L-$layer SAGT, we can acquire the final node representations $\mathbf{H}^{\text{out}} \in \mathbb{R}^{(k+l)\times d}$. Then, we utilize a graph pooling layer to obtain a global news representation $\mathbf{x}$.

$$\mathbf{x} = \mathrm{Pool}\left(\left\{\mathbf{h}_1^{\text{out}}, \mathbf{h}_2^{\text{out}}, \ldots, \mathbf{h}_{k+l}^{\text{out}}\right\}\right) \tag{8}$$

To classify the authenticity of the news, we input the news representation $\mathbf{x}$ into a fully connected network with a softmax activation function to get the predicted result as:

$$\hat{y}_i = \mathrm{Softmax}\left(\mathbf{x}_i \mathbf{W}_{cls} + \mathbf{b}_{cls}\right) \tag{9}$$

where $\mathbf{W}_{cls}$ and $\mathbf{b}_{cls}$ are the learnable parameters. For the fake news detection task, we use the standard cross-entropy loss function as below:

$$\mathscr{L}_{CE} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|} y_i \cdot \log \hat{y}_i \tag{10}$$

where $|\mathcal{D}|$ is the number of news, y_i and $\hat{y}$ respectively denote the ground-truth label and the predicted probability of the i-th news from the classifier.

4.5 Adversarial Contrastive Learning

Since fake news creators often fool detection models through well-designed forgery techniques, a simple classifier based on cross-entropy loss is unable to detect carefully faked fake news. For this reason, we propose a supervised adversarial contrastive learning auxiliary task to achieve effective and robust detection.

4.5.1 *Adversarial Data Augmentation.*

To avoid model detection, fake news publishers often utilize camouflage strategies to make the fake news closer to real news instances, thus confusing the detection model. The ultimate goal is to make the representations of fake news closer to that of real news in the hidden space. Therefore, we propose an adversarial data augmentation approach to model this behavior in the hidden space and generate adversarial samples to help the model achieve robust and efficient detection. Ideally, the adversarial data augmentation perturbs the original news representation to maximize the cross-entropy loss, thus confusing the detection model.

Specifically, following the classical adversarial training methods [26, 27], we utilize the Fast Gradient Value method (FGM) [11] to estimate the gradient adversarial perturbation as a noise vector for each news representation:

$$\mathbf{x}_i^{adv} = \mathbf{x}_i + \delta = \mathbf{x}_i + \varepsilon * \frac{\nabla_{\mathbf{x}_i}\mathscr{L}_{CE}\left(\mathbf{x}_i, y_i\right)}{\left\|\nabla_{\mathbf{x}_i}\mathscr{L}_{CE}\left(\mathbf{x}_i, y_i\right)\right\|_2} \tag{11}$$

where $\mathbf{x}_i^{adv}$ is the adversarial view that shares the same label with the original news, δ is an adversarial perturbation, ε is the norm parameter is the norm parameter to control the normalized gradient as a valid perturbation. The gradient represents the first-order differential of the classification loss function $\mathscr{L}_{CE}$ for a specific target sample, indicating the direction in which the classification loss increases rapidly. We set the perturbation δ less than the norm parameter ε to ensure that the perturbation is imperceptible. Gradient-based adversarial data augmentation can maintain news semantics and simulate malicious forgery by fake news creators.

4.5.2 *Contrastive Training.*

To ensure effective and robust detection, we propose an auxiliary supervised adversarial contrastive learning task to align the representations of original and adversarial samples. The core idea is to minimize the distance between representations of original and adversarial samples from the same class closer while maximizing the distance between representations from different classes. The corresponding adversarial contrastive loss $\mathcal{L}_{ACL}$ is expressed in the following manner:

$$\mathscr{L}_{ACL} = \frac{-1}{|\mathcal{P}\left(\mathbf{x}\right)|}\sum_{\mathbf{x}_p \in \mathcal{P}(\mathbf{x})} \log \frac{\exp\left(\cos\left(\mathbf{x}, \mathbf{x}_p / \tau\right)\right)}{\sum_{\mathbf{h}_n \in \mathcal{N}(\mathbf{x})} \exp\left(\cos\left(\mathbf{x}, \mathbf{x}_n / \tau\right)\right)} \tag{12}$$

where $\mathbf{x}$ is the anchor, $\mathbf{x}_p$ is the positive sample with the same label as the anchor $\mathbf{x}$, and $\mathbf{x}_n$ is the negative sample with the label different from the anchor $\mathbf{x}$. $\mathcal{P}\left(\mathbf{x}\right)$ represents the positive set, comprising both the original samples and adversarial samples of the same class within the batch. On the other hand, $\mathcal{N}\left(\mathbf{x}\right)$ denotes the negative set, encompassing samples of different classes within the batch. τ is a temperature parameter and $\cos\left(\cdot\right)$ represents the cosine similarity function.

The design of the auxiliary adversarial contrastive learning task has two benefits. First, it pulls the original and adversarial news representations close to each other in the embedding space, thereby enhancing robustness. Second, it improves the uniformity of representations within the same class and the disparity of representations between different classes, thereby leading to more efficient classification.

Finally, we combine the cross-entropy loss and supervised adversarial contrastive loss to train our model:

$$\mathscr{L} = \mathscr{L}_{\text{CE}} + \lambda \mathscr{L}_{\text{ACL}} \tag{13}$$

where λ represents a hyper-parameter used to regulate the degree of adversarial contrastive learning.

5 EXPERIMENTS

In this part, we conduct experiments on three public datasets to answer the following research questions:

- **RQ1:** How does SAGT-ACL perform compared to previous multimodal fake news detection methods?
- **RQ2:** How effective are various model components in improving the performance of SAGT-ACL?
- **RQ3:** How does SAGT-ACL perform under different hyper-parameter settings?
- **RQ4:** How does SAGT-ACL perform in capturing intra and inter-modality semantic incongruity features?

5.1 Experiment Setup

5.1.1 Datasets. To validate the effectiveness of SAGT-ACL, we perform experiments on three public datasets: Weibo [6], Twitter [2], and Gossipcop [14]. Details of datasets are presented below:

Weibo dataset is proposed by Jin et al. [6] and is extensively utilized in the Chinese multimodal fake news detection task. The dataset comprises news articles sourced from Xinhua News Agency[1] and Weibo[2] platform. The training set contains 3,783 real news and 3,675 fake news, and the test set contains 1,685 news. **Twitter** dataset is released for MediaEval Verifying Multimedia Use task [2]. The news in the dataset is collected from the Twitter[3] platform. The training set comprises 5,139 fake news and 4,031 real news, and the test set contains 1,406 news. **Gossipcop** dataset is proposed by Shu et al. [14]. The real news in the Gossipcop dataset is collected at the famous trusted media website E!Online[4], and the fake news is collected from the fact-checking website Gossipcop[5]. The training set contains 7,974 real news and 2,036 fake news. The test set has 2,830 fake news.

5.1.2 Comparison Methods. To validate the performance of SAGT-ACL on the multimodal fake news detection task, We compare SAGT-ACL with several typical multimodal methods, which can be roughly divided into three categories, as below:

Joint representation methods: **SpotFake** [18] extracts text and image features using pre-trained BERT [7] and VGG19 [16], respectively, and concatenates them for classification. **SpotFake+** [17] extracts text and image features using pre-trained XLNet [29] and VGG19, respectively, and concatenates them for classification. **EANN** [23] concatenates text and image features as multimodal news representations. Furthermore, EANN introduces an auxiliary event classification task to help the model obtain better news representations. **MVAE** [8] obtains news representations by concatenating unimodal features. MVAE introduces an auxiliary news reconstruction task to learn better multimodal representation.

Similarity representation methods: **SAFE** [31] utilizes an image caption model to convert the image to text and compare the similarity of the original text with the generated text. **MCNN** [28] maps different modality features into the same embedding space and compares the similarity between them. **FND-CLIP** [33] utilizes the CLIP score to measure cross-modal similarity. The CLIP score also guides the model's use of different modality data.

Alignment representation methods: **MCAN** [24] employs a co-attention network to align multimodal features. **CARMN** [19] proposes a cross-modal attention residual network to align multimodal features and design a multi-channel CNN to mitigate the noise generated during the modal fusion process. **HMCAN** [13] employs a hierarchical attention model that takes into account the hierarchical semantics of the text as well as the contextual semantics between different modalities. **BTIC** [30] employs a self-attention mechanism to align multimodal features and trains the model using supervised contrastive loss and cross-entropy loss. **CMMTN** [22] uses a multi-modal masked transformer network to align the multimodal features and mask the irrelevant context between modalities.

5.1.3 Implementation Details. For our proposed SAGT-ACL model, the implementation details are as follows. In the feature extraction component, we use the pre-trained BERT [7] to get the token embeddings, and the textual embedding dimension of BERT is set to $d_t = 768$. We use the “bert-base-chinese” model for Chinese data and the “bert-base-uncased” model for English data. For the input image, we resize it to 224 × 224 and employ "ViT-B/16" [4] pre-trained on ImageNet to get the patch embeddings. and the visual embedding dimension of ViT is set to $d_v = 768$. To prevent overfitting, we freeze the parameters of BERT and ViT during the training process. The common embedding space dimension to which we project the text and image features is $d = 768$. In the SAGT component, the number of graph transformer layers L is set to 2, and the number of attention heads h is set to 8. In the ACL component, the contrastive coefficient λ is set to 0.3. In the training phase of the model, we set the mini-batch size as 128, the learning rate as $5e-4$, and the training epoch as 100 with an early stopping mechanism to mitigate overfitting. Meanwhile, we utilize the Adam algorithm [9] to optimize the parameters. For all other comparison methods, we adopt the parameter settings from the original paper to ensure optimal performance.

5.2 Performance Comparison (RQ1)

We compare our proposed SAGT-ACL method with twelve typical multimodal fake news detection methods on three datasets. The overall results are shown in Table 1, from which we have the following observations:

- The SAGT-ACL model outperforms other state-of-the-art methods in terms of both ACC and F1 metrics on all three datasets. This can be attributed to two key factors. Firstly, SAGT-ACL can comprehensively mine consistent and incongruent features in multimodal news. Secondly, SAGT-ACL can identify elaborate fake news through adversarial contrastive learning, thereby enhancing the effectiveness and robustness of the model.
- Joint representation methods (e.g., EANN, MVAE, SpotFake, SpotFake+) perform weaker compared to other methods. It is because they completely ignore the semantic associations between modalities, which can provide sufficient clues for fake news detection.
- The similarity representation methods perform better than the joint representation methods in terms of ACC and F1 metrics on all three datasets, which is attributed to their ability to capture semantic associations between modalities. However, they perform weaker than aligned representation methods in most of the metrics. Because they can only capture coarse-grained associations between modalities, but ignore fine-grained semantic interactions.
- Almost all alignment representation methods demonstrate superior performance compared to both joint representation methods and similarity representation methods on all three datasets. This proves that fine-grained semantic interactions within and between modalities are important for fake news detection. Nevertheless, they ignore the intra and

[1] http://www.xinhuanet.com
[2] https://weibo.com
[3] https://twitter.com/
[4] https://www.eonline.com/
[5] https://wwwgossipcop.com/

Table 1: Performance comparison between SAGT-ACL and other methods on three datasets. We evaluate the performance using ACC and F1. Bold numbers in the table indicate the best performance.

	Method	Weibo		Twitter		Gossipcop	
		ACC	F1	ACC	F1	ACC	F1
Joint Representation Methods	EANN [23]	0.779	0.781	0.761	0.764	0.753	0.758
	MVAE [8]	0.782	0.784	0.752	0.756	0.774	0.771
	SpotFake [18]	0.794	0.791	0.788	0.789	0.781	0.783
	SpotFake+ [17]	0.796	0.797	0.791	0.786	0.789	0.792
Similarity Representation Methods	SAFE [31]	0.811	0.813	0.801	0.799	0.831	0.825
	MCNN [28]	0.823	0.816	0.812	0.809	0.821	0.809
	FND-CLIP [33]	0.859	0.856	0.856	0.851	0.849	0.856
Alignment Representation Methods	MCAN [24]	0.868	0.865	0.866	0.871	0.847	0.841
	CARMN [19]	0.865	0.846	0.865	0.861	0.851	0.849
	HMCAN [13]	0.876	0.874	0.871	0.873	0.858	0.842
	BTIC [30]	0.884	0.883	0.879	0.875	0.863	0.869
	CMMTN [22]	0.889	0.882	0.881	0.879	0.853	0.859
Our Method	SAGT-ACL	**0.933**	**0.931**	**0.934**	**0.932**	**0.912**	**0.911**

inter-modality semantic incongruity, resulting in a decrease in performance. CMMTN utilizes a mask-attention mechanism to eliminate irrelevant components between modalities and diminish noise in modal interactions, thereby enhancing overall performance. In addition, BTIC utilizes supervised contrastive learning to optimize the embedding space, leading to relatively favorable outcomes.

5.3 Ablation Study (RQ2)

In this subsection, we perform an ablation experiment to verify the effectiveness of each key component of the model. Specifically, we design the following variants of SAGT-ACL by removing partial components from the model:

- w/o inter-modality edges: We remove the inter-modality edges and keep only the intra-modality edges during the construction of the multimodal graph.
- w/o intra-modality edges: we exclude the intra-modality edges and retain only the inter-modality edges during the creation of the multimodal graph.
- w/o SA: We replace the signed attention mechanism in the SAGT module with a traditional attention mechanism.
- w/o ACL: We remove the auxiliary adversarial contrastive learning task and only retain the classification task.
- w/o ADA: we conduct a simplified supervised contrastive learning process without building adversarial samples by adversarial data augmentation.

We compare the performance of SAGT-ACL and its variants on three datasets. The experiment results are presented in Table 2. The performance of SAGT-ACL drops after removing each component, proving that each component contributes to the model. Specifically, removing either the inter-modality or inter-modality edges resulted in a significant decrease in SAGT-ACL performance, suggesting that both intra-modality and inter-modality semantic interactions contribute to understanding the news. SAGT-ACL w/o SA performs significantly weaker than SAGT-ACL, which suggests that capturing inconsistent or even conflicting semantics between features can be helpful for fake news detection. Furthermore, SAGT-ACL significantly outperforms both SAGT-ACL w/o ACL and SAGT-ACL w/o ADA. This suggests the value of both contrastive learning and adversarial data augmentation. Contrastive learning can explore the intrinsic relationships to help identify the fundamental differences between classes. Meanwhile, adversarial data augmentation can enhance the robustness and effectiveness of the model in the face of elaborate fake news.

Table 2: Results of the ablation study on three datasets. Bold numbers in the table indicate the best performance.

Method	Weibo		Twitter		Gossipcop	
	ACC	F1	ACC	F1	ACC	F1
SAGT-ACL	**0.933**	**0.931**	**0.934**	**0.932**	**0.912**	**0.911**
w/o inter-modality edges	0.907	0.905	0.904	0.909	0.901	0.893
w/o intra-modality edges	0.922	0.921	0.919	0.921	0.909	0.893
w/o SA	0.913	0.909	0.911	0.915	0.898	0.891
w/o ACL	0.901	0.903	0.906	0.894	0.873	0.877
w/o ADA	0.909	0.911	0.915	0.912	0.888	0.887

5.4 Sensitivity Analysis (RQ3)

In this section, we conduct experiments to validate the effect of different hyperparameter settings on the performance of SAGT-ACL.

5.4.1 The number of SAGT layers L. The hyperparameter L indicates the number of SAGT layers. We test the performance of SAGT-ACL when L = 1, 2, 3, and the results are shown in Figure 3. We can obtain the following conclusions:

The SAGT-ACL achieves significant performance improvement when L changes from 1 to 2 and achieves the optimal result when

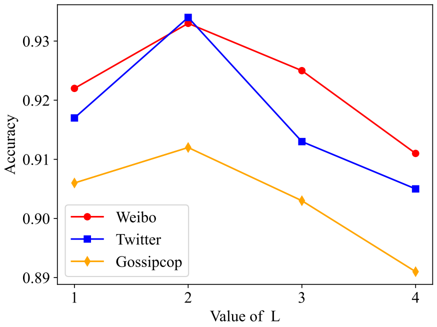


Figure 3: The influence of different SAGT layers on model performance.

$L = 2$. This indicates that single-layer SAGT is not sufficient to capture complex semantic associations in the multimodal graph.

However, the effect of SAGT-ACL decreased when L increased from 2 to 4. This suggests that a moderate L can help the model capture fine-grained semantic interactions in multimodal news, but too large L can cause over-smoothing problems on the graph.

5.4.2 The contrastive coefficient λ. The hyperparameter λ determines the extent of the auxiliary adversarial contrastive learning task in model training. We perform experiments to see how the contrastive coefficient λ affects the performance of SAGT-ACL. Specifically, we evaluate the accuracy of SAGT-ACL on three datasets as the parameter λ takes values ranging from 0.0 to 0.5. The experiment results are shown in Figure 4, from which we can draw the following conclusions. There is a significant improvement when λ ranges from 0.0 to 0.10. Note that $\lambda = 0.0$ indicates that the model focuses only on the classification task. This suggests that the auxiliary adversarial contrastive learning task can indeed improve the effectiveness of the model. The performance grows with λ increasing and peaks at the best when $\lambda = 0.30$ on all three datasets, which indicates that the optimal value of the hyperparameter λ is 0.30.

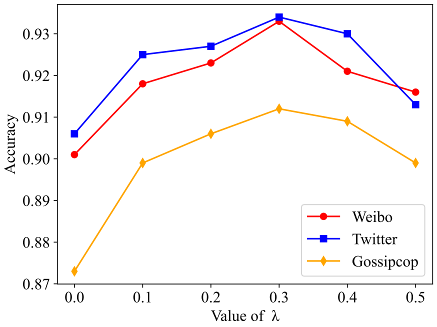
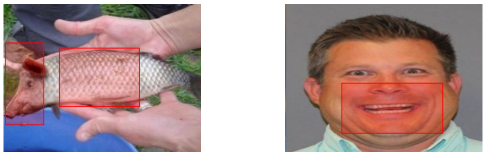


Figure 4: The influence of different contrastive coefficients on model performance.

As the parameter λ is incremented from 0.30 to 0.50, the model's performance exhibits a decline on all three datasets. This phenomenon can be attributed to the notion that a moderate λ value facilitates the acquisition of dependable representations that aid in the classification task; conversely, an elevated λ value may lead the model astray from its primary classification object.

5.5 Case Study (RQ4)

To intuitively demonstrate the effectiveness of the SAGT module in mining intra and inter-modality incongruity features, we select some typical cases and analyze them. Specifically, we visualize the incongruity features in multimodal news learned by the SAGT module. In the news shown in Figure 5 (a), SAGT captures fine-grained inconsistency features within image modalities, such as the body of a fish and the head of a pig being contradictory. In the news shown in Figure 5 (b), SAGT captures inter-modality inconsistent features, where the death event described in the text is contradictory to the smile in the image.



Text: New species of fish found in Arkansas.

(a)

Text: An employee of the Jefferson County morgue died this morning after being accidentally cremated by one of his coworkers.

(b)

Figure 5: The visualization of features captured by the SAGT module. Red highlights indicate inconsistent features captured by the SAGT module.

6 CONCLUSION

This study introduces a new model called Signed Attention-based Graph Transformer with Adversarial Contrastive Learning (SAGT-ACL) for multimodal fake news detection. SAGT-ACL focuses on comprehensively mining common and incongruent features in multimodal news data. In addition, SAGT-ACL designs a novel auxiliary adversarial contrastive learning task for robust and effective multimodal fake news detection. The experimental results show that SAGT-ACL achieves high accuracies of 0.933, 0.934, and 0.912 on Weibo, Twitter, and Gossipcop datasets, respectively, underscoring the efficacy of the proposed approach.

In future research, we intend to expand our current work in two main directions: first, by integrating additional modalities (including video and audio) for fake news detection, and second, by devising self-supervised tasks to efficiently utilize the extensive unlabeled news data in social media platforms.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant (61972047, 62372060) and the NSFC-General Technology Basic Research Joint Funds under Grant (U1936220).

REFERENCES

[1] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A Survey on Multimodal Disinformation Detection. In *Proceedings of the International Conference on Computational Linguistics*. 6625–6643.

[2] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 71–86.

[3] Jinyin Chen, Chengyu Jia, Haibin Zheng, Ruoxi Chen, and Chenbo Fu. 2023. Is multi-modal necessarily better? Robustness evaluation of multi-modal fake news detection. *IEEE Transactions on Network Science and Engineering* 10, 6 (2023), 3144–3158.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*. 1–21.

[5] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open* 3 (2022), 133–155.

[6] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the ACM International Conference on Multimedia*. 795–816.

[7] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186.

[8] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *Proceedings of the International World Wide Web Conference*. 2915–2921.

[9] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*. 1–15.

[10] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* 56, 2 (2023), 1–40.

[11] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial Training Methods for Semi-Supervised Text Classification. In *Proceedings of the International Conference on Learning Representations*. 1–11.

[12] Syed Tufael Nabi, Munish Kumar, Paramjeet Singh, Naveen Aggarwal, and Krishan Kumar. 2022. A comprehensive survey of image and video forgery techniques: variants, challenges, and future directions. *Multimedia Systems* 28, 3 (2022), 939–992.

[13] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 153–162.

[14] 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 (2020), 171–188.

[15] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[16] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations*. 1–14.

[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 *Proceedings of the AAAI Conference on Artificial Intelligence*. 13915–13916.

[18] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *Proceedings of the IEEE International Conference on Multimedia Big Data*. 39–47.

[19] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing and Management* 58, 1 (2021), 102437.

[20] Ludovic Trottier, Philippe Giguere, and Brahim Chaib-Draa. 2017. Parametric exponential linear unit for deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Machine Learning and Applications*. 207–214.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of the Conference on Neural Information Processing Systems* 30 (2017), 5998–6008.

[22] Jinguang Wang, Shengsheng Qian, Jun Hu, and Richang Hong. 2023. Positive Unlabeled Fake News Detection Via Multi-Modal Masked Transformer Network. *IEEE Transactions on Multimedia* 26 (2023), 234–244.

[23] 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 *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 849–857.

[24] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2560–2569.

[25] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2020), 4–24.

[26] Yatie Xiao, Chi-Man Pun, and Bo Liu. 2020. Adversarial example generation with adaptive gradient search for single and ensemble deep neural network. *Information Sciences* 528 (2020), 147–167.

[27] Yue Xing, Qifan Song, and Guang Cheng. 2022. Phase Transition from Clean Training to Adversarial Training. In *Proceedings of the Conference on Neural Information Processing Systems*. 9330–9343.

[28] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing and Management* 58, 5 (2021), 102610.

[29] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the Conference on Neural Information Processing Systems*. 1–18.

[30] Wenjia Zhang, Lin Gui, and Yulan He. 2021. Supervised Contrastive Learning for Multimodal Unreliable News Detection in COVID-19 Pandemic. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 3637–3641.

[31] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: similarity-aware multi-modal fake news detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 354–367.

[32] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. 2019. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 836–837.

[33] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. 2825–2830.