Abstract—Fake news detection (FND) has attracted much research interests in social forensics. Many existing approaches introduce tailored attention mechanisms to fuse unimodal features. However, they ignore the impact of cross-modal similarity between modalities. Meanwhile, the potential of pretrained multimodal feature learning models in FND has not been well exploited. This paper proposes an FND-CLIP framework, i.e., a multimodal Fake News Detection network based on Contrastive Language-Image Pretraining (CLIP). FND-CLIP extracts the deep representations together from news using two unimodal encoders and two pair-wise CLIP encoders. The CLIP-generated multimodal features are weighted by CLIP similarity of the two modalities. We also introduce a modality-wise attention module to aggregate the features. Extensive experiments are conducted and the results indicate that the proposed framework has a better capability in mining crucial features for fake news detection. The proposed FND-CLIP can achieve better performances than previous works on three typical fake news datasets.

Index Terms—Fake news detection, multimodal learning, CLIP, multimodal fusion

I. INTRODUCTION

Fake News Detection (FND) using machine learning is an efficient way of detecting the widespread dissemination of fake news, which can help readers identify the bias and misinformation and eliminate the negative spreading. Early works on fake news detection focus on analyzing text-only or image-only content [1] and demonstrated good performance on unimodal fake news detection. However, modern news and posts contain rich information, e.g., containing text and image at the same time. In some fake news, real images may be combined with completely fabricated rumors, and correct words may be used to describe a manipulated image. In these cases, unimodal FND methods are not sufficient for finding correlations across different modalities.

In recent years, many works have been conducted to detect anomalies in news and posts by aggregating multimodal features [2]–[4]. Researchers prefer using features from various sources, such as texts, images, comments, up-vote ratios, and even spreading graphs, to evaluate the truthfulness of a post. However, these additional modalities are not always available at the same time. Therefore, we focus on FND with only text and image modalities [3], [5], [6].

Multimodal FND involves analyzing both textual and visual content of news posts. However, these modalities are not always strongly correlated, and not every modality contains valid information that can be used for detection. This presents a major challenge for developing effective FND methods. Figure 1 shows two examples of posts in the Weibo dataset. The text and image of the first post are related, while the second post, whose text contains little event information, has a weaker correlation between text and image. This weak correlation can negatively affect the performance of the network during mul-
timodal fusion. Chen et al. [4] proposed a method to address this issue by calculating the correlation and generating fused features. They trained a variational autoencoder to compress the images and texts, and contrastively learned to minimize the Kullback-Leibler divergence for news with correct image-text pairs. The cross-modal ambiguity score was then used to reweight the multimodal features. This method has achieved good performance in multimodal detection. However, several open questions still need to be addressed. For example, it is unclear how the similarity of features from different modalities can be accurately calculated and how it affects the decision-making process in FND.

In this paper, we propose FND-CLIP, in which we use the pretrained Contrastive Language-Image Pretraining (CLIP) [7] model to address cross-modal ambiguity. Image and text are encoded by fine-tunable ResNet and BERT encoders, as well as pretrained CLIP encoders. Unimodal features are generated by concatenating the CLIP-generated features with fine-tunable counterparts. Fused features consist of the two CLIP outputs. Three projection heads shrink their sizes to distill important features. We calculate the cosine similarity as the cross-modal similarity score, which guide the usage of fused feature. An attention layer outputs three scores to measure the significance of these features in fake news detection. The classifier processes summarized features to distinguish fake news from real ones.

The primary contributions of this work can be summarized as follows: 1) We propose FND-CLIP, a multimodal fake news detection model. We employ a CLIP-based learning approach to extract semantic information and explicitly measure correlation between texts and images, which is used as weight parameter. 2) We propose a method to dynamically adjust the usage of unimodal and fused features. Specifically, we introduce an attention layer that adaptively outputs channel-wise scores to measure the significance of each modality and adjust its usage accordingly. 3) Experiments on three well-known datasets show that FND-CLIP outperforms state-of-the-art methods. Analysis of experimental results illustrates the advantages and thoughtfulness of our approach.

II. RELATED WORKS

Fake News Detection. Several methods have been proposed to extract valuable features from images and texts in news articles for FND. Earlier methods in this field focused on designing advanced, yet black-box, attention mechanisms for multimodal feature fusion [8]. Some of these methods suggest better aligning extracted features from different modalities before feeding them into the classifier. This is accomplished by using auxiliary tasks, such as event classification in EANN [3], variational auto-encoder based representation in MVAE [2], or correlation calculations in CAFE [4]. Singhal et al. [9] proposed Spotfake, which employs VGG and BERT to extract features, and then further refined the approach in Spotfake+ [5] to detect fake news in full-length articles. SAFE [10] calculates the relevance between textual and visual information in news articles. LIIMR [11] suppresses information from weaker modalities and extracts relevant information from the strong modality on a per-sample basis. MCAN [12] stack multiple co-attention layers to fuse the multimodal features.

Instead of focusing solely on network design, some methods leverage more information from datasets. For example, Qi et al. [13] argue that image feature extractors are not capable of understanding visual entities, such as celebrities, landmarks, and text, within images. Consequently, they manually extract this kind of information as linguistic assists. Allein et al. [14] proposed DistilBert, which employs latent representations of news articles and user-generated content to guide model learning. Shu et al. [15] proposed dEFEND, which features a sentence-comment co-attention subnetwork that exploits both news content and user comments to jointly detect fake news. Han et al. [16] proposed GNNCL, which utilizes GNNs to differentiate between the propagation patterns of fake and real news on social media.

Although these methods have achieved decent performance in multimodal FND, challenges persist regarding explicitly measuring the correlation between images and texts and efficiently and effectively exploiting different modalities to make decisions.

Multimodal Learning. Multimodal machine learning has gained much attention in the past decade [17]. In multimodal tasks, using priors and features from different modalities is crucial, and algorithms or deep networks with only a unimodality is not effective. Several generic technologies have been developed for learning joint representations of image content and natural language. For instance, the CLIP model [7] acts as a bridge between computer vision and natural language processing. It was trained on a diverse set of image-text pairs to predict the most relevant text snippet given an image, without directly optimizing for the task. CLIP can identify the most similar paired images and texts for matched image-text pairs. Although CLIP-based multimodal learning has been applied in various downstream tasks [18], [19], its application in FND has not yet been explored.

III. PROPOSED METHOD

A. Approach Overview

For multimodal FND, we focus on the samples that contain both text and image. Let each sample be \( x = (x_{\text{Txt}}, x_{\text{Img}}) \). Denote the ground-truth label as \( y \). When \( y = 0 \), \( x \) is a real news, otherwise it is fake. A rich set of features are first extracted from \( x_{\text{Txt}} \) and \( x_{\text{Img}} \). These features are then fused and projected into a single value of \( \hat{y} \), i.e., real or fake.

\[
\hat{y} = F_{\text{cls}}(F_{\text{Mix}}(F_{\text{Txt}}(x_{\text{Txt}}), F_{\text{Img}}(x_{\text{Img}}))),
\]

The procedure is depicted in (1), where \( F_{\text{Txt}} \) and \( F_{\text{Img}} \) are unimodal feature extractors, \( F_{\text{Mix}} \) is a feature fusing model and \( F_{\text{cls}} \) is the classification head.

Instead of applying sophisticated and black-box feature-fusing networks, we propose a simple yet effective method. We extract unimodal features using pretrained networks, then we use the CLIP model to measure the cross-modal similarity.
Fused Feature

CLIP Guided Features Generation and Aggregation

Classification

\[ f_{\text{fused}} = f_{\text{BERT}} + f_{\text{CLIP}} \]

which predicts the emotion, noise and other features irrelevant to image and text matching. When collaborating it with the unimodal features, we can better scrutinize the news from different aspects.

After feature extraction and alignment, we use a light-weight network to implement \( L_{\text{cls}} \) which predicts \( \hat{y} \).

**B. Network Specification**

Figure 2 illustrates the network architecture of the proposed FND-CLIP. The whole pipeline consists of three phases, i.e., features extraction, CLIP-guided features generation and aggregation, and classification.

**Unimodal feature generation.** We use a pretrained BERT model to extract the feature \( f_{\text{BERT}} \) from \( x_{\text{Txt}} \). We also use the ResNet \([20]\) to obtain deep representations \( f_{\text{ResNet}} \) from the image \( x_{\text{Img}} \). Meanwhile, we use the CLIP to encode text and image and obtain the features \( f_{\text{CLIP-T}} \) and \( f_{\text{CLIP-I}} \). For each branch, we concatenate the features extracted from different encoders to enhance the unimodal representation.

\[
\begin{align*}
    f_{\text{Txt}} &= \text{concat}(f_{\text{BERT}}, f_{\text{CLIP-T}}) \\
    f_{\text{Img}} &= \text{concat}(f_{\text{ResNet}}, f_{\text{CLIP-I}}).
\end{align*}
\]

**CLIP-guide feature generation.** Since the extracted text feature \( f_{\text{CLIP-T}} \) and the image feature \( f_{\text{CLIP-I}} \) have significant cross-modal semantic gaps, it is difficult for the network to learn their intrinsic semantic correlation. Therefore, the two features are then concatenated by

\[
    f_{\text{Mix}} = \text{concat}(f_{\text{CLIP-T}}, f_{\text{CLIP-I}}). \tag{3}
\]

The multimodal feature is a supplement to the unimodal features, it is used to enhance the semantic representations of them. Previous works often use a single network to mine both the coarse and the fine features from one modality. Alternatively, we use three pretrained models of CLIP, BERT, and ResNet for unimodal tasks CLIP uses large-scale image-text pairs to learn the extraction of semantics, which eliminates the emotion, noise and other features irrelevant to image and text matching. When collaborating it with the unimodal features, we can better scrutinize the news from different aspects.

Next, we use individual projection heads to adjust the dimension and remove the redundant information respectively. There are three projection heads in Figure 2, namely \( P_{\text{Txt}}, P_{\text{Img}} \) and \( P_{\text{Mix}} \). Each projection head contains two sets of a full connected layer with a BatchNorm layer, a ReLU activation function and a dropout layer. The projection heads share the same architecture but not the weights.

If we merely combine the CLIP-based features as the multimodal feature, reliable information cannot be provided since there is more ambiguity information in the features. To eliminate the ambiguity, we further design a fusion adjustment module, which is shown in Figure 3. The module adjusts the intensity of the fused features by measuring the cosine similarity between the text features and the image features provided by CLIP. The cosine similarity is calculated and further standardized into \([0, 1]\) in (4).

\[
    \text{sim} = \frac{f_{\text{CLIP-T}} \cdot (f_{\text{CLIP-I}})^T}{\| f_{\text{CLIP-T}} \| \cdot \| f_{\text{CLIP-I}} \|}. \tag{4}
\]

With the projection heads and the fusion adjustment module, three channels of the image feature \( m_{\text{Txt}} \), the text feature \( m_{\text{Img}} \) and the fused feature \( m_{\text{Mix}} \) can be generated, respectively. The generation is shown in (5), where \( M(\cdot) \) is a linear map function.

\[
\begin{align*}
    m_{\text{Txt}} &= P_{\text{Txt}}(f_{\text{Txt}}) \\
    m_{\text{Img}} &= P_{\text{Img}}(f_{\text{Img}}) \\
    m_{\text{Mix}} &= M(\text{sim}) \cdot P_{\text{Mix}}(f_{\text{Mix}}). \tag{5}
\end{align*}
\]
Feature aggregation. We further design a modality-wise attention mechanism to reweight the channels of $m_{	ext{Txt}}$, $m_{	ext{Img}}$ and $m_{	ext{Mix}}$ before feature aggregation. Inspired by the Squeeze-and-Excitation Network (SE-Net) [21], we design a modality-wise attention module as Figure 4. First, we concatenate the three features to a dimension of $L \times 3$, where $L$ is the feature length. The average pooling and maximum pooling are applied feature-wise, obtain a $1 \times 3$ vector. Then, the initial weight obtained in the previous step is sent into two $3 \times 3$ fully connected layers. With a GELU and Sigmoid based normalization, we obtain the attention weights $\text{att} = \{\text{att}_{\text{Txt}}, \text{att}_{\text{Img}}, \text{att}_{\text{Mix}}\}$. The weights are multiplied with $m_x$, $m_y$, and $m_{\text{mix}}$, and a sum process is performed, obtaining an aggregated feature $m_{\text{Agg}}$ of $L \times 1$.

$$m_{\text{Agg}} = \text{att}_{\text{Txt}} \cdot m_{\text{Txt}} + \text{att}_{\text{Img}} \cdot m_{\text{Img}} + \text{att}_{\text{Mix}} \cdot m_{\text{Mix}}.$$  (6)

Classification Finally, we feed the aggregated representation $m_{\text{Agg}}$ into a classifier $F_{\text{cls}}$ to predict the label $\hat{y}$. The classifier is a two-layer fully-connected network. In the proposed FND-CLIP scheme, we use cross-entropy as the objective function.

$$L_{\text{CE}} = y \log (\hat{y}) + (1 - y) \log (1 - \hat{y}).$$  (7)

C. Training Details

When selecting the BERT pretrained models, we use the “bert-base-chinese” model on Chinese data and the “bert-base-uncased” model on English data. We set the length of the input text as 300 words. Visual features are extracted by a pre-trained ResNet-101, in which the size of the input image is 224 $\times$ 224 and the same as that of CLIP. Since CLIP is pre-trained for English text, we use Google Translation API\(^1\) to translate Chinese texts to English. The pre-trained CLIP model is “ViT-B/32”. To meet the upper bound of input size in CLIP, we use a summary generation model\(^2\) to generate summary statements for the text longer than 50. We fine-tune ResNet in training stage, and freeze BERT and CLIP due to their difficulty in training on small datasets. We use the Adam optimizer with default parameters. The learning rate is $1 \times 10^{-3}$ and weight decay is 12. The batch size is set as 64. We trained the model for 50 epochs and chose the epoch that has the best testing accuracy to avoid over-fitting.

\(^1\)https://translate.google.com
\(^2\)https://huggingface.co/t5-large

IV. EXPERIMENTAL RESULTS

A. Setup

We use three real-world datasets collected from social media, namely, Weibo [23], Gossipcop, and Politifact [24]. During experiments, the unimodal news posts with no image or no text are removed. If a news post contains a text with multiple associated images, we arbitrarily select one image.

Weibo [23] is a widely used Chinese dataset in fake news detection. The training set contains 3,749 real news and 3,783 fake news, and the test set contains 1,996 news. Politifact and Gossipcop datasets are two English datasets collected from the political and entertainment domains of FakeNewsNet [24] repository, respectively. Politifact contains 244 real news and 135 fake news in the training set, and 75 real news and 29 news in the test set. Gossipcop contains 10,010 training news, including 7,974 real news and 2,036 fake news. The test set has 2,285 real news and 545 fake news.

B. Comparisons

We compare FND-CLIP with state-of-the-art methods. The comparison results are presented in Table I. The ‘-’ symbol in the table indicates that the results are not available from the original paper. FND-CLIP achieves the highest accuracy of 90.7%, 94.2%, and 88.0% on the three datasets, respectively, which is 0.7%, 6.8%, and 1.3% higher than the best results achieved by the other methods. FND-CLIP also ranks either 1\(^{st}\) or 2\(^{nd}\) in precision, recall, and accuracy in all tests.

Many methods, such as EANN and Spotfake, fuse features by concatenating or using attention mechanisms. However, these methods often suffer from a lack of correlation information due to the extracted features not being in the same semantic space. In contrast, CAFE uses cross-modal alignment to train encoders that can map texts and images into the same semantic space. However, due to the limited datasets and rough labels, the encoding effect is not optimal, resulting in a significant semantic gap between text and image features. In comparison, FND-CLIP achieves better performance for several reasons. First, the pre-trained CLIP encoders can generate semantically information-rich text and image features, which provide complementary information for unimodal features and generate fused features that accurately represent the correlations between text and image. Additionally, the modality-wise attention mechanism adaptively determines the weights of text, image, and fused features, which generates a better-aggregated feature and improves the classification accuracy. Overall, the results of our comparative study demonstrate the effectiveness of FND-CLIP in addressing the challenges of feature fusion for multimodal FND.

C. Ablation Study

To evaluate the impacts of key components in FND-CLIP, we remove different components and train the models from scratch on the three datasets, which are listed as follows.

(1) FND-CLIP w/o A. We remove the modality-wise attention module and direct aggregate three features to obtain the aggregated feature; (2) FND-CLIP w/o F. We remove

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE I
PERFORMANCE COMPARISON BETWEEN FND-CLIP AND OTHER METHODS ON THREE DATASETS. OUR METHOD ACHIEVES THE HIGHEST ACCURACY, AND THE PRECISION, RECALL AND F1-SCORE ARE ALSO HIGHER THAN MOST OF THE COMPARED METHODS.

| Method                  | Accuracy | Fake News | Real News |
|-------------------------|----------|-----------|-----------|
|                         |          | Precision | Recall | F1 score | Precision | Recall | F1 score |
| EANN [3]                | 0.827    | 0.847     | 0.812    | 0.829    | 0.807     | 0.843  | 0.825    |
| MVAE [2]                | 0.824    | 0.854     | 0.769    | 0.809    | 0.802     | 0.875  | 0.837    |
| Spotfake [9]            | 0.892    | 0.902     | 0.964    | 0.932    | 0.847     | 0.656  | 0.739    |
| SAFE [10]               | 0.762    | 0.831     | 0.724    | 0.774    | 0.695     | 0.811  | 0.748    |
| LIIMR [11]              | 0.900    | 0.882     | 0.823    | 0.847    | 0.908     | 0.941  | 0.925    |
| MCA [12]                | 0.899    | 0.913     | 0.889    | 0.901    | 0.884     | 0.909  | 0.897    |
| CAFE [4]                | 0.840    | 0.855     | 0.830    | 0.842    | 0.825     | 0.851  | 0.837    |
| FND-CLIP                | **0.907**| **0.914** | **0.901**| **0.908**| **0.914** | **0.901**| 0.907    |
|                         |          |           |          |          |           |        |          |
| Spotfake+ [5]           | 0.846    | -         | -        | -        | -         | -      | -        |
| dEFEND [15]             | 0.904    | -         | -        | -        | 0.902     | 0.956  | 0.928    |
| GNNCL [16]              | 0.803    | -         | -        | -        | 0.806     | 0.801  | 0.801    |
| Politifact              | 0.832    | -         | -        | -        | 0.836     | 0.832  | 0.829    |
| LSTM-ATT [22]           | 0.874    | 0.851     | 0.840    | 0.840    | 0.889     | 0.903  | 0.896    |
| SAFE [10]               | 0.741    | 0.875     | 0.636    | 0.737    | 0.647     | 0.880  | 0.746    |
| DistilBert [14]         | 0.864    | 0.724     | 0.778    | 0.750    | 0.895     | 0.919  | 0.907    |
| CAFE [4]                | 0.942    | 0.897     | 0.897    | 0.897    | 0.960     | 0.960  | 0.960    |
| FND-CLIP                | **0.907**| 0.914     | 0.901    | 0.908    | **0.914** | 0.901  | 0.907    |

the fusion module and use two unimodal features to classify news; (3) FND-CLIP w/o C. We remove all CLIP-related modules and only use BERT and ResNet to extract text and image features; (4) FND-CLIP MC-only: We remove the unimodal feature extractor, BERT and ResNet, and only use the multimodal CLIP fused feature; (5) FND-CLIP I-only: We remove all text-related features and only use image feature extracted by ResNet to classify; (6) FND-CLIP T-only: We only use the BERT-extracting feature to implement the detection task without any visual information.

**Effectiveness of Each Component.** The results in Table II show the impacts of each component. 1) FND-CLIP outperforms FND-CLIP w/o C, indicating that CLIP can effectively provide discernable features for fake news detection and can significantly improve detection accuracy. As only the intra-modal features are used for classification, the final feature cannot represent the intrinsic relationships between images and texts. 2) FND-CLIP outperforms FND-CLIP w/o F, indicating that the fused features by CLIP are important for detection. Although the unimodal branches also contain the CLIP-coded features, they are unable to represent the inter-modal characteristics. Meanwhile, FND-CLIP w/o F outperforms FND-CLIP w/o C, indicating that the complement to unimodal features using CLIP-coded features is effective. 3) FND-CLIP outperforms FND-CLIP w/o A, indicating that modality-wise attention is useful in balancing the modalities.

**Contributions from Different Modalities.** The second set of experiments is to evaluate the classification performance of different modalities in fake news detection. The results in Table II show the contributions of different modalities. 1) FND-CLIP I-only performs worst, indicating that only using images is insufficient for detecting fake news. 2) FND-CLIP MM-only achieves accuracy of 81.7% on Weibo, it even performs worse than FND-CLIP T-only. This indicates that the correlation information of images and texts is helpful for detection, however, only use of CLIP is not valid, the fused feature should be used along with the unimodal features. In addition, the CLIP-based features focus on the semantics of the text, and the BERT-based features contain emotional information. This complimentary characteristic is also helpful for the detection task. 3) FND-CLIP T-only achieves the second-best results, indicating that only using the text feature can basically complete the classification task. However, FND-CLIP outperforms FND-CLIP T-only, which hints that the visual feature can provide useful information.

**V. Conclusions**

In this paper, we propose a novel method for multimodal fake news detection, called FND-CLIP. Our approach leverages CLIP model to extract aligned multimodal features and guide the learning of the network for different modalities. Additionally, we propose a modality-wise attention mechanism to better aggregate the text, image, and fused features. From the extensive experiments on the well-known FND datasets, we find that the CLIP-based multimodal feature representation and fusion can well collaborate with the ResNet-based and
BERT-based unimodal features, which achieves state-of-the-art results in most cases. The results also indicate that the proposed modality-wise attention mechanism can improve the classification results by removing the noises and redundancies from the aggregated features. Apart from the performances gained by the proposed FND-CLIP, the outputs are still in binary forms, which are still unable to explain what elements in the news are vital for abnormal detection. In future works, we hope to develop more reliable fake news detection systems.

## References

[1] Nadia K Conroy, Victoria L Rubin, and Yimin Chen, “Automatic deception detection: Methods for finding fake news,” ASIST, vol. 52, no. 1, pp. 1–4, 2015.

[2] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, et al., “Mvae: Multimodal variational autoencoder for fake news detection,” in WWW, 2019, pp. 2915–2921.

[3] Yaqing Wang, Fenglong Ma, Zhiwei Jin, et al., “Eann: Event adversarial neural networks for multi-modal fake news detection,” in SIGKDD, 2018, pp. 849–857.

[4] Yixuan Chen, Dongsheng Li, Peng Zhang, et al., “Cross-modal ambiguity learning for multimodal fake news detection,” in WWW, 2022, pp. 2897–2905.

[5] Shivangi Singhal, Anubha Kabra, Mohit Sharma, et al., “Spotfake+: A multimodal framework for fake news detection via transfer learning,” in AAAI, 2020, vol. 34, pp. 13915–13916.

[6] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, et al., “Booststrapping multi-view representations for fake news detection,” in AAAI, 2023.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., “Learning transferable visual models from natural language supervision,” in ICML, 2021, pp. 8748–8763.

[8] Gaurav Bhatt, Aman Sharma, and Shivam andothers Sharma, “Combining neural, statistical and external features for fake news stance identification,” in WWW, 2018, pp. 1353–1357.

[9] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, et al., “Spotfake: A multi-modal framework for fake news detection,” in BigMM, 2019, pp. 39–47.

[10] Xinyi Zhou, Jindi Wu, and Reza Zafarani, “Safe: Similarity-aware multimodal fake news detection,” in PAKDD, 2020, pp. 354–367.

[11] Shivangi Singhal, Tanisha Pandey, Suksham Mrig, et al., “Leveraging intra and inter modality relationship for multimodal fake news detection,” in WWW, 2022, pp. 726–734.

[12] Yang Wu, Pengwei Zhan, Yunjian Zhang, et al., “Multimodal fusion with co-attention networks for fake news detection,” in IJCNLP, 2021, pp. 2560–2569.

[13] Peng Qi, Juan Cao, Xirong Li, et al., “Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues,” in ACM MM, 2021, pp. 1212–1220.

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

[15] Kai Shu, Limeng Cui, Suhan Wang, et al., “Defend: Explainable fake news detection,” in SHOKDD, 2019, pp. 395–405.

[16] Yi Han, Shanika Karunasekera, and Christopher Leckie, “Graph neural networks with continual learning for fake news detection from social media,” arXiv preprint arXiv:2007.03316, 2020.

[17] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE TPAI, vol. 41, no. 2, pp. 423–443, 2018.

[18] Tianyi Wei, Dongdong Chen, Wenbo Zhou, et al., “Hairclip: Design your hair by text and reference image,” arXiv preprint arXiv:2112.05142, 2021.

[19] Marcos V Conde and Kerem Turgutlu, “Clip-art: contrastive pre-training for fine-grained art classification,” in CVPR, 2021, pp. 3956–3960.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[21] Jie Hu, Li Shen, and Gang Sun, “Squeeze-and-excitation networks,” in CVPR, 2018, pp. 7132–7141.

[22] Jun Lin, Glennia Tremblay-Taylor, Guanyi Mou, et al., “Detecting fake news articles,” in Big Data, 2019, pp. 3021–3025.

[23] Zhiwei Jin, Juan Cao, Han Guo, et al., “Improving fake news detection via transfer learning,” in PAKDD, 2020, pp. 354–367.