Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.050344

ARTICLE

Cross-Modal Consistency with Aesthetic Similarity for Multimodal False Information Detection

Weijian Fan[1,*] **and Ziwei Shi**[2]

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, 100024, China

[2]Software Research Institute, China United Network Communication Group Co., Ltd., Beijing, 100024, China

*Corresponding Author: Weijian Fan. Email: fanwj@cuc.edu.cn

Received: 03 February 2024 Accepted: 02 April 2024 Published: 15 May 2024

ABSTRACT

With the explosive growth of false information on social media platforms, the automatic detection of multimodal false information has received increasing attention. Recent research has significantly contributed to multimodal information exchange and fusion, with many methods attempting to integrate unimodal features to generate multimodal news representations. However, they still need to fully explore the hierarchical and complex semantic correlations between different modal contents, severely limiting their performance detecting multimodal false information. This work proposes a two-stage detection framework for multimodal false information detection, called ASMFD, which is based on image aesthetic similarity to segment and explores the consistency and inconsistency features of images and texts. Specifically, we first use the Contrastive Language-Image Pre-training (CLIP) model to learn the relationship between text and images through label awareness and train an image aesthetic attribute scorer using an aesthetic attribute dataset. Then, we calculate the aesthetic similarity between the image and related images and use this similarity as a threshold to divide the multimodal correlation matrix into consistency and inconsistency matrices. Finally, the fusion module is designed to identify essential features for detecting multimodal false information. In extensive experiments on four datasets, the performance of the ASMFD is superior to state-of-the-art baseline methods.

KEYWORDS

Social media; false information detection; image aesthetic assessment; cross-modal consistency

1 Introduction

The widespread use of online social media has become the primary platform for people to access and share information. While offering convenience, this digital transformation of public space has also brought about negative consequences such as the proliferation of false information, political polarization, and hate speech [1,2]. The 2020 US presidential election and global events like the COVID-19 outbreak and the Russia-Ukraine war have all been marred by the spread of false information [3,4]. In the current post-truth era, where online public opinion is increasingly influential, the need for trust and consensus is more pressing than ever [5,6]. Therefore, developing quantitative

analysis and automatic detection methods for false information has become an urgent necessity and has recently attracted a lot of research attention.

While some studies [7–9] have attempted to address the issue of false information detection, their focus on single-mode detection methods based on text content and social backgrounds is limited. In the era of media convergence, information dissemination is no longer limited to text, but also includes different modalities of content such as videos and images [10,11]. The inclusion of images or videos in social media posts significantly increases their reach and impact, making the authenticity of such information even harder to determine [12,13]. Existing multimodal detection methods like Recurrent Neural Network with an attention mechanism (att-RNN) [12] and Multimodal Knowledge-aware Event Memory Network (MKEMN) [14] require a substantial amount of social background, which is often not available during the early stages of fake news dissemination, particularly at the release stage [15]. Therefore, the extraction of clues from multimodal content presents a practical and viable direction for multimodal false information detection.

Recently, many researchers have shifted their focus to multimodal false information detection and proposed many features that can effectively integrate intra-modal and inter-modal features and facilitate detection, including exploring the similarity between news text and images from a semantic perspective [16] and improving the performance of fake news detection by extracting visual entities to enhance text and visual representations [17]; Combining multimodal content and social context [12,14]; Attention-based cross-modal research [1] focuses on highly correlated fragments from another modality and infers clues from these relevant information; Enhancing understanding of internal relationships between objects in different local area through adversarial sample generation [18].

Although existing multimodal false information detection methods have achieved good results, there are some shortcomings: Firstly, researchers often assume that the text and images in real information are consistent, which leads to excessive attention to content with high intrinsic correlation in the modeling process; secondly, existing methods often overlook important features in irrelevant content in cross-modal interactions. These issues severely limit its performance in detecting multimodal false information.

To address these issues, this work proposes a multi-stage detection framework for multimodal false information detection, which is called ASMFD. This framework focuses on high similarity consistency features and considers low similarity inconsistency information in multimodal interaction processes. Specifically, in the first stage, we utilize the advantages of Contrastive Language-Image Pre-training (CLIP) [19] in multimodal feature representation to fine-tune the relationship between learning text and images in a label-aware manner. In the second stage, the cross-modal correlation matrix between text and image is calculated based on the pre-trained CLIP model in the first stage. Then, Named Entity Recognition (NER) technology is used to extract entities from the text content and retrieve relevant information on the platform to obtain candidate-related graphs. The cross-modal correlation matrix is divided into consistency and inconsistency matrices by calculating the aesthetic similarity between images and related images. Finally, the fusion module is used to determine the critical features for identifying multimodal false information.

The main contributions of this work are as follows:

a) We propose the ASMFD model, which focuses on high similarity consistency features and considers low similarity inconsistency information in multimodal interaction processes. This method is more suitable for identifying false information on social platforms as it allows for the dynamic selection of dependent features for judgment.

b) We explore the correlation matrix by dividing it into two aspects: Consistency and inconsistency. Meanwhile, we introduce image aesthetic similarity as the consistency threshold. To our knowledge, this is the first work to introduce image aesthetic assessment in multimodal false information detection. The experiment has proven that the attribute similarity of image aesthetics can improve the performance of the model to a certain extent.
c) We constructed an auxiliary task to encourage the model to choose more important features to recognize different information.
d) Experiments on four public datasets have shown that the proposed ASMFD can achieve optimal performance across all baselines.

2 Related Work

2.1 Multimodal False Information Detection

Due to the rapid dissemination of fake news using out-of-context real images and text/titles, researchers have begun to address this important social issue. Zhou et al. [15] first utilized Text CNN (Convolutional Neural Network) and VGG models to extract features from text and related images, then calculated the similarity to predict fake news. Similarly, Singhal et al. [20] used Bidirectional Encoder Representation from Transformers (BERT) as a text encoder to extract text features, VGG19 as an image encoder to extract visual features, and finally fused them for classification. Most of these existing methods cannot perform well on invisible events. They proposed a method in which they simultaneously trained an event-filtering network and a multimodal fake news detector. This network aims to eliminate the characteristics of specific events while preserving common features between different events. Silva et al. [21] proposed a framework to jointly preserve do-main-specific and cross-domain knowledge in news records to detect fake news from different domains using a model combining domain-specific and cross-domain features. In addition, it proposes an unsupervised method to select a subset of unlabeled, valuable news records to be manually labeled. In Wang et al.'s work [22], meta-learning and neural processing methods were integrated to achieve high performance even on events with limited labeled data.

The emergence of large-scale visual language models for learning information from images and text has proven helpful for computer vision tasks. Models like CLIP [19] and ALIGN (A Large-scale ImaGe and Noisy-text embedding) [23] have been trained in a comparative manner on large-scale datasets of approximately 400 million and 1B image text pairs, respectively, allowing them to learn rich representations between images and text. Zhou et al. [24] used the CLIP model, image encoder (ResNet), and text encoder (BERT) to extract multimodal features and use attention from different modalities to obtain the final classification features. Some works also use social media-related information to detect fake news. Abdelnabi et al. [25] proposed a consistency-checking network that measures the similarity between images and corresponding texts and utilizes internet search results (images and text) to improve the performance of CLIP-based models further.

2.2 Image Aesthetic Assessment

Evaluating the aesthetic quality of images involves many factors that affect preferences. Although many of these factors are difficult to quantify, some known aesthetic attributes can affect preferences. With the introduction of the AADB and EVA datasets, researchers can transform from the initial binary classification problem to a regression problem of predicting overall aesthetic scores and scores for multiple attributes. The AADB dataset [26] includes overall aesthetic ratings and ratings for 11 image attributes. Explainable Visual Aesthetics (EVA) [27] includes overall aesthetic and four

attribute scores. Based on these two datasets, researchers have developed and designed many methods. Kong et al. [26] developed a multitasking neural network by fine-tuning AlexNet and trained a Siamese network to predict aesthetic ratings. Hou et al. [28] applied distance-based loss of square bulldozers for training and compared different deep networks, including AlexNet, VGG, and a residual network. They found that fine-tuning the VGG model achieved the best performance. Li et al. [29] proposed a multitasking model that can learn image aesthetics and personality traits. The multitasking network proposed by Celona et al. [30] can predict aesthetic scores, style, and composition attributes. Pan et al. [31] proposed a neural network architecture for adversarial learning inspired by generative adversarial networks. This is a multi-task deep CNN that simultaneously learns aesthetic scores and attributes. Li et al. [32] proposed a layered image aesthetic attribute prediction model. The Topic Perceived Visual Attribute Inference model can predict six attributes of the AADB dataset [33].

Compared with previous studies, the ASMFD model proposed in this paper leverages the advantages of the visual language model CLIP in multimodal feature representation, fine-tuning the relationship between learning text and images through label perception. Meanwhile, ASMFD does not directly model cross-modal consistency and inconsistency features, but instead uses image aesthetic attribute similarity as a threshold for choosing consistency and inconsistency features. This model is the first work to introduce the image aesthetic assessment task in multimodal false information detection tasks.

3 Methodology

This section will introduce the proposed ASMFD architecture for multi-modal false information recognition. The ASMFD is based on image aesthetic similarity to segment and explores the consistency and inconsistency features of images and texts, which can effectively aggregate different modal features in the false information detection task.

As shown in Fig. 1, ASMFD has two stages: The pre-training stage for CLIP-based multimodal feature extraction and aesthetic evaluation and the recognition stage for multimodal false information. Specifically, in the first stage, we utilize the advantages of CLIP in multimodal feature representation to fine-tune the relationship between learning text and images in a label-aware manner. Simultaneously, based on existing publicly available image aesthetic datasets and pre-trained image aesthetic attribute rating networks. In the second stage, the cross-modal correlation matrix between text and image is calculated based on the pre-trained CLIP model in the first stage. Then, we use Named Entity Recognition (NER) methods to extract entities from the text content, retrieving relevant information on the platform to obtain related images. The cross-modal correlation matrix is divided into consistency and inconsistency matrices by calculating the aesthetic similarity between images and related images. Finally, the fusion module is used to determine the key features for identifying multimodal false information.

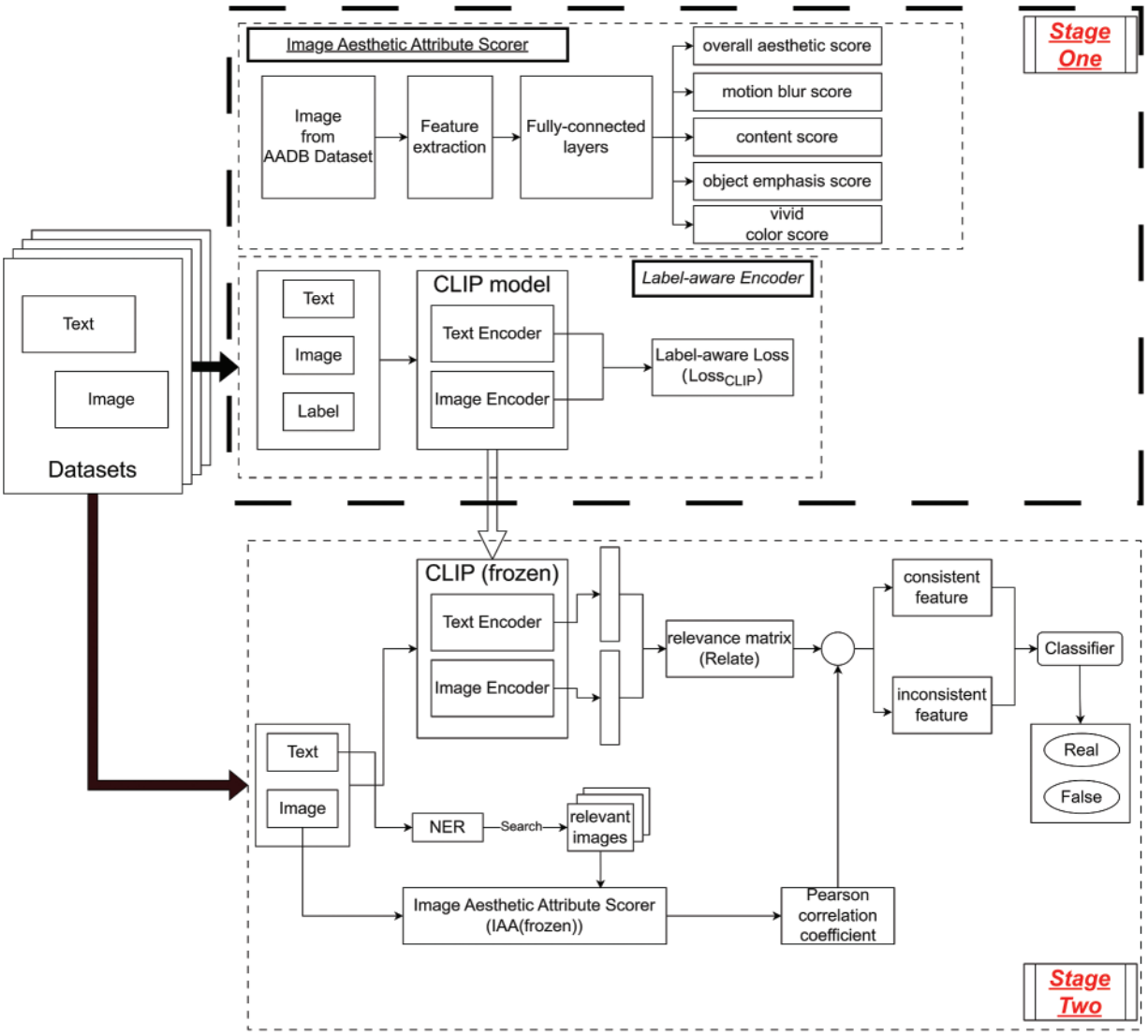

Figure 1: Illustration of the proposed ASMFD architecture

3.1 Problem Formulation

Multimodal information includes both textual and visual modalities. In this article, we mainly focus on the posting text T and an accompanying image I of social media information, which are formally written as $A = (T, I)$, without considering other information such as comments and videos. We follow the definition of most work on the false information detection task and treat the multimodal false information detection task as a binary classification task. As shown in Eq. (1), our goal is to design a recognizer f that can model and aggregate the feature of two modalities to classify multimodal information as false ($y = 0$) or real ($y = 1$).

$$f(A) \rightarrow y \in \{0, 1\} \tag{1}$$

In addition, in the proposed ASMFD, we have also introduced an image aesthetic assessment module. Similarly, we follow the common design of this task and consider the image aesthetic assessment task as a regression task. As shown in Eq. (2), we aim to train an image aesthetic attribute scorer f_{IAA}, which outputs scores for k different aesthetic attributes *score* based on the input image I.

$$f_{IAA}(I) \rightarrow \left(score_y | score_1, score_2, \ldots, score_k\right) \tag{2}$$

where $score_y$ represents the overall aesthetic score of image I, and $score_k$ represents the *k-th* attribute score. The range of values for all aesthetic scores is $[-1,1]$.

3.2 Stage One: Pretraining Models

In this stage, we train all layers of the original CLIP model in an end-to-end manner using the false information datasets. Meanwhile, the aesthetic evaluation dataset was used to train an image aesthetic mathematical scorer.

3.2.1 Label-Aware Encoder

In previous work, researchers found that using the self-supervision method as the warm-up stage for model training may achieve better performance in downstream tasks [2,16]. Following this discovery, we finetuned the original CLIP model to learn the relationships between text and images through label perception. Specifically, we used multimodal alignment loss in the fine-tuning stage. When the information is real, the backbone model can bring the feature distance between the image and the corresponding text closer, and when the news is fake, it can distance them further. This training objective helps us to segment and utilize consistent and inconsistent features in the second stage.

Firstly, we perform three enhancement operations on the original image: Random cropping, horizontal flipping, and standardization. Then pass them into CLIP to obtain the feature embeddings for the original image, text, and enhanced images. In the training stage, the inputs are an image and its corresponding text, formally as I and T, and the enhanced images of the original image are represented as $I_e, e = 1, 2, 3$. The embeddings generated through CLIP are $\left[\widehat{I}, \widehat{T}, \widehat{I}_e\right]$. In contrastive learning, we treat samples with real labels as positive pairs and samples with false labels as negative pairs.

For real label pairs, we hope that the trained CLIP model can enable image embedding $\widehat{I}$, the enhanced images $\widehat{I}_e, e = 1, 2, 3$ and the text $\widehat{T}$ to get as close as possible. Therefore, we use two combinations to construct the positive pairs: Origin image and enhanced image pairs, origin image and text pairs. As shown in Eqs. (3) and (4), the contrast loss corresponding to the *e-th* image enhancement of the input sample can be calculated.

$$Loss_{I-A}^{real} = -\log \frac{\exp\left(\frac{sim\left(\widehat{I}, \widehat{T}\right)}{\tau}\right)}{\sum_{i,k=1}^{S}\left(\sum_{a=1}^{3}\exp\left(sim\left(\widehat{I}_i, \widehat{I}_{k,a}\right)\right) + \exp\left(sim\left(\widehat{I}_i, \widehat{T}_k\right)\right) + \exp\left(sim\left(\widehat{T}_i, \widehat{T}_k\right)\right)\right)} \tag{3}$$

$$Loss_{I-E}^{real} = -\log \frac{\exp\left(\frac{sim\left(\widehat{I}, \widehat{I}_e\right)}{\tau}\right)}{\sum_{i,k=1}^{S}\left(\sum_{t=1}^{3}\exp\left(sim\left(\widehat{I}_i, \widehat{I}_{k,t}\right)\right) + \exp\left(sim\left(\widehat{I}_i, \widehat{T}_k\right)\right) + \exp\left(sim\left(\widehat{T}_i, \widehat{T}_k\right)\right)\right)} \tag{4}$$

where S represents the number of training dataset; i, k represent the *i-th* and *k-th* sample of dataset, and $i \neq k$; $\widehat{I}_{k,t}$ represents the enhanced images feature of *k-th* image; $sim\,(\cdot)$represents the cosine similarity function. The τ is the temperature coefficient, it was directly optimized from training in the original CLIP model. In this work, we directly set the value of τ is 0.07.

Then, we can calculate the total loss of the real information as Eq. (5).

$$Loss^{real} = Loss^{real}_{I-T} + Loss^{real}_{I-E} \tag{5}$$

For the false pairs, the corresponding loss is calculated in the same way as for real pairs. But for negative pairs, our goal is to separate café. Therefore, the final loss calculation for negative pairs is shown in Eq. (6).

$$Loss^{false} = -\log \frac{\exp\left(\frac{sim\left(\widehat{I}, \widehat{I}_e\right)}{\tau}\right)}{\sum_{i,k=1}^{S}\left(\sum_{t=1}^{3}\exp\left(sim\left(\widehat{I}_i, \widehat{I}_{k,t}\right)\right) + \exp\left(sim\left(\widehat{I}_i, \widehat{I}_k\right)\right) + \exp\left(sim\left(\widehat{T}_i, \widehat{T}_k\right)\right) + \exp\left(sim\left(\widehat{I}_i, \widehat{T}_k\right)\right)\right)} \tag{6}$$

The final training CLIP model's label-aware contrastive loss is calculated by weighting the real and false pair losses, as shown in Eq. (7).

$$Loss_{CLIP} = Loss^{real} + \lambda Loss^{false} \tag{7}$$

where λ is a trade-off parameter to measure the false and real terms.

3.2.2 Image Aesthetic Attribute Scorer

With the release of AADB datasets in the field of aesthetic assessment and the continuous deepening of related work, it provides a good guarantee for this work to calculate the similarity of images similarity from the perspective of aesthetic attributes. In this section, we designed a multitask learning CNN architecture to predict the overall aesthetic score and multiple scores of related attributes of an image efficiently and accurately. Due to the AADB dataset containing relatively few image samples, we finetuned the pre-trained CNN model, i.e., VGG16. Many studies on the aesthetic assessment task have also experimentally proven that using pre-trained networks is better than scratch.

As shown in Fig. 2, the image aesthetic attribute scorer designed in this work takes an image as input and extracts inter-feature representations using the first five blocks of VGG16 convolutional layers. After the last convolutional block, we added an average pooling layer and two fully connected layers to obtain the final feature embedding. we also use the Dropout to prevent overfitting. Our designed model consists of multiple units in the output layer, one for predicting overall aesthetic score and the other for predicting attribute scores. In the AADB dataset, it has 11 attributes and a total score. To better adapt to the context of social media information, we chose four at-tributes (motion blur, content, object emphasis, and vivid color) and the total score. So, the final output layer has five output units.

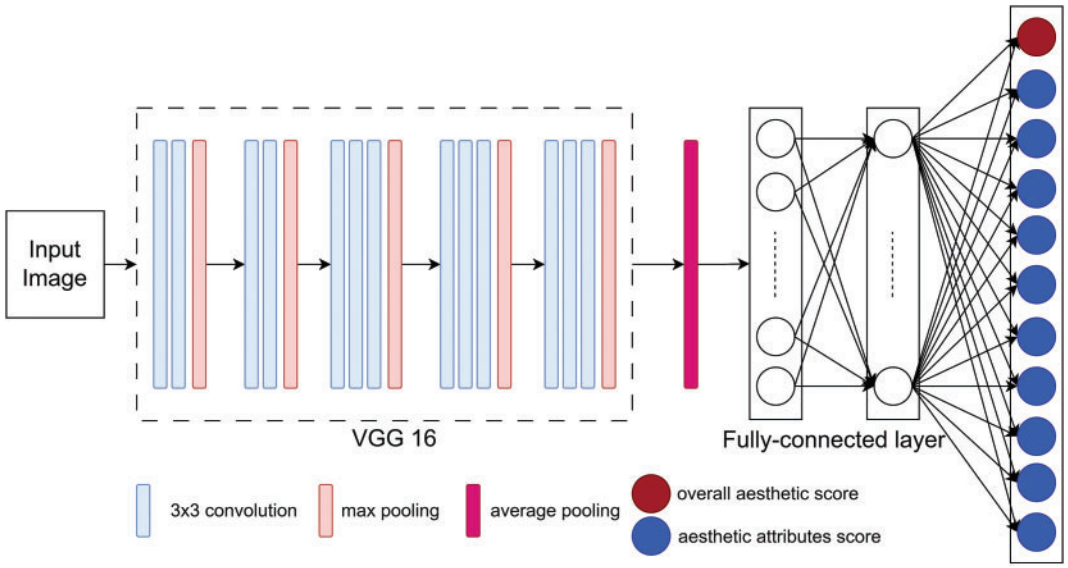

Figure 2: The framework of image aesthetic attribute scorer

3.3 Stage Two: Detection

After pretraining two models in stage one, we can further train the model for multimodal false information detection. This stage mainly consists of four modules: Feature encode module, feature interaction module, feature aggregation module, and detection module.

3.3.1 Feature Encode

In this module, to extract better features of multimodal information, we use the CLIP model pretrained in stage one as the feature encoder for multimodal information, which can effectively model the relationship features between text and image in a label-aware contrastive loss. As shown in Eqs. (8) and (9), we pass the image and corresponding text, i.e., $A_i = (T_i, I_i)$ obtains $\left[\widehat{T}_i, \widehat{I}_i\right]$ through a full connected layer after the text encoder and image encoder from the pre-trained CLIP model.

$$\widehat{T}_i = FC_T\left(CLIP_{frozen} - TE\left(T_i\right)\right) \tag{8}$$

$$\widehat{I}_i = FC_I\left(CLIP_{frozen} - IE\left(I_i\right)\right) \tag{9}$$

where $CLIP_{frozen}$ represents the parameters of the CLIP model are frozen and un-trained in whole stage two; FC is the full connected layer.

3.3.2 Feature Interaction

In the feature interaction module, we use cosine similarity to evaluate the correlation scores in different modalities following the previous work [10]. The correlation feature matrix between the two models can be calculated for text feature $\widehat{T}_i = \{t_1, t_2, \ldots, t_N\}$ and image features $\widehat{I}_i = \{i_1, i_2, \ldots, i_N\}$, as shown in Eq. (10).

$$Relate_{j,k} = \frac{t_j \cdot i_k}{\left\|t_j\right\| \cdot \left\|i_k\right\|}, j, k \in [1, N] \tag{10}$$

where $Relate_{j,k}$ is the correlation score of the word t_j and image region i_k, the range of value is $[-1,1]$.

Previous research aggregates high correlation features of different modalities through correlation matrices and use them for inference [10]. However, this operation may lead them to overlook the importance of low correlation features in detecting false information. In the process of feature aggregation, low correlation features have little impact or even loss due to lower attention scores. In real-life scenarios, false information often places greater emphasis on adding highly relevant multimodal content to gain public trust, thereby presenting the illusion of consistency. The previously overlooked low correlation parts may contain important indicators for detecting false information. Therefore, in this work, we attempt to divide the correlation matrix into label-aware consistency features and inconsistency features based on the similarity of image aesthetic attributes of relevant images.

Firstly, as shown in Fig. 3, we use existing NER techniques to recognize the person, location, and product of the input. Then, based on the NER results of these three types of information, we search on the corresponding social platform to obtain relevant images I_S.

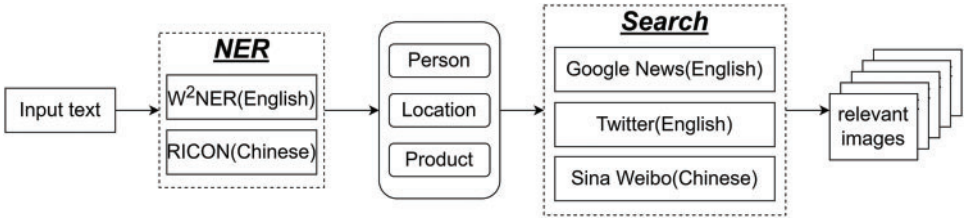

Figure 3: The process of searching relevant images by input text

Then, as shown in Eq. (11), we use the trained aesthetic attribute scorer in stage one to calculate the correlation coefficient between the input and retrieved images as the segmentation threshold η_{IAA}.

$$\eta_{IAA} = \left| \frac{\sum_{si=1}^{SI} corr\left(IAA\left(I\right), IAA\left(I_{si}\right)\right)}{SI} \right| \tag{11}$$

where IAA is pretrained image aesthetic attribute scorer in stage one; SI is the number of relevant images $I_S \in [1, 5]$; $corr\left(\cdot\right)$ is the Pearson correlation coefficient function.

Finally, we divide the correlation matrix $Relate$ into consistent features R_c and inconsistent features R_{ic} based on the threshold η_{IAA}, as shown in Eqs. (12) and (13).

$$R_c = \left[Relate_{j,k} \geq \eta_{IAA}\right] \tag{12}$$

$$R_{ic} = \left[Relate_{j,k} < \eta_{IAA}\right] \tag{13}$$

3.3.3 Feature Aggregation

After obtaining the consistency and inconsistency correlation matrices, we aggregate features for different types.

For the consistency correlation matrix, we follow the attention mechanism to aggregate features according to attention scores, and the calculation is shown in Eq. (14).

$$c_j^c = \sum_{k=1}^{N} \sigma\left(Relate_{j,k}\right) * i_k + t_j, j, k \in R_c \tag{14}$$

where c_j^c is the complementary information from the image modality.

For the inconsistency correlation matrix, due to the small value of attention scores, we use the element-wise addition $\oplus$ to obtain the aggregate feature $c^{ic}_{j,k}$, and the calculation is shown in Eq. (15).

$$c^{ic}_{j,k} = t_j \oplus i_k, j, k \in R_{ic} \tag{15}$$

To fully utilize the important information hidden in consistent and inconsistent features, we first use two independent Multilayer Perceptron (MLP) networks to map the consistency and inconsistency aggregation matrices and calculate the feature matrices of different categories through the mapped vectors. Then, we assign weights V to each part of the features to evaluate which feature is more important in the final evaluation. The calculation of the process is shown in Eqs. (16)–(18).

$$V_{cor} = W_{cor}\left(\sum t_j sigmoid\left(MLP\left(c^{c}_{j}\right)\right)\right) + b_{cor} \tag{16}$$

$$V_{icor} = W_{icor}\left(\sum t_j sigmoid\left(MLP\left(c^{ic}_{j,k}\right)\right)\right) + b_{icor} \tag{17}$$

$$V = sigmod\left(\langle V_{cor}, V_{icor}\rangle\right) \tag{18}$$

where $\langle\rangle$ is the concatenation operation; W and b are learnable parameters.

3.3.4 Detection

The detection module is used to predict the credibility threshold of multimodal information by transforming its features into two categories using a two-layer perceptron.

$$\hat{y} = W_2\left(ReLU\left(W_1 V \circ \left\langle \sum t_j sigmoid\left(MLP\left(c^{c}_{j}\right)\right), \sum t_j sigmoid\left(MLP\left(c^{ic}_{j,k}\right)\right)\right\rangle + b_1\right)\right) + b_2 \tag{19}$$

We follow the multimodal false information detection task, use the Eq. (20) as the detection loss.

$$Loss = -\left[y \log \hat{y} + (1-y)\log\left(1-\hat{y}\right)\right] \tag{20}$$

where y is the ground-truth label.

In addition, to better enable the network to learn the importance V of consistent and inconsistent feature matrices in label aware, we set up an auxiliary task to guide the learning process. Specifically, we designed a new label for the importance weight V based on the original label. When the label is true, the new label is [0.7, 0], and when the label is false, the new label is [0, 0.7]. The reason we set this is because we hope that the ASMFD can consider more inconsistent features when predicting false information. The corresponding loss function is shown in Eq. (21).

$$Loss_{aux} = \|y_{aux} - V\|_2 \tag{21}$$

where y_{aux} is the new label of importance weight V corresponding to the original ground-truth label y.

4 Experiment and Parameter Setup

4.1 Datasets

To verify the effectiveness and generalization of the proposed ASMFD for identifying multimodal false information, we conducted extensive experiments on four public datasets.

The first dataset is the FakeNewsNet dataset [34], which is an English dataset. FakeNewsNet consists of two parts, GossipCop and PolitiFact. They all include the title, images, comments, and content of the news. The original PolitiFact contained 748 news articles, but only 286 of them contained images. To adapt to our task setting, we only used news containing images, called PolitiFact-2.

The second dataset is the Twitter dataset [35], which is also an English dataset. It was released along with the MediaEval-2016 task. Each tweet in this dataset consists of textual content, image or vide, and corresponding social context. We removed tweets containing videos from the dataset.

The third dataset is the Weibo dataset [12], which is a Chinese dataset collected from the Chinese social network platform Weibo. It contains 9,528 pieces of information including text, images, and social metadata.

The specific distribution of the data is described in detail in Tables 1–3. We followed the preprocessing steps of the original paper and the same data segmentation method in the experiments to ensure the comparability of our experimental results.

Table 1: FakeNewsNet dataset

SubDataset	GossipCop	PolitiFact	PolitiFact-2
False	2,505	396	107
Real	9,302	352	179
Comments	31,023	43,487	13,979
Images	11,807	286	286

Table 2: Twitter dataset

Label	Number
False	10,717
Real	7,332

Table 3: Weibo dataset

Label	Number
False	4,749
Real	4,779

In addition, in this work, we used the AADB dataset [26] to train our own aesthetic attribute scorer. AADB dataset is an image aesthetic assessment benchmark that includes 10,000 images collected from the Flickr website with a size of 256 $\times$ 256 RGB image. Each image has an overall aesthetic rating provided by 5 different raters. The rating ranges from 1 to 5, with 5 being the most aesthetically pleasing rating. In this dataset, each image also contains 11 aesthetic attributes labeled as Negative, Null and Positive, as shown in Table 4. Similarly, we trained according to the official dataset partition [26], with 500 images for validation, 1,000 images for testing, and the rest for training.

Table 4: AADB dataset

Attribute	Negative	Null	Positive
Balancing element	1,750	3,876	2,832
Color harmony	797	1,819	5,842
Content	2,327	935	5,196
Depth of field	1,734	4,248	2,476
Lighting	2,788	2,212	3,458
Motion blur	700	7,361	397
Object emphasis	3,256	1,214	3,988
Repetition	0	6,775	1,683
Rule of thirds	2,699	3,250	2,509
Symmetry	0	7,687	771
Vivid color	3,378	1,864	3,216

4.2 Evaluation Metric

Following the setting of the multimodal false information detection task, we adopt four commonly used evaluation metric: Accuracy (*Acc*), precision (*P*), recall (*R*), and F1 score (*F1*).

4.3 Compare Models

To validate the effectiveness of ASMFD, we compare it with multiple multimodal false information detection models. Here are the detailed descriptions of these baseline models: a) TextGCN [36]. It represents text as a graph structure and uses graph convolutional networks for feature extraction and classification; b) att-RNN [12]. Using the Recurrent Neural Networks (RNNs) with attention mechanism to fuse multimodal features such as image features, text features, and social context features for rumor detection; c) EANN [16]. Encoding event invariant features is beneficial for detecting false in-formation in new events; d) MVAE [37]. Detecting fake news using bimodal variational autoencoder and a binary classifier; e) SpotFake+ [38]. Extracting text features using XLNet, extracting image features using VGG, and predicting information authenticity through concatenated features; f) SAFE [15]. By using pre-trained model to convert images into text and calculate their similarity to detect false information; g) CAFE [11]. Using cross modal alignment mechanism to map different modal features to the same shared space and evaluate the inconsistency of different modal features; h) HMCAN [10]. Extracting multimodal features using BERT and ResNet50, and aggregating features based on contextual attention networks.

4.4 Implementation Details

The proposed model is trained, tested, and evaluated inside the AutoDL platform. We chose the NVIDIA GeForce RTX 4090 as the experiment machine and used Python to write the experiment code. In stage one, we use the CLIP ViT-B/32 [19] as the backbone network. We trained two models separately for Chinese and English, using the training set from the corresponding language datasets. In Eq. (7), we respectively set λ as 0.1 and 0.15 for Chinese and English datasets. In IAA task, we use the Glorot uniform [39] to initialize the weights. We use Adam and set the learning rate is 0.001. The weight decay of 10^{-4}. In stage two, for English and Chinese, we choose the W2NER [40] and

RICON [41] models as NER method. In stage two, we finetuned each dataset separately and test the performance of the finetuned model. Batch size is 64 across all dataset splits.

5 Results and Discussions

In this section, we conducted experiments on four publicly available datasets to evaluate the effectiveness of ASMFD in multimodal false information recognition. Specifically, we propose the following four research questions (RQs) to guide the experiments.

RQ1: Does ASMFD perform better than baseline and state-of-the-art methods in identifying multimodal false information?

RQ2: What are the effects of consistency and inconsistency features on the final performance of the ASMFD?

RQ3: What impact will the performance of the aesthetic evaluation module have?

RQ4: Does the setting of the auxiliary task help improve model performance?

5.1 Overall Performance (RQ1)

In the previous sections, we provided a detailed introduction to the proposed ASMFD model, four common datasets for multimodal false information detection, and performance evaluation metrics. To answer RQ1, we conducted a comprehensive experiment on four datasets: GossipCop, PolitiFact-2, Twitter, and Weibo. Tables 5–8 show the performance of all models on the corresponding dataset.

Table 5: Comparison of performance on the GossipCop dataset

Models	Acc	P	R	F1
EANN	0.844	0.83	0.844	0.837
SpotFake+	0.753	0.848	0.753	0.798
SAFE	0.810	0.572	0.462	0.671
CAFE	0.804	**0.899**	0.804	0.849
HMCAN	0.846	0.832	0.845	0.839
ASMFD (ours)	**0.862**	0.853	**0.871**	**0.857**

Table 6: Comparison of performanc on the PolitiFact-2 dataset

Models	Acc	P	R	F1
EANN	0.852	0.857	0.852	0.854
SpotFake+	0.741	0.76	0.742	0.751
SAFE	0.857	0.871	0.856	0.863
CAFE	0.895	0.896	0.895	0.895
HMCAN	0.912	0.913	0.912	0.912
ASMFD (ours)	**0.964**	**0.966**	**0.964**	**0.965**

Table 7: Comparison of performance on the Twitter dataset

Models	Acc	P	R	F1
att_RNN	0.664	0.669	0.672	0.670
EANN	0.715	0.724	0.765	0.744
SpotFake+	0.771	0.777	0.769	0.773
SAFE	0.784	0.778	0.787	0.782
CAFE	0.806	0.807	0.801	0.804
HMCAN	0.897	**0.912**	0.875	0.893
ASMFD (ours)	**0.906**	0.903	**0.910**	**0.906**

Table 8: Comparison of performance on the Weibo dataset

Models	Acc	P	R	F1
TextGCN	0.726	0.741	0.851	0.792
att_RNN	0.772	0.848	0.753	0.798
EANN	0.788	0.786	0.791	0.788
MVAE	0.740	0.728	0.722	0.730
SpotFake+	0.782	0.791	0.804	0.797
SAFE	0.823	0.818	0.897	0.856
CAFE	0.840	0.841	0.804	0.822
HMCAN	0.885	0.832	0.845	0.838
ASMFD (ours)	**0.891**	**0.853**	**0.871**	**0.862**

We first experimented with five comparative models in English news with long content, and found that ASMFD outperformed existing methods in accuracy, recall, and F1 scores. Especially in the PolitiFact-2 dataset, it achieved an accuracy of 0.964, and other detailed results are shown in Tables 5 and 6. However, in the GossipCop dataset, the accuracy of ASMFD is slightly lower than that of the CAFE model, but there is a significant improvement in the other three indicators. We believe that this is due to the unbalance distribution of real and false news samples in the GossipCop dataset.

Then, we conducted experimental analysis on six comparative models in the English social media Twitter dataset, and the results are shown in Table 7. ASMDF also has the best performance compared to other models, exceeding 0.9 in all indicators. This proves that the algorithm proposed in this article can be applied not only to long text news, but also to short text information on social media.

Finally, we experimented with all comparative models on the Chinese dataset Weibo, and the results are shown in Table 8. Due to the powerful ability of pre trained models, all methods achieve an accuracy of over 0.7. In addition, the fine-grained attention mechanism used by HMCAN is superior to SAFE, CAFE and SpotFake+, which only consider the global characteristics of single mode attention weighting. Compared to HMCAN, the model proposed in this paper further refines the fine-grained attention mechanism, considering high correlation features while designing an aggregation method for label sensitive low correlation features, achieving the best accuracy of 0.891.

5.2 Study on Consistency Feature (RQ2)

To answer RQ2, we designed three variant models of ASMDF and conducted experiments on two datasets:

- w/o consistent: Only passing the inconsistent feature to detection module.
- w/o inconsistent: Only passing the consistent feature to detection module.
- w/o interaction: Using the traditional attention to fuse the consistent feature in-stead of the feature interaction module and feature aggregation module.

As shown in Table 9, the performance of w/o interaction is the best in both datasets. This indicates that our operations in feature interaction is necessary for detecting the false information. We find that w/o inconsistent has higher accuracy than w/o consistent. This shows that high correlation features are more important than low correlation features. But the performance of w/o inconsistent has decreased, proving that inconsistent feature is effective in improving model performance.

Table 9: Accuracy of ASMDF variant models on Weibo and PolitiFact-2

Variant model	PolitiFact-2	Weibo
w/o consistent	0.925	0.883
w/o inconsistent	0.943	0.885
w/o interaction	0.895	0.875
ASMFD	0.964	0.891

5.3 Study on Image Aesthetic Assessment (RQ3)

In proposed ASMFD, we use the pretrained image aesthetic attribute scorer as the threshold to obtain the consistent and inconsistent feature. This is a first exploratory work to introduce aesthetic assessment task into false information detection. To answer RQ3, we constructed two experiments to analyze the effectiveness of introducing this task:

- EX1: Using models which have different performance on the AADB dataset.
- EX2: Comparative analysis of the impact of using all attributes and selecting four attributes in this work on model performance.

In EX1, we trained a low performance model using the aesthetic attribute scorer in stage one, which has lower Spearman's rank correlation coefficient than the models used in ASMFD. The performance is shown in Table 10 labeled EX1 (low). We found that even though the performance of the aesthetic evaluation model has decreased significantly, its impact on false detection tasks is minimal. We speculate that this may be due to some differences in the goal between the two tasks, and not all aesthetic attributes are related to false information detection. Therefore, we continued with EX2.

In EX2, we use all attributes in AADB to train the aesthetic attribute scorer. The performance is shown in Table 10 labeled EX2 (all). We found that the scorer trained with all attributes showed an improvement in Spearman's rank correlation coefficient, but its performance decreased in false information detection task, further confirming our hypothesis: Not all aesthetic attributes in aesthetic assessment task are related to images in false information detection. To better utilize aesthetic

attributes, we will also explore the role of image aesthetic attributes in false information detection in future work.

Table 10: Spearman's rank correlation coefficient (Sc) of EX1 and EX2 on AADB and accuracy (Acc) of EX1 and EX2 on Weibo and Twitter datasets

SubDataset	AADB (Sc)	Twitter (Acc)	Weibo (Acc)
EX1 (low)	0.6782	0.904	0.885
EX2 (all)	0.7067	0.897	0.882
ASMFD	0.7041	0.906	0.891

5.4 Study on Auxiliary Task Settings (RQ4)

In stage two, to enable the ASMFD model to pay more attention to inconsistent features, we specifically designed an auxiliary task and set the label of the auxiliary task based on the ground-truth label. As shown in Table 11, we conducted some experiments on the setting of auxiliary labels and whether to use auxiliary tasks to answer RQ4.

Table 11: Accuracy of auxiliary task settings on Weibo and Twitter

Settings	Twitter	Weibo
w/o auxiliary task	0.898	0.876
Auxiliary label with [0.5, 0]	0.902	0.884
Auxiliary label with [1, 0]	0.617	0.726
ASMFD	0.906	0.891

As shown in Table 11, after removing the auxiliary task, the performance of the model decreased by 0.08 and 0.15 on Twitter and Weibo datasets, respectively. This indicates that using the auxiliary task can help improve model performance. Next, we explored the setting of different auxiliary labels. In ASMFD, our label mapping is [0.7, 0]. It can be observed that when the label is set to [0.5, 0], the performance of the model only slightly decreases, but when the label is set to [1, 0], the performance of the model will significantly decrease. We believe that the reason for this phenomenon is that although non consistent features can help the model improve in detection tasks, they cannot be independently used as features for final discrimination. In future work, we will further explore the specific roles of consistency and inconsistency features.

6 Conclusions

In this work, we propose a multi-stage detection framework ASMFD for multi-modal false information detection, which is based on image aesthetic similarity to segment and explore the consistency and inconsistency features of images and texts. To our knowledge, this is the first work to introduce image aesthetic evaluation into multimodal false information detection tasks. We introduce image aesthetic similarity as a consistency segmentation threshold in the multimodal interaction process. While focusing on high similarity consistency features, we also consider inconsistent information with low similarity. This method is more suitable for identifying false information on social platforms as

it allows for the dynamic selection of dependent features for judgment. Experiments on four public datasets have shown that our proposed ASMFD can achieve optimal performance across all baselines. More detailed experiments have also demonstrated that the attribute similarity of image aesthetics can improve the performance of the model to a certain extent.

However, our proposed model has some limitations: Our experiments found that the current setting of image aesthetic evaluation tasks does not perfectly fit the field of multimodal false information recognition; not all aesthetic attributes are helpful for this task. In future work, we will construct an image aesthetic attribute evaluation dataset that is more suitable for the evaluation criteria in this field based on the characteristics of news and social media platform information to further enhance the applicability and generalization of this framework in multimodal false information detection tasks. At the same time, we will further consider algorithm complexity and platform transferability, enhance the supporting role of algorithms in decision-making scenarios, and improve their usability in practical scenarios [42–45].

Acknowledgement: All authors sincerely thank all organizations and institutions that have provided data and resources.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: W. J. Fan; data collection: W. J. Fan; analysis and interpretation of results: W. J. Fan, Z. W. Shi; draft manuscript preparation: W. J. Fan, Z. W. Shi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, W. J. Fan, upon reasonable request.

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

References

[1] W. Fan, Y. Wang, and H. Hu, "Mimicking human verification behavior for news media credibility evaluation," *Appl. Sci.*, vol. 13, no. 17, pp. 9553, 2023. doi: 10.3390/app13179553.

[2] V. S. Devi and S. Kannimuthu, "Author profiling in code-mixed WhatsApp messages using stacked convolution networks and contextualized embedding based text augmentation," *Neural Process. Lett.*, vol. 55, no. 1, pp. 589–614, 2023. doi: 10.1007/s11063-022-10898-3.

[3] D. Orso, N. Federici, R. Copetti, L. Vetrugno, and T. Bove, "Infodemic and the spread of fake news in the COVID-19-era," *Eur J. Emerg. Med.*, vol. 27, no. 5, pp. 327–328, 2020. doi: 10.1097/MEJ.0000000000000713.

[4] F. Sufi, "Social media analytics on russia-ukraine cyber war with natural language processing: Perspectives and challenges," *Information*, vol. 14, no. 9, pp. 485, 2023. doi: 10.3390/info14090485.

[5] W. Fan and Y. Wang, "Cognition security protection about the mass: A survey of key technologies," (in Chinese), *J. Commun. University China (Sci. Technol.)*, vol. 29, no. 3, pp. 1–8, 2022. doi: 10.16196/j.cnki.issn.1673-4793.2022.03.009.

[6] F. Miró-Llinares and J. C. Aguerri, "Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'," *Eur. J. Crminol.*, vol. 20, no. 1, pp. 356–374, 2023. doi: 10.1177/1477370821994059.

[7] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc.WWW '11*, New York, NY, USA, 2011, pp. 675–684. doi: 10.1145/1963405.1963500.
[8] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th Int. Conf. Data Min.*, Dallas, TX, USA, 2013, pp. 1103–1108. doi: 10.1109/ICDM.2013.61.
[9] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on Sina Weibo," in *Proc. MDS '12*, New York, NY, USA, 2012, pp. 1–7. doi: 10.1145/2350190.235020.
[10] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. SIGIR '21*, New York, NY, USA, 2021, pp. 153–162.
[11] Y. Chen *et al.*, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. WWW '22*, New York, NY, USA, 2022, pp. 2897–2905.
[12] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proc. MM '17*, New York, NY, USA, 2017, pp. 795–816.
[13] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Trans. Multimed.*, vol. 19, no. 3, pp. 598–608, 2017. doi: 10.1109/TMM.2016.2617078.
[14] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *Proc. MM '19*, New York, NY, USA, 2019, pp. 1942–1951.
[15] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-Aware multi-modal fake news detection," in *Proc. PAKDD 2020*, Singapore, 2020, pp. 354–367. doi: 10.1007/978-3-030-47436-2_27.
[16] Y. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. KDD '18*, New York, NY, USA, 2018, pp. 849–857.
[17] P. Qi *et al.*, "Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues," in *Proc. MM '21*, New York, NY, USA, 2021, pp. 1212–1220.
[18] S. Huang, W. Fu, Z. Zhang, and S. Liu, "Global-local fusion based on adversarial sample generation for image-text matching," *Inf. Fusion*, vol. 103, no. 8, pp. 102084, 2024. doi: 10.1016/j.inffus.2023.102084.
[19] A. Radfor *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.
[20] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," in *Proc. BigMM*, Singapore, 2019, pp. 39–47.
[21] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 1, pp. 557–565, 2021. doi: 10.1609/aaai.v35i1.16134.
[22] Y. Wang, F. Ma, H. Wang, K. Jha, and J. Gao, "Multimodal emergent fake news detection via meta neural process networks," in *Proc. KDD '21*, New York, NY, USA, 2021, pp. 3708–3716.
[23] C. Jia *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
[24] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multimodal fake news detection via CLIP-guided learning," in *Proc. ICME*, Brisbane, Australia, 2023, pp. 2825–2830.
[25] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources," in *Proc. CVPR*, 2022, pp. 14940–14949.
[26] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Comput. Vis.-ECCV 2016*, Amsterdam, The Netherlands, 2016, pp. 662–679.
[27] C. Kang, G. Valenzise, and F. Dufaux, "EVA: An explainable visual aesthetics dataset," in *Proc. ATQAM/-MAST'20*, New York, NY, USA, 2020, pp. 5–13.
[28] L. Hou, C. P. Yu, and D. Samaras, "Squared earth movers distance loss for training deep neural networks on ordered-classes," in *Proc. NIPS 2017*, Long Beach, CA, USA, 2017, pp. 1–6.
[29] L. Li, H. Zhu, S. Zhao, G. Ding, H. Jiang and A. Tan, "Personality driven multi-task learning for image aesthetic assessment," in *Proc. ICME*, Shanghai, China, 2019, pp. 430–435.

[30] L. Celona, M. Leonardi, P. Napoletano, and A. Rozza, "Composition and style attributes guided image aesthetic assessment," *IEEE Trans. Image Process.*, vol. 31, no. 11, pp. 5009–5024, 2022. doi: 10.1109/TIP.2022.3191853.
[31] B. Pan, S. Wang, and Q. Jiang, "Image aesthetic assessment assisted by attributes through adversarial learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 679–686, 2019. doi: 10.1609/aaai.v33i01.3301679.
[32] L. Li, J. Duan, Y. Yang, L. Xu, Y. Li and Y. Guo, "Psychology inspired model for hierarchical image aesthetic attribute prediction," in *Proc. ICME*, Taipei, Taiwan, 2022, pp. 1–6.
[33] L. Li *et al.*, "Theme-aware visual attribute reasoning for image aesthetics assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4798–4811, Sep. 2023. doi: 10.1109/TCSVT.2023.3249185.
[34] S. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, Jun. 2020. doi: 10.1089/big.2020.0062.
[35] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou and Y. Kompatsiaris, "Detection and visualization of misleading content on Twitter," *Int. J. Multimed. Inf. Retr.*, vol. 7, no. 1, pp. 71–86, Mar. 2018. doi: 10.1007/s13735-017-0143-x.
[36] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 7370–7377, 2019. doi: 10.1609/aaai.v33i01.33017370.
[37] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. WWW '19*, New York, NY, USA, 2019, pp. 2915–2921.
[38] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty and P. Kumaraguru, "SpotFake+: A multimodal framework for fake news detection via transfer learning (Student Abstract)," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 10, pp. 13915–13916, 2020. doi: 10.1609/aaai.v34i10.7230.
[39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 249–256.
[40] J. Li *et al.*, "Unified named entity recognition as word-word relation classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, pp. 10965–10973, 2022. doi: 10.1609/aaai.v36i10.21344.
[41] Y. Gu, X. Qu, Z. Wang, Y. Zheng, B. Huai and N. J. Yuan, "Delving deep into regularity: A simple but effective method for Chinese named entity recognition," arXiv preprint arXiv:2204.05544, 2022.
[42] W. Fan and Y. Wang, "Multidisciplinary fusion perspective analysis method for false information recognition," *Adv. Electr. Comput. Eng.*, vol. 24, no. 1, pp. 61–70, 2024. doi: 10.4316/AECE.2024.01007.
[43] W. Lin *et al.*, "A deep neural collaborative filtering based service recommendation method with multi-source data for smart cloud-edge collaboration applications," *Tsinghua Sci. Technol.*, vol. 29, no. 3, pp. 897–910, 2024. doi: 10.26599/TST.2023.9010050.
[44] H. Liu, L. Qi, S. Shen, A. A. Khan, S. Meng and Q. Li, "Microservice-driven privacy-aware cross-platform social relationship prediction based on sequential information," *Softw. Pract. Exp.*, vol. 54, no. 1, pp. 85–105, 2024. doi: 10.1002/spe.3240.
[45] Y. Shen, S. Shen, Q. Li, H. Zhou, Z. Wu and Y. Qu, "Evolutionary privacy-preserving learning strategies for edge-based IoT data sharing schemes," *Digit. Commun. Netw.*, vol. 9, no. 4, pp. 906–919, 2023. doi: 10.1016/j.dcan.2022.05.004.