IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 March 2025, accepted 8 May 2025, date of publication 19 May 2025, date of current version 27 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3571624

RESEARCH ARTICLE

FGMFN: Fine-Grained Multiscale Cross-Modal Sentiment Analysis in Advertisements

HAN WANG[1], PENG CHEN[2], AND XIANGYU DU[1]
[1]School of Design, Xuzhou University of Technology, Xuzhou 221008, China
[2]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

Corresponding author: Han Wang (wanghan@xzit.edu.cn)

This work was supported in part by the General Project of Philosophy and Social Science Research for Higher Education Institutions in Jiangsu Province under Grant 2024SJYB0834, and in part by the General Project of Educational Science Research at Xuzhou University of Technology under Grant YGJ2424.

ABSTRACT Cross-modal sentiment analysis in advertising has gained significant attention due to its potential in brand communication and consumer behavior analysis. However, traditional methods struggle to handle the multi-scale features and redundant objects in advertising images effectively, resulting in limited emotion recognition accuracy. To address the challenges of insufficient multi-scale features and target redundancy in multi-modal sentiment analysis of advertisements, we introduce a novel framework, Fine-Grained Multiscale Cross-Modal Feature Network (FGMFN). The model is designed to process multi-scale feature inputs, facilitate efficient sentiment fusion between images and text. FGMFN employs a multi-scale network to extract key features from advertising images, and uses visual features to guide the textual data representation. Additionally, to reduce textual ambiguity caused by strong intra-class similarity in advertising contexts, we introduce a multi-task learning approach combining image-text matching loss with image-text mutual information loss. This strategy narrows the gap between visual features and sentiment semantics, improving the model's generalization capabilities. Finally, we construct a fine-grained image-text sentiment analysis dataset (YTB-ADS), which, in contrast to traditional coarse-grained datasets with high intra-class similarity, better serves the specific needs of advertising sentiment analysis. Experimental results show that FGMFN outperforms existing methods on the YTB-ADS dataset, as well as on the publicly available Twitter-2015 and Twitter-2017 datasets, confirming the model's superior performance in advertising sentiment analysis tasks.

INDEX TERMS Multi-modal sentiment analysis, deep learning, feature extraction, cross modal attention.

I. INTRODUCTION

With the rapid advancements in internet and information technologies, the online advertising market has experienced unprecedented growth, particularly driven by social media and short-video platforms. Consequently, advertising formats have become more diverse and interactive. Static visual advancements, such as posters, have become a key channel for brand communication by integrating images and text [1], [2]. In this context, visual ads, as a central medium of communication, capture a large audience's attention and effectively influence their purchase intentions through concise and intuitive image-text designs [3]. Therefore, sentiment analysis in advertisements is particularly crucial for advertisers and marketers [4], [5].

Sentiment analysis is generally categorized into two types: unimodal and multimodal [5]. In contrast, multi-modal sentiment analysis provides a more comprehensive understanding of users' emotions by integrating information from multiple modalities. This is particularly important in evaluating advertising effectiveness, as unimodal analysis often fails to capture the complex emotional responses of viewers to advertisements [6], [7], [8]. Data from social networks and advertisements usually include multiple modalities, such as text and images. These modalities interact and complement each other, providing a rich platform for

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

emotional expression. However, previous multimodal fusion methods still encounter three significant challenges when applied to advertising sentiment analysis.

First, advertisements often contain numerous non-essential background elements, whereas images in traditional datasets typically emphasize key objects [10]. Traditional text-image sentiment analysis methods in related fields often overlook the need for redundant feature filtering, hindering the model's comprehension of advertisement content [9]. As a result, extracting salient features from advertising images becomes a significant challenge. Furthermore, previous methods fail to account for the multi-scale information inherent in advertisements, leading to underutilization of these features and, consequently, less precise emotion recognition. To tackle the multi-scale and target redundancy challenges in advertising images, we employ a multi-scale network [11] for salient feature extraction. To refine image-text fusion, we utilize image features to dynamically guide the textual representation, applying this methodology to emotion recognition for more flexible and adaptable inputs.

Second, in contrast to natural scenes, advertising scenes exhibit strong intra-class similarity. A single text may correspond to multiple negative samples that closely resemble real images, introducing ambiguity into the optimization objectives during training. To address this challenge, we propose a multi-task learning strategy. By integrating image-text matching loss [12] and image-text mutual information loss, this approach enhances the model's generalization ability, enabling it to learn related tasks more effectively.

Third, in traditional datasets, text typically exhibits lower intra-class similarity compared to advertising image-text datasets, making these datasets less suitable for sentiment analysis tasks in the advertising domain. To address this challenge, we developed the YTB-ADS dataset specifically for advertising image-text sentiment analysis, minimizing intra-class similarity and enhancing generalization in sentiment recognition tasks.

The main contributions of this paper can be summarized as follows:

- To address the challenges of insufficient multi-scale features and target redundancy in advertising multimodal sentiment analysis, we propose a novel Fine-Grained Multiscale Cross-Modal Feature Network (FGMFN). This cross-modal approach is specifically designed to process multi-scale feature inputs, while utilizing visual features to dynamically guide textual representations, thereby facilitating effective feature fusion for effective feature fusion.
- To mitigate textual feature ambiguity caused by high intra-class similarity in advertising images, we introduce both image-text matching loss and image-text mutual information loss. These losses serve to bridge the gap between visual features and emotional semantics, enhancing the alignment between the two modalities.
- We also introduce a new advertising image-text sentiment analysis dataset, YTB-ADS. Experimental results show that FGMFN performs competitively on both the custom-built YTB-ADS dataset and the publicly available Twitter-2015, Twitter-2017 datasets.

II. RELATED WORK

Previous studies have explored targeted advertisement recommendations for internet and television users by analyzing the correlation between advertisement content and social interactions [13], [14]. These studies utilized both visual and textual features of advertisements, along with user behavior data and click-through rates, to develop recommendation systems. Content-based multimedia feature analysis plays a crucial role in the design and production of multimedia content, serving as a foundational element in tailoring effective advertising strategies [15].

Recent advancements in advertisement emotion recognition have focused on the fusion and analysis of multimodal data to more effectively capture the emotional responses elicited by advertisement content. Traditional emotion recognition approaches primarily relied on single-modal feature extraction and basic fusion techniques. For instance, Zadeh et al. [16] introduced the Tensor Fusion Network (TFN), which facilitates the comprehensive modeling of unimodal, bimodal, and trimodal data. They also proposed the Memory Fusion Network (MFN) [17], which models relationships both within and across modalities by utilizing single-perspective and cross-perspective viewpoints. Majumder et al. [18] enhanced feature extraction by employing a hierarchical GRU model, which enables richer interactions among modalities. Building on the success of Transformer models in single-modal tasks, researchers have extended these methods to multimodal sentiment analysis. For example, Xu et al. [19] applied Cross-Modal Attention to align features across modalities. Despite these significant advancements, such methods often struggle to capture the complex and dynamic interactions of multimodal content in advertisements. As a result, they may fall short of accurately identifying the nuanced emotional responses of audiences, limiting their effectiveness in real-world sentiment analysis tasks in advertising contexts.

Recent research has increasingly focused on fine-grained emotion analysis to explore how different types of advertisements influence the emotional responses of audiences. For example, Zhu et al. [20] introduced the Image-Text Interaction Network (ITIN), which uses Convolutional Neural Network (CNN) [21] to extract visual features and Recurrent Neural Network (RNN) [22] to extract textual features. These features are then integrated through a joint architecture to capture the interactions between visual and textual modalities, enabling emotion prediction in social media posts. Yang et al. [23] proposed the TPMSA model, which improves emotion classification by leveraging large-scale unlabeled data for pretraining. Similarly, Ye et al. [24] developed the Sentiment-Aware Multimodal Pretraining (SMP) method, which integrates both textual and visual

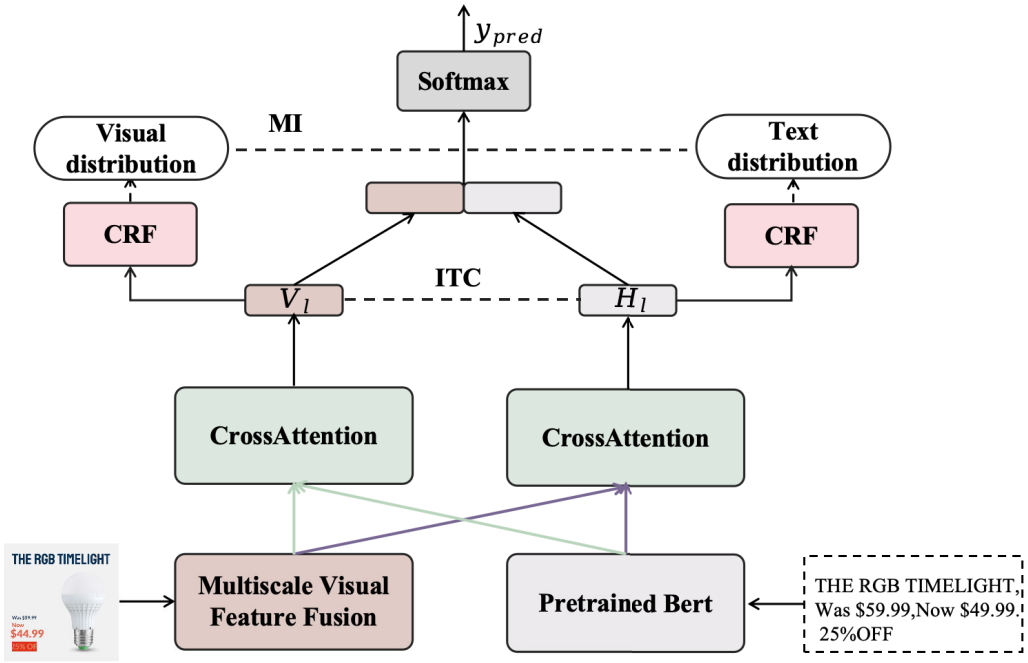


FIGURE 1. Overall architecture of the proposed FGMFN.

data to extract fine-grained emotional representations. These approaches provide valuable insights into the nuanced emotional responses evoked by advertisements, particularly by addressing the interplay of multimodal features and utilizing pretraining strategies to enhance emotion recognition.

Numerous methods have been developed in the field of cross-modal interactions to enhance the alignment and correlation of multimodal features. Tsai et al. [25] employed a cross-modal attention mechanism to align features across modalities, enabling more precise integration of multimodal data. Lu et al. [26] proposed an image-text consistency-driven approach that further explores inter-modal correlations, enhancing feature alignment. Xiao et al. [27] introduced the Multimodal Attention Graph Convolutional Network (MAGCN), which combines graph convolution and self-attention mechanisms to dynamically model relationships among modalities. To address fine-grained semantic alignment, Yu et al. [28] developed the Entity-Sensitive Attention Fusion Network (ESAFN), which explicitly models local interactions between images and text to capture nuanced relationships. Additionally, Xin et al. [29]incorporated multitask learning frameworks to supervise and optimize cross-modal alignment, leading to significant improvements in analysis accuracy. These approaches collectively advance the understanding of cross-modal interactions by addressing key challenges, such as feature alignment, semantic consistency, and dynamic modeling of inter-modal relationships. However, their application to complex multimodal scenarios, such as advertisement emotion recognition, requires further exploration to tackle domain-specific challenges, including audience response variability and the unique nature of advertisement content.

III. METHODOLOGY

In this section, we present the design details of the FGMFN for multimodal sentiment analysis. As shown in Figure 1, the structure of our proposed method comprises three main components. First, features are extracted independently from the image and text modalities. Next, a cross-attention mechanism facilitates multimodal fusion. Finally, multi-scale visual features are aligned with textual representations using a multi-task loss function, ensuring coherent integration and improved sentiment prediction.

A. UNIMODAL EMBEDDING

We provide an overview of the unimodal embedding strategy from two fundamental perspectives: visual embedding and text embedding.

1) VISUAL EMBEDDING

For input images, we employ a CNN model pre-trained on the ImageNet dataset [30], with its parameters fine-tuned for the current task [32]. The global feature representation of the image is defined as:

$$v^{(g)} = \text{CNN}\,(I, \theta_I) \tag{1}$$

where I represents the input image, $v^{(g)}$ denotes the global features extracted by the CNN, and θ_I refers to the parameters of the CNN.

2) TEXT EMBEDDING

Given the remarkable performance of the BERT model [33] in text comprehension, we leverage the pre-trained BERT to extract features from the text modality. Specifically, we use the vector corresponding to the first token from the final layer

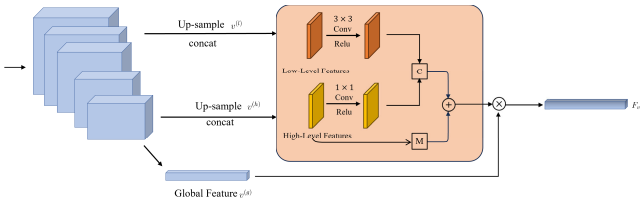


FIGURE 2. The CNN module utilizing multiscale visual feature fusion to integrate features from multiple receptive fields.

of the BERT model to represent the entire sentence, as defined by:

$$h_t = \text{BERT}\,(I_t, \theta_t) \tag{2}$$

where, I_t denotes the input text and θ_t represents the parameters of the BERT model.

B. MULTISCALE VISUAL FEATURE

While $v^{(g)}$ captures a substantial portion of the image's information, relying solely on global features poses notable limitations, particularly in the context of advertisement images. Unlike natural images, advertisements often contain numerous distinct objects, and global feature representations may introduce redundancy, thereby hindering the generation of optimal image embeddings. Moreover, advertisement images frequently exhibit strong multi-scale characteristics, with objects appearing at varying scales. Features extracted at a single scale are inadequate for effectively capturing such diversity. As the depth of convolutional layers increases, pooling operations in CNNs tend to discard critical information about smaller objects, leading to the omission of finer details from global features. Inspired by existing methods [32], we propose Multi-Scale Visual Feature Fusion to enhance image representation by integrating global features with local features extracted from multiple convolutional layers. Specifically, feature maps from the first three layers are upsampled and concatenated with deeper feature maps to form low-level feature representations. Concurrently, feature maps from the final two layers retain high-level semantic information. This process is formally defined as:

$$\begin{aligned}
\left\{v^{(m)}\right\}_{m=1}^{5}, v^{(g)} &= \text{CNN}\,(I, \theta_I) \\
\{F_m\}_{m=1}^{5} &= \text{Upsample}\left(\left\{v^{(m)}\right\}_{m=1}^{5}\right) \\
v^{(l)} &= \text{Cat}\,(F_1, F_2, F_3) \\
v^{(h)} &= \text{Cat}\,(F_4, F_5)
\end{aligned} \tag{3}$$

where, $v^{(m)}$ denotes the feature map produced by each layer of the convolutional network. The upsampling operation adjusts feature maps from various layers to a uniform spatial resolution, resulting in upsampled feature maps F_m. This ensures that all feature maps, whether from low-level or high-level layers, share consistent dimensions. Channel-wise concatenation, represented as Cat (x, y), merges feature vectors x and y along the channel dimension. Finally, we obtain the low-level features $v^{(l)}$, high-level features $v^{(h)}$ and the global features $v^{(g)}$.

To improve the representation of small targets in advertising images, we propose a multi-scale feature fusion network that extracts joint information from both low-level and high-level features. In the multi-scale feature fusion stage, we apply feature transformation on both sets of features. Initially, the low-level features are downsampled using a 3×3 convolution to align with the dimensions of the high-level feature maps. A Rectified Linear Unit (ReLU) activation function is then applied to introduce non-linearity, resulting in the transformed low-level feature, denoted as $\widehat{v}^{(l)}$. For the high-level features, a 1×1 convolution is applied to adjust the dimensions, followed by ReLU activation, producing $\widehat{v}^{(h)}$. These two transformed feature maps are subsequently concatenated along the channel dimension. To preserve critical information from the high-level features, we compute the mean of $v^{(h)}$ and add it to the joint low-level and high-level features through a residual connection. This process is expressed as follows:

$$v^{(lh)} = \text{Mean}\left(v^{(h)}\right) \oplus \text{Cat}\left(\widehat{v}^{(l)}, \widehat{v}^{(h)}\right) \tag{4}$$

where Mean (x) denotes the mean of the vector x, $\widehat{v}^{(l,h)}$ represents the joint feature representation of the high-level and low-level image features, and $\oplus$ denotes vector addition.

C. CROSS-ATTENTION GUIDED FEATURE

After extracting the image and text features from their respective encoders, a multi-head attention mechanism is used to construct a cross-modal encoder that effectively handles the complex interactions between the image and text modalities. The mechanism is defined as follows:

$$\begin{aligned}
&\text{MultiheadAttn(Query, Key)} \\
&= \text{Concat}\,(\text{head}_1, \text{head}_2, \cdots, \text{head}_{num})\, W_o \\
&\text{head}_c = \text{attention}(Q_c, K_c, V_c) \\
&\text{attention}(Q_c, K_c, V_c) = \text{softmax}\left(\frac{Q_c K_c^T}{\sqrt{d}}\right) V_c
\end{aligned} \tag{5}$$

The cross-modal encoder consists of multiple stacked sub-encoders, each implementing a cross-modal attention mechanism (CrossAttn). This mechanism involves bidirectional attention computations: text features serve as the query, while image features function as both the key and value, and vice versa. Formally, this process is expressed as:

$$\begin{aligned}
\widehat{h}_l^k &= \text{Cross Attn}\left(h_i^{k-1}, \left\{v_1^{k-1}, v_2^{k-1}, \ldots, v_m^{k-1}\right\}\right) \\
\widehat{v}_l^k &= \text{Cross Attn}\left(v_i^{k-1}, \left\{h_1^{k-1}, h_2^{k-1}, \ldots, h_n^{k-1}\right\}\right)
\end{aligned} \tag{6}$$

where $\widehat{h}_l^k$ and $\widehat{v}_l^k$ represent the cross-modal features generated by the attention mechanism. In the text modality, h_i^{k-1} corresponds to the text features from the previous layer, while $\left\{v_1^{k-1}, v_2^{k-1}, \ldots, v_m^{k-1}\right\}$ represents image context features from the previous layer. In the image modality, the roles are

reversed: the image features serve as the query, and the text features act as both the key and value.

D. OPTIMIZATION FUNCTION

1) TASK LOSS

The fused features z_s is obtained by concatenating v_l and h_l, z_s is passed through a multi-layer perceptron with an activation function, followed by a softmax classifier to generate the final prediction, as defined in Equation.7.

$$\hat{y} = \text{Softmax}(\tanh(FC(z_s))) \tag{7}$$

where $\hat{y}$ represents the final output of the model. For the multimodal emotion recognition task, formulated as a classification problem, the cross-entropy loss function is used to quantify the error between the predicted probability distribution $\hat{y}$ and the ground truth labels y:

$$\mathcal{L}_{\text{task}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right) \tag{8}$$

where n denotes the batch size and C represents the total number of classes. The term $y_{i,c}$ represents the binary ground truth label for sample i in category c, while $\hat{y}_{i,c}$ is the predicted probability that sample i belongs to class c.

2) IMAGE-TEXT CONTRASTIVE LEARNING LOSS

To address the bias and instability commonly observed in multimodal embedding spaces, the Image-Text Contrastive Loss (ITC Loss) [12] is introduced within the cross-sample fusion module. This mechanism plays a crucial role in ensuring robust multimodal representations by capturing latent semantic relationships between images and text. For instance, ITC aligns features corresponding to visual elements with their associated textual descriptions in advertisements. The goal is to model the interdependencies between images and text, rather than merely matching descriptive labels to visual content. The ITC loss fosters semantic alignment by drawing features of images and text from the same data closer together in the feature space, while pushing apart features from different samples.

Formally, let L_n and I_n denote the normalized text and image features of the n-th sample in a training batch. ITC seeks to minimize the cosine similarity between L_n and I_n (positive pairs) while maximizing similarity between mismatched pairs such as L_n and I_m, where I_m corresponds to a different sample in the batch. Given a batch size of N, the ITC loss is defined as:

$$\begin{aligned} l_{itc} &= \frac{1}{2}(l_1 + l_2) \\ l_1 &= -\frac{1}{N}\Sigma_{n=1}^{N}\log\frac{\exp\left(LI^T/e^{\tau}\right)}{\Sigma_{j=1}^{N}\exp\left(LI^T/e^{\top}\right)} \\ l_2 &= -\frac{1}{N}\Sigma_{n=1}^{N}\log\frac{\exp\left(IL^T/e^{\tau}\right)}{\Sigma_{j=1}^{N}\exp\left(IL^T/e^{\tau}\right)} \end{aligned} \tag{9}$$

where, τ is a learnable temperature parameter that scales the logits to adjust the sharpness of the similarity distribution.

3) MUTUAL INFORMATION LOSS

In multimodal tasks, long visual contexts can lead to an overemphasis on visual features, potentially diminishing the contribution of textual information during gradient backpropagation. To address these challenges, we introduce a Mutual Information Loss into the model, which minimizes the KL divergence between two probability distributions. The Mutual Information Loss fosters alignment between these distributions, ensuring that both the visual and textual modalities contribute equally and meaningfully to the overall model learning process.

One is derived from textual representations H_l encoded by a BERT model, and the other from visual representations V_l obtained through multi-scale encoding. These are then fed into the Conditional Random Field(CRF) [34] layer. For a label sequence $y = (y_1, y_2, \ldots, y_n)$, taking the text representation H_l as an example, we define the probability of the label sequence y as:

$$\begin{aligned} s(H_l, y) &= \sum_{j=0}^{n} M_{y_j, y_j+1} + \sum_{j=1}^{n} P_{j,y_j} \\ p(y \mid H_l) &= \text{Softmax}(s(H_l, y)) \end{aligned} \tag{10}$$

where $M_{y_j,y_{j+1}}$ represents the randomly initialized transition matrix from label y_j to y_{j+1}, while P_{j,y_j} is the emission matrix for label y_j, derived from a linear transformation of H_l.

The mutual information loss is mathematically equivalent to computing the cross-entropy loss between these two distributions:

$$\begin{aligned} L_{mi} &= KL(p(y \mid H_l) \| p(y \mid V_l)) \\ &= \sum_{y \in Y} p(y \mid V_l)\log(p(y \mid H_l)) \end{aligned} \tag{11}$$

where $p(y \mid V_l)$ and $p(y \mid H_l)$ represent the probability distributions over the class labels y, conditioned on the visual representations V_l and the textual representations H_l.

Finally, by combining the task loss L_{task}, image-text contrastive loss L_{itc}, and mutual information loss L_{mi}, along with the trade-off parameters α and β, the final objective function is expressed as:

$$L = L_{\text{task}} + \alpha L_{\text{itc}} + \beta L_{\text{mi}} \tag{12}$$

IV. EXPERIMENTS

In this section, we first describe the process of constructing the manually annotated YouTube advertisement dataset and outline the specific guidelines for sentiment analysis. We then highlight the advantages of the proposed FGMFN method through a comprehensive analysis supported by comparative experiments and ablation studies, providing insights into its crucial role in enhancing the model's overall effectiveness.

A. DATASETS

Most existing public datasets are not specifically designed for the advertising domain and suffer from class imbalance. To overcome these limitations, we developed a high-quality

TABLE 1. Statistics of the YTB-ADS dataset (after resampling).

Modality	Positive	Negative
Multimodal sentiment	11745	11372

dataset named YTB-ADS. A brief overview of the datasets is given below.

We collected 5,000 advertising videos from YouTube using online crawling tools, covering categories such as food and beverages, movie trailers, healthcare, and apparel. During video collection, we randomly selected advertisements from each category to reduce dataset bias and improve the generalizability of multimodal representations. For each video, we performed data augmentation by randomly selecting five frames to generate image-text advertisements with consistent labels. Only frames containing more than five words of text were retained. After manual annotation, we balanced the distribution of positive and negative sentiment categories through random resampling. Table 1 and Table 2 display the statistics and content related to the YTB-ADS dataset.

For the YTB-ADS dataset, we conducted a binary classification task by collecting user comments related to each advertisement on YouTube. To assess the sentiment of these comments, we utilized the SentiStrength tool [35], a sentiment analysis software, to compute sentiment intensity scores. These scores range from -5 to 5, with higher values indicating more positive sentiment and lower values indicating more negative sentiment. Advertisements with an average sentiment intensity score greater than 2.5 were categorized as exhibiting positive sentiment, while those with scores of -2.5 or below were categorized as exhibiting negative sentiment. To ensure sentiment label accuracy, we invited five students specializing in sentiment analysis—three master's students and two undergraduates—as reviewers. Additionally, one master's student was assigned as the verifier, responsible for making final decisions on entries marked as ''uncertain'' by the reviewers. Finally, all reviewers discussed ambiguous cases collaboratively to reach a consensus.

To ensure the reproducibility of our methods, we also conducted independent experiments using the Twitter-2015 and Twitter-2017 datasets [36]. Both datasets serve as valuable resources for sentiment analysis. The Twitter-2015 dataset consists of tweets on 15 distinct topics, each labeled as either positive or negative, making it particularly suitable for binary sentiment analysis tasks. In contrast, the Twitter-2017 dataset covers 17 topics and includes a more diverse range of sentiment annotations, supporting multi-class sentiment classification.

B. IMPLEMENTATION DETAILS

To ensure result comparability and reliability, all experiments were conducted under consistent conditions. The experiments utilized Python 3.10 and the PyTorch 1.11.0 framework, and were trained on an RTX 3090 GPU. For optimization, the AdamW optimizer was employed with a weight decay of 0.01. To identify the optimal hyperparameter configuration, a random search was first conducted over a logarithmically uniform distribution of learning rates ranging from 1×10^{-5} to 1×10^{-3}. Additionally, the auxiliary training loss hyperparameters were inspired by MISA [48], and a grid search was conducted in Section IV-F to determine the optimal values for α and β. Table 3 provides the detailed hyperparameter settings for the Twitter-2015, Twitter-2017, and YTB-ADS datasets.

C. BASELINES

In this study, we evaluated the effectiveness of our proposed multimodal sentiment analysis model by comparing it against several representative baseline models. These baseline models are grouped into three categories:

Image-based methods:

- **Res-Target** [37] a visual feature extraction technique employed in sentiment analysis tasks, utilizing deep residual networks (ResNet). It effectively captures subtle emotional nuances from visual data, enabling a refined analysis of image-based sentiment cues.
- **VGG16** [38] is a deep convolutional neural network renowned for its ability to learn complex image features. It is extensively used in multimodal sentiment analysis.

Text-based methods:

- **BiLSTM** [39] incorporates contextual information from both directions along the temporal axis, excelling in the capture of long-term dependencies in text sequences.
- **TextCNN** [40] is a text classification method that leverages convolutional neural networks to extract local features from text through convolutional kernels of various sizes.
- **RoBERTa** [41] is a Transformer-based, pre-trained language model that refines the BERT architecture by enhancing its training process.

Multimodal methods:

- **MIMN** [42] introduces a model comprising two interactive memory networks, fostering deep interaction between visual and textual modalities. This interaction facilitates the capture of semantic information directly relevant to the task at hand, enhancing multimodal understanding and improving performance.
- **ESAFN** [28] is designed to optimize attention mechanisms for entities and the fusion network, refining the interactions within each modality.
- **VilBERT** [43] is a pre-trained vision-language model that captures intricate semantic relationships between visual and linguistic inputs. By jointly processing these modalities, VilBERT enhances the model's ability to predict emotional tendencies, making it highly effective in multimodal sentiment tasks.

TABLE 2. Example of YTB-ADS dataset content.

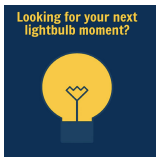
Image				
Text	the RGB TIMELIGHT, Was $59.99, Now $49.99. 25% OFF	Looking for your next lightbulb moment	THIS APP WILL RUIN YOUR LIFE	What's more beautiful than golden hour?
Label	Positive	Positive	Negative	Negative

TABLE 3. Hyperparameter settings for Twitter-2015, Twitter-2017 and YTB-ADS datasets datasets.

Setting	Twitter-2015	Twitter-2017	YTB-ADS
Epoch	100	20	100
Batch Size	32	64	64
Optimizer	Adam	Adam	Adam
Learning Rate	2×10^{-5}	2×10^{-5}	5×10^{-5}
Temperature	0.1	0.1	0.1
Dropout Rate	0.1	0.2	0.5
α	0.8	0.1	0.5
β	0.05	0.05	0.05

- **TomBERT** [36] is a sophisticated multimodal BERT architecture with a three-branch design, specifically tailored for target-oriented tasks.
- **ARFN** [44] employs the Yolov5 algorithm to precisely localize regions of interest in images, combining textual and visual features through a multimodal interaction layer.
- **AoM** [45] addresses modality noise interference by jointly modeling fine-grained sentiment aggregation and aspect-level semantic information.
- **LXMERT-MMSA** [46] employs a cross-modality attention mechanism and Transformer encoders to deeply fuse text and image features, enhancing sentiment polarity prediction.
- **MTVAF** [47] mitigates modality gaps and irrelevant visual noise by transforming image data into text descriptions and employing dynamic attention for effective cross-modal fusion.

D. QUANTITATIVE RESULTS AND ANALYSIS

1) OVERALL PERFORMANCE

In Table 4, we compare our method with the baselines on the Twitter-2015, Twitter-2017, and YTB-ADS datasets to demonstrate superior performance.

For the Twitter-2015 dataset, unimodal models, such as Res-Target and VGG16, performed relatively poorly in both Accuracy and Macro-F1, especially in sentiment analysis tasks. Relying solely on image information was inadequate for capturing emotional cues effectively. For instance, VGG16 achieved only 64.32% accuracy and a Macro-F1 score of 60.09%. These results suggest that, while images can serve as useful supplementary information in sentiment analysis, they alone are insufficient for reliably predicting sentiment polarity in tweets. In contrast, text-based methods, such as RoBERTa, demonstrated stronger performance on this dataset, achieving an accuracy of 74.15%. This highlights the advantages of leveraging large-scale, pre-trained language models for textual analysis. However, even a powerful text model like RoBERTa could not fully capture the complex semantic relationships between images and text. Multimodal methods, which combine both image and text data, showed superior performance. Specifically, FGMFN achieved an accuracy of 79.92% and a Macro-F1 score of 75.28%, significantly outperforming other multimodal models.

For the Twitter-2017 dataset, A similar trend was observed. Image-based models continued to perform poorly, while text-based models, such as RoBERTa, maintained a distinct advantage in sentiment prediction tasks. FGMFN continued to demonstrate strong performance on the Twitter-2017 dataset, experiencing a slight decline compared to its results on the Twitter-2015 dataset, yet still outperforming other multimodal baselines. Furthermore, it is worth noting that, in terms of Macro-F1, ARFN slightly outperformed FGMFN. We attribute this to ARFN's ability to extract adjective-noun pairs from images as external knowledge. The integration of this supplementary semantic information likely provides ARFN with an edge, enabling it to capture emotional cues within the Twitter-2017 dataset more accurately.

For the YTB-ADS Dataset, Image-based models continued to show limited effectiveness, with VGG16 achieving an accuracy of only 57.82%, while Res-Target performed even worse, underscoring the limited auxiliary role of images in this dataset. In contrast, the text model RoBERTa significantly outperformed the image models, showcasing the strength of text-based approaches in sentiment analysis. FGMFN maintained its superior performance on the YTB-ADS dataset, achieving an impressive accuracy of 70.54%, far surpassing all other methods. This result underscores the effectiveness of FGMFN in sentiment analysis tasks within advertising contexts, highlighting its ability to leverage both visual and textual features for more accurate predictions.

TABLE 4. Performance of the proposed method and baseline models.

Modality	Models	Twitter-2015		Twitter-2017		YTB-ADS
		Acc(%) ↑	Macro-F1(%) ↑	Acc(%) ↑	Macro-F1(%) ↓	Acc(%) ↑
Image	Res-Target	59.88	46.48	58.59	53.98	48.71
	VGG16	64.32	60.09	66.45	62.66	57.82
Text	BiLSTM	70.30	63.43	61.67	57.97	65.12
	TextCNN	71.17	64.21	64.75	61.46	64.61
	RoBERTa	74.15	68.86	68.15	65.23	67.83
Image and Text	MIMN	71.84	65.69	65.88	62.99	65.72
	ESAFN	73.38	67.37	67.83	64.22	67.47
	VilBERT	73.69	69.53	67.86	64.93	68.62
	TomBERT	77.15	71.15	70.34	68.03	69.42
	LXMERT-MMSA	77.83	72.01	71.28	68.59	69.84
	ARFN	78.14	73.68	71.14	**69.58**	68.03
	AoM	78.50	73.70	70.58	68.43	68.43
	MTVAF	78.63	74.52	70.93	68.95	69.27
	FGMFN (Ours)	**79.92**	**75.28**	**71.64**	69.16	**70.54**

TABLE 5. Comparison of model efficiency: Parameters, FLOPs, Inference latency and training time.

Dataset	Method	Parameters (M)	FLOPs (G)	Inference Latency (ms)	Training Time (h)
YTB-ADS	MIMN	50	1.80	60	1.2
	ESAFN	55	2.00	65	1.5
	TomBERT	160	3.50	75	2.1
	ARFN	120	2.80	70	1.7
	FGMFN (Ours)	130	3.00	68	1.9
Twitter-2015	MIMN	48	1.70	58	1.1
	ESAFN	53	1.90	62	1.4
	TomBERT	155	3.40	72	1.9
	ARFN	115	2.70	68	1.7
	FGMFN (Ours)	125	2.90	66	1.8

2) EFFICIENCY ANALYSIS

To assess efficiency, we present additional results in Table 5, detailing parameter count, computational load, inference latency, and training time. Efficiency metrics were measured on a single RTX 3090, with the first iteration excluded to ensure steady-state accuracy. Experimental results show that FGMFN achieves performance comparable to existing multimodal models while maintaining a balance between accuracy and efficiency, ensuring manageable computational overhead with performance improvements.

3) CONFUSION MATRIX

The confusion matrices in Figure 3 illustrate the per-class prediction performance of the proposed FGMFN model on the Twitter-2015, Twitter-2015 and YTB-ADS datasets. As depicted in the confusion matrices, the model exhibits strong discriminative capability across different sentiment categories. Although some misclassifications lean toward the majority classes due to class imbalance, the overall performance remains robust.

E. ABLATION STUDY

To further evaluate the effectiveness of each component in our proposed model, we conducted a series of ablation experiments across three datasets to quantify the contribution of each module. The configurations for these experiments are as follows:

- w/o MSVF: the Multiscale Visual Feature Fusion is excluded;
- w/o CA: the cross-attention-guided feature module was removed and replaced with a six-layer Transformer structure;
- w/o ITC: the image-text contrastive learning constraints on multimodal features are removed;
- w/o MI: the mutual information constraints that align visual and textual modalities are omitted.

As shown in Table 6, the results indicate that FGMFN outperforms all other configurations across all three datasets, confirming that each module is essential for achieving optimal performance. Specifically, the w/o MSVF configu-

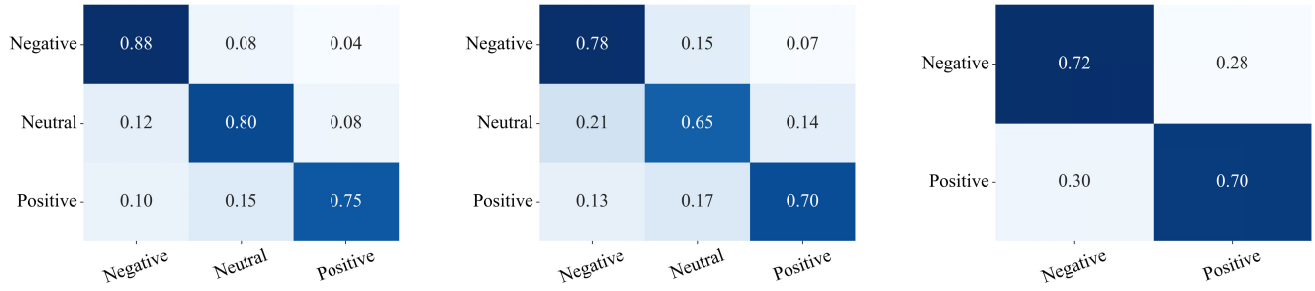


FIGURE 3. Confusion matrices of the proposed method evaluated on (from left to right) Twitter-2015, Twitter-2017, and YTB-ADS datasets.

TABLE 6. Ablation study results of different configurations on Twitter-2015, Twitter-2017, and YTB-ADS datasets.

Models	Twitter-2015		Twitter-2017		YTB-ADS
	Acc2	F1	Acc2	F1	Acc2
FGMFN	**79.92**	**75.28**	**71.64**	**69.16**	**70.54**
w/o MSVF	79.15	74.31	71.61	69.84	69.74
w/o CA	78.80	73.90	70.32	67.53	69.14
w/o ITC	77.34	72.77	69.77	67.65	68.27
w/o MI	76.80	70.61	70.69	68.45	67.36

ration, which removes multiscale visual context information, results in a performance drop, underscoring the importance of incorporating contextual visual cues for improved sentiment analysis. This contextual information facilitates more accurate sentiment transfer between text and images, enhancing the nuanced understanding of multimodal data. Similarly, the w/o CA configuration also leads to a noticeable performance degradation. This finding suggests that the cross-attention mechanism is crucial for effective inter-modal feature alignment and the reduction of modality-specific biases. The w/o ITC configuration results in a slight performance degradation, indicating that the contrastive learning mechanism is essential for improving the alignment of information across modalities. Specifically, contrastive learning mitigates modality-specific biases, enhancing the model's ability to handle complex multimodal interactions and produce more accurate sentiment predictions. Finally, the w/o MI configuration exhibits the most significant performance decline, emphasizing the critical role of mutual information maximization in capturing task-relevant information between visual and textual modalities. This mechanism preserves and enhances sentiment-related features during modality fusion, underscoring its importance in overall performance improvement.

F. DIFFERENT LOSS WEIGHTS

The detailed weights of each training loss are provided in the experimental parameter settings. To investigate the impact of the weight coefficients α and β on model performance more clearly, we conducted a systematic experimental analysis using grid search to explore various weight combinations.

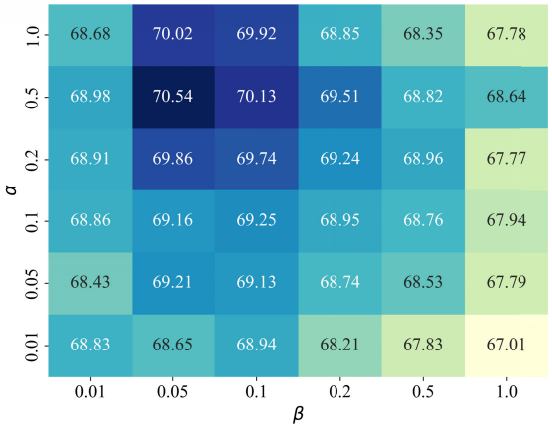


FIGURE 4. Experimental results of FGMFN with different $\alpha$and β values.

This experiment was conducted on the custom-built dataset, YTB-ADS, with binary classification accuracy as the evaluation metric. Figure 4 illustrates the specific values used in the experiment and their corresponding results.

The experimental results show that the model performs optimally when $\alpha = 0.5$ and $\beta = 0.05$. As shown in Figure 4, relatively larger values for α and smaller values for β lead to better model performance. This is because the primary task of the model is sentiment recognition, and the introduction of the image-text matching loss helps further align the emotional feature space, thereby facilitating better integration of multimodal information. The incorporation of mutual information loss aims to balance the contributions of image and text features. However, assigning a higher weight to the mutual information loss relative to the contrastive learning loss shifts the model's focus toward adjusting the spatial relationships between features, deviating from the primary goal of multimodal sentiment analysis. This shift negatively affects the model's overall performance.

G. CASE STUDY

To further validate our approach, we selected representative samples from Table 7 and conducted a comparative analysis with RoBERTa and AoM.

TABLE 7. Predictions of different methods on four test samples.

Image				
Text	(a) What do health heroes look like? Dr Lucille Corti died AIDS 1996, Dr Lukwiya died Ebola 2000	(b) In a devastating speech, iconic actor Robert Niro declare that Donald Trump will be failure.	(c) Kelsey Plum and Devin Booker share a beautiful moment.	(d) Breaking news:Renowned singer Taylor will be attending the Travis Kelce Game vs Broncos in KC Thursday!
Ground Truth	(Dr Lucille Corti, Positive) (AIDS, Negative) (Dr Lukwiya, Positive) (Ebola, Negative)	(Robert Niro, Neutral) (Donald Trump, Negative)	(Kelsey Plum, Positive) (Devin Booker, Positive)	(Taylor, Positive) (Travis Kelce, Neutral) (Broncos, Neutral)
RoBERTa	×(Lucille Corti, Neutral) ✓(AIDS, Negative) ×(Dr Lukwiya, Neutral) ✓(Ebola, Negative)	×(Robert, Negative) ✓(Donald Trump, Negative)	✓(Kelsey Plum, Positive) × -	✓(Taylor, Positive) ×(Game, Neutral) ✓(Broncos, Neutral)
AoM	×(Corti, Positive) ✓(AIDS, Negative) ×(Lukwiya, Neutral) ✓(Ebola, Negative)	×(Robert Niro, Negative) ✓(Donald Trump, Negative)	×(Kelsey Plum, Neutral) ✓(Devin Booker, Positive)	✓(Taylor, Positive) ×(Travis Kelce, Neutral) ×(Broncos, Positive)
Ours	✓(Dr Lucille Corti, Positive) ✓(AIDS, Negative) ✓(Dr Lukwiya, Positive) ✓(Ebola, Negative)	✓(Robert Niro, Neutral) ✓(Donald Trump, Negative)	✓(Kelsey Plum, Positive) ✓(Devin Booker, Positive)	✓(Taylor, Positive) ✓(Travis Kelce, Neutral) ✓(Broncos, Neutral)

We first observed that both RoBERTa and AoM primarily rely on textual information for judgment, which can introduce harmful textual biases. For instance, in cases (a) and (d), both models show significant sentiment prediction biases towards "Dr. Lukwiya" and "Travis Kelce." RoBERTa fails to effectively leverage image information, relying exclusively on textual content for sentiment classification, which results in inaccurate predictions. Although AoM has some multimodal fusion capability, it lacks refined cross-modal feature integration and deep coordination between textual and visual features, leading to errors in both sentiment and aspect predictions. Compared to these methods, our proposed FGMFN alleviates textual feature ambiguity caused by category similarity in advertisement images through an innovative text-image matching mechanism and mutual information loss design, improving the alignment between visual and sentiment semantics. This enables FGMFN to better handle challenging cases such as "Dr. Lukwiya" and "Travis Kelce," accurately capturing the true sentiment.

Furthermore, in cases (b) and (c), although AoM correctly identifies the target aspects "Robert Niro" and "Kelsey Plum," it still misclassifies the sentiment, highlighting its limitations in multimodal information fusion. By leveraging a refined cross-modal feature fusion strategy, FGMFN effectively captures semantic associations between text and images, greatly enhancing cross-modal information integration and improving sentiment analysis accuracy.

V. CONCLUSION

This paper presents a novel framework for advertising sentiment recognition that effectively captures the emotional interplay between advertisement images and their accompanying text. We introduce FGMFN, specifically designed to address challenges such as limited multi-scale features and target redundancy in advertising. Additionally, we curate YTB-ADS, a fine-grained advertisement image-text sentiment analysis dataset. Experiments show that combining a multi-scale network for visual feature extraction with visually guided text-based feature fusion significantly enhances sentiment recognition accuracy. Furthermore, integrating image-text matching loss and mutual information loss effectively reduces textual ambiguities caused by intra-class similarity.

Extensive evaluations on the YTB-ADS dataset and public benchmarks (Twitter-2015 and Twitter-2017) confirm that multimodal feature fusion is crucial for advancing advertising sentiment analysis. Results show that FGMFN achieves state-of-the-art performance in controlled experiments. Beyond its technical advancements, FGMFN's real-world applicability is a key strength. Designed for practical deployment, it offers efficient inference latency and manageable training time on standard hardware, making it well-suited for dynamic advertising platforms.

FGMFN can be integrated into ad placement systems for real-time sentiment analysis, supporting adaptive content strategies based on instant consumer feedback. Additionally, it serves as a monitoring tool for ad campaigns, providing detailed insights into consumer reactions through continuous multimodal content evaluation. By utilizing advertisement data from real-world scenarios, the model's relevance is further strengthened. Future work will emphasize system-level optimizations and API development to enable seamless

integration into commercial advertising systems, bridging the gap between research and practical implementation.

REFERENCES

[1] L. M. Lodish, M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson, and M. E. Stevens, "How TV advertising works: A meta-analysis of 389 real world split cable TV advertising experiments," *J. Marketing Res.*, vol. 32, no. 2, pp. 125–139, 1995.

[2] N. Vedula, W. Sun, H. Lee, H. Gupta, M. Ogihara, J. Johnson, G. Ren, and S. Parthasarathy, "Multimodal content analysis for effective advertisements on Youtube," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1123–1128.

[3] J. Xu, F. Huang, X. Zhang, S. Wang, C. Li, Z. Li, and Y. He, "Sentiment analysis of social images via hierarchical deep fusion of content and links," *Appl. Soft Comput.*, vol. 80, pp. 387–399, Jul. 2019.

[4] B. Wang, J. Wang, and H. Lu, "Exploiting content relevance and social relevance for personalized ad recommendation on Internet TV," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, no. 4, pp. 1–23, Aug. 2013.

[5] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, "Online video recommendation based on multimodal fusion and relevance feedback," in *Proc. 6th ACM Int. Conf. Image Video Retr.*, Jul. 2007, pp. 73–80.

[6] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172948–172964, 2019.

[7] Y. Yu, H. Lin, J. Meng, and Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, p. 41, Jun. 2016.

[8] A. Kumar, K. Srinivasan, W.-H. Cheng, and A. Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102141.

[9] M. Wang and T. Song, "Remote sensing image retrieval by scene semantic matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2874–2886, May 2013.

[10] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 309–325, Jan. 2023.

[11] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, and X. Sun, "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[12] J. Li, R. R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 9694–9705.

[13] Y. Tanahashi and K.-L. Ma, "Design considerations for optimizing storyline visualizations," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2679–2688, Dec. 2012.

[14] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 2, no. 1, pp. 1–19, Feb. 2006.

[15] P. Pham and J. Wang, "Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis," in *Proc. 22nd Int. Conf. Intell. User Interfaces*, Mar. 2017, pp. 67–78.

[16] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, *arXiv:1707.07250*.

[17] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1.

[18] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.

[19] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3777–3786.

[20] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Trans. Multimedia*, vol. 25, pp. 3375–3385, 2023.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[22] J. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, Jun. 1990.

[23] B. Yang, L. Wu, J. Zhu, B. Shao, X. Lin, and T.-Y. Liu, "Multimodal sentiment analysis with two-phase multi-task learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2015–2024, 2022.

[24] J. Ye, J. Zhou, J. Tian, R. Wang, J. Zhou, T. Gui, Q. Zhang, and X. Huang, "Sentiment-aware multimodal pre-training for multimodal sentiment analysis," *Knowl.-Based Syst.*, vol. 258, Dec. 2022, Art. no. 110021.

[25] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.

[26] X. Lu, Y. Ni, and Z. Ding, "Cross-modal sentiment analysis based on CLIP image-text attention interaction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, 2024.

[27] L. Xiao, X. Wu, W. Wu, J. Yang, and L. He, "Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4578–4582.

[28] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 429–439, 2020.

[29] Y. Xin, J. Du, Q. Wang, K. Yan, and S. Ding, "MmAP: Multi-modal alignment prompt for cross-domain multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 14, pp. 16076–16084.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[31] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R^3net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[32] A. Radoi and M. Datcu, "Multilabel annotation of multispectral remote sensing images using error-correcting output codes and most ambiguous examples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2121–2134, Jul. 2019.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[34] Z. Luo, S. Huang, and K. Q. Zhu, "Knowledge empowered prominent aspect extraction from product reviews," *Inf. Process. Manage.*, vol. 56, no. 3, pp. 408–423, May 2019.

[35] M. Thelwall, "The heart and soul of the Web? Sentiment strength detection in the social Web with SentiStrength," in *Cyberemotions: Collective Emotions in Cyberspace*. Cham, Switzerland: Springer, 2017, pp. 119–134.

[36] J. Yu and J. Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5408–5414.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[39] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 207–212.

[40] A. Rakhlin, "Convolutional neural networks for sentence classification," *GitHub*, vol. 6, p. 25, Jul. 2016.

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[42] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect-based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 371–378.

[43] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019.

[44] L. Jia, T. Ma, H. Rong, and N. Al-Nabhan, "Affective region recognition and fusion network for target-level multimodal sentiment classification," *IEEE Trans. Emerg. Topics Comput.*, pp. 1–11, 2023.
[45] R. Zhou, W. Guo, X. Liu, S. Yu, Y. Zhang, and X. Yuan, "AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis," 2023, *arXiv:2306.01004*.
[46] Z. Yin, Y. Du, Y. Liu, and Y. Wang, "Multi-layer cross-modality attention fusion network for multimodal sentiment analysis," *Multimedia Tools Appl.*, vol. 83, no. 21, pp. 60171–60187, Jan. 2024.
[47] Y. Li, H. Ding, Y. Lin, X. Feng, and L. Chang, "Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis," *Artif. Intell. Rev.*, vol. 57, no. 4, p. 78, Mar. 2024.
[48] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -Specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Jan. 2020, pp. 113–1122.

HAN WANG received the bachelor's degree in literature from East China University of Science and Technology, in 2009, and the master's degree in literature from Southeast University, in 2012. She is currently a Lecturer with Xuzhou University of Technology, China. Her research interests include the applications of computer vision and advertising design.

PENG CHEN received the M.Eng. degree from China University of Mining and Technology, in 2017. He is currently an Assistant Researcher with Xuzhou University of Technology, China. His research interests include cross-modal learning and multimodal sentiment analysis.

XIANGYU DU received the Master of Fine Arts (MFA) degree from Kookmin University, South Korea, in 2013. He is currently a Lecturer with Xuzhou University of Technology, China. His research interests include user experience design and multisensory design.