Dual-Perspective Fusion Network for Aspect-Based Multimodal Sentiment Analysis

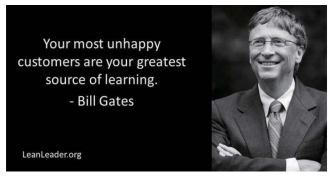
Di Wang, *Member, IEEE*, Changning Tian, Xiao Liang, Lin Zhao, Lihuo He, and Quan Wang

***Abstract*—Aspect-based multimodal sentiment analysis (ABMSA) is an important sentiment analysis task that analyses aspect-specific sentiment in data with different modalities (usually multimodal data with text and images). Previous works usually ignore the overall sentiment tendency when analyzing the sentiment of each aspect term. However, the overall sentiment tendency is highly correlated with aspect-specific sentiment. In addition, existing methods neglect to explore and make full use of the fine-grained multimodal information closely related to aspect terms. To address these limitations, we propose a dual-perspective fusion network (DPFN) that considers both global and local fine-grained sentiment information in multimodal data. From the global perspective, we use text-image caption pairs to obtain a global representation containing information about the overall sentiment tendencies. From the local fine-grained perspective, we construct two graph structures to explore the fine-grained information in texts and images. Finally, aspect-level sentiment polarities can be obtained by analyzing the combination of global and local fine-grained sentiment information. Experimental results on two multimodal Twitter datasets show that the proposed DPFN model outperforms state-of-the-art methods.**

***Index Terms*—Aspect-based sentiment analysis, multimodal sentiment analysis, graph neural network.**

I. Introduction

With the rapid development of social media, multimedia data containing multiple modalities (e.g., text, image, video, audio) have grown explosively in recent years on various social networking sites, such as Facebook and Twitter. Multimodal sentiment analysis (MSA) aims to automatically capture people's standpoints or attitudes on issues by analyzing data with different modalities. It has a wide range of applications, such as helping businesses determine whether users like their products or helping governments understand users' attitudes on a hot issue. However, people usually express different views on different aspects in a tweet. For example, in the comment "*The food is delicious although the service is poor*", the sentiment polarity of two aspects "**food**" and "**service**" are positive and negative, respectively. Therefore, it is more important to mine the sentiment polarity of different aspect entities in data. Aspect-based multimodal sentiment analysis (ABMSA) is an emerging task aimed at identifying the sentiment polarity of each aspect entity from multimodal data (usually text-image pairs). Fig. 1(a) and (b) provide two examples of aspect-based sentiment analysis (ABSA). Given a tweet containing both text and images and one or more aspect words, the objective of ABMSA is to accurately determine the sentiment polarity of each aspect word.



Your most unhappy customers are your greatest source of learning. - [**Bill Gates**]$_{\text{Pos}}$ #Leadership

(a)

U.S. U-16 boys beat [**England**]$_{\text{Pos}}$, fall to [**Italy**]$_{\text{Neg}}$ to finish sixth at rneo Delle Nazioni #soccer

(b)

Fig. 1. Here are two examples of the ABMSA task, where the aspect words are in bold within the brackets. In (a), the aspect word is **Bill Gates**, and the ground truth sentiment is Positive. In (b), the aspect words are **England** and **Italy**, and the ground truth sentiments are Positive and Negative, respectively.

Many methods [1], [2], [3], [4], [5], [6], [7], [8], [9] have been proposed around the challenging ABMSA task. Although these approaches have made some progress, some critical challenges remain unsolved as follows:

- *Global sentiment information is ignored:* Existing methods [1], [2], [3], [4], [5], [6], [7], [8], [9] for aspect-level sentiment analysis often consider sentiment information only at the word level, which can result in prediction errors when the sentiment polarity of aspect words is unclear. For example, the aspect word "**Bill Gates**" in Fig. 1(a) is neutral, but the text in the tweet "*Your most unhappy customers are your greatest source of learning*" and the image caption

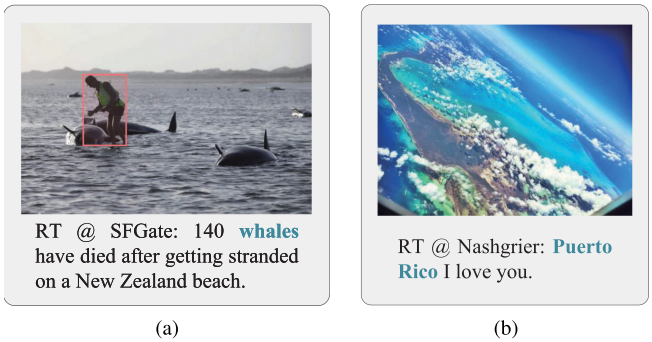
Manuscript received 23 March 2023; revised 12 June 2023 and 24 July 2023; accepted 19 September 2023. Date of publication 2 October 2023; date of current version 23 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62072354, 61972302, 62172222, and 62072355, in part by the Key Research and Development Program of Shaanxi Province of China under Grant 2022GY-057, in part by the Key Industry Innovation Chain Projects of Shaanxi Province of China under Grant 2021ZDLGY07-04, in part by the Foundation of National Key Laboratory of Human Factors Engineering under Grant 6142222210101, in part by the Fundamental Research Funds for the Central Universities under Grants QTZX23084, QTZX23105, and QTZX23108, and in part by the Science and Technology Program of Guangzhou under Grant SL2022A04J00303. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Ichiro Ide. *(Corresponding author: Di Wang.)*

Di Wang, Changning Tian, Xiao Liang, Lihuo He, and Quan Wang are with the Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, Xidian University, Xi'an 710071, China (e-mail: wangdi@xidian.edu.cn; cntian@stu.xidian.edu.cn; ecoxial2012@outlook.com; lhhe@mail.xidian.edu.cn; Qwang@xidian.edu.cn).

Lin Zhao is with the Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: linzhao@njust.edu.cn).

The source code is publicly available at https://github.com/cntian0/DPFN.

Digital Object Identifier 10.1109/TMM.2023.3321435



(a) (b)

Fig. 2. Two examples from the Twitter-2015 dataset. The salient object is circled by a red box, and the aspect word is indicated in bold blue font. It shows that the sentiment information related to aspect words in images may not be accurately captured by object detection. (a) The extracted salient object is a person while the aspect word is the whale. (b) No significant object is extracted.

in the tweet "*a man in a suit and tie smiling*" are positive in overall sentiment. An intuitive idea is to account for the global sentiment information of the entire tweet (including text and image caption), as the sentiment polarity of ambiguous aspect words may be consistent with the overall sentiment of the tweet.

- *Syntactic knowledge of sentences is overlooked:* In the ABMSA task, it is crucial to find the corresponding words that describe the aspect terms. Existing methods [2], [3], [4], [5], [6], [7], [8], [9] utilize attention mechanisms to adaptively search for the associated dependencies between words in a sentence. However, this method may not be accurate because it ignores the fact that phrases should be considered as a whole and sentiment judgments can be easily influenced by other words. For example, the aspect words "**England**" and "**Italy**" in Fig. 1(b) can be easily mispredicted as having the same sentiment if the dependency relationships corresponding to the predicates "*beat*" and "*fall to*" in the sentence are not considered. Syntactic knowledge of the sentence provides direct relationships between words. Therefore, utilizing syntactic knowledge to establish connections between words in the sentence and aspect terms can help achieve accurate aspect-level sentiment analysis.
- *Difficulties in extracting visual information related to aspect words from images:* The ABMSA task can benefit from visual information to achieve better performance, but irrelevant visual information in images can negatively affect sentiment analysis results. Previous works [5], [6], [8], [10] have used pretrained object detectors to extract salient objects from images to mitigate noise from irrelevant areas. However, due to the complexity of visual scenes, object detectors often struggle to extract objects that are closely related to aspect words; as shown in Fig. 2(a), the salient object extracted from the image is a person, and the aspect word is "**whale**". Furthermore, object detectors have limited categories, and sometimes they may not be able to extract any object at all, as shown in Fig. 2(b). Therefore, extracting information from images that are closely related to aspect words remains a major challenge.

In this article, we propose a novel aspect-based multimodal sentiment analysis method named **D**ual-**P**erspective **F**usion **N**etwork (DPFN) to address the above problems. The proposed DPFN extracts the global sentiment information and the local fine-grained sentiment information from multimodal data and obtains the final aspect-level sentiment polarity from both global and local perspectives. The proposed DPFN consists of three modules: the global semantic extraction module, local syntax enhancement module, and language-guided fusion module. Specifically, for the global semantic extraction module, we convert images into captions and input them together with tweet texts into a pretrained global text encoder, extracting global semantic information from both modalities. In this way, the semantic gap between image and text modalities is eliminated. Additionally, the local syntax enhancement module utilizes dependency parsing to capture syntax information associated with aspect terms and utilizes GCN to enhance the local text encoder in extracting fine-grained semantic information. The language-guided fusion module utilizes cross-modal attention to guide the fusion of fine-grained visual features based on text features and learn fine-grained correspondences between image patches and words. Finally, we combine the global semantic features, locally syntax-enhanced fine-grained text features, and language-guided fine-grained visual features to achieve aspect-based sentiment analysis. In summary, our main contributions are as follows:

- A novel aspect-based multimodal sentiment analysis method named DPFN is proposed. It analyses aspect-level sentiment polarity from both global and local perspectives rather than considering only word-level granularity as existing methods. The global sentiment information helps in accurately identifying the sentiment of each aspect term.
- Syntax knowledge is utilized to enhance aspect-related semantic feature extraction. Text information is employed to guide the image semantic feature extraction.
- Extensive experiments on two benchmark datasets, Twitter-2015 and Twitter-2017, demonstrate that the proposed DPFN achieves significant improvements compared with the state-of-the-art methods.

The rest of this article is organized as follows. Section II reviews some closely related works. Section III details the proposed DPFN method. Section IV presents the experimental results and analysis. Finally, the conclusions are reached in Section V.

II. Related Work

A. Sentiment Analysis

Traditional sentiment analysis tasks are oriented to the text modality only. The goal of a multimodal sentiment analysis task is to discover the sentiment expressed in multimodal data. For example, some work [11], [12], [13] combines images and text to analyze sentiment, and others [14] extract multimodal data from video data for sentiment analysis. Due to the outstanding performance of NLP tasks, transformer-based models have

been widely applied to text sentiment analysis tasks [15], [16]. By fine-tuning pretrained language models such as BERT [17], GPT [18], XLNet [19], SKEP [20], and SentiLARE [21] on specific sentiment analysis datasets, these models have achieved impressive results. For multimodal sentiment analysis tasks, recent approaches [22], [23], [24], [25], [26], [27], [28] use features from different modalities as transformer input for modality fusion. For example, Tsai et al. [23] employed cross-modal attention to establish interactions between different modalities. Furthermore, transformer-based models are often used in fine-grained sentiment analysis methods [29], [30], [31], [32] to extract information related to aspect terms from text. However, attention mechanisms are also susceptible to the influence of other opinion targets in a sentence due to their superior long- and short-term capturing capabilities. Moreover, due to lacking syntactic understanding, fixed collocations may be separated by the attention mechanism, which can affect sentiment analysis.

B. Aspect Relations Modelling in Sentiment Analysis

The text-based aspect-level sentiment analysis task analyses the sentiment polarity of aspectual targets by modeling aspectual relationships. For example, Tang et al. [33] modeled the relationship between aspect words and other words in text through deep memory networks. Graph-based models have been widely used in text sentiment analysis, as evidenced by works [34], [35], [36], [37] that utilize graph neural networks (GNNs) [38] to model the relationships between words in a sentence. Specifically, Zhang et al. [34] proposed a graph convolutional network (GCN) [39] built upon a dependency tree of the sentence to introduce syntactic constraints and long-range word dependencies. Given the uncertainty of sentence structures, Tang et al. [36] and Li et al. [37] designed the dependency graph enhanced dual-transformer (DGEDT) and dual graph convolutional network (DualGCN) models, respectively, which combine the syntax and semantics of the sentence by utilizing continuous dependency graphs to guide text encoder learning. Although achieving promising results, these methods have yet to be applied to multimodal sentiment analysis.

C. Multimodal Interactions for Aspect-Based Sentiment Analysis

As a fine-grained multimodal task, most methods for ABMSA [1], [2], [3], [4], [5], [6], [7], [8], [9] employ transformer-based models for fusing information from different modalities, with a focus on leveraging image information to improve performance. Previous studies [1], [2], [4] utilized ResNet [40] to extract global image features without considering the fine-grained information related to aspect terms, leading to including more noisy information. To address this problem, JML [4] proposed a relationship detection module to determine the relevance of image information for sentiment analysis. HIMT [5] and VLP-MABSA [6] utilized object detection [41] to extract salient objects from an image, excluding redundant information such as the background. It is worth mentioning that Khan et al. [3] converted images into image captions to reduce the semantic gap between different modalities. Zhao et al. [9] proposed a knowledge enhancement framework for improving task performance by extracting adjective-noun pairs in images, improving the performance of the model by extending it to existing attention-based models. ITM [8] improves the model performance by jointly performing coarse-to-fine-grained image-target matching and target sentiment classification. Recently, a powerful baseline model, FITE-DE-Large [7], has achieved good results on the ABMSA task due to the superiority of the model and a powerful text pretraining model, BERTweet-Large [42].

III. Methodology

A. Task Definition

Given a set of multimodal samples $\mathcal{M}$, for each sample $m \in \mathcal{M}$, which contains a sentence $S = (w_1, w_2, \ldots, w_N)$ containing N words, an accompanying image I associated with the sentence, and an aspect term T, which is a subsequence of the sentence S. The aspect term is generally an entity mentioned in the sentence, which is associated with a sentiment label y, where $y \in \{negative, neutral, positive\}$. The goal of aspect-based multimodal sentiment analysis is to learn a target-oriented sentiment classifier that can correctly predict the sentiment labels of aspect terms in unseen samples. For example, by entering a text such as "*The waiter at this restaurant has a great attitude, but the food tastes terrible*" with a picture, the model is can predict that the target "**waiter**" is positive, while the target "**food**" is negative.

B. Overview

Our DPFN overall structure is shown in Fig. 3, which comprises three main parts: a global semantic extraction module (GSEM), local syntax enhancement module (LSEM), and language-guided fusion module (LGFM). In the global semantic extraction module, the image is first converted into a caption through a caption converter [3] to bridge the visual language gap. Then, the caption and tweet text are concatenated into a sentence as the input of a pretrained text encoder [17] to extract global semantic information. The details of the GSEM are introduced in Section C.1.

In the local syntax enhancement module, we use only text as input to the pretrained local text encoder [17]. In addition, we use dependency parsers [43] to obtain the dependencies between each word to construct a text adjacency matrix and construct TextGCN to enhance the extraction of semantic information related to aspect terms. The details of LSEM are introduced in Section C.2.

In the language-guided fusion module, the image is divided into patches and converted into patch embeddings, which are used as inputs to the pretrained image encoder [44] to obtain fine-grained visual features. Additionally, we use cross-attention to extract image features related to the text and fuse the dependency relationships of the text and image patches to construct the fusion adjacency matrix. Then, we construct the FusionGCN to achieve language-guided fine-grained feature fusion. The details of LSEM are introduced in Section C.3.

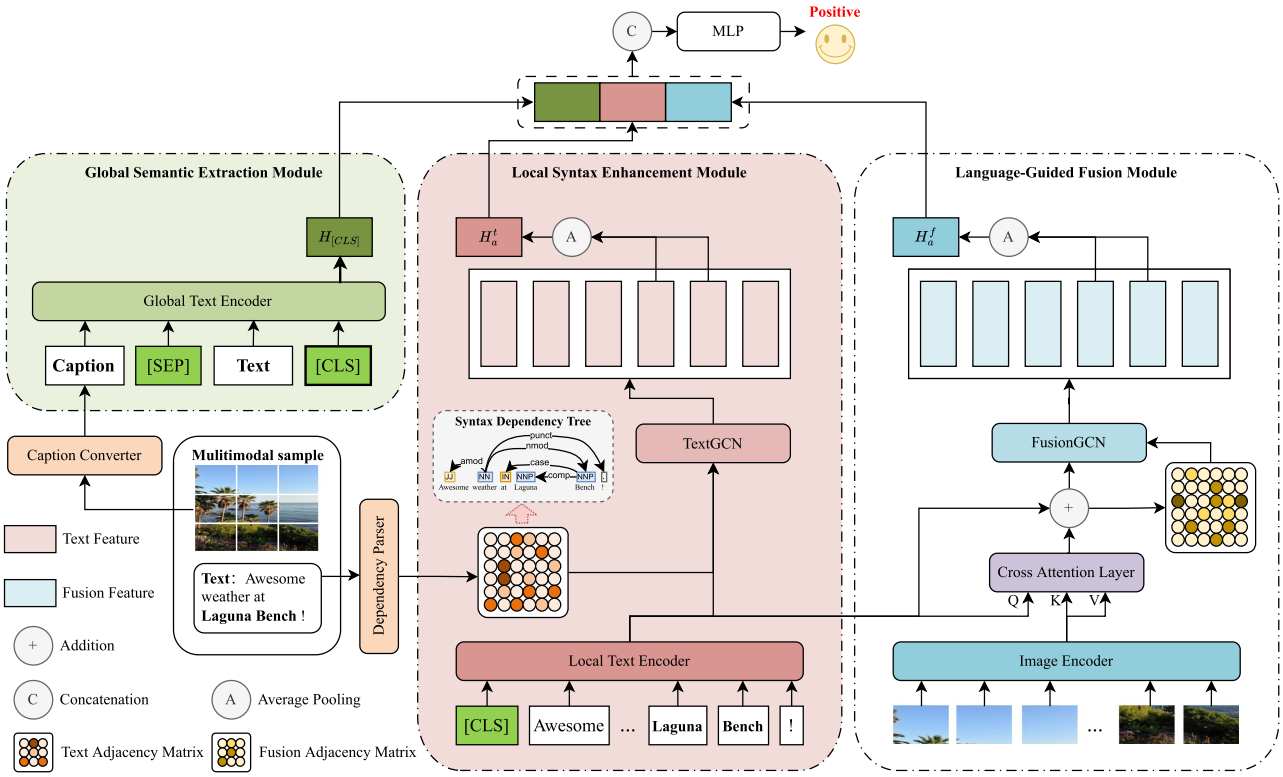


Fig. 3. Overall architecture of DPFN mainly consists of three modules: the Global Semantic Extraction Module (GSEM), the Local Syntax Enhancement Module (LSEM), and the Language-Guided Fusion Module (LGFM). The input includes tweet texts and their corresponding images.

The final sentiment analysis results are obtained by fusing the features extracted by the above three modules.

C. Dual-Perspective Fusion Network

The overall architecture of the DPFN is a joint analysis of aspect-level sentiment from a global perspective and a local fine-grained perspective. Algorithm 1 shows the DPFN during the training stage.

1) Global Semantic Extraction Module: In the global semantic extraction module, we analyze the sentiment polarity of the entire tweet by leveraging both the image and text. Inspired by the work of CapTrBERT [3], we utilize an image converter to generate image captions. This approach allows us to capture the overall sentiment expressed by the image while also making the integration of text and image more interpretable. We use the same image converter as CapTrBERT to process an input image $I \in \mathbb{R}^{3\times H\times W}$, which is converted into a textual representation C. To combine the text and image captions, we use the special token $[SEP]$ to form a sentence pair and prefix the sentence pair with the special token $[CLS]$ to indicate global information. The tokens X that are input to the global text encoder take the following form:

$$X = \{[CLS], t_1^S, t_2^S, \ldots, t_{N_S}^S, [SEP], t_1^C, t_2^C, \ldots, t_{N_C}^C\} \quad (1)$$

where t_i^S and t_j^C represent the tokens of sentence S and sentence C, respectively. Sentence S represents the original text of the input, while sentence C refers to the image caption converted from the image. The number of tokens in sentence S is denoted by N_S, and the number of tokens in sentence C is denoted by N_C.

Given BERT's outstanding performance in various NLP domains, we adopted it as a global text encoder to represent X as hidden representations as follows:

$$H^g = BERT^g(X) \quad (2)$$

where $H^g \in \mathbb{R}^{N_g\times d}$ represents the generated hidden feature of the sentence pair, d is the dimension of the hidden feature, and N_g represents the length of the sentence pair token. Additionally, we have $N_S + N_C < N_g$. To extract the global sentiment feature, we adopt the hidden feature $H_{[CLS]}^g$ that corresponds to the special token $[CLS]$.

Unlike CapTrBERT, which directly uses sentences spliced from text and caption to predict sentiment polarity at the aspect level, our approach uses the spliced sentences to predict the global sentiment polarity.

2) Local Syntax Enhancement Module: Applying attention mechanisms to locate opinion words related to aspect terms can be susceptible to interference from irrelevant opinion words in a sentence, leading to inaccurate sentiment analysis. To overcome this issue, we leverage syntactic information to establish connections between aspect terms and their related opinion words, thereby improving task performance.

Text Representation: To extract text features for the local syntax enhancement module, we use the original text S as input to the text encoder. Similar to the global semantic extraction module, we use BERT as the text encoder to obtain hidden features

Algorithm 1: Dual-Perspective Fusion Network.

Input: text S and image I in multimodal data, aspect words T, sentiment labels y.
Output: prediction sentiment polarity $\hat{y}$.

```
Initialize model parameters M(θ, x).
Convert image I to image caption C.
Generate text syntactic dependency probability matrix A^t using LAL-Parser.
for epoch = 1, 2, ..., end do
    for mini-batch in dataLoader do
        H^g_[cls] ← (C, S) by GSEM. // Compute global sentiment features.
        H^t_a ← (S, A^t) by LSEM. // Compute graph-based aspect-level text features.
        H^f'_a ← (S, I, A^t) by LGFM. // Compute graph-based aspect-level fusion features.
        O ← (H^g_[cls], H^t_a, H^f'_a) by aggregation operation. // Compute the final sentiment features.
        Compute the loss J using (17).
        Update the network parameters θ using BP algorithm.
    end for
end for
```

of the text:

$$h_1, h_2, \ldots, h_{N_S}^S = BERT^l(\{t_1^S, t_2^S, \ldots, t_{N_S}^S\}) \tag{3}$$

where t_i^S denotes the token of the original input text S. $h_i \in \mathbb{R}^d$ is the i-th output of the encoder, where $i \in [1, N_S]$ and d is the dimension of the hidden layer.

To match the wordpiece-based BERT representation with the word-based syntactic dependency results, we sum the token features belonging to the same word to obtain the word-level representation. Let $span_k$ denote all the tokens of the k-th word and the word-level features are obtained as follows:

$$h_k^l = sum(\{h_{span_k}\}) \tag{4}$$

where, $h_k^l \in \mathbb{R}^d$ represents the hidden feature obtained by aggregating token features extracted by the local text encoder for the k-th word, where k ranges from 1 to N, N is the length of sentence S, and we have $N < N_S$.

Text Graph Convolutional Network (TextGCN): The purpose of constructing a text graph is to enable aspect words to focus on the information expressed by opinion words that positively affect their sentiment analysis through semantic dependencies between words. Inspired by the work of DualGCN [37], we use the probability matrix of all dependency arcs in the dependency analyzer [43] as the adjacency matrix $A^t \in \mathbb{R}^{N \times N}$ of the text graph. This probability matrix mitigates dependency resolution errors compared to the discrete output obtained through the dependency analyzer, thus capturing rich structural information.

In TextGCN, the text hidden features $H^l = \{h_1^l, h_2^l, \ldots, h_N^l\}$ obtained from the local text encoder are used as the initial nodes in the text graph. The graph-based representations of each node are then obtained by applying a graph convolution operation, as shown in the following equation:

$$h_i^{(k)} = \sigma\left(\sum_{h_j \in \mathcal{N}_i} A_{ij}^t W_t^{(k)} h_j^{(k-1)} + b^{(k)}\right) \tag{5}$$

where $h_i^{(k)}$ is the hidden representation of the i-th node in the k-th layer text graph, $\mathcal{N}_i$ represents the node set associated with the i-th node, $W_t^{(k)}$ is the weight matrix of layer k, $b^{(k)}$ is the bias, and σ is the activation function. The graph-based text representation $H^t = \{h_1^t, h_2^t, \ldots, h_N^t\}$ is obtained after applying K^t graph convolution layers.

3) Language-Guided Fusion Module: While information from image modalities can enhance ABMSA task performance, irrelevant visual information in images can negatively impact the sentiment analysis results. To address this issue, we propose a language-guided fusion module that can extract visual information related to aspectual terms in images more effectively.

Image Representation: Previous works have primarily used ResNet [40] to obtain information about the whole image or have employed object detection [41] to extract fine-grained information from the image. However, these approaches have limitations in fully utilizing the fine-grained information in images. To address this issue, we propose segmenting the input image I into $N_v \times N_v$ image patches, denoted by $p_1, p_2, \ldots, p_{N_v \times N_v}$, and feeding these patches into a visual encoder. We use ViT [44] as the visual encoder and add a linear layer to obtain the same feature dimension as the text features:

$$h_1^p, h_2^p, \cdots = ViT(\{p_1, p_2, \cdots\}) \tag{6}$$

where $h_k^p \in \mathbb{R}^d$ is the k-th output of the image encoder, $k \in [1, N_v \times N_v]$, and d is the size of the hidden layer of image features.

Fusion Graph Convolutional Network (FusionGCN): Previous research has demonstrated the beneficial effect of incorporating image information in sentiment analysis. However, textual information has been shown to better express sentiment information compared to visual semantic information. To effectively leverage visual information while reducing the semantic gap between different modalities, we propose a language-guided fusion approach. Specifically, we use multi-head cross attention (MHCA) to fuse the visual node features H^p and the textual node features H^l. First, we calculate the attention weights between H^l and H^p to obtain the matching degree between the two modalities. Then, we aggregate all the visual nodes into a weighted combination of their feature vectors according to the attention weights, ensuring that relevant visual information related to aspect terms is effectively extracted and combined with the textual features. The formula is as follows:

$$MHCA(H^l, H^p) = concat(h_1^f, h_2^f, \ldots, h_N^f)W_C \tag{7}$$

where $W_C \in \mathbb{R}^{Nd \times d}$, and h_i^f is calculated as follows:

$$h_i^f = softmax\left(\frac{[W_Q h_i^l]^\top [W_K h_i^p]}{\sqrt{d}}\right)[W_V h_i^p]^\top \tag{8}$$

where W_Q, W_K, and W_V are the weight matrices of query, key, and value, respectively, and d is the length of the hidden layer features.

After acquiring the image information that corresponds to the text, we integrate the text information with the image information to obtain the text-image fusion features H^f:

$$H^f = LN(H^l + MHCA(H^l, H^p)) \tag{9}$$

where $H^f \in \mathbb{R}^{N\times d}$ is the text-image fusion feature. LN indicates the layer normalization operation.

The fused features of each fine-grained node contain both textual and image patch information associated with the node, but they do not combine information from the associated image patches. This results in each image patch being isolated and not fully utilizing the fine-grained image features. To address this, we relate image patches at the fine-grained level based on semantic information. We build a fusion graph on the fusion features $H^f = \{h_1^f, h_2^f, \ldots, h_N^f\}$ and aggregate related features through the edges in the fusion graph. First, we calculate the similarity between each element in the fused features to match the associated image patches:

$$x_{i,j} = cos(h_i^f, h_j^f) \tag{10}$$

where $x_{i,j}$ represents the similarity between the i-th node and the j-th node in the fusion features, and the similarity $x_{i,j}$ between every two nodes is combined to obtain the similarity matrix $\mathcal{S}$.

Then, we compute the Hadamard product of the text graph's adjacency matrix A^t with the similarity matrix $\mathcal{S}$ and perform L_2 normalization to obtain the adjacency matrix $A^f \in \mathbb{R}^{N\times N}$ of the fusion graph:

$$A^f = ||A^t \circ \mathcal{S}||_2 \tag{11}$$

Similar to TextGCN, we use the fused hidden features $H^f = \{h_1^f, h_2^f, \ldots, h_N^f\}$ as the initial nodes in the fusion graph and allow relevant node information to interact through a graph convolution operation:

$$h_i^{(k)} = \sigma\left(\sum_{h_j\in\mathcal{N}_i} A_{ij}^f W_f^{(k)} h_j^{(k-1)} + b^{(k)}\right) \tag{12}$$

where $h_i^{(k)}$ is the hidden representation of the i-th node in the k-th layer of the fusion graph, and the graph-based fusion representation $H^{f'} = \{h_1^{f'}, h_2^{f'}, \ldots, h_N^{f'}\}$ is obtained after K^f layer graph convolution.

Finally, the aspect nodes information in the graph-based text representation and the graph-based fusion representation obtained by TextGCN and FusionGCN are averaged to obtain the aspect-level text features and fusion features from a fine-grained perspective. The hidden features of aspect terms are represented using H_a^t and $H_a^{f'}$, $f(\cdot)$ represents the averaging pooling function, and a_m is the number of words in the aspect term:

$$H_a^t = f(h_{a_1}^t, h_{a_2}^t, \ldots, h_{a_m}^t) \tag{13}$$

$$H_a^{f'} = f(h_{a_1}^{f'}, h_{a_2}^{f'}, \ldots, h_{a_m}^{f'}) \tag{14}$$

TABLE I
BASIC STATISTICS FOR THE TWO TWITTER DATASETS

	Twitter-2015			Twitter-2017		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Positive	928	303	317	1508	515	493
Neutral	1883	670	607	1638	517	573
Negative	368	149	113	416	144	168
Total Aspect	3179	1122	1037	3562	1176	1234
#Sentence	2101	727	674	1746	577	587

#sentence indicates the average number of sentences.

4) Perspective Fusion and Loss Function: After obtaining global sentiment features and local fine-grained sentiment features, the information from these two parts is fused to obtain global sentiment-aided aspect-level sentiment features. Specifically, we concatenate the global sentiment features $H_{[cls]}^g$, the graph-based aspect-level text features H_a^t and the graph-based aspect-level fusion features $H_a^{f'}$:

$$O = [H_{[cls]}^g, H_a^t, H_a^{f'}] \tag{15}$$

The resulting representation O is fed into a linear layer, followed by a softmax function, to obtain a sentiment probability distribution p:

$$p(y|O) = softmax(W_u O + b_u) \tag{16}$$

where W_u and b_u are the learnable weights and bias terms, respectively.

To optimize all the parameters in the model, our goal is to minimize the standard cross-entropy loss function, as shown below:

$$\mathcal{J} = -\frac{1}{|D|}\sum_{j=1}^{|D|} logp(y|O) \tag{17}$$

IV. EXPERIMENTS

A. Experimental Settings

Downstream Datasets: To evaluate the effectiveness of our model, we use two benchmark datasets annotated by Yu and Jiang (2019) [2] for goal-oriented multimodal sentiment classification, namely, Twitter-2015 and Twitter-2017. These two Twitter datasets contain tweets posted during the 2014-2015 and 2016-2017 periods. Multimodal tweets consist of text, images posted with the tweet, and aspect terms in the tweet. Each aspect term is assigned a label from the $\{negative, neutral, positive\}$ set. The basic statistics of these two datasets are shown in Table I.

Implementation Details: We use pretrained BERT and ViT as text and visual encoders, respectively. The batch size is set to 32, and the learning rate is 5e-5. The hidden size of the model is set to 768, which is the same as the standard BERT. We freeze all parameters of the visual encoder ViT and use only its extracted features. ViT is not involved in model training to reduce the learnable parameters and thus avoid model overfitting. During training, we train the model for 20 epochs and monitor its performance on the validation set. After training is completed, we

take the model with the last epoch as our final model and evaluate its performance on the test set. We implement our model with PyTorch and train it on an RTX 3090 GPU.

Evaluation Metrics: We evaluated our model on the ABMSA task and used Accuracy (Acc) and Macro-F1 score (F1) as evaluation metrics to measure performance.

B. Baseline

We compare our model with several existing competing models. These models can be classified as visual-based, text-based, and text-visual-based according to the modality used. A brief description of these models is provided below.

Visual-based:

- Res-Target [40]: Directly connect the aspect words features extracted using BERT [17] with the visual features extracted by ResNet [40].

Text-based:

- MGAN [45]: Using a multigrained attention network to understand aspects.
- BERT: A pretrained BERT model to capture the interaction between aspects and text.
- BERT+BL: Increase the number of layers on top of the original BERT model.
- BERT-Pair-QA [46]: Constructing an auxiliary sentence from aspect words transforms the original task into a sentence pair classification task.
- DualGCN [37]: A dual graph convolutional network model is proposed that considers both syntactic structure and semantic association complementarity.

Visual-text-based:

- Res-MGAN: Connects the visual features extracted by ResNet [40] with the hidden representation of MGAN [45] to achieve a simple combination of text and visual content.
- MIMN [1]: A multi-interactive memory network is proposed to learn the interactive influence of cross-modal data and the self-influence of unimodal data.
- ESAFN [47]: An entity-sensitive attention fusion network to capture dynamic features of aspect-text and aspect-image.
- Res-BERT+BL: Connects the visual features extracted by ResNet [40] with the hidden representation of BERT+BL.
- TomBERT [2]: A goal-oriented multimodal BERT architecture.
- CapTrBERT [3]: Converting images into image captions with aspect words to build auxiliary sentences for sentiment classification.
- HIMT [5]: A generic hierarchical interactive multimodal transformer model.
- JML-MASC [4]: A joint multimodal learning method with assisted cross-modal detection.
- VLP-MABSA [6]: A task-specific vision-language pretraining model.
- ITM [8]: A coarse-to-fine-grained image-target matching model to accurately capture fine-grained and coarse-grained image-target matching.
- ITM-ViT [8]: Replace the visual encoder of ITM with ViT.

TABLE II
EXPERIMENTAL RESULTS COMPARISON ON TWO PUBLICLY AVAILABLE DATASETS

Model	Twitter-2015		Twitter-2017	
	Acc	**Macro-F1**	**Acc**	**Macro-F1**
	Visual			
Res-Target	59.9	46.6	58.6	54.0
	Text			
MGAN	71.2	64.2	64.8	61.5
BERT	74.3	70.0	68.9	66.1
BERT+BL	74.3	70.0	68.9	66.1
BERT-Pair-QA	74.4	67.7	63.1	59.7
DualGCN	75.5	70.4	64.7	60.5
	Text + Visual			
Res-MGAN	71.7	63.9	66.4	63.0
MIMN	71.8	65.7	65.9	63.0
ESAFN	73.4	67.4	67.8	64.2
Res-BERT+BL	75.0	69.2	69.2	66.5
TomBERT	77.2	71.8	70.5	68.0
CapTrBERT	78.0	73.2	72.3	70.2
HIMT	78.1	73.7	71.1	69.2
JML-MASC	78.7	-	72.7	-
VLP-MABSA	78.6	73.8	73.8	71.8
KEF-TomBERT	78.7	73.8	72.1	70.0
ITM	78.3	74.2	72.6	72.0
ITM-ViT	77.3	72.1	71.3	70.1
FITE	78.5	73.9	70.9	68.7
FITE-DE-Large	78.8	74.8	73.9	**73.0**
Ours(DPFN)	**79.9**	**76.0**	**73.9**	72.6

- KEF-TomBERT [9]: A knowledge enhancement framework that improves task performance by exploiting adjective-noun pairs in images. We compare the results of its enhancement on TomBERT [2].
- FITE [7]: Extracting face information from images and translating them into sentiment text to extract visual sentiment information.
- FITE-DE-Large [7]: Replace the text encoder in FITE from BERT to BERTweet-Large [42].

C. Experimental Results

Table II presents the performance of different approaches on Twitter-2015 and Twitter-2017. The results indicate that visual-based methods exhibit the lowest performance. This suggests that with the current technology and methods, the effectiveness of image information in sentiment analysis tasks might be less than that of textual information. Most multimodal methods perform better than unimodal baseline methods, suggesting that image information can complement textual information and improve sentiment classification performance to some extent. Therefore, sentiment information extraction from images is a significant factor in enhancing performance. Methods that utilize significant region features in images as visual information (e.g., FITE [7], ITM [8], KEF-TomBERT [9], VLP-MABSA [6], HIMT [5]) outperform those that use the entire image information (e.g., TomBERT [2] and CapTrBERT [3]), implying that noise in images can hinder sentiment analysis,

and using fine-grained information can alleviate this limitation. FITE-DE-Large [7] achieves the highest performance among previous approaches. The model converts face information in images into sentiment text and aligns aspect targets with facial expressions to obtain visual information associated with aspect words. Another reason for the higher performance achieved by FITE-DE-Large is the use of a more powerful text pretraining model, BERTweet-Large [42].

The main reason that FITE-DE-Large outperforms our model in the Macro-F1 metrics for the Twitter-17 dataset is that they use a more powerful text pretraining model. FITE [7] is the model for changing the text encoder of FITE-DE-Large from BERTweet-Large to BERT. Compared to FITE, our model performs significantly better across all metrics for both datasets. Specifically, our model exceeds its performance in the Macro-F1 metric by 2.8% and 5.7% and in the accuracy metric by 1.8% and 4.2%, respectively. To ensure the fairness of the comparison, we changed the visual encoder of ITM [8] from Faster R-CNN to ViT (denoted by ITM-ViT [8]). Our method outperforms ITM-ViT. The performance gap exhibited when using ViT is even greater compared to using Faster R-CNN. We posit that the reason behind this observation is that while ViT is a robust visual pretraining model, models utilizing Faster R-CNN are specifically designed to extract object-level information from images. ViT may not accurately capture object-level information in images. These results demonstrate the effectiveness of our model on the ABMSA task. Our approach analyses the task from different perspectives and combines information from different modalities to derive the final sentiment polarity. From a global perspective, we translate the images into captions and feed them along with the original text into the text encoder to derive the overall sentiment features. This not only eliminates the semantic gap between different modalities but also fully extracts the objects and information in the images. From a local fine-grained perspective, we construct a text graph to fully extract the fine-grained information in the text and reduce the influence of noise in the text. Additionally, we fuse the text modality with the imaging modality and construct a fusion graph to better fuse the information of different modalities and extract the useful fine-grained information in the image.

D. Ablation Study

To further investigate the effectiveness of each component of our proposed method, we conducted an extensive ablation study, and the results are reported in Table III. We performed six sets of ablation experiments. We removed the text graph construction and discontinued the graph convolution operation on the text features (**w/o TextGCN**). As a result, the model could not learn information related to aspect words from text features using graph convolution operations, which led to a degradation in performance on both datasets. Similarly, we removed the graph convolution operation on fused features (**w/o FusionGCN**). This prevented the model from learning scene information by associating related fine-grained image patches. The performance of the model significantly degraded when this module was removed.

TABLE III
EXPERIMENTAL RESULTS OF ABLATION STUDY

Model	Twitter-2015		Twitter-2017	
	Acc	**Macro-F1**	**Acc**	**Macro-F1**
w/o Local	72.7	67.1	58.7	56.5
w/o Global	77.8	72.6	71.9	69.5
w/o TextGCN	77.6	71.4	71.9	70.3
w/o FusionGCN	78.6	73.8	72.2	69.4
w/o Sim	78.3	74.2	70.9	69.0
w/ LSEM	77.4	72.5	71.5	70.1
DPFN	**79.9**	**76.0**	**73.9**	**72.6**

This experiment demonstrated that the similarity between feature nodes in the fusion graph was used as an important factor in the association strength of each node. Experiments by removing similarity information from the fusion graph (**w/o Sim**) further demonstrated that linking related image patches together can provide more favorable information for aspect-level fusion features. To validate our idea of considering sentiment polarity from a global perspective and from a local fine-grained perspective, we removed the global perspective (global semantic extraction module) (**w/o Global**) and the local fine-grained perspective (local syntax enhancement module and language-guided fusion module) (**w/o Local**) from the model to verify the effectiveness of these two perspectives on the overall model. It is clear from the results that removing either one drastically reduced the model's performance, which shows that our starting point of considering aspect-level sentiment from a different perspective was correct. To verify the performance of the model consisting of only the local syntax enhancement module, we kept only the local syntax enhancement module (w/ LSEM). It can be seen from the results that the performance with only the local syntax enhancement module was lower than that of the complete model. This shows that the reasonable use of beneficial information in images can improve the performance of the model. Overall, all modules in our DPFN model contributed positively to the task, and all achieved optimal performance.

E. Case Study

To further analyze the robustness of our method to error sensitivity, we visualized some prediction results of different methods, as shown in Fig. 4. These methods contain CapTrBERT, VLP-MABSA, our model without TextGCN (denoted by **w/o TextGCN**), and our model without global coarse-grained perspective (denoted by **w/o Global**). As seen in the results, CapTrBERT made incorrect predictions in all the tested examples. This is because image captions can only represent the overall information of the image and do not capture the fine-grained information related to aspects. VLP-MABSA uses object detection to extract image features. In the third test example, object detection was only able to detect people, and the predicted sentiment target was a "**Tel Aviv**" view, so VLP-MABSA made an incorrect prediction. In the second example, the incorrect opinion word "*fun*" is noticed without using TextGCN, so **w/o TextGCN** obtained the incorrect prediction result. In the last example, the

Image				
Text	Chilean pride! Gabriela Mistral honoured on Google today. chile#	RT @ thehill: Warren pokes fun at Scott Brown, Sarah Palin:	Hello, Israel! Tel Avivand the view to Jaffa	RT @ jonathanchait There' s a weird new meme that Clinton' s campaign strategy is tearing America apart.
Target	**chile**	**Sarah Palin**	**Tel Aviv**	**Clinton**
GT	**Positive**	**Neutral**	**Positive**	**Negative**
Pred	CapTrBERT: Neutral ✘ VLP-MABSA: Positive ✓ w/o TextGCN: Positive ✓ w/o Global: Positive ✓ Ours: Positive ✓	CapTrBERT: Negative ✘ VLP-MABSA: Neutral ✓ w/o TextGCN: Positive ✘ w/o Global: Neutral ✓ Ours: Neutral ✓	CapTrBERT: Neutral ✘ VLP-MABSA: Neutral ✘ w/o TextGCN: Positive ✓ w/o Global: Positive ✓ Ours: Positive ✓	CapTrBERT: Neutral ✘ VLP-MABSA: Neutral ✘ w/o TextGCN: Negative ✓ w/o Global: Positive ✘ Ours: Negative✓

Fig. 4. Results of the different methods in the four test cases.

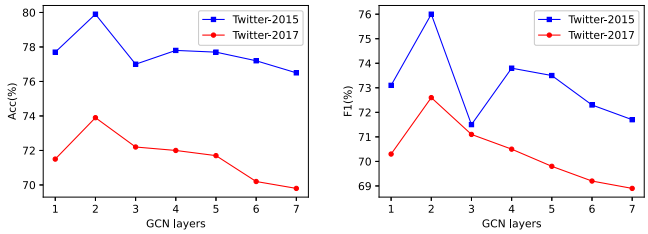


Fig. 5. Effect of the number of GCN layers on the results.

opinion target "**Clinton**" was smiling and giving a thumbs up, which expresses a positive sentiment. However, from the global perspective of the image, it should express a negative sentiment, so our model made an incorrect prediction without considering the global information. In all cases, our complete model made the correct predictions. This not only shows that our model outperforms others but also demonstrates the importance of each component of our model.

F. Impact of the TextGCN and FusionGCN Layer Number

We investigated the effect of the number of layers of GCN on the model on both Twitter-2015 and Twitter-2017 datasets. The results were evaluated for the number of layers 1-7, and the number of layers remained consistent for TextGCN and FusionGCN. The results shown in Fig. 5 indicate that the model with two GCN layers performed best. This is because when the number of layers was too small, the nodes did not propagate very far, such that the aspect words did not capture the information of the relevant opinion words. When the number of layers was too large, a large amount of noise was introduced, making the model prediction results inaccurate.

V. Conclusion

In this article, we analyze the task of aspect-based multimodal sentiment analysis from different perspectives and propose the DPFN model. The overall sentiment tendency is analyzed from the global perspective. The fine-grained sentiment of aspectual words is analyzed from a local perspective by combining textual and image information. Finally, the final aspect-level affective polarity is obtained from global and local perspectives. Experimental results show that considering the overall sentiment information can improve the performance of this task. Our proposed approach achieves state-of-the-art performance in the ABMSA task on two benchmark datasets, Twitter-15/17. In our model, syntactic information is used to capture information related to aspectual words. We also use linguistic information to guide the extraction of visual information and perform fine-grained feature fusion. In the future, we will further explore how to mine sentiment-related information from different modalities.

References

[1] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 371–378.
[2] J. Yu and J. Jiang, "Adapting BERT for target-oriented multimodal sentiment classification," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 5408–5414.
[3] Z. Khan and Y. Fu, "Exploiting BERT for multimodal target sentiment classification through input space translation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3034–3042.
[4] X. Ju et al., "Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4395–4405.
[5] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1966–1978, Jul.-Sep. 2023.
[6] Y. Ling, J. Yu, and R. Xia, "Vision-language pre-training for multimodal aspect-based sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2149–2159.
[7] H. Yang, Y. Zhao, and B. Qin, "Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 3324–3335.
[8] J. Yu, J. Wang, R. Xia, and J. Li, "Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, 2022, pp. 4482–4488.
[9] F. Zhao et al., "Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 6784–6794.
[10] T. Zhu et al., "Multimodal sentiment analysis with image-text interaction network," *IEEE Trans. Multimedia*, vol. 25, pp. 3375–3385, 2023.

[11] Q.-T. Truong and H. W. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 305–312.
[12] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in *Proc. 9th ACM Int. Conf. Web Search Data Mining*, 2016, pp. 13–22.
[13] Q. You, L. Cao, H. Jin, and J. Luo, "Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1008–1017.
[14] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[15] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," *Int. J.*, vol. 2, no. 6, pp. 282–292, 2012.
[16] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 8, no. 4, 2018, Art. no. e1253.
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
[18] Z. Yang et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[19] Z. Yang et al., "XlNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.
[20] H. Tian et al., "SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4067–4076.
[21] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "SentiLARE: Sentiment-aware language representation learning with linguistic knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 6975–6988.
[22] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang.*, 2020, pp. 1–7.
[23] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558—-6569.
[24] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[25] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2020, pp. 2359–2369.
[26] D. Wang et al., "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Trans. Multimedia*, early access, Jun. 16, 2022, doi: 10.1109/TMM.2022.3183830.
[27] J. Zeng, J. Zhou, and T. Liu, "Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities," *IEEE Trans. Multimedia*, early access, Sep. 19, 2022, doi: 10.1109/TMM.2022.3207572.
[28] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, 2021.
[29] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.
[30] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 452–461.
[31] B. Huang, Y. Ou, and K. M. Carley, "Aspect level sentiment classification with attention-over-attention neural networks," in *Proc. 11th Social, Cultural, Behav. Model. Int. Conf.*, 2018, pp. 197–206.
[32] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Targeted sentiment classification with attentional encoder network," in *Proc. Artif. Neural Netw. Mach. Learn.*, 2019, pp. 93–103.
[33] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 214–224.
[34] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 4568–4578.
[35] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3229–3238.
[36] H. Tang, D. Ji, C. Li, and Q. Zhou, "Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6578–6588.
[37] R. Li et al., "Dual graph convolutional networks for aspect-based sentiment analysis," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6319–6329.
[38] Z. Wu et al., "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[42] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proc. Conf. Empirical Methods Natural Lang. Process. Syst. Demonstrations*, 2020, pp. 9–14.
[43] K. Mrini et al., "Rethinking self-attention: Towards interpretability in neural parsing," in *Proc. Findings Assoc. Comput. Linguistics*, 2020, pp. 731–742.
[44] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[45] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3433–3442.
[46] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 380–385.
[47] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 429–439, 2020.

Di Wang (Member, IEEE) received the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2016. She is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. She has authored or coauthored several scientific articles in her research areas which include machine learning and multimedia information retrieval, refereed journals including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and conferences including the SIGIR and IJCAI.

Changning Tian received the B.S. degree in software engineering from the Lanzhou University of Technology, Lanzhou, China, in 2021. He is currently working toward the M.S. degree in computer science and technology with Xidian University, Xi'an, China. His research interests include machine learning and multimodal sentiment analysis.

Xiao Liang received the M.S. degree from the Graduate School of Information, Production and Systems, Waseda University, Tokyo, Japan, in 2018. She is currently working toward the Ph.D degree with the School of Computer Science and Technology, Xidian University, Xi'an, China. Her research interests include computer vision and multimodal information processing.

Lin Zhao received the Ph.D. degree in pattern recognition and intelligent systems form Xidian University, Xi'an, China, in 2017. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science & Technology, Nanjing, China. His research interests include human related computer vision tasks, such as human pose estimation and tracking, 3D human shape reconstruction, and anomaly detection.

Lihuo He received the B.Sc. degree in electronic and information engineering and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2008 and 2013, respectively. He is currently an Associate Professor with Xidian University. His research interests include computational vision, pattern recognition, and artificial intelligence.

Quan Wang received the B.Sc., M.Sc., and Ph.D. degrees in computer science and technology from Xidian University, Xi'an, China. He is currently a Professor with the School of Computer Science and Technology, Xidian University. His research interests include input and output technologies and systems, image processing, and image understanding.