# Weakly Supervised Interaction Discovery Network for Image Sentiment Analysis

Lifang Wu, Heng Zhang, Ge Shi and Sinuo Deng



November 10, 2021

# Weakly Supervised Interaction Discovery Network for Image Sentiment Analysis

Lifang Wu, Heng Zhang, Ge Shi ✉, and Sinuo Deng

Faculty of Information Technology, Beijing University of Technology, Beijing, China
tinkersxy@gmail.com

**Abstract.** Visual sentiment is subjective and abstract, and it is very challenging to locate the sentiment features from images accurately. Some researchers devote themselves to extracting visual features but ignore the relation features. However, sentiment reaction is a comprehensive action of visual content, and regions may express different emotions and contribute to the image sentiment. This paper takes the abstract sentiment relation as the starting point and proposes the Weakly Supervised Interaction Discovery Network that couples detection and classification branch. Specifically, the first branch detects sentiment maps with the cross-spatial pooling strategy, which generates the representations of emotions. Then, we employ a stacked Graph Convolution Network to extract the interaction feature from the above features. The second branch utilizes both interaction and visual features for robust sentiment classification. Extensive experiments on six benchmark datasets demonstrate that the proposed method exceeds the state-of-the-art methods for image sentiment analysis.

**Keywords:** visual sentiment analysis · sentiment classification · convolutional neural networks · graph convolution network.

## 1 Introduction

The development of social networks has attracted many researchers to dig and explore the emotional information in social media. There have been some applications, such as advertising, education, and other fields. Among them, because images can express human emotions and ideas more intuitively, image sentiment analysis has gradually become an essential part of the field of emotion computing.

The sentiment is a vague and subjective abstract concept, and analyzing image sentiment is a challenging task. In the existing research of image sentiment analysis, researchers devote themselves to accurately extracting the sentimental information contained in the image. In 2016, some researchers applied deep learning to sentiment feature extraction and achieved a good performance. Later, Yang et al. [14] put forward the concept of "Affective Region." They extract features from regions with strong sentiment in the image and supplement fine-grained features utilizing feature fusion strategy such as concatenation or pooling. Later, with the help of a saliency detection tool, Wu et al. [12] propose to extract the local sentiment feature from the salient region and achieve remarkable performance improvement. The above two methods assume that the image sentiment comes from the corresponding local region, ignoring the emotional connection between different regions. Further, Yang et al. [13] propose the Weakly Supervised Coupled Networks (WSCNet) to obtain the emotional regions of the image and get the "Sentiment Map" to describe the emotion interaction. However, they adopt the weighted sum strategy to represent the relationship among emotions, which simplifies the emotion interaction. As shown in Fig. 1, in the emotional response of human beings, areas with corresponding emotion makes a non-negligible contribution to sentiment, and there is still hard-to-describe sentiment interaction information among the regions.

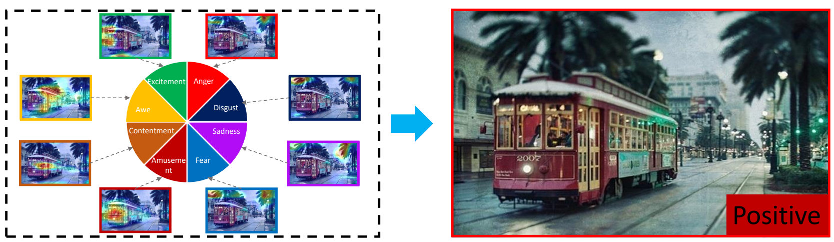


**Fig. 1.** Examples of image sentiment. We label sentiment regions with heat maps, and extract the sentiment interaction features according to the emotional relations of Mikel's emotion model.

To solve this problem, we propose an end-to-end model to extract and utilize emotional interaction characteristics. Inspired by the work of Yang et al. [13], instead of dragging the boxes of objects in the image, we use an emotion map to describe the weight of the region in the image that stimulates the corresponding emotion. Specifically, for the feature maps obtained in the convolutional neural network, a cross-spatial pooling strategy is adopted to obtain the spatial information and the regions in the highlight image to stimulate corresponding emotions. Then, we use a topological diagram to define and describe the emotional relationship, take the regional features corresponding to multiple emotions as nodes, and take the weight between categories in the emotion model as the corresponding "sentiment map". Then, GCN is used to update and aggregate the node features and obtain the emotional interaction features. Finally, we integrate the visual features representing the scene information with the interactive features representing the sentiment interaction information to realize the prediction of emotion categories. Our method only needs the image-level labels of the sentiment in the training process, making full use of the existing image emotion image datasets.

Our contributions are as follows:

Firstly, we propose to describe the sentiment interaction at the cognitive level with a graph and define the emotional relationship with the results of psychological research. For an input image, we transform it into a graph and demonstrate the effectiveness of sentiment relation knowledge.

Secondly, we propose a model that makes full use of sentiment interaction features rather than visual features. We use the weakly supervised sentiment region detection method to identify the interaction semantics extracted from the visual semantics at the cognitive level, which effectively utilizes sentiment interaction features and reduces data preprocessing steps.

Finally, our method achieves state-of-the-art. It is proved that the extensive use of cognitive features and visual features can promote image sentiment analysis.

## 2 Related Work

### 2.1 Visual Sentiment Prediction

The sentiment is an abstract and subjective concept. Accurate extraction of emotional features in images is one of the difficulties in image sentiment analysis. Some researchers are working to bridge the "affective gap" between visual content and sentiment. Inspired by psychological research, Machajdik and Hanbury [6] realized the task of image emotion classification by utilizing the handcraft features, such as color, texture, composition, and so on. Based on [11], Sun et al. proposed the sentiment region based on the object proposal method and realized sentiment classification using corresponding depth features.

Further, Yang et al. [14] proposed to utilize the Affective Region (AR) with the help of instance segmentation tools. By fusing the features of AR and original images, they get a better classification performance. Wu et al. [12] utilize a saliency detection algorithm to enhance local features and improve classification performance in a large margin.

These methods focus on extracting features at the visual level and ignoring affective relationships at the cognitive level. Later, with the help of instance segmentation, Wu et al. proposed to leverage the sentimental interaction information among objects. However, the gap between object and sentiment limits the improvement of this method.

Different from previous works, we propose an end-to-end approach. This model aims to extract the sentimental relation features from sentiment semantics, which avoids the loss of information between object semantics and sentiment semantics and omits data preprocessing steps.

### 2.2 Weakly Supervised Detection

Compared with the objective existence of objects, sentiment is subjective and fuzzy. Therefore, the semantic level information extraction of objects cannot fully express the emotional information in the image. Besides, few works concentrating on end-to-end CNN frameworks for weakly supervised object detection without additional localization information.

There is a work concentrating on weakly supervised sentiment detection [13], they proposed joint sentiment detection and classification and achieved an improvement with the "Sentiment Map". Specifically, they proposed the cross-spatial pooling strategy to summarize the feature maps of the network into "sentiment map", which is a weighted sum of the features of different emotions. This method effectively overcomes the problem of weak feature differentiation caused by the fuzziness of human emotion, but there are still some limitations in the exploration of emotional relationships. In the absence of sentimental domain knowledge, the response values of feature maps may not fully represent the interaction among sentiments. We utilize the convolutional network and cross-spatial pooling strategy to detect the sentiment regions and design an interaction feature extraction method based on GCN to make use of the relational information in the sentiment regions.

### 2.3 Graph Convolutional Network

Graph neural networks were proposed by Gori et al. [5] and further developed by Scarselli et al. [8]. However, due to the limitations of methods and computer technology, this method needs many computing resources on massive data, which is challenging to realize. Further, Bruna et al. [2] proposed the graph convolution networks, which attracts the attention of researchers in various fields.

Different from the CNN model, the graph describes the relations among nodes by building a relational model. Chen et al. [4] proposed to utilize the inter-dependent object information from labels in multi-label image classification. However, this method needs the annotation of image objects, which requires a lot of human resources.

In this paper, we employ the graph structure to capture and explore the sentimental interaction information. Specifically, we couple the sentiment detection and classification tasks. With the help of weakly supervised sentiment region detection, we employ a stacked GCN model to capture the sentiment interaction feature in images.

## 3 Method

This section aims to develop an algorithm to extract the sentimental relation information with only image-level labels. An overview of our proposed Weakly Supervised Interaction Discovery Network is illustrated in Fig. 2. The proposed WSINDNet learns both detection and classification tasks with two branches. We employ the detection branch to generate the sentiment regions and utilize GCN to leverage the sentiment interaction information, which is then fed into the classification branch to fuse the holistic as the relational representations.

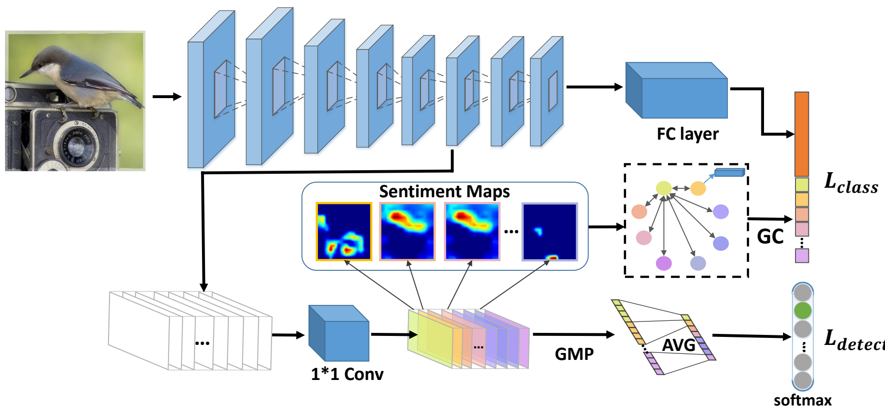


**Fig. 2.** Overview of the Weakly Supervised Interaction Discovery Network. During training, the model needs only image-level sentimental labels and we adopt a joint training strategy to couple the detection and classification branches.

### 3.1 Sentiment Map Detection

Unlike the target of the location task, the response map of emotions is independent due to the fuzziness of human emotions. The same area in the image may express multiple emotions, and one emotion can also correspond to different regions. Inspired by the work of Yang et al. [13], we also adopt the cross-spatial pooling strategy to realize the weakly supervised sentiment region detection.

Specifically, we employ VGGNet as the backbone, and utilize the feature maps to generate the sentiment regions. Let $\{(x_i, y_i)\}_{i=1}^{N}$ be a collection of training examples, where $x_i$ is an image, and $y_i$ is the corresponding sentiment label. For each instance, let $F \in \mathbb{R}^{w \times h \times n}$ be the feature maps of the last convolution layer, where $w$ and $h$ are the spatial size (width and height) of the feature maps, and n is the number of feature channels. As shown in Fig. 2, the $1 \times 1$ convolution layer is used as k detectors to capture the high response regions for each emotion category which results in $F' \in \mathbb{R}^{w \times h \times kC}$. By summarizing all the information to a image-level score, the cross-spatial pooling strategy achieves the weakly supervised region detection regardless of the input size:

$$v_c = \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{G}_{\max}\left(f_{c,i}\right), c \in \{1, \cdots, C\} \tag{1}$$

where $f_{c,i}$ is the $i$-th feature map of $c$-th emotion. $G_{\max}$ represents the Global Max Pooling (GMP), which is utilized to identify the discriminative parts of the feature maps and generate a $1 \times 1 \times kC$ vector. $k$ represents an average pooling operation to maximize the discriminative feature and results in the vector $\mathbf{v} \in \mathbb{R}^C$, which is fed into a $C$-class soft-max layer to supervise the detection performance.

$$L_{detect} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbf{1}\left(y_i = c\right)\log v_c \tag{2}$$

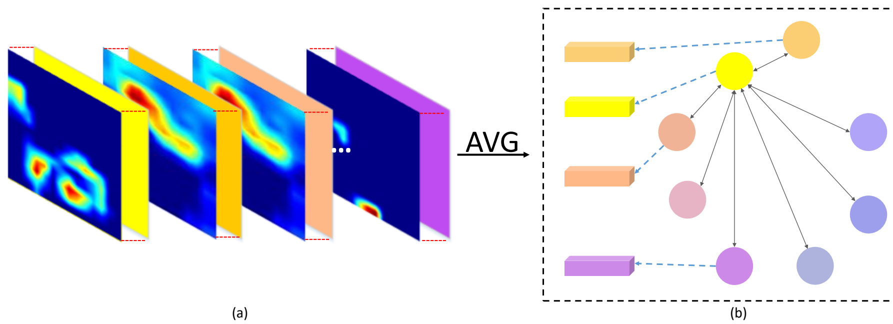


**Fig. 3.** An example of sentiment graph: (a) An object segmentation result, where the object are distinguished by different color overlay. (b) A sentiment graph structure, where the nodes represent objects of corresponding color and edges reflect the similarity of nodes in the sentiment space.

Unlike Yang et al., we aim to leverage the sentiment interaction and did not sum the feature map of multiple emotions into a sentiment map. We retain the $F'$ by average pooling and get corresponding heatmaps of emotions, which is repeat to $w \times h \times n$ and make Hadamard product with $F$ to highlight the features of connected regions. Then, As shown in Fig. 3, we employ an average pooling operation to generate the feature vectors corresponding to each emotion category.

### 3.2 Sentiment Interaction Extraction

**Sentiment Graph** In the introduction, we introduced the contribution of sentiment interaction to human emotion. As an abstract and subjective logical response, sentiment is difficult to capture and extract. In addition to the visual features, we propose the Sentiment Graph to define and extract the relation features in the image. Specifically, we construct a unique undirected graph with emotion categories as nodes, describe the relationship between emotions with an adjacency matrix, and extract the sentiment interaction features with stacked GCNs.

To accurately describe the emotional relationship between nodes, we take the distance between emotions in Mikels' emotional model to measure emotional similarity and reciprocal emotional distance as the adjacency matrix in the Sentiment Graph. At the same time, although the sentiment distance can effectively represent emotional similarity, it cannot reflect the difference between the positive and negative, so we put forward the method in Formula 3 to calculate the sentiment relationship:

$$A_{ij} = \begin{cases} \frac{1}{dis_{ij}} + 1, \text{ if } S_i * S_j < 0 \\ \frac{1}{dis_{ij}}, \text{ otherwise} \end{cases} \mathbf{1}\left(y_i = c\right) \log v_c \tag{3}$$

where $A_{ij}$ is a element of adjacency matrix $A$, $S_i$, $S_j$ represent the polarities of $i$-th, $j$-th emotion and $dis_{ij}$ is the sentiment distance between them.

**Interaction Extraction** To simulate the sentimental interaction, we select GCN to propagate and aggregate the representation of objects under the supervision of sentiment relations. Specifically, we employ the stacked GCNs, in which the input of each layer is the output $H^l$ from the previous layer, and output the new node feature $H^{l+1}$. The feature of the first GCN layer $H^0$ is generated from the detection branch introduced above.

Formula (4) shows the feature update process of layer $l$, where $\tilde{A}$ describes the relationship among nodes. $H^l$ is the output of the previous layer $l-1$, and $H^{l+1}$ is the output of the current layer, $W^l$ is the weight matrix of the current layer, and $\sigma$ is the nonlinear activation function.

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^l W^l) \tag{4}$$

In addition, $\tilde{D}$ is the degree matrix of $\tilde{A}$, and obtained by Equation (5).

$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij} \tag{5}$$

### 3.3 Joint Training Strategy

**Sentiment Classification Branch** Like previous works [11, 12, 14, 15, 17], we select VGGNet as the backbone to capture the deep feature of images. To highlight the effect of sentiment interaction and make a fair comparison with previous works, we keep this branch as simple as possible. Previous studies have demonstrated the sentimental feature extraction capability of VGGNet with 16 layers [10]. We select it as the backbone to supplement the global context information missing in the interactive features. Besides, we changed the last fully connected layer from 4096 to 2048 and get the image feature $F_h$.

From the perspective of image representation, the original convolutional feature $F_h$ represents the holistic feature of the image. The sentiment interaction feature provides fine-grained features that contain sentiment interaction information. We use the concatenate operation to fuse the two features, which results in $F = [F_h; F_s]$. $F_s$ represents the sentiment interaction feature generated from the GCN model. The classification is carried out by minimizing the following loss function:

$$L_{class} = -\frac{1}{N}\sum_{i=1}^{N}(y_i * log\hat{y}_i + (1-y_i) * log(1-\hat{y}_i)) \tag{6}$$

**Joint Training** We adjust the joint training strategy to couple the sentiment detection and classification task. By minimizing the collective loss function, we can detect the sentiment regions with the image-level label and leverage the sentiment regions to extract the sentiment interaction information to facilitate classification. The joint loss function is described in Formula 7.

$$L = L_{class} + \lambda L_{detect} \tag{7}$$

## 4 Experiment

### 4.1 Datasets

We evaluate our method on five public datasets: **FI** [17],**Flickr** [1], **Emotion-ROI** [7], **Twitter I** [16] and **Twitter II** [1]. FI is collected from Flickr and Instagram. The researchers select eight emotion categories (i.e., amusement, anger, awe, contentment, disgust, excitement, fear, sadness) as keywords to query images, and they get about 90,000 raw images with noise. Then, they employ 225 Amazon Mechanical Turk (AMT) workers to annotate the emotion and result in 23,308 images. Flickr contains 484,258 images from Flickr, which are labeled by a corresponding adjective and noun pairs (ANP) automatically. Though Flickr has an enormous data scale, automatic labeling makes it less reliable. Emotion-ROI contains 1,980 images with six emotion labels (i.e., anger, disgust, fear, joy, sadness, surprise), which are annotated manually with 15 regions that can evoke human emotions. Twitter I and Twitter II are annotated with sentiment labels (positive and negative) by AMT workers, consisting of 1296 and 603 images. Specifically, following [13], we conduct training and testing on the three subsets of Twitter I: 'Five agree', 'At least four agree', and 'At least three agree', which are filtered based on the annotation results. For example, 'Five agree' means that the five AMT workers label the same sentiment to a given image.

### 4.2 Baselines

To demonstrate the performance of our proposed method, we compare our approach against several previous works, including methods using traditional features, CNN-based methods, and CNN-based methods with a local feature branch.

- The global color histograms (**GCH**) extract 64-bin RGB histogram as image representation, and the local color histogram features (**LCH**) [9] divide image into 16 blocks and calculate 64-bin RGB histogram for each block.
- Borth et al. [1] introduced semantic information by **SentiBank**, and they filtered 1,200 ANPs as the presentation of sentiment semantic.
- **DeepSentibank** [3] employs CNN to realize both ANPs prediction and sentiment classification. We utilize a pre-trained DeepSentiBank to extract 2089-dimension deep features from the last fully-connected layer and realize sentiment prediction by LIBSVM.

- You et al. [16] build a potentially cleaner dataset and proposed **PCNN** trained with weakly supervised data and achieve a generalization improvement.
- Yang et al. [14] proposed the 'Affective Regions' with the help of instance segmentation, which contain reach sentiment and object information, and the design three fusion strategy to fuse the 'Affective Regions' and image feature.
- Wu et al. [12] proposed to utilize the salient regions in sentiment analysis by salient detection algorithm and achieved a significant performance improvement.

### 4.3 Implementation Details

Following previous works [14], we select VGG-16 [7] as backbone and initialize it with pre-trained model on ImageNet. We randomly crop and resize the input images into 224 $\times$ 224 with random horizontal flipped for data enhancement. SGD is selected as the optimizer, and Momentum is 0.9. The initial learning rate is 0.01, which drops by a factor of 10 per 20 epoch. We set the hyper-parameters $\lambda$ to be 0.1, which is tuned on the FI validation set.

To make a fair comparison, we adopted the same split for five datasets with Yang et al. [14]. At the same time, we also convert the dataset labels with emotion categories to the sentiment labels. For example, EmotionROI has six emotion categories: anger, disgust, fear, joy, sadness, and surprise. Images with labels of anger, disgust, fear, sadness are relabeled as negative, and those with joy and surprise are labeled as positive. Also, different datasets have different emotion categories. We select the detection branch training parameters of FI, the larger manually annotated dataset, as the initialization weight of other datasets.

**Table 1.** Sentiment classification performance on FI, Flickr, Twitter I, Twitter II, EmotionROI. Results with bold indicate the best accuracy compared with other algorithms.

| Method | FI | Flickr | Twitter I | | | Twitter II | EmotionROI |
|---|---|---|---|---|---|---|---|
| | | | Twitter I-5 | Twitter I-4 | Twitter I-3 | | |
| GCH | - | - | 67.91 | 97.20 | 65.41 | 77.68 | 66.53 |
| LCH | - | - | 70.18 | 68.54 | 65.93 | 75.98 | 64.29 |
| SentiBank | - | - | 71.32 | 68.28 | 66.63 | 65.93 | 66.18 |
| DeepSentiBank | 61.54 | 57.83 | 76.35 | 70.15 | 71.25 | 70.23 | 70.11 |
| VGGNet [10] | 70.64 | 61.28 | 83.44 | 78.67 | 75.49 | 71.79 | 72.25 |
| PCNN | 75.34 | 70.48 | 82.54 | 76.50 | 76.36 | 77.68 | 73.58 |
| Yang [14] | 86.35 | 71.13 | 88.65 | 85.10 | 81.06 | 80.48 | 81.26 |
| Wu [12] | 88.84 | 72.39 | 89.50 | 86.97 | 81.65 | 80.97 | 83.04 |
| Ours | **89.18** | **74.53** | **91.25** | **87.96** | **84.34** | **81.32** | **83.33** |

### 4.4 Classification Results

As shown in Table 1, we evaluate the performance of our proposed WSIDNet against different methods on five datasets. Compared with the hand-crafted features, CNN-based models can extract the sentimental feature from images. Our proposed method performs favorably against the state-of-the-art methods for sentiment classification, e.g., about 2.14% improvement on Flickr and 1.75% on TwitterI-5 datasets, which illustrates that WSIDNet can learn more discriminative representation for sentiment classification.

### 4.5 Ablation Study

The ablation study results on five datasets are shown in Table 2. Compared with backbone, our WSIDNet further improves the classification performance by 6.13% and 7.59% on FI and TwitterI-3 datasets. These results suggest that the sentiment interaction feature can effectively provide the information at the cognitive level.

**Table 2.** The model performance comparison across image datasets.

| Method | FI | Flickr | Twitter I | | | Twitter II | EmotionROI |
|---|---|---|---|---|---|---|---|
| | | | Twitter I-5 | Twitter I-4 | Twitter I-3 | | |
| Backbone | 83.05 | 70.12 | 84.35 | 82.26 | 76.75 | 76.99 | 77.02 |
| WSIDNet | **89.18** | **74.53** | **91.25** | **87.96** | **84.34** | **81.32** | **83.33** |

### 4.6 Detection Results

Fig. 4(a) shows the detected sentiment maps for a joy image from the EmotionROI generated by the sentiment detection branch. Compared with the Emotion Stimuli Map that human annotates, the heat maps of amusement, excitement, and contentment have a high consistency with ground truth. In particular, awe tends to come from the natural landscape, so it is different from other emotions that express positive categories. In the classes of negative emotion, anger and fear focus on the boy's clenched fist, while disgust comes from the soil. Besides, the scene of the boy leaving expresses more sadness.

We also display some poor results of the detection branch. As shown in Fig. 4(b), the detection results of the negative sentiment are concentrated in the debris accumulation in the image, which is significantly different from the marked area in the ground truth. This indicates that the image does not necessarily contain all the emotion categories, and the performance of weakly supervised sentiment detection is still limited.

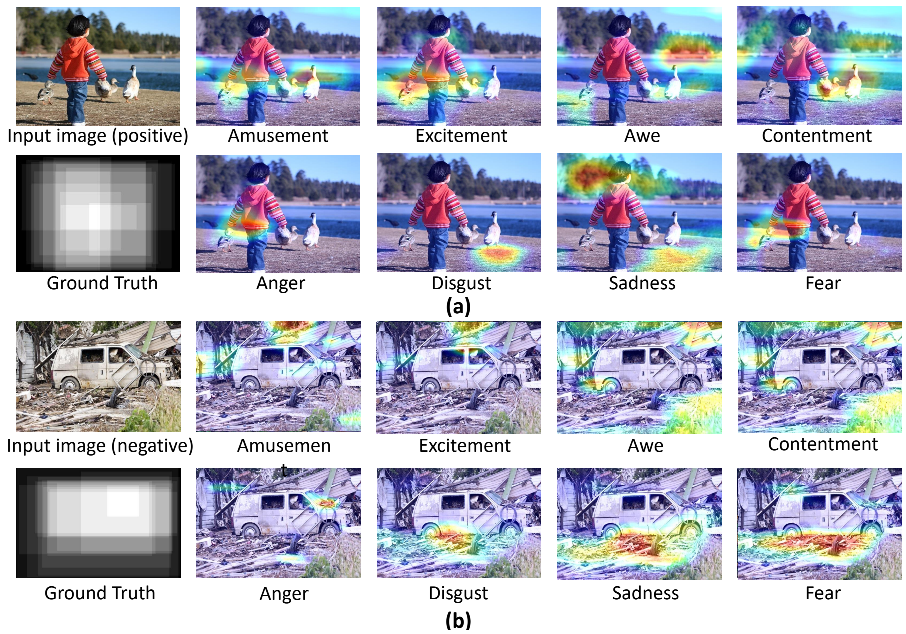


**Fig. 4.** Detected sentiment maps of the proposed WSIDNet on the EmotionROI. (a) detection results the detection branch (b) Poor results of the sentiment region detection branch with wrong and incomplete detection locations

## 5 Conclusion

This paper addresses fuzzy and subjective emotion in visual emotion analysis by utilizing sentiment interaction information. In particular, we propose a Weakly Supervised Interaction Discovery Network, an end-to-end model to couple the detection and classification task. Firstly, we adjust the cross-spatial pooling operation to realize automatic detection of sentiment and design a "sentiment graph" to model the sentiment relation, which takes the emotion as nodes and defines the adjacency matrix with the sentiment distance. Then, we employ a stacked GCN model to aggregate and update node features to obtain the expression of sentiment interaction. We evaluated the model's performance on five public datasets, and our approach exceeded the best available. Also, how to make more effective use of cognitive object interaction information is still a challenging problem.

## 6 Acknowledgement

This work was supported in part by Beijing Municipal Education Committee Science Foundation (KM201910005024), Beijing Postdoctoral Research Fundation (Q6042001202101).

## References


1. Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.
2. Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and deep locally connected networks on graphs. In *2nd International Conference on Learning Representations, ICLR 2014*, pages 61–80, 2014.
3. Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.
4. Zhao Min Chen, Xiu Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.
5. Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.
6. Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010.
7. Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. Where do emotions come from? predicting the emotion stimuli map. In *ICIP*, 2016.
8. Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
9. Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 715–718, 2010.
10. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.
11. Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *ICME*, 2016.
12. Lifang Wu, Mingchao Qi, Meng Jian, and Heng Zhang. Visual sentiment analysis by combining global and local information. *Neural Processing Letters*, pages 1–13, 2019.
13. Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018.
14. Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525, 2018.
15. Quanzeng You, Hailin Jin, and Jiebo Luo. Visual sentiment analysis by attending on local image regions. In *AAAI*, 2017.
16. Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.
17. Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, 2016.