ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230622

DCNet: Weakly Supervised Saliency Guided Dual Coding Network for Visual Sentiment Recognition

Xinyue Zhang[a, b], **Jing Xiang**[c], **Hanxiu Zhang**[c], **Chunwei Wu**[c], **Hailing Wang**[c] **and Guitao Cao**[b, c;*]

[a]School of Computer Science and Technology, East China Normal University, Shanghai, China
[b]Shanghai Institute of AI for Education, East China Normal University, Shanghai, China
[c]MoE Engineering Research Center of SW/HW Co-design Technology and Application, East China Normal University, Shanghai, China

Abstract.

Visual sentiment recognition is a challenging task with scientific significance in probing vision-processing mechanisms. Recent approaches mainly focused on using overall images or precise annotations to learn emotional representations, yet neglected to capture abstract semantics from regional information, or led to a heavy annotation burden. In this paper, we propose an end-to-end weakly supervised framework, called **Dual Coding Network** (DCNet), which models a dual coding process for both shallow features and high-level regional information. On the one hand, with the help of the fine-grained module (FG), visual features (e.g. texture features) are utilized to enhance the learning of distinguished representation. On the other hand, the DCNet innovatively leverages saliency information to imitate the neural decoding of perceived visual sentiment contents in human brain activity. Specifically, the saliency information guides the generation of sentiment-specific pseudo affective maps (SAMG), which serve as weak annotations. Then the DCNet couples fine-grained features with pseudo affective maps, and obtains semantic vectors for final sentiment prediction. Extensive experiments show that the proposed DCNet outperforms the state-of-the-art performance on five benchmark datasets.

1 Introduction

Visual information in social media offers useful information and enables people to share their instant psychological and physiological status [11]. Visual sentiment recognition plays a considerable role in understanding the sentimental response when humans see specific visual content. Therefore, probing visual sentiment recognition in-depth could benefit various potential applications, such as opinion mining [25], and affective computing [33], to name a few.

Nowadays, since the success of deep learning, numerous deep learning approaches have been proposed to predict image emotions [29, 28]. As shown in (a) of Figure 1, global-level strategies predict visual sentiment with entire images. Lin et al. [9] fed the whole image into a multi-source domain adaptation method to predict the sentiment results. Peng et al. [15] built their global-level architecture with a transfer learning strategy. Rather than excavating sentimental information from the entire images, some researchers discover affective regions in the images. As shown in (b) of Figure 1, some studies adopt weak supervision to capture classification-specific information, which predicts final results relying on object scores. For example, Zhang et al. [28] predicted the final sentimental classification via class activation mapping (CAM) technology [32]. Different from weak supervision based methods, proposal based research relied on precise annotations to localize affective regions by calculating the affective scores of different substantial image regions with the proposed arithmetic formula (shown in (c) of Figure 1). For instance, Zhang et al. [30] exploited the object detection proposals that are potentially reflected sentiments to predict emotional results.

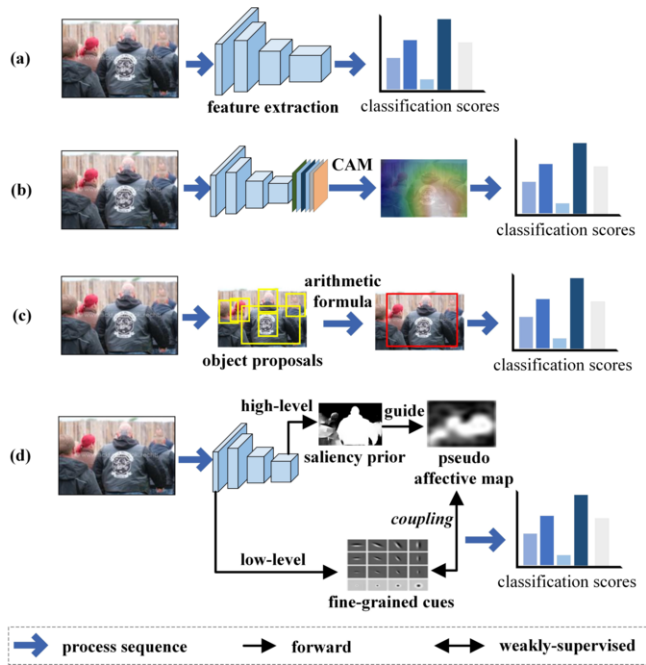

Figure 1. Illustration of different visual sentiment recognition methods. (a) global level visual sentiment recognition method. (b) class activation mapping (CAM) [32] based weakly supervised architecture. (c) detection based local-level architecture. (d) our proposed dual coding architecture.

However, several issues exist when using the aforementioned deep learning methods to address the visual sentiment recognition task, which is explained as follows. First, learning visual features from integral images merely expresses the dominant emotional tendency in

* Corresponding Author. Email: gtcao@sei.ecnu.edu.cn.

the image, while ignoring the hierarchical nature of human cognitive neuroscience. Second, weak supervision based methods merely pay attention to attractive objects, leading to unsatisfactory predictions with coarse affective regions. Third, proposal based methods need a complex and exhaustive process to obtain and select the appropriate annotations.

Recognizing this, we seek solutions to solve the above questions. Psychological research has shown that visual features (e.g. low-level features) and people's prior experiences (e.g. contour of targets) influence human perception and recognition of visual stimuli [10]. For example, when we see an image, we naturally perceive the contours of the affective regions in our mind. Then the concept of affective regions is visually encoded in the brain, and are serving as compasses for the prediction of human emotions. Inspired by the research result, we believe that the visual sentiment should be decoded using a combination of both the actual visual semantic features that are presented, and the prior salient semantic information that is associated with the sentiment. Specifically, we propose a weakly supervised framework, called Dual Coding Network (DCNet) to learn the discriminative representation for visual sentiment analysis, which is shown in (d) of Figure 1. First, different from the global level strategy that analyzes the sentiment from the entire image, we focus on discovering precise affective regions in an automatic manner. Second, unlike existing weak supervision methods that ignore subjective objects and background, we introduce saliency clues to imitate human-like visual processes and obtain prior experiences. Concretely, we leverage saliency-guided pseudo maps to reason higher-level sentiment. The saliency information is derived from a local attention mechanism, which guides the generation of pseudo affective maps. Then DCNet indicates the probability of evoking the emotion in each receptive field, and generates saliency-guided pseudo affective maps as soft pseudo labels. Besides, based on the fact that low-level features stimulate association and touch emotion to a certain extent, we consider fine-grained cues as the other coding process. In the fine-grained process, each pixel of shallow features corresponds to a small area of perceptual field overlap, which helps to capture more details. Finally, DCNet utilizes weakly supervised coupling as a bridge to connect the dual coding process, and avoids the time-consuming selecting process for appropriate annotations that evoke sentiment. The coupling operation addresses the third challenge in an end-to-end manner for human-like intelligence.

In summary, our main contributions are listed as follows:

(1) We propose a novel weakly supervised based dual coding network (DCNet) for visual sentiment recognition. DCNet couples pseudo affective maps derived from high-level visual semantics and fine-grained cues facilitated by shallow feature details, which achieves a consistent multi-level joint representation.

(2) We innovatively utilize saliency information as guidance for generating affective regions in the field of visual sentiment recognition. The saliency information enhances the representation of prominent regions and objects, for neural decoding of perceived visual sentiment categories in human brain activity.

(3) We have conducted experiments on benchmark datasets of different sizes. Experimental results show that our proposed network can effectively improve sentiment perception performance.

2 Methodology

2.1 Overview of the Proposed Architecture

As shown in Figure 2, the proposed network is an end-to-end architecture. We adopt ResNet-101 [5] as our backbone network (denoted as $D_i(i = 1, 2, 3, 4, 5)$) to extract multi-level features from the original image. To be concrete, we remove the average pooling and fully connected layers, and replace them with the proposed two modules. We denote the output features of the feature extraction branch as $d_i(i = 1, 2, 3, 4, 5)$. In the dual coding process, on the one hand, the saliency-guided affective map generation module (SAMG) consists of two stages: the saliency guidance stage and the pseudo affective map generation stage. Initial information derived from the saliency guidance stage is followed by the second stage. The pseudo affective map generation stage learns the predicted score of affective regions via the prior classification loss L_{prior} with the help of the ground truth labels. On the other hand, the lowest level features d_1 and the highest level features d_5 are fed into the fine-grained module (FG) to efficiently leverage low-level valuable details. Then the output is passed through a global average pooling (GAP) layer and fully connected layers to learn the predicted results via the second classification loss L_{cls}.

2.2 Saliency Guided Pseudo Affective Map Generation

Visual stimuli theory deems saliency as stimulus features noticed by the natural vision system and task demands, which affect observers' attention. However, the saliency map only focuses on foreground objects, yet ignores other sentiment-related areas. In SAMG, the pseudo affective map is guided by the saliency map for the final classification. The combination overcomes the above challenges, and provides visual sentiment recognition with local information from the perspective of weakly-supervised and local representation.

2.2.1 Stage-I: Saliency Prior Enhancement

In stage I, we employ channel attention to suppress noise background interference and facilitate intra-class consistency. Specifically, given a set of highest-level feature vectors $d_5 \in R^{C\times H\times W}$ (where C is the channel, W is the width, and H is the height of feature map). Then reshape $H \times W$ dimension of d_5 to $d_5^R \in R^{C\times M}$. Next, we leverage matrix product operation to multiply d_5^R and transpose matrix, and obtain $d_5^{RP} \in R^{C\times C}$. In order to obtain the final attention map, we conduct the proposed C&S operation in advance, which can be formulated as:

$$Att_i = \varrho(d_5^{RP}) - d_5^{RP}, \tag{1}$$

where ϱ utilizes a threshold function $Max(\cdot, -1)$ to control pixels of feature maps in a stable range, thus enhancing the convergence rate to a local minimum. Then the attention map Att_i is generated by subtracting the results of the matrix product operation. ϱ suppresses inconspicuous pixels and pays much attention to significant ones. Next, the attention maps are generated through a Softmax layer:

$$P(Att_i^q|Att_i^p) = \frac{e^{(Att_i^q \cdot Att_i^p)}}{\sum_{c\in C} e^{(Att_i^q \cdot Att_i^p)}}, \tag{2}$$

where $P(Att_i^q|Att_i^p)$ represents the effect of p-th channel on q-th channel ($q, p \in C$) in feature map Att_i. Finally, the output of stage I is calculated as:

$$OA_i = \delta \cdot \textstyle\sum_{c\in C}(d_5 \otimes P(Att_i^q|Att_i^p)) + d_5, \tag{3}$$

where $\otimes$ denotes the matrix multiplication, and δ denotes a learnable parameter initialized to zero.

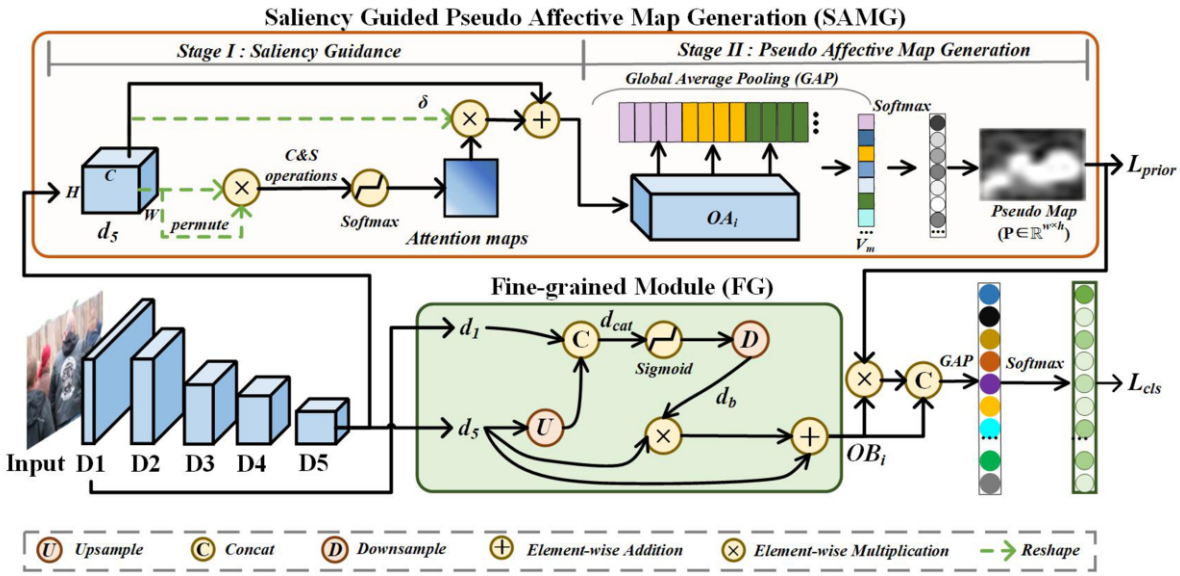

Figure 2. Architecture of the proposed visual sentiment recognition architecture. The dual coding process consists of the saliency-guided affective map generation module (SAMG) derived from high-level visual semantics and the fine-grained module (FG) originated from low-level feature details. We finally use weakly-supervised coupling to connect the dual coding process. The dual coding network are trained cooperatively to improve the final classification scores.

2.2.2 *Stage-II: Pseudo Affective Map Generation*

Inspired by class activation map technology and cross-spatial pooling strategy [18], we use saliency information generated in stage I to guide the generation of the pseudo map reflecting sentiment in stage II. Given training samples $\{(OA_i, L_i)\}_{i=1}^{n}$, where OA_i represents input feature maps, $L_i \in \{1, 2, ..., M\}$ denotes the corresponding affective label among *M* categories. In particular, we first use a downsampling layer with 1 × 1 convolution kernels to obtain rough sentiment classification of different image regions. Then we adopt the global average pooling (GAP) to obtain a sentiment category for feature maps in the form of a vector $v_m, m \in \{1, 2, ..., M\}$, which assigns M units with specific weight for corresponding sentiment category as follows.

$$v_m = \sum_{i=1}^{m} GAP(f_{q,i}), \tag{4}$$

where *GAP* represents GAP operation, and $f_{q,i}$ represents the i-th pseudo map for the q-th channel.

Different affective regions reflect different emotions. The pseudo map $P \in R^{W \times H}$ is calculated through weighted linear summation as Equation (5):

$$P = \sum_{k=1}^{K} v_m f_k(x, y), \tag{5}$$

where $f_k(x, y)$ denotes the activation of the k-th feature map produced at pixel (x, y). The activation stems from each v_m value, which is the weight of each sentiment. A higher pixel value represents the greater contribution of the related affective region to the network prediction. In the meanwhile, it implies a stronger response in human recognition.

2.3 *Fine-grained Module*

Neurophysiological research has revealed that a significant amount of visual processing is dedicated to the analysis of low-level features, such as texture information [7]. Given that visual processing is sensitive to low-level information, preserving the low-level information of objects is critical to the excavation of sentiment. In this section, we describe the proposed fine-grained module to reserve structures leveraging low-level information for the affective regions. As illustrated in the green box of Figure 2, the FG module is introduced to generate prompts and cues for clearer affective regions. We integrate low-level features containing rich texture information with high-level semantic information containing rich location information to remedy the mistake information introduced by low-level features with background information. Specifically, we use two 1 × 1 convolution layers to change the channel of d_1 to 56, and change the channel of d_5 to 224. Then we concatenate the output of d_1 and upsampling result of d_5 as Equation (6).

$$d_{cat} = Conv_{1\times1}(d_1) \uplus \hat{f}(Conv_{1\times1}(d_5)), \tag{6}$$

where $Conv_{1\times1}$ denotes 1 × 1 convolution operation, $\uplus$ represents concatenate operation, and $\hat{f}(\cdot)$ means the upsampling operation.

Next, we leverage two 3 × 3 convolution layers, one 1 × 1 convolution layer, and the Sigmoid function to obtain the features d_b:

$$d_b = \frac{1}{1+e^{-Conv_{3\times3}^{1\times1}(d_{cat})}}, \tag{7}$$

where $Conv_{3\times3}^{1\times1}(d_{cat})$ denotes the convolutional operation with 1×1 and 3×3 convolutional kernels.

In order to reinforce the representation of high-level semantic information, we fully make use of fine-grained cues as well as excavate the interaction of low and high features by multiplying downsampling d_b and d_5 containing high-resolution features. Finally, we use a skip-connection convolution layer to fuse the results and obtain the final features OB_i.

$$OB_i = d_5 \otimes \check{f}(d_b) \oplus d_5, \tag{8}$$

where $\otimes$ denotes the matrix multiplication, $\check{f}(\cdot)$ represents the downsampling layer, and $\oplus$ denotes the element-wise addition layer.

2.4 Weakly-supervised Coupling Classification

Analyzed from the perspective of image representation, the SAMG module and the FG module highlight the local cues and underlying features of the image, respectively, which provide effective information for further classification of sentiment. The original convolutional features are represented as a whole, and the resultant maps generated by the two modules are used as local feature representations of the convolutional features. We combine affective regions and fine-grained cues to form intact features containing rich sentiment information, based on which the output maps of the two modules are multiplied at the pixel level.

Concretely, we adopt Hadamard product to multiply the pseudo-labeled map of affective regions AM_i and the output maps of FG module BM_i, which is shown below. Then the coupled features are input into the Softmax function to obtain the final prediction.

$$\bar{F} = AM_i \otimes BM_i \uplus BM_i, \tag{9}$$

where $\otimes$ denotes the element-wise multiplication, and $\uplus$ denotes the concatenate operation.

Traditional theories of object recognition have emphasized the role of shape information in high-level vision. Hence, we want to stress the edge details in low-level features by introducing related boundary-aware loss. Two loss functions are defined to jointly supervise the network in an end-to-end manner, one is a prior classification loss function L_{prior}, and the other is a loss function emphasizing edge information. L_{prior} helps the end-to-end network find the labels closest to the affective region in the classification task, and boundary loss transfers the supervised information from the low-level information in the FG to the current coupling branch. More specifically, the prediction loss calculates the distance between the prediction pseudo map and ground truth, which is constructed using a M-class Softmax function with the input vector $v_m \in R^M$. Denoting the equivalence discrimination between prediction category p and ground truth label y_i by condition function $f(e)$. If $y_i = p$, $f(e) = 1$, conversely, if $y_i \neq p$, $f(e) = 0$.

$$L_{prior} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} f(e) log v_m, \tag{10}$$

Boundary-aware aims at penalizing distorted and blurry errors around edges. Boundary-aware loss contains two loss functions:

(1) Gradient based Fine-grained Loss: we first use Sobel filters to obtain the gradient of the pseudo affective map. Then we place the gradient norm as an L_1 loss function between the actual and target gradient norms as:

$$L_G = \frac{1}{n}\sum_{i=1}^{n}(ln(\nabla_x \|Q\|_1 + \alpha_1) + ln(\nabla_y \|Q\|_1 + \alpha_2)), \tag{11}$$

where $\|\cdot\|_1$ represents $L1$-norm, Q represents $p_i - g_i$. p_i denotes estimate gradient and g_i denotes ground truth gradient. ∇_x and ∇_y are the spatial derivative of $p_i - g_i$ calculated at i-th pixel with respect to x direction and y direction, respectively. $\alpha_1, \alpha_2 > 0$ is controllable parameters.

(2) Surface Normal Loss: we use the surface normal loss to further measure the details and structures of affective pseudo maps, regarding the normal to the surface of affective estimate and ground truth. The surface normal of the predicted pseudo affective map can be denoted as $u_i^p \equiv [-\nabla_x(p_i), -\nabla_y(p_i), 1]^\top$, and the surface normal of ground truth can be represented as $u_i^g \equiv [-\nabla_x(g_i), -\nabla_y(g_i), 1]^\top$.

$$L_N = \frac{1}{n}\sum_{i=1}^{n}(1 - \frac{\langle u_i^p, u_i^g\rangle}{\sqrt{\langle u_i^p, u_i^p\rangle}\sqrt{\langle u_i^g, u_i^g\rangle}}), \tag{12}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product.

The overall loss function L_{cls} is defined as follows:

$$L_{cls} = \lambda_1 L_G + \lambda_2 L_N + \lambda_3 L_{prior} \tag{13}$$

where λ_1, λ_2, λ_3 are the trade-off parameters, and we use stochastic gradient descent to optimize the joint loss function L_{cls}.

3 Experiments

3.1 Datasets

In this section, we evaluate the proposed network on five visual affective classification datasets, including one big-scale public affective dataset Flickr and Instagram (FI-8) [27], and four widely-used affective datasets: EmotionROI (6 classes) [16], Emotion-6 (6 classes) [13], IAPS-Subset [12], and Twitter II [1].

Flickr and Instagram: The FI-8 dataset consists of images from Flickr and Instagram, containing 60,745 images and 42,856 images, respectively. Each sample in the FI-8 dataset consists of an image and an annotated emotion label, which contains 8 categories (i.e. anger, amusement, awe, contentment, disgust, excitement, fear, and sadness). FI-8 dataset is the largest dataset currently available in the visual sentiment recognition domain.

EmotionROI: The EmotionROI dataset was selected from Flickr and each image is labeled with 15 affective regions that evoke emotions, each normalized to between 0 and 1 and forming an emotional stimulus map. EmotionROI consists of 1980 images and 6 basic human emotion categories, which are anger, disgust, fear, joy, sadness, and surprise.

Emotion-6: This dataset is created for emotion prediction and contains 8350 images, which are derived from 150K images taken from Google and Flickr. The emotion labels of this dataset are divided according to six basic human emotions, including anger, disgust, fear, joy, sadness, and surprise.

IAPS-Subset and Twitter II: The IAPS-Subset and Twitter II datasets were collected from Twitter and social media platforms, containing 395 and 603 images, respectively. Each image is labeled by an emotional polarity (i.e., negative or positive).

3.2 Experiment Setting Details

The DCNet is built in the widely adopted PyTorch framework [14], and is constructed in a ResNet-101 network using initialized model weights pre-trained by a large-scale dataset via the ImageNet dataset [17]. The input size of each training image is 448 × 448. We first apply the random resized crop to crop the input image randomly, then adopt random horizontal flips to flip the image horizontally as an implicit data enhancement, which aims to reduce the overfitting problem in data deficient scenarios and improve the model generalization. All datasets are randomly divided into 80% training set and 20% test set. We leverage Stochastic Gradient Descent (SGD) for the specific training process for model updating. To improve the computational efficiency, we set the momentum decay and weight decay to 0.9 and 5×10^{-4}, respectively. we set the initial learning rate size to be 1×10^{-4}, and reduce it by a factor of 100 every 10 iterations. The batch size of training is set to 14. In addition, we set the values of λ_1, λ_2, and λ_3 on ResNet-101 to 0.5, 1, and 1, respectively. All our experiments were conducted on Nvidia Tesla P100-PCIE with 16 GB memory in total.

3.3 Results on FI-8

In order to verify the effectiveness of the DCNet for visual sentiment recognition, we compare DCNet with state-of-the-art methods on widely used image sentiment datasets. Table 1 shows the comparison experiment results on the big-scale dataset FI-8. The methods in Table 1 are centered around the visual sentiment recognition task, and the comparison results show that the proposed model achieves an accuracy of 71.65% on the FI-8 dataset, a result that outperforms other methods. We further analyzed the performance differences between different methods. The relative performance of our proposed method is improved by 21.64% compared to the classification result using the ResNet-101 network alone. Besides, unlike Zhou's CAM-ResNet101 [32] architecture that is mainly implemented through the GAP technique, DCNet is guided by regions of interest to generate final affective regions evoked emotions. The comparison result shows that the DCNet is 3.11% higher than CAM-ResNet101. Additionally, different from She et al. [18], we introduce saliency prior to a dual coding process, rather than straightly using a weak supervision strategy. The result demonstrates that the performance of DCNet is 1.49% higher than the classification result of She's. Results in Table 1 illustrate that the DCNet effectively improves the accuracy of visual sentiment recognition on FI-8 dataset.

Table 1. Classification accuracy comparison on FI-8 dataset.

Methods	FI-8
Self-Attention [21]	24.01
Zhao et al. [31]	46.13
Sentibank [1]	49.23
DeepSentibank [2]	51.54
ImageNet-AlexNet [8]	38.26
ImageNet-VGG16 [19]	41.22
ImageNet-ResNet101 [5]	50.01
Fine-tuned AlexNet	52.16
Fine-tuned VGG16	54.75
Fine-tuned Inception-v3	56.90
Fine-tuned ResNet101	57.16
Yang et al. [22]	66.79
SPN [34]	66.57
WILDCAT [3]	67.03
MAP [6]	68.13
Zhou et al. [32]	68.54
She et al. [18]	70.07
DCNet	**71.65**

Additionally, we use the confusion matrix to visualize the classification results on the FI-8 dataset, which is shown in Figure 3. The confusion matrix counts the number of observations that the classification model classifies in the wrong category and the right category, and then presents the results in a table. The DCNet performs well in expressing Amusement emotion. We infer the reason may be that the amusement category has the second-largest training sample size compared to other datasets. However, the other categories are most likely to be confused with Contentment and Disgust, the reason is probably that compared to the other categories the contentment category has more training samples, and the disgust category has a stronger visual similarity, which includes the more noisy and similar data and causes the classifier's prediction conducive to the contentment and disgust categories.

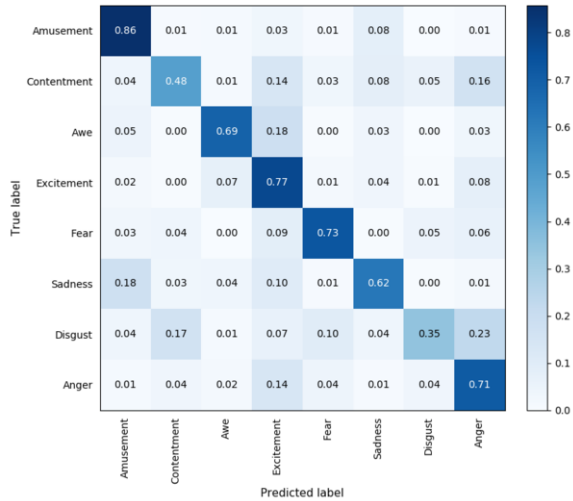

Figure 3. Confusion matrix on FI-8 dataset.

3.4 Results on Small-scale Datasets

We further conduct a series of experiments on four relatively small image sentiment datasets, and the results of the comparison experiments are shown in Table 2. Both Emotion-6 and EmotionROI datasets contain six basic sentiment classifications, and both IAPS-Subset and Twitter II are binary classification task datasets with the performance shown in the first, second, third, and fourth columns of Table 5, respectively. The classification accuracy of the Emotion-6 dataset shows the DCNet achieves 58.92% accuracy for visual sentiment recognition, which is 3.32% higher than Zhang's integration method [30]. The DCNet also outperforms on EmotionROI dataset, which achieves 59.60% accuracy. It is 1.35% higher than She's weakly supervised method [18]. On the IAPS-Subset dataset, Zhang's architecture [28] was built on the class activation mapping network and used multiscale feature information. Compared with Zhang's method, DCNet achieves a considerable performance of 95.90% by taking into account the dual coding process for affective regions. On the Twitter II dataset, the DCNet has a better classification performance than She's framework of 1.15% improvement.

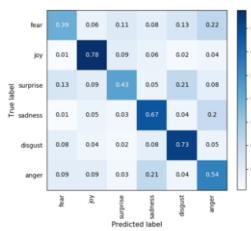

Figure 4. Confusion matrix on Emotion-6.

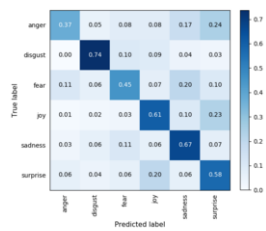

Figure 5. Confusion matrix on EmotionROI.

Here we select to visualize the confusion matrix of ResNet-101 based DCNet on two representative small-scale datasets, which are

Table 2. Classification accuracy comparison on small-scale datasets

Methods	Emotion-6	Methods	EmotionROI	Methods	IAPS-Subset	Methods	Twitter II
Peng et al. [15]	45.2	Zhao et al. [31]	34.84	SentiBank	81.79	DeepSentiBank	70.23
BCPNN [24]	45.4	DeepSentibank [2]	42.53	DeepSentiBank	85.63	VGGNet	71.79
Gao et al. [4]	46.1	Yang et al. [22]	52.40	PCNN [26]	88.84	DenseSIFT+VLAD	77.17
ACPNN [24]	46.9	SPN [34]	52.70	VGGNet	88.51	WILDCAT [3]	78.81
Zhou et al. [32]	48.7	WILDCAT [3]	55.05	Fine-tuned VGGNet	89.37	Zhou et al. [32]	79.13
Yang et al. [22]	52.4	Zhou et al. [32]	55.72	Yang et al. [23]	92.39	Sun et al. [20]	80.91
Zhang et al. [30]	55.6	She et al. [18]	58.25	Zhang et al. [28]	95.83	She et al. [18]	81.35
DCNet	**58.92**	DCNet	**59.60**	DCNet	**95.90**	DCNet	**82.50**

the Emotion-6 dataset (depicted in Figure 4) and the confusion matrix on the EmotionROI dataset (shown in Figure 5). Figure 4 shows that the DCNet has a better performance in expressing the Joy and Disugust emotions, yet the other categories are most likely to be confused with Fear and Surprise. Figure 5 implies the DCNet performs better in Disgust and Sadness categories, yet underperforms in Anger and Fear emotions. The improvement on small-scale datasets also proves that the DCNet can accurately identify affective regions and clarify sentiment categories.

3.5 Ablation Studies

We conducted a series of ablation experiments to show the effectiveness of the DCNet, and boundary loss functions on the performance of visual sentiment recognition. We conduct the ablation experiments on the large dataset FI-8 and the classic small-scale Emotion-6 dataset.

Table 3. Impact of the backbone network on visual sentiment recognition accuracy

Backbone network	FI-8	Emotion-6
AlexNet	59.22	47.23
VGG-16	65.08	53.29
ResNet-50	70.66	56.41
ResNet-101	71.65	58.92

The performance comparison using different networks as DCNet's backbone networks is provided in Table 3. For the visual sentiment recognition results using ResNet-101 network as the backbone network, we obtain the highest accuracy on both datasets, which is 1.93%, and 3.51% higher than the second highest ResNet-50 on two datasets, respectively. While the results in the case of VGG-16 and AlexNet network as the backbone network were not desirable. We speculate that the high network level and complexity with sufficient training samples ensure the feature extraction capability of the DCNet.

Table 4. Impact of the modules on visual sentiment recognition accuracy

Module	FI-8	Emotion-6
M *w/o* S_1	68.97	55.98
M *w/o* S_2	69.37	54.21
M *w/o FG*	70.01	55.05
M	71.65	58.92

To demonstrate the effectiveness of the dual coding process, we further conduct ablation experiments shown in Table 4. It can be seen that the experimental results are 2.68% and 2.94% lower with the removal of stage I (M *w/o* S_1) than with stage I. It is 2.28% and 4.71% lower with the absence of affective region prediction results (M *w/o* S_2) on two datasets. The experimental results are 1.64% and 3.87% lower than with the absence of fine-grained cues (M *w/o* FG). Thus, it can be seen that the affective region prediction information and fine-grained information can enhance the effects on the visual sentiment recognition task. Also, the saliency guidance provided by stage I in SAMG can facilitate obtaining affective regions, and simulate human eye fixation. Saliency information provides an important reference for the subsequent affective region pseudo maps.

Table 5. Impact of the loss functions on visual sentiment recognition accuracy

Backbone	L_{Prior}	L_G	L_N	FI-8	Emotion-6
√	√			70.02	55.40
√		√		70.23	55.81
√			√	70.45	56.02
√		√	√	71.27	57.15
√	√	√	√	71.65	58.92

This paper also provides the performance of the visual sentiment recognition task using boundary-aware loss, and the results are shown in Table 5. Combining boundary-aware loss on FI-8 and Emotion-6 datasets is higher than without and with only partial loss functions. The performance of the gradient based fine-grained loss L_G and the surface normal loss L_N is comparable. The introduction of boundary aware loss function ($L_G + L_N$) improves the recognition accuracy to 71.27% and 57.15% on two datasets, respectively. The L_G and L_N are complementary to the classification loss. Hence, integrating the three loss functions ($L_G + L_N + L_{Prior}$) can further improve the classification accuracy.

Table 6. Impact of the feature integration in low levels on visual sentiment recognition accuracy

D_1	D_2	D_3	FI-8	Emotion-6
√			71.65	58.92
	√		71.28	58.89
		√	71.02	57.96
√	√		70.83	58.09
√		√	71.20	57.93
	√	√	70.31	56.72
√	√	√	70.54	57.02

The DCNet makes full use of both low and high-level features, and we conduct further ablation experiments to confirm the range of applicable features at low levels, and the results are shown in Table 6. The experiments were conducted on FI-8 and Emotion-6 with

ResNet-101 as the backbone network. The results using single-layer low-level features (row 1 to row 3) are overall better than those using multi-layer low-level features (row 4 to row 7). From the experimental results, better performance is obtained by using only D_1 as low-level features for the fine-grained process, and the performance of the FG module is mainly dependent on D_1 features.

3.6 Visualization Results

To show the prediction performance of the pseudo affective maps in DCNet, we provide qualitative results on the EmotionROI dataset. Since we have applied random horizontal flipping and cropped a random patch from input images as a form of data augmentation, the obtained pseudo affective maps are part of the original images. As shown in Figure 6, we use the yellow dashed box to outline the area corresponding to our pseudo affective maps for better comparison.

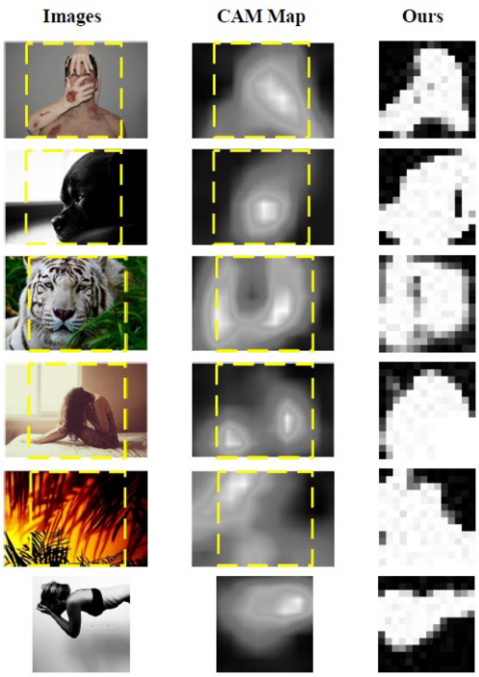

Figure 6. Prediction results using different methods on EmotionROI.

Figure 6 shows samples of pseudo affective maps predicted by our method (column 3) and precise affective regions generated by weakly supervised CAM [32] (column 2). As can be seen in rows 1 to 3, in contrast to the CAM-generated maps, DCNet outlines the affective regions, and further provides object details, such as the shape of fingers and animal faces. Besides, under low-contrast circumstances (demonstrated in row 4), DCNet-generated affective maps outperform CAM-generated maps. To be concrete, the CAM-based method produces two focal points, resulting in a blurry affective map that is differing significantly from the ground truth, yet DCNet efficiently focuses on the sentiment-related regions and generates appropriate affective maps. In addition, when the background is complex, our affective maps still perform well (shown in row 5). Our affective map accurately localizes the emotional area, highlights the brighter parts of the picture, yet suppresses the black background. Moreover, as depicted in row 6, our affective maps provide effective predictions for images in different sizes. While the CAM-based approach merely generates a general range of regions related to sentiment, DCNet outlines the affective regions in greater detail.

4 Conclusion

Based on the fact that visual recognition is highly correlated with visual features and people's prior experiences (e.g. object location and shape), in this paper, we propose a novel dual coding network for visual sentiment recognition that couples pseudo affective maps derived from high-level visual semantics and fine-grained cues originating from low-level feature details. Extensive experimental results show that our proposed network can effectively improve sentiment perception performance. And abundant ablation studies verify the effectiveness of our dual coding process. In future work, we will try to deploy the proposed network in real-world applications, such as sentiment evaluation in education, and visual question answering, which stress the regional information. Also, we hope that our research will be of great value to both vision and cognitive neuroscience researchers.

5 Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 61871186 and 61771322.

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 *Proceedings of the 21st ACM international conference on Multimedia*, pp. 223–232, (2013).

[2] 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).

[3] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord, 'Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 642–651, (2017).

[4] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng, 'Deep label distribution learning with label ambiguity', *IEEE Transactions on Image Processing*, **26**(6), 2825–2838, (2017).

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).

[6] Xiaohao He, Huijun Zhang, Ningyun Li, Ling Feng, and Feng Zheng, 'A multi-attentive pyramidal model for visual sentiment analysis', in *2019 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, (2019).

[7] Akshay Vivek Jagadeesh, *Texture and Object Representation for Human Visual Perception*, Stanford University, 2022.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks', *Communications of the ACM*, **60**(6), 84–90, (2017).

[9] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua, 'Multi-source domain adaptation for visual sentiment classification', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2661–2668, (2020).

[10] Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark, 'Effects of language on visual perception', *Trends in cognitive sciences*, **24**(11), 930–944, (2020).

[11] Ziyu Ma, Fuyan Ma, Bin Sun, and Shutao Li, 'Hybrid multimodal fusion for dimensional emotion recognition', in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 29–36, (2021).

[12] Jana Machajdik and Allan Hanbury, 'Affective image classification using features inspired by psychology and art theory', in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 83–92, (2010).

[13] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury, 'Contemplating visual emotions: Understanding and overcoming dataset bias', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 579–595, (2018).

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., 'Pytorch: An imperative style, high-performance deep learning library', *Advances in neural information processing systems*, **32**, (2019).

[15] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher, 'A mixed bag of emotions: Model, predict, and transfer emotion distributions', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 860–868, (2015).

[16] Kuan-Chuan Peng, Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen, 'Where do emotions come from? predicting the emotion stimuli map', in *2016 IEEE international conference on image processing (ICIP)*, pp. 614–618. IEEE, (2016).

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., 'Imagenet large scale visual recognition challenge', *International journal of computer vision*, **115**, 211–252, (2015).

[18] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L Rosin, and Liang Wang, 'Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection', *IEEE Transactions on Multimedia*, **22**(5), 1358–1371, (2019).

[19] Karen Simonyan and Andrew Zisserman, 'Very deep convolutional networks for large-scale image recognition', *arXiv preprint arXiv:1409.1556*, (2014).

[20] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen, 'Discovering affective regions in deep convolutional neural networks for visual sentiment prediction', in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, (2016).

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', *Advances in neural information processing systems*, **30**, (2017).

[22] Jufeng Yang, Dongyu She, and Ming Sun, 'Joint image emotion classification and distribution learning via deep convolutional neural network.', in *IJCAI*, pp. 3266–3272, (2017).

[23] 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).

[24] Jufeng Yang, Ming Sun, and Xiaoxiao Sun, 'Learning visual sentiment distributions via augmented conditional probability neural network', in *Thirty-first AAAI conference on artificial intelligence*, (2017).

[25] Chunyong Yin, Sun Zhang, and Qingkui Zeng, 'Hybrid representation and decision fusion towards visual-textual sentiment', *ACM Transactions on Intelligent Systems and Technology*, **14**(3), 1–17, (2023).

[26] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, 'Robust image sentiment analysis using progressively trained and domain transferred deep networks', in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, p. 381–388. AAAI Press, (2015).

[27] 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 *Proceedings of the AAAI conference on artificial intelligence*, volume 30, (2016).

[28] Haimin Zhang and Min Xu, 'Weakly supervised emotion intensity prediction for recognition of emotions in images', *IEEE Transactions on Multimedia*, **23**, 2033–2044, (2021).

[29] Jing Zhang, Mei Chen, Han Sun, Dongdong Li, and Zhe Wang, 'Object semantics sentiment correlation analysis enhanced image sentiment classification', *Knowledge-Based Systems*, **191**, 105245, (2020).

[30] Jing Zhang, Xinyu Liu, Mei Chen, Qi Ye, and Zhe Wang, 'Image sentiment classification via multi-level sentiment region correlation analysis', *Neurocomputing*, **469**, 221–233, (2022).

[31] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun, 'Exploring principles-of-art features for image emotion recognition', in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 47–56, (2014).

[32] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, 'Learning deep features for discriminative localization', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, (2016).

[33] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian, 'Multimodal sentiment analysis with image-text interaction network', *IEEE Transactions on Multimedia*, (2022).

[34] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao, 'Soft proposal networks for weakly supervised object localization', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1841–1850, (2017).