Multi-Dimension Full Scene Integrated Visual Emotion Analysis Network

Weiye Peng
College of Computer Science and Software Engineering, Shenzhen University,
National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University,
Shenzhen 518060, PR China
pengweiye2022@email.szu.edu.cn

Sheng-hua Zhong*
College of Computer Science and Software Engineering,
Shenzhen University, Shenzhen 518060, PR China
csshzhong@szu.edu.cn

Abstract

Visual Emotion Analysis (VEA) aims to research what emotions are evoked in the viewer by watching visual content. Existing methods often operate on pixel-level, neglecting the complexity and abstract process of emotion. To tackle this challenge, we propose a novel multi-dimension full scene integrated network, focusing on scene feature, background style information and facial region. For making the proposed network more relevant to effective information, we leverage the channel attention mechanism and we design a multi-dimension loss function to distinguish the basic emotion categories mixed with each other. Experiments show our proposed method outperforms the state-of-the-art approaches on four public visual emotion datasets, especially where it outperforms existing state-of-the-art methods by a large margin, +4.22% on the Emotion6 dataset at six classes classification accuracy. Moreover, ablation study and visualization prove the effectiveness of our method.

***Keywords:** Visual Emotion Analysis,Emotion Prediction, Attention Mechanism, Multi-Dimension Loss.*

1. Introduction

Visual Emotion Analysis (VEA) aims to investigate the emotions elicited by visual content in emotional images [24]. With the rapid growth of social media, the number of images posted by users on it has skyrocketed. VEA provides an alternate way to understand users' behaviors and emotions for smart recommendation and opinion mining. Despite the substantial progress in computer vision tasks, ranging from conventional natural image classification (*e.g.* ImageNet-1k and CIFAR-10) to more fine-grained tasks like CUB-200, VEA presents a more great challenge. The biggest challenge of VEA is the affective gap which means a disparity between pixel-level information in the image and the abstract emotional information perceived by users. Existing methods addressing the affective gap can be categorized into traditional and deep learning approaches. Traditional methods rely on image statistical information and hand-crafted descriptors such as covariance matrices and color histograms. Deep learning methods leverage the potent representation capabilities of Convolutional Neural Networks (CNN) to extract global and local image features.

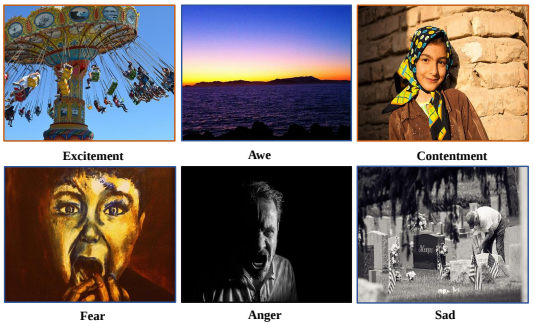

Figure 1. The images with distinct categories from emotion dataset FI. It is evident that emotional images encompass a plethora of content, such as human faces, background styles, objects and so on.

However, these methods often operate on pixel-level, neglecting the complexity and abstract process of emotion. In this work, we try to fit the affective gap in a more holistic and definite aspect. Specifically, we focus on three types of features: scene feature, background style information and facial region. The most influential factor in determining the emotional impact of an image is its scene feature, providing overarching information about the image.

Drawing from Face Superiority Effect [18], our brain

*Sheng-hua Zhong is the corresponding author of this paper.

exhibits a robust ability to recognize face. When faces present in emotional images, they significantly influence the conveyed emotion, as mentioned in [17]. Therefore, facial information within images emerge as a crucial feature for emotions prediction. While not all images contain explicit semantic or facial information, such as the majority of scenic images, they still convey intense emotional messages. For instance, images of blue sky, white clouds, and dark clouds share semantic similarity but express divergent emotional tones. Termed background style information, this information encompasses elements like colors, shapes, and textures, being an indispensable component for VEA. Existing methods handle basic emotion categories discretely and independently, yet this way overlooks the inherent correlations within emotions, rendering it unable to distinctly differentiate between basic emotion categories. For instance, in the Valence-Arousal-Dominance(VAD) emotion model, Anger is closely related to Disgust, while being far apart from Surprise. Consequently, the penalty in the loss function for misclassifying Anger as Disgust or Surprise should differ theoretically. However, this gap is disregarded in discrete dimension. Motivated by this, we integrate discrete and continuous dimensions, proposing a novel multi-dimension cross entropy loss function.

Our contributions can be summarized as follows:

- We propose a novel framework called multi-dimension full scene integrated visual emotion analysis network based on theory and experience. We focus on scene feature, background style information and facial region in images for VEA and design three specific branches to extract features of emotional images. Our method outperforms the state-of-the-art on four public visual emotion datasets.
- We utilize the channel attention mechanism to make the proposed network automatically learn the more emotionally inclusive information in scene feature and background style information, through cross-channel interaction and channel weight assignment.
- We design a novel multi-dimension cross entropy loss function combining discrete and continuous dimensions, improving the accuracy of the classification outcomes.

2. Related Works

Traditional Methods: Traditional methods base on image statistical information and hand-crafted descriptors. Machajdik *et al*. [12] extracted specific image features and combined them to predict emotions, which consisted of color, texture, composition, and content. Zhao *et al*. [26] used the principles of art including balance, emphasis, harmony and so on, to complete both classification and regression tasks. Borth *et al*. [1] constructed the visual sentiment ontology and proposed visual concept detectors, *i.e*. Sentibank, for predicting emotions. While effective on numerous small datasets, these approaches are limited in capturing all significant aspects of emotions.

Deep Learning Methods: More recently, CNN-based methods with powerful representation ability have also been employed in VEA. She *et al*. [16] designed a weakly supervised coupled network (WCSNet) for joint sentiment detection and classification with sentiment map. It discovered emotion regions to predict emotions in an end-to-end manner, and performed better than the basic CNN methods. Zhang *et al*. [25] proposed a novel CNN model to explore discriminative representations consisting of content and style representation for image emotions recognition. Zhao *et al*. [27] developed polarity-consistent deep attention network which integrated attention into a CNN backbone with an emotion polarity constraint for fine-grained visual emotion regression. Yang *et al*. [22] proposed a stimuli-aware VEA method consisting of stimuli selection, feature extraction and emotions prediction with three specific networks.Recently, Xu *et al*. [20] proposed a novel multi-level dependent attention network called MDAN. It consisted of multi-head attention and level-dependent class activation map, classifying emotions both globally and locally in a simultaneous manner. It investigated fine-grained emotions prediction with competitive performance. Feng *et al*. [4] proposed a sentiment-oriented pretraining method that was based on the human visual sentiment perception mechanism, as a means of addressing the lack of high-level concepts related to sentiment. Part of the framework they proposed employ self-supervised learning, but the entirety of the approach remains a supervised learning paradigm.

Learning from these methods, we are more concerned with holistic and definite information. In this paper, we propose a three-branch network, which integrates scene feature, background style information and facial region,with channel attention mechanism and multi-dimension cross entropy loss function to fit affective gap.

3. Methodology

In this section, we introduce a novel multi-dimension full scene integrated network for emotions prediction from mining the full scene of the image. As shown in Fig. 2, our proposed framework mainly consists of three branches, namely Fore-Net, BS-Net, and Face-Net. We employ Fore-Net to extract scene feature, utilize BS-Net to extract background style information and capture facial expression with Face-Net from the image.

3.1. Feature Extraction

We leverage three specialized branches, Fore-Net, BS-Net, and Face-Net to extract corresponding features, respectively.

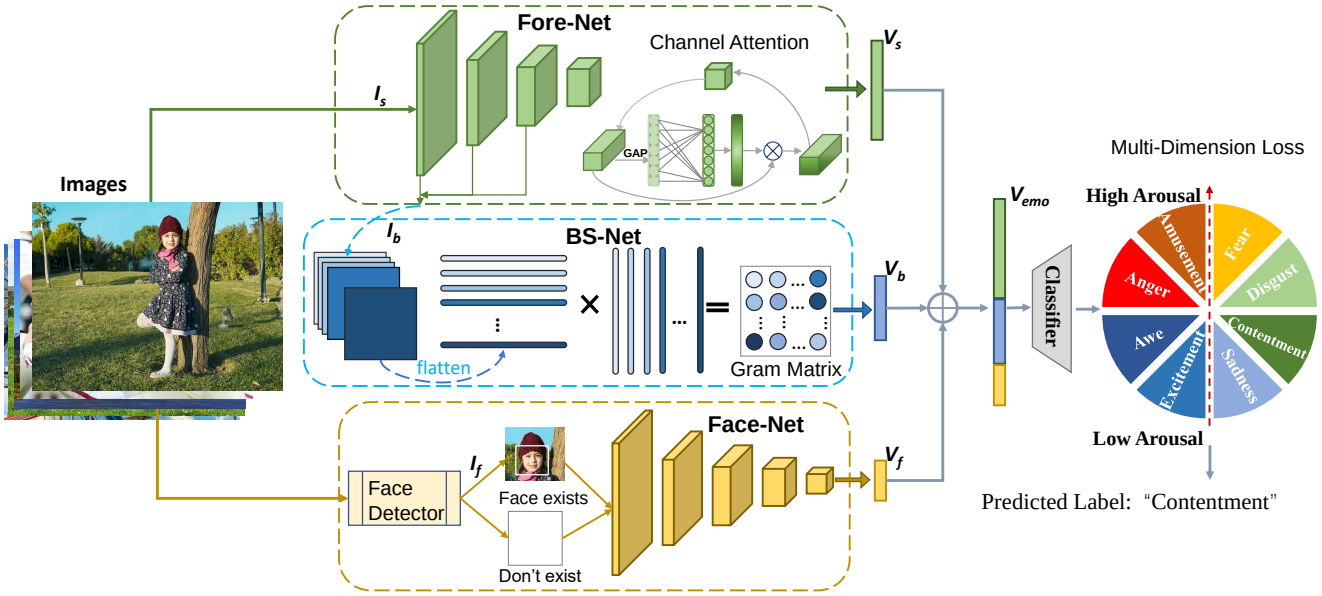

Figure 2. The proposed network architecture. We construct a three-branch network which consists of Fore-Net, BS-Net, and Face-Net. Given the same input image, three branches network work respectively for extracting feature. These features will be integrated into the hybrid vector of various emotion information to predict emotions. Besides, we leverage the channel attention mechanism and design a multi-dimension loss function to improve the performance of proposed network.

(1) Fore-Net: In this work, we use ResNet-101 as the backbone of Fore-Net, initialize it with pre-trained parameters from ImageNet and fine-tune it on emotion dataset. Specifically, it consists of five convolutional blocks and a Global Average Pooling (GAP) layer, which is described in Eq. (1). In Fore-Net we can get the feature V_s by:

$$F_s = FCN_{res101}(I_s), \tag{1}$$
$$V_s = G_{\text{avg}}(F_s), \tag{2}$$

where FCN_{res101} is the fully convolutional network in ResNet-101,I_s denotes the input image and G_{avg} is the following GAP layer. $F_s \in \mathbb{R}^{d_1 \times w \times h}$ represents the feature maps extracting from the last convolutional layer and $V_s \in \mathbb{R}^{d_1}$ represents the extracted scene vector. Specifically, w, h are the spatial size of the feature map while $d_1 = 2048$ is the dimension of the scene vector.

Particularly, another important component of Fore-Net is the channel attention mechanism following [19]. It utilizes a local cross-channel interaction strategy to automatically cover the range of channel interactions through an adaptive method. In this work, we embed it into each convolutional block of backbone, making the network better focus on important features and suppress unimportant ones. It is worth noting that BG-Net also uses this channel attention. The calculation method is simplified as follows:

$$F_a = G_{avg}(F_{in}), \tag{3}$$
$$W = \sigma(C1D_k(F_a)), \tag{4}$$

where F_{in} is the feature extracted by CNN backbone, with dimension of $c * w * h$, representing channel, width, and height and G_{avg} denotes GAP layer. In Eq. (4), $C1D_k$ is 1D convolution with kernel size of k, σ means sigmoid activation function, and W represents the weight matrix of channel with dimension of $c * 1 * 1$. The kernel size k is adaptively determined by:

$$k = \psi(c) = \left| \frac{log_2(c)}{\gamma} + \frac{b}{\gamma} \right|_{odd}, \tag{5}$$

where $\psi(c)$ is a mapping of channel c which can be calculated as the rightmost, with the hyper-parameter of γ and b. The value $|t|_{odd}$ indicates the nearest odd number of t. Finally, we utilize the weights matrix to weight the feature F_{in} and get the result of F_{out}:

$$F_{out} = W * F_{in}, \tag{6}$$

where F_{out} stay the same with F_{in} in dimension.

(2) BS-Net: BS-Net stands for Background Style Net and is used to extract background style information like colors, textures, and shapes. Following [5], the Gram Matrix operation on the output of the convolutional layer gives

some information about the overall style with the image. Specifically, we use the same backbone as Fore-Net, and obtain the feature extracted by the 1st, 2nd, and 3rd convolutional block, namely I_{b1}, I_{b2}, and I_{b3}. Then we perform Gram Matrix operation on the shallow features shown in Eq. (8). The operation method is as followed:

$$X_i = I_{bi}, \tag{7}$$

$$G_i = X_i {X_i}^{\mathrm{T}}, \tag{8}$$

where $i \in \{1, 2, 3\}$, denotes different depth of fully convolutional blocks and X_i represents the feature maps extracted by different layer from backbone. We combine the obtained gram feature G_i as background style feature V_b with the dimension of $d_2 = 832$, which is the sum of the first three output dimensions of ResNet-101:

$$V_b = concate\left[G_1, G_2, G_3\right], \tag{9}$$

(3) Face-Net: The size of faces appearing in emotional images is relatively small, which does not necessitate the use of a very large network. Many previous works[22] have also employed a shallow deep networks. So we adopt ResNet-18 as the backbone to construct Face-Net, with pretrained parameters from FER2013 dataset [6] and fine-tune it on emotion dataset. We choose face detection library Dlib [9] as face extractor. Specifically, Dlib first detects the face in each image and then crops the facial region with size of 48x48 for further processing as facial information if face exists. In case where Dlib fails to detect a face, it outputs nothing. When multiple faces are present in an image, we only select the face with the largest area as the feature I_f. When Dlib does not detect a face, we use a zero vector to replace the facial information shown in Eq. (10). Then we put the facial feature to the backbone. Specifically, Face-Net consists of five convolutional blocks and a global average pooling layer. The feature V_f extracted by Face-Net is from:

$$V_f = \begin{cases} G_{avg}(FCN_{res18}(I_f)), V_f \in R^{d_3} & , I_f \text{ exists} \\ \overrightarrow{\mathbf{0}} & , \text{else} \end{cases} \tag{10}$$

where FCN_{res18} is the fully convolutional networks in ResNet-18 and G_{avg} is the following GAP layer. Besides, $V_f \in \mathbb{R}^{d_3}$ represents the feature with dimension $d_3 = 512$.

3.2. Emotions Prediction

We introduce feature fusion module and multi-dimension cross entropy loss function here.

(1) Feature Fusion: We integrate the previously extracted features, which serve as the ultimate emotion feature V_{emo} for classification. These features will be fed into a fully connected layer to obtain the emotions prediction.

$$V_{emo} = concate\left[V_s, V_b, V_f\right], \tag{11}$$

where $V_{emo} \in \mathbb{R}^{d_1+d_2+d_3}$ denotes final emotion feature.

(2) Multi-Dimension Cross Entropy Loss: We combine basic cross entropy loss function based on normal discrete dimension and continuous cross entropy loss function as a novel multi-dimension loss function. The continuous loss function focuses on the Arousal dimension which means the intensity of emotion, and we employ it as a criterion to subdivide emotional categories into high-arousal and low-arousal. Specifically, Fear, Amusement, Anger, and Disgust are high arousal categories, while Awe, Contentment, Excitement, and Sadness belong to low arousal categories according [14]. We augment the basic cross entropy loss with an auxiliary continuous loss function of emotion. Our approach imposes greater penalties for misclassifications both in emotion categories and emotion arousal, aiming to enhance the performance of model in capturing subtle variations in emotion intensity. We compute the predicted probabilities as:

$$p_{emo}(i) = \frac{\exp(y_{emo}(i))}{\sum_i \exp(y_{emo}(i))}, \tag{12}$$

where $y_{emo}(i)$ is the output from last fully connected layer, $p_{emo}(i)$ is the predicted probabilities. Then we represent the arousal probabilities as follow:

$$high-arousal: p_{arl}\left(1\right) = \sum_{i=1}^{C/2} p_{emo}\left(i\right), \tag{13}$$

$$low-arousal: p_{arl}\left(0\right) = \sum_{i=C/2}^{C} p_{emo}\left(i\right), \tag{14}$$

where p_{arl} can be viewed as the prediction probability of high-arousal and low-arousal. The high-arousal emotions are correspond to the former $C/2$ positions in p_{emo} while the low-arousal emotions are the rest $C/2$ respectively. We calculate these two loss functions separately:

$$\mathcal{L}_{emo} = -\sum_i q_{emo}(i) \log(p_{emo}(i)), \tag{15}$$

$$\mathcal{L}_{arl} = -\sum_{j=0}^{1} q_{arl}\left(j\right) \log\left(p_{arl}\left(j\right)\right), \tag{16}$$

where $q_{emo}(i)$ and $q_{arl}(j)$ are real labels. The whole multi-dimension cross entropy loss can be formed as:

$$\mathcal{L}_{cls\text{-}mul} = \mathcal{L}_{emo} + \lambda \mathcal{L}_{arl}, \tag{17}$$

where λ is a hyper-parameter balancing the importance between the two losses and is further researched in ablation study. Finally, the whole network is optimized through Eq. (17).

Table 1. Comparison with previous work on six public emotion datasets with top-1 accuracy (%). The bold numbers indicate the best, while the underscored numbers indicate the second best.

Method	Emoset	FI	Emotion6	Artphoto	Abstract	IAPSa
Traditional Methods						
Sentibank [1]	-	49.23	35.24	53.96	50.68	73.58
PrinciplesofArt [26]	-	46.13	34.84	63.65	62.13	58.86
DeepSentibank [2]	-	51.54	42.53	68.54	66.48	75.88
Visual Backbone Methods						
AlexNet [10]	59.85	59.85	44.19	67.03	61.96	72.24
VGG-16 [11]	65.52	65.52	49.75	68.16	62.41	74.78
ResNet-101 [7]	76.49	66.16	51.60	69.36	63.56	75.09
CAE [3]	75.77	66.76	59.56	74.33	70.41	84.89
EVA [**?**]	76.26	66.93	<u>63.71</u>	75.63	70.91	<u>86.87</u>
VEA SOTA Methods						
WSCNet [16]	76.32	67.88	58.25	72.86	64.45	82.25
StyleNet [25]	77.11	68.85	59.60	-	-	-
PDANet [27]	76.59	68.05	59.34	74.62	67.13	80.92
Stimuli-aware [22]	**78.40**	**72.42**	61.62	-	-	-
MDAN [20]	76.41	69.39	59.85	<u>78.12</u>	<u>72.34</u>	85.96
Probing-oriented [4]	76.88	70.70	60.41	75.78	69.90	86.16
Ours	<u>78.01</u>	<u>71.03</u>	**67.93**	**82.10**	**78.18**	**89.74**

4. Experiments

In this section, we evaluate the proposed method on six public visual emotion datasets. We compare our proposed method with the state-of-the-art methods, along with ablation study on architecture of our proposed method. Besides, we show more details like confusion matrix, failure cases analysis, and visualization result.

4.1. Datasets

We evaluate our framework on six public visual emotion datasets, including the Emoset [21], FI [23], Emotion6 [15], ArtPhoto [12], Abstract [12] and IAPSa [13]. Emoset dataset is the largest well-labeled emotion dataset with 3.3 million images in total, and 118,102 of these images carefully labeled by human annotators. It is labeled with eight emotion categories as amusement, anger, awe, contentment, disgust, excitement, fear and sad. FI dataset is collected from the Flickr and Instagram with 23,164 images and it has eight categories in keeping with Emoset. Emotion6 dataset contains 1,980 images with six emotion categories, anger, disgust, fear, joy, sad, and surprise. Moreover, we also evaluate the proposed method on three small scale datasets, namely, Artphoto, Abstract and IAPSa datasets. They consist of 806, 280, and 395 photos, respectively.

4.2. Implementation Details

Our framework is implemented using PyTorch. We train the proposed network using stochastic gradient descent (SGD) for 60 epochs on an NVIDIA TESLA V100-PCIE GPUs with 16 GB onboard memory. Batch size, learning rate, weight decay and momentum are set to 32, 0.0005, 0.0005 and 0.9, respectively. The learning rate is drop by a factor of 10, every 10 epochs. The two largest datasets, Emoset and FI are randomly split into training set (80%), validation set (5%) and test set (15%) following the same configuration in [21]. The others are split into 80% and 20% for training and test, respectively. In addition, a 224 × 224 size is randomly cropped from each original image, and a horizontal flip is applied to the cropped image.

4.3. Comparison with the State-of-the-art Methods

To verify the effectiveness of our framework, we first compare it with the state-of-the-art methods on the large-scale datasets Emoset, FI, and Emotion6, as shown in Table 1. The compared methods can be divided into traditional methods and deep learning ones. For traditional methods, they are Sentibank [1], Principles-of-Art *et al*. [26], and DeepSentibank [2]. Besides, we also conduct experiments on typical visual backbones like Alexnet [10], VGG-16 [11], ResNet-101 [7], CAE [3] and EVA [**?**] which are initialized with the pre-trained parameters on ImageNet and then fine-tuned on each dataset respectively. For deeep learning methods, they are WCSNet [16], PDANet [27], Stimu-Aware *et al*. [22], MDAN [20] and Probing-oriented [4]. As seen in Table 1, our framework demonstrates competitive performance and outperforms the state-of-the-art methods on Emotion6 which indicates our proposed network holds certain advantage in VEA.

We further evaluate the effectiveness of the proposed method on three small-scale datasets, including Artphoto, Abstract, and IASPa. We use 5-fold cross-validation and report the average result follow the setup in [28]. As the emotion category of anger contains only 8 and 3 samples

in the IAPSa and Abstract, respectively, it is insufficient to perform the 5-fold cross-validation. Thus, we remove the category of anger on these two datasets following [23]. It is worth mentioning that the missing data represented by "-" in Table 1 is caused by lacking both classification results and open source codes. Our method outperforms the state-of-the-art methods on Abstract, Artphoto, and IAPSa particularly. The result shows our network has robustness which performs comparable on both large-scale and small-scale datasets.

4.4. Ablation study

We perform ablation study on the network architecture and hyper-parameter λ.

(1)Network architecture analysis: We perform ablation study on the proposed network to reveal the effect of each branch and channel attention mechanism(Att) on the classification performance. We remove one of the module from the entire network at a time separately to show the contribution of each module. It is noteworthy that, since the Fore-Net and the BS-Net share a common backbone network, removing the global feature means that this branch network does not output global features but only background style information, and vice versa for removing the background style information. As shown in the Table 2, the overall performance is degraded after removing the any module, showing that the four modules are indeed useful. The performance decreases most when Fore-Net is removed, indicating that Fore-Net is most important, consistent with our expectation. In addition, we find that removing other module like BS-Net or Face-Net performs differently on different datasets. In Fig. 3, we visualize and reconstruct the parts that the different branches focused on. As we can see, when a face appears in the image, Face-Net will focus more on the part related to face. We restore the extracted image features of BS-Net to show what it extracted, and the entire network can focus on the most critical information of the image. Visualization result shows that the structure of our proposed method is reasonable.

Table 2. Ablation study of structure on Emoset, FI, and Emotion6 datasets with classification accuracy (%).

Component				Dataset		
Fore	BS	Face	Att	Emoset	FI	Emotion6
✓	✓		✓	<u>77.60</u>	69.25	63.64
✓		✓	✓	76.56	<u>69.59</u>	63.89
	✓	✓	✓	64.55	52.56	42.17
✓	✓	✓		76.21	68.05	<u>64.55</u>
✓	✓	✓	✓	**78.01**	**71.03**	**67.93**

Table 3. Hyper-parameter analysis of λ on each dataset with classification accuracy (%).

λ	Emoset	FI	Emotion6	Artphoto	Abstract	IAPSa
0	77.84	<u>70.87</u>	<u>67.19</u>	76.63	<u>76.36</u>	<u>89.61</u>
0.1	**78.01**	**71.03**	**67.93**	**82.10**	**78.18**	**89.74**
0.2	<u>77.88</u>	70.14	66.92	<u>80.86</u>	76.11	88.31
0.3	77.66	70.01	64.39	78.40	74.55	88.05

(2)Hyper-parameter analysis: Parameter λ controls the proportion of the continuous part in the loss function Eq. (17), which is a decisive hyper-parameter for the classification results. As shown in Table 3, when λ is not equal to zero, performance has definitely improved. The loss function $\mathcal{L}_{arl}$ does not necessitate a significant weight, and the best performance is achieved when $\lambda = 0.1$.

5. Failure Cases Analysis

We present the classification results through a confusion matrix in Fig. 4, revealing all classification results. It can be observed that a significant number of samples labeled as Amusement are misclassified as Contentment and Excitement. Conversely, many samples labeled as Contentment and Excitement are also classified as Amusement. This indicates that there are notable commonalities among the categories of Amusement, Contentment, and Excitement, which degrades the model's performance to distinguish between these classes. Additionally, a small subset of Excitement samples is inaccurately categorized as Awe. Focusing on these specific categories, we analyze the possible reason behind the misclassification. Then we show some of the most representative and misclassification examples from Emoset dataset in Fig. 5.

Take the first row for an example, the image from last column is labeled with Amusement but is classified as Contentment. Amusement implies a quality of being witty, playful, and entertaining. Contentment entails a stable sense of peace, ease, and satisfaction. The last image depicts several individuals sitting on a grassy lawn, smiling. The background includes sunlight and trees, evoking a warm, comfortable, and peaceful ambiance. In contrast to the former two images, it is observed that the third image lacks the pronounced rhythmic and dynamic qualities present in the first image. Instead, it exhibits a highly similarity with the second image, and the proposed network tends to predict it with Contentment understandably.

Examining the last image in the second row, it is labeled with Excitement but is classified as Awe. Excitement arises from the novelty, uniqueness, or desirability of something, manifesting in increased heart rate, breathing, and perspiration. Awe is experienced when one encounters something greater or more powerful than oneself, evoking feelings of being overwhelmed and magnificent. The main content of

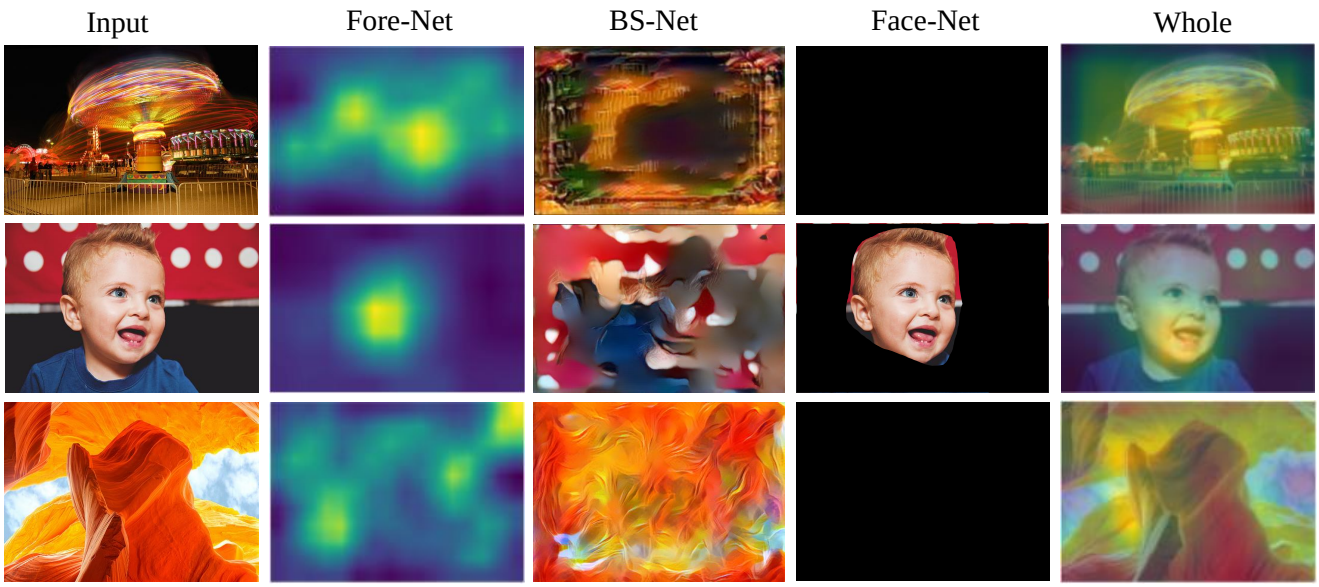

Figure 3. Visualization of each branch of our proposed network.

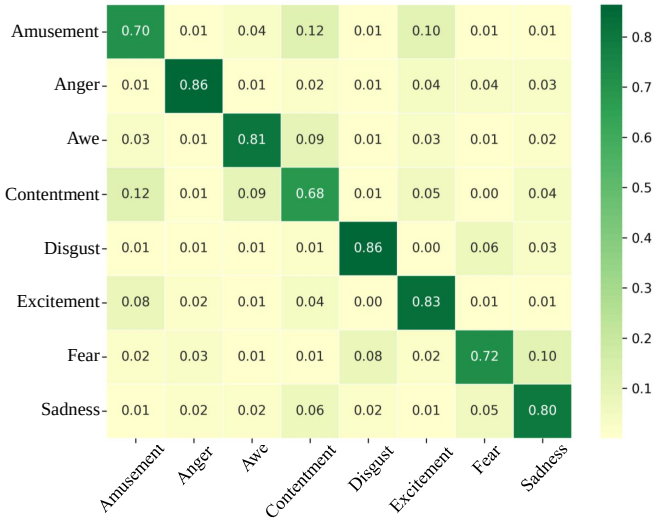

Figure 4. Confusion Matrix of overall prediction performance on Emoset dataset. The scale is on the right.

this image is people walking on a mountain, with the background consisting of mountains and the sky, imparting a solemn and magnificent atmosphere. The scene feature, background style, and facial region of this image differ significantly from the dynamic and rhythmic qualities of the first image. However, it bears considerable semantic similarity to the second image, justifying the network's prediction of Awe.

These instances of failure cases are primarily attributed to the complexity and subjectivity inherent in emotions. As observed from the preceding discussion, it is evident that the fundamental emotional categories are not entirely mutually exclusive. There exist subtle commonalities between them. Moreover, one image may elicit not just one but multiple emotions. Consequently, employing a multi-label approach for annotating emotion images and developing corresponding multi-label learning methods to address this issue may represent a future direction of development.

6. Classification Performance Visualization

We employ the t-SNE [8] algorithm to visualize the classification performance of our method on the Emoset dataset in Fig. 6. For comparison, we also present the t-SNE visualization results of the existing method, MDAN. It can be observed that the sample points in our results exhibit greater cohesion and clearer boundaries between categories. Our method demonstrates superior classification performance compared to MDAN.

7. Conclusion

In this paper, we propose a novel multi-dimension full scene integrated visual emotion analysis network, consisting of Fore-Net, BS-Net and Face-Net, focusing on scene feature, background style information and facial region. Moreover, we leverage channel attention mechanism to capture relationships between different channels, enhancing the feature representation capability. Besides, we devise a novel multi-dimension cross entropy loss function combining discrete and continuous dimension. Experiments show our approach outperforms the state-of-the-art methods on four public visual emotion datasets, demonstrating the robustness and effectiveness of our method. Our framework with comprehensive consideration and incorporating previously overlooked information, strives to encompass factors

Figure 5. The examples of failure cases. The former two columns of each row are the images and the corresponding true labels. In the third column, the first label is the true one, while the latter is the incorrectly predicted one marked with red.

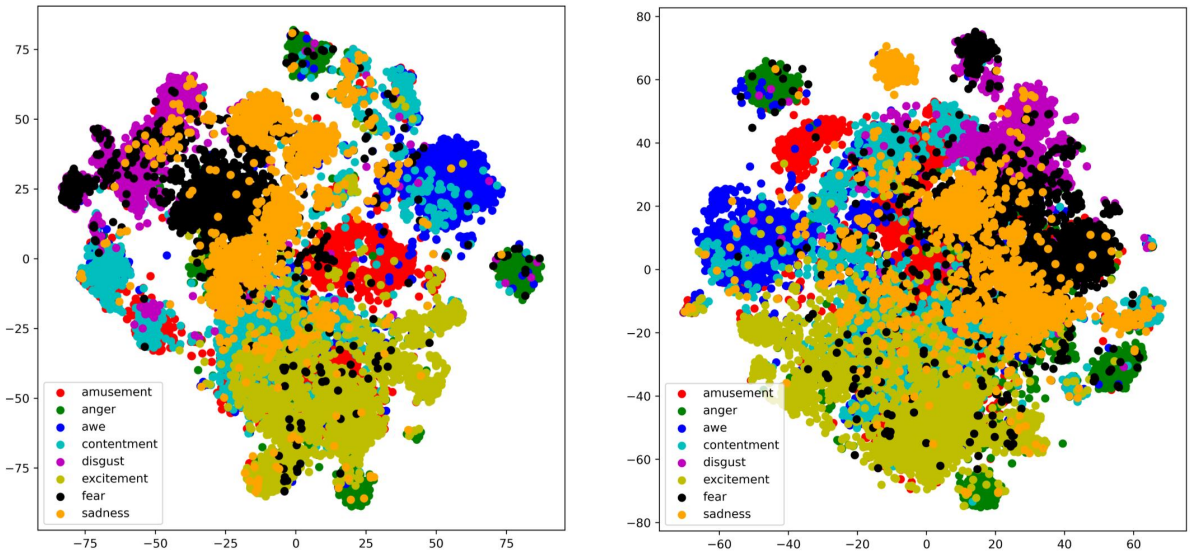

Figure 6. T-SNE visualization on proposed method and MDAN with Emoset dataset.

that influence the arousal of emotions as much as possible. Moreover, ablation study and visualization show that our method has effectiveness on emotions prediction. In failure cases analysis, we find that the emotions conveyed by emotional images may not be singular but could be multiple. This insight suggests that using a multi-label approach might be suitable for studying VEA. Besides, we are still limited by the scale of the dataset, and we may explore ways to combine VEA and image generation in the future.

Acknowledgement

This research was funded by the National Natural Science Foundation of China (62472291), Natural Science Foundation of Guangdong Province (2025A1515012154, 2023A1515012685), Open Fund of National Engineering Laboratory for Big Data System Computing Technology (Grant No. SZU-BDSC-OF2024-14).

References

[1] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, page 223–232, New York, NY, USA, 2013. Association for Computing Machinery. 2, 5

[2] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks, 2014. 5

[3] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang. Context autoencoder for self-supervised representation learning. *Int J Comput Vis*, 132, 208–223 (2024), 2024. 5

[4] T. Feng, J. Liu, and J. Yang. Probing sentiment-oriented pre-training inspired by human sentiment perception mechanism. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2850–2860, 2023. 2, 5

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 262–270, Cambridge, MA, USA, 2015. MIT Press. 3

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *ICONIP*, pages 117–124, 2013. 4

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[8] G. E. Hinton and S. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. 7

[9] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009. 4

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 5

[11] S. Liu and W. Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015. 5

[12] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, page 83–92. Association for Computing Machinery, 2010. 2, 5

[13] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 2005. 5

[14] S. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia, jul 2018. Association for Computational Linguistics. 4

[15] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 614–618, 2016. 5

[16] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang. Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia*, 22(5):1358–1371, 2020. 2, 5

[17] H. Sun, C. Pi, and W. Xie. Semi-supervised facial expression recognition by exploring false pseudo-labels. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 234–239, 2023. 2

[18] J. W. Tanaka and M. J. Farah. The holistic representation of faces. *Perception of faces, objects, and scenes: Analytic and holistic processes*, 2003. 1

[19] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539. IEEE Computer Society, 2020. 3

[20] L. Xu, Z. Wang, B. Wu, and S. Lui. Mdan: Multi-level dependent attention network for visual emotion analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9469–9478, 2022. 2, 5

[21] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20326–20337. IEEE Computer Society, 2023. 5

[22] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021. 2, 4, 5

[23] Q. You, J. Luo, H. Jin, and J. Yang. Building a large scale dataset for image emotion recognition: the fine print and the benchmark. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 308–314. AAAI Press, 2016. 5, 6

[24] J. Zhang, H. Sun, Z. Wang, and T. Ruan. Another dimension: Towards multi-subnet neural network for image sentiment analysis. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1126–1131, 2019. 1

[25] W. Zhang, X. He, and W. Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, 2020. 2, 5

[26] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 47–56. Association for Computing Machinery, 2014. 2, 5

[27] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer. Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 192–201. Association for Computing Machinery, 2019. 2, 5

[28] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu. Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3595–3601, 2017. 5