sensors

MDPI

Article

ARMNet: A Network for Image Dimensional Emotion Prediction Based on Affective Region Extraction and Multi-Channel Fusion

Jingjing Zhang [1,2,3,*], Jiaying Sun [1,3], Chunxiao Wang [1,3,4], Zui Tao [1,3] and Fuxiao Zhang [1,2,3]

[1] Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture and Tourism, Communication University of China, Beijing 100024, China; an123xinying@cuc.edu.cn (J.S.); cunxiao_cuc@cuc.edu.cn (C.W.); taozui19981214@163.com (Z.T.); fuxiao@cuc.edu.cn (F.Z.)

[2] School of Computer and Cyber Sciences, Communication University of China, Beijing 100024, China

[3] Beijing Key Laboratory of Modern Entertainment Technology, Communication University of China, Beijing 100024, China

[4] Center for Ethnic and Folk Literature and Art Development, Ministry of Culture and Tourism, Beijing 100007, China

* Correspondence: zjj_cuc@cuc.edu.cn

Abstract: Compared with discrete emotion space, image emotion analysis based on dimensional emotion space can more accurately represent fine-grained emotion. Meanwhile, this high-precision representation of emotion requires dimensional emotion prediction methods to sense and capture emotional information in images as accurately and richly as possible. However, the existing methods mainly focus on emotion recognition by extracting the emotional regions where salient objects are located while ignoring the joint influence of objects and background on emotion. Furthermore, in the existing literature, when fusing multi-level features, no consideration has been given to the varying contributions of features from different levels to emotional analysis, which makes it difficult to distinguish valuable and useless features and cannot improve the utilization of effective features. This paper proposes an image emotion prediction network named ARMNet. In ARMNet, a unified affective region extraction method that integrates eye fixation detection and attention detection is proposed to enhance the combined influence of objects and backgrounds. Additionally, the multi-level features are fused with the consideration of their different contributions through an improved channel attention mechanism. In comparison to the existing methods, experiments conducted on the CGnA10766 dataset demonstrate that the performance of valence and arousal, as measured by Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), has improved by 4.74%, 3.53%, 3.62%, 1.93%, 6.29%, and 7.23%, respectively. Furthermore, the interpretability of the network is enhanced through the visualization of attention weights corresponding to emotional regions within the images.

Keywords: image emotion prediction; dimensional emotion space; attention mechanism; eye fixation detection; multi-channel fusion

check for updates

Citation: Zhang, J.; Sun, J.; Wang, C.; Tao, Z.; Zhang, F. ARMNet: A Network for Image Dimensional Emotion Prediction Based on Affective Region Extraction and Multi-Channel Fusion. *Sensors* **2024**, *24*, 7099. https://doi.org/10.3390/s24217099

Academic Editor: Stefania Perri

Received: 29 August 2024
Revised: 1 November 2024
Accepted: 2 November 2024
Published: 4 November 2024

1. Introduction

Emotion analysis is a critical research area that aims to enhance human–computer interaction and enable intelligent sensing of emotions [1–3]. Effective emotion sensing involves accurately perceiving and interpreting the subtle emotional states depicted in images. In mainstream research, image emotion prediction is usually performed based on dominant affective categories, which include six or eight basic emotion categories [4,5]. However, discrete affective spaces can only represent basic affective categories and are insufficient for subtle emotion representation. With the increasing demand for subtle emotional sensing and representation, image emotion analysis based on dimensional emotion space gradually attracts extensive attention from researchers.

In contrast, dimensional emotion spaces use precise numerical values to represent emotions, making them better suited for describing subtle emotional details [6,7], such as the Pleasure–Activation–Dominance (PAD) model and the valence–arousal–dominance (VAD) model. In this way, dimensional emotion spaces provide an infinite range of emotions and preserve intermediate emotion states, offering a more comprehensive emotional sensing capability [8]. Therefore, dimensional emotion prediction has broader applications in multimedia and other fields, such as intelligent advertising, multimedia retrieval, and public opinion analysis.

Currently, most of the public dimensional emotion annotated image datasets, such as International Affective Picture System (IAPS) [9], Nencki Affective Picture System (NAPS) [10], Geneva Affective Picture Database (GAPED) [11], and Open Affective Standardized Image Set (OASIS) [12], are built based on 2D or 3D emotion models. These datasets are small, generally containing around 1000 images per dataset. In 2017, Kim et al. established the first large-scale dataset named CGnA10766 [13], which included 10,766 images based on dimensional emotion annotations. When the dataset contains a considerable number of samples, deep-learning techniques can be used to solve complex problems [13–15]. In addition, Kim et al. proposed the first deep learning-based prediction model combining the different levels of features and proved that these features are related to emotions [13]. In 2019, Zhao et al. developed a polarity-consistent deep attention network (PDANet) that integrates spatial and channel-wise attention into a convolutional neural network (CNN) with an emotion polarity constraint. Experiment results demonstrated that the PDANet outperforms the state-of-the-art approaches [16]. In 2021, Li et al. proposed a spatial and channel-wise attention-based emotion prediction (SCEP) model that uses the results from saliency detection for spatial attention and leverages spatial and channel-wise attention, multi-layer characteristics [17]. Alarcão et al. extracted 30 discriminant handcrafted features and analyzed the impact of using the discriminant handcrafted features on three well-known CNNs to identify the feature's contribution [8].

Dimensional emotion spaces use precise numerical values to represent emotions. This high-precision representation of emotion requires dimensional emotion prediction methods to capture emotional information in images as accurately and richly as possible. However, existing methods mainly extract affective regions with the salient objects, which starts from the object level [18,19], but they only focus on a few affective regions with the salient objects, neglecting the joint influence of objects and background. Most image emotion research fuses multi-level features through skip connections [20,21], which means that these methods do not take into account the differences between the low-level features and the high-level features [22]. Therefore, they fused multi-level features without considering their different contributions to emotion prediction, which makes it challenging to discriminate between valuable and useless features and cannot enhance the utilization of effective features.

An image prediction network based on dimensional emotion, joint affective region, and multi-channel fusion (ARMNet) is designed to solve the above problems. The main contributions of this paper are as follows:

(1) A method for extracting union affective regions, combining eye fixation detection and attention detection, is proposed to expand the effective emotional area. This method can extract the joint affective regions composed of the objects and the background, which has high contributions to emotion prediction.

(2) An improved channel attention mechanism is proposed, which increases the gating mechanism and fuses the multi-level features to consider the different contributions from multi-level features through attention-based weight adaptive adjustment.

2. Related Work

2.1. Image Emotion Analysis Based on Specific Affective Regions

The key to image emotion analysis is to extract the appropriate discriminant features [16] and those specific affective regions that have a strong influence when triggering emotion [18,23]. Based on the definition of local regions, there are two main methods that exist to extract the specific affective regions. One of the methods extracts specific affective regions based on segmentation or object detection. Xiong et al. segmented local sentiment regions by considering the similarities of colors and textures [24]. Yang et al. used the EdgeBoxes tool to generate thousands of candidate regions and then selected the affective regions automatically by calculating the emotion and object scores of each candidate region [25]. Rao et al. used Faster-RCNN to generate candidate boxes with emotions instead of objects [20]. Another method extracts specific affective regions using the spatial attention mechanism. She et al. used a cross-spatial pooling strategy in the detection branch to generate the spatial attention weight map [23]. Yao et al. extracted polarity- and emotion-specific attended representations by utilizing polarity-specific attention and specific affective attention in lower and higher layers, respectively [26]. Zhao et al. proposed integrating spatial and channel attention into CNN so that both spatial and channel attention could be considered [16]. Li et al. took the results of the salient object detection as an attention distribution to pay attention to every feature entry of multi-layer feature maps [17]. They suppressed the irrelevant regions via a progressive attention process over multiple layers.

Likewise, to highlight the degree of human attention to the information-rich regions, more methods have been proposed using saliency object detection to locate the emotional areas of images [18,19]. The eye fixation detection module can also predict the likelihood of where the human eyes are staring. Wang et al. quantified the performance of human eye detection networks based on deep learning [27]. The results showed that the Salicon [28] has relatively superior comprehensive performance. However, when the number of salient regions in an image is relatively small, current methodologies often neglect a substantial amount of non-salient emotional information.

2.2. *Image Emotion Analysis Based on Multi-Level Features Fusion*

Some studies tried to model this phenomenon using multi-level image features to predict emotion [20,22,29]. For example, Rao et al. [20] proposed a multi-level depth representation network (MldrNet) with a backbone and four branches, which fused the multi-level depth representation with the Mean function. Zhu et al. proposed a BI-GRU framework for visual emotion recognition based on the assumption that features are at different levels [21]. Nagappan et al. presented a multi-stream feature extraction method that captures object and scene features for emotion prediction, integrating various deep image features using a multi-layer perceptron (MLP) and multi-task learning [30]. Rapolu et al. introduced a deep convolutional neural network (CNN) fusion technique that utilizes a differential CNN for extracting emotional features and a supplementary CNN for capturing central object details, significantly enhancing image emotion recognition performance compared to existing state-of-the-art methods [31]. However, most methods fused multi-level features by simply concatenating multi-level features into one. This simplistic fusion approach, characterized by a straightforward skipping connection, overlooks the distinct contributions of high-level and low-level features. However, these features' roles in emotion prediction are not equivalent; thus, enhancing the multi-layer feature fusion method is essential.

3. Method

In this paper, an image-dimension emotion prediction network, joint affective region, and multi-channel fusion named ARMNet is designed to predict the valence and arousal (VA) values of images, as shown in Figure 1. The ARMNet comprises three components: the union affective region extraction module, the improved channel attention module, and the VA values prediction module. Firstly, the union affective region extraction module utilized a pre-trained ResNet-101 [32] to extract image features, followed by the combined detection of eye fixation and spatial attention to generate the union affective regions (details in Section 3.1). Secondly, the improved channel attention module fuses the multi-level features by considering their different contributions through attention-based weight adjustment (details in Section 3.2). Finally, the ARMNet predicts the emotion of the image based on the valence–arousal space through three fully connected layers (details in Section 3.3). The training of the entire framework is performed in an end-to-end manner.

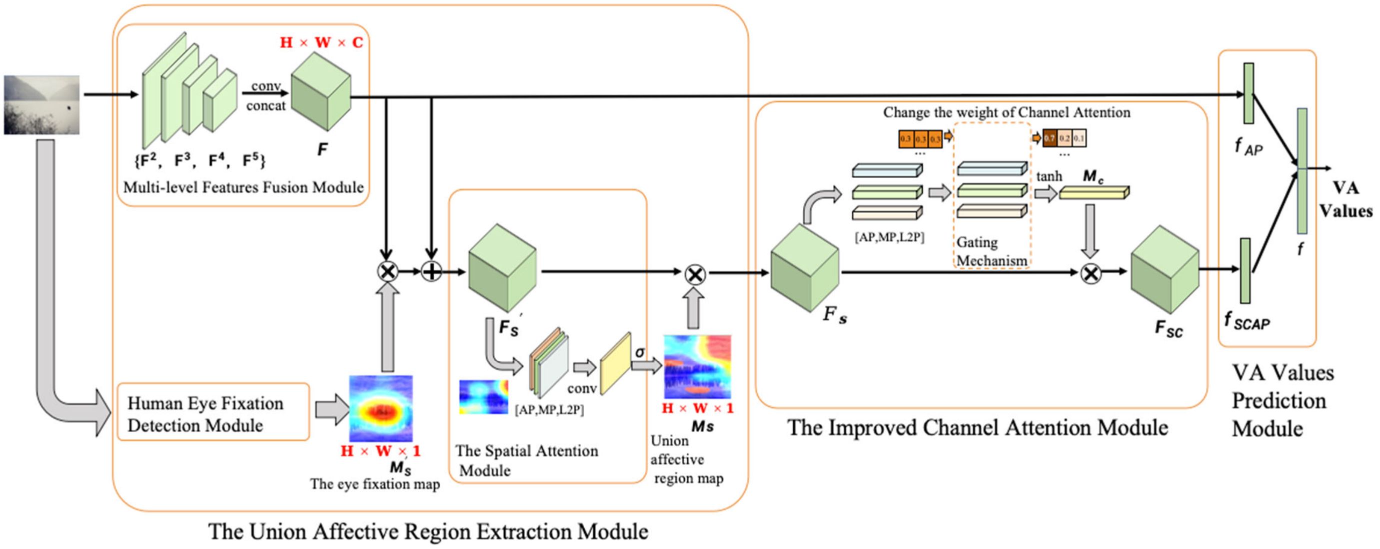
H × W × C
conv
concat
{F², F³, F⁴, F⁵}
F
Multi-level Features Fusion Module
Human Eye Fixation Detection Module
H × W × 1
Ms
The eye fixation map
Fs'
conv
[AP,MP,L2P]
σ
The Spatial Attention Module
H × W × 1
Ms
Union affective region map
Change the weight of Channel Attention
0.3 0.3 0.3
0.7 0.2 0.1
[AP,MP,L2P]
Gating Mechanism
tanh
Mc
Fs
Fsc
The Improved Channel Attention Module
f AP
f SCAP
f
VA Values
VA Values Prediction Module
The Union Affective Region Extraction Module

Figure 1. Illustration of the ARMNet.

3.1. The Union Affective Region Extraction Module

3.1.1. The Multi-Level Features Fusion Module

To extract the image features, the pre-trained ResNet101 is selected as the backbone of the multi-level features fusion module. The features of Conv2, Conv3, Conv4, and Conv5 branches are defined as $\{F^2, F^3, F^4, F^5\}$, as shown in Equation (1). The Conv1 is not included due to its large memory cost.

$$F^l = \left[f_1^{c^l}, f_2^{c^l}, \ldots, f_{x^2}^{c^l}\right] \in R^{c^l x^2} \tag{1}$$

where c^l is the number of channels, l is the number of branches, and x is the length of the feature. After the dimension in convolution mode is decreased, the spatial size of features is identical to x^2. $\{F^2, F^3, F^4, F^5\}$ are concatenated along the channel dimension to obtain the multi-level feature F, as shown in Equation (2) and illustrated in Figure 2.

$$F \in R^{(c^2+c^3+c^4+c^5)x^2} \tag{2}$$

3.1.2. The Human Eye Fixation Detection Module

To generate the union effective region, the Salicon [28] network is directly integrated into the eye fixation detection module. Then, the raw image is injected into the pre-trained Salicon network to obtain the eye fixation map M_S'.

3.1.3. The Spatial Attention Module

The spatial attention module can attend to a significant amount of non-salient emotional regions to generate the union affective region. In the Convolutional Block Attention Module (CBAM), the spatial attention module adopts both *AvgPool* and *MaxPool* to improve the scale of the feature set [33]. But, *AvgPool* and *MaxPool* cannot capture the details of spatial information, which leads to the problem that CBAM has insufficient information when guiding attention learning. Sermanet et al. [34] proposed the concept of *L2Pool* and claimed that its generalization ability is better than $MaxPool$. The calculation process is shown in Equation (3).

$$L2Pool(f_{h\times w}) = \left(\frac{1}{c}\sum\nolimits_{i=1}^{c} f_i^2\right)^{\frac{1}{2}} \tag{3}$$

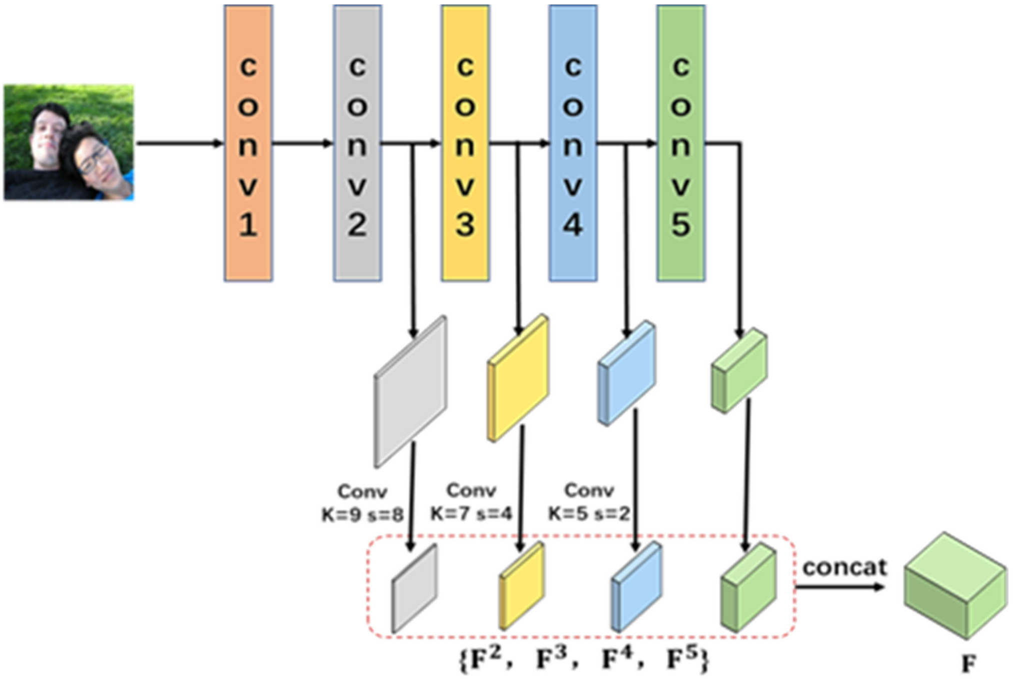

Figure 2. The multi-level features fusion module.

To enrich the feature description and extract useful intermediate features, $L2Pool$ was added to calculate the two-dimensional spatial information as an additional feature descriptor of CBAM, as shown in Figure 3 and Equation (4).

$$F_S = M_S \otimes F_S' \tag{4}$$

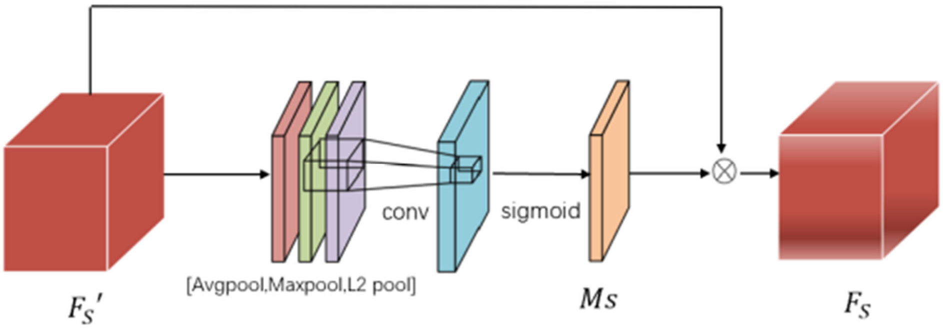

Figure 3. Spatial attention module.

For the ARMNet, after the eye fixation detection module processes the image, the multi-layer feature is weighted by the eye fixation map to obtain saliency information, as shown in Equation (5).

$$F_S' = (M_S' + 1) \odot F \tag{5}$$

F_S' is fed into the attentional module to generate the spatial attention map and the channel attention map. First, F_S' is received by the spatial attention module. Then, three context descriptions, $AvgPool(F_S')$, $MaxPool(F_S')$, and $L2Pool(F_S')$, were derived using $AvgPool$,

$MaxPool$ and $L2Pool$ operations, respectively. And these Spatial context descriptors are concatenated along the channel dimension noted as $\hat{F_S'}$. Then, the union affective region map $M_S \in R^{H\times W\times 1}$ is generated via a convolution layer and a Sigmoid activation function successively, as shown in Equation (6).

$$Ms = \sigma\left(conv\left(\hat{F_S'}\right)\right) \tag{6}$$

3.2. The Improved Channel Attention Module

The improved channel attention module is proposed to fuse the multi-level features by considering their different contributions through attention-based weight adjustment. The structure of the enhanced channel attention module is demonstrated in Figure 4. $L2Pool$ operations and cross-channel feature normalization are implemented, and the fully connected layer is replaced by a gating mechanism [35], as shown in Equation (7).

$$L2Pool(F_S) = \frac{\alpha_c\left\{\left[\sum_{i=1}^{H}\sum_{j=1}^{W}\left(x_c^{i,j}\right)^2\right]+\epsilon\right\}^{\frac{1}{2}}}{(H\times W)^{\frac{1}{2}}} \tag{7}$$

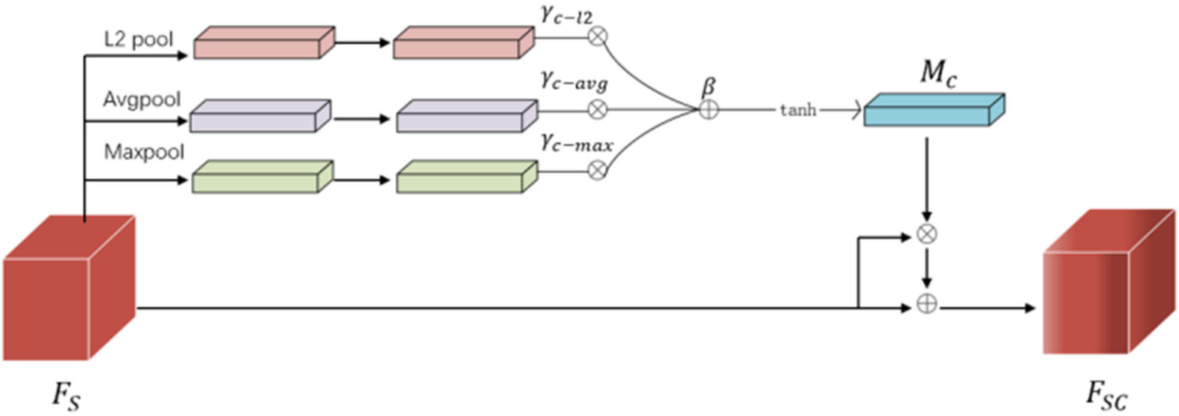

Figure 4. Enhanced channel attention module.

The channel information is aggregated through $AvgPool$, $MaxPool$ and $L2Pool$ and then can be expressed as $\widehat{S_{c-Avg}}$, $\widehat{S_{c-Max}}$, and $\widehat{S_{c-l2}}$ by normalization. ϵ is a very small constant, which avoids the problem of taking the division at 0, as shown in Equations (8)–(10).

$$\widehat{S_{c-l2}} = \frac{L2Pool(F_S)}{\left\{\left(\sum_{i=1}^{C}[L2Pool(F_S)]^2\right)+\epsilon\right\}^{\frac{1}{2}}} \tag{8}$$

$$\widehat{S_{c-Avg}} = \frac{AvgPool(F_S)}{\left\{\left(\sum_{i=1}^{C}[AvgPool(F_S)]^2\right)+\epsilon\right\}^{\frac{1}{2}}} \tag{9}$$

$$\widehat{S_{c-Max}} = \frac{MaxPool(F_S)}{\left\{\left(\sum_{i=1}^{C}[MaxPool(F_S)]^2\right)+\epsilon\right\}^{\frac{1}{2}}} \tag{10}$$

Appropriate cross-channel interaction can preserve performance while drastically decreasing model complexity. By analyzing the channel attention module in the Squeeze-and-Excitation Network (SENet) [36], Efficient Channel Attention Network (ECANet) [37] empirically shows that avoiding dimensionality reduction is critical for learning channel attention. Therefore, referring to the gating mechanism of the Gated channel transformation

(GCT) module [35], the fully connected layer is replaced, and the trainable parameters γ_{c-Avg}, γ_{c-Max}, γ_{c-l2}, and β are designed in this module.

γ_{c-Avg}, γ_{c-Max}, and γ_{c-l2} control the activation state of the three-channel descriptors and assign different weights to the three descriptions, respectively. Finally, the weight map $M_c \in R^{C\times 1}$ of channel attention is obtained, as shown in Equation (11).

$$M_c = 1 + tanh\left[\frac{\left(\gamma_{c-Avg}\widehat{S_{c-Avg}} + \gamma_{c-Max}\widehat{S_{c-Max}} + \gamma_{c-l2}\widehat{S_{c-l2}} + \beta\right)}{3}\right] \tag{11}$$

3.3. The VA Values Prediction Module

To predict the emotion of the image based on the valence–arousal space, the VA values prediction module is proposed. This section utilizes a residual structure where the input of the network consists of both the global features of the image and the features of the joint emotion map branch.

For the global features of the image, as illustrated in Figure 1, we use the outputs of the last residual block of the pre-trained ResNet101 for each stage. The channel dimensions are denoted as $\{c^2, c^3,\ c^4, c^5\}$ for the Conv2, Conv3, Conv4, and Conv5 outputs, respectively. The multi-level feature $F \in R^{(c^2+c^3+c^4+c^5)x^2}$ is obtained by a multi-layer feature fusion module, and the global semantic vector f_{AP} is obtained by using the average pooling on F.

For the features of the joint emotion map branch, the union feature F_{SC} is obtained by calculating F_S' with the weight M_S of the spatial attention first and then calculating the result with the weight M_C of the channel attention. In addition, average pooling is used to obtain the semantic vector f_{SCAP}, as shown in Equation (12) and Equation (13) separately.

$$F_{SC} = M_C \otimes (M_S \odot F_S') \tag{12}$$

$$f_{SCAP} = AvgPool(F_{SC}) \tag{13}$$

As mentioned above, f_{AP} and f_{SCAP} are concatenated to generate the semantic vector f. Finally, f is input into the full connection layers to predict the image emotion based on valence–arousal space.

4. Experiment and Results Analysis

4.1. Implementation Details

All our experiments are carried out on two NVIDIA RTX 2080Ti GPUs using the PyTorch 1.12.1. Images are resized to 300 × 300. To reduce overfitting, images are randomly horizontally flipped and randomly cropped to 224 × 224 patches as a way of data augmentation. We employed the SGD optimizer to fine-tune all layers with 0.0001 as the learning rate with a 0.0005 weight decay, a 0.9 momentum, and a batch size of 32 for 400 epochs. The valence–arousal labels are normalized to [0, 1], and the dataset is randomly divided into 70% training set, 20% testing set, and 10% validation set. The number of network parameters is 124M, and the average training time is about 20 h.

4.2. Datasets

CGnA10766 dataset [13] is composed of 10,766 images searched from Flickr, including eight emotion categories (Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, and Sad). The Amazon Mechanical Turk (AMT) assigned the valence–arousal values for all the images in the CGnA10766 dataset ranging from 1 to 9. The valence is ranged from most negative (1 to 3), negative (3 to 5), positive (5 to 7), and most positive (7 to 9). The arousal is ranged from most calm (1 to 3), calm (3 to 5), exciting (5 to 7), and most exciting (7 to 9). Each image has been evaluated by at least five annotators, and the average value has been finally assigned to each image. The distribution of images in the CGnA10766 dataset is shown in Figure 5 [13]. Compared with other image emotion datasets, the emotion distribution of CGnA10766 is more extensive [17].

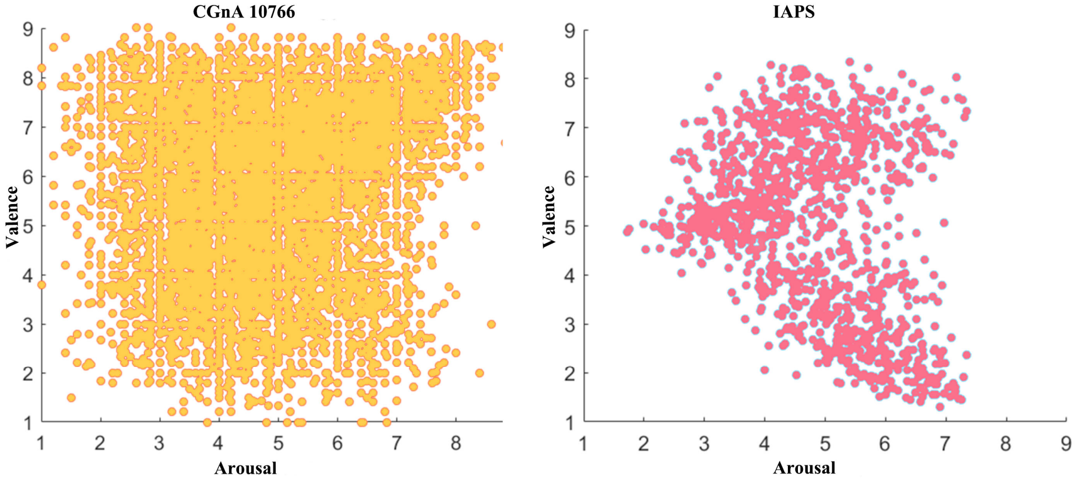

Figure 5. Emotion distribution of CGnA10766 and IAPS based on VA emotion space.

4.3. Performance Comparison

Three commonly used indicators of the regression model are used to evaluate the model performance: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) [16]. The proposed network structure ARMNet is compared with other visual emotion regression models, including a pre-training ResNet101, PDANET [16] (we obtained experimental results on the CGnA dataset using this paper's open-source code), Vision Transformer (ViT) [38], and SCEP [17] with a ResNet101 backbone. The proposed ARMNet performs the best in every indicator. As shown in Table 1, comparing with the best results, experiments on the CGnA10766 dataset show that the performance of ARMNet is improved by 4.74%, 3.53%, 3.62%, 1.93%, 6.29%, and 7.23%, respectively. Experiments show that ARMNet has better robustness and performance by combining the attention mechanism promoted by CBAM with the human eye fixation prediction module.

Table 1. Comparison experiment results ("V" and "A" represent "valence" and "arousal", respectively. "↓" indicates that the value should be as smaller as possible, "↑" indicates that the value should be as larger as possible. The bold numbers indicate the best results).

	MSE_V ↓	**MSE_A ↓**	**MAE_V ↓**	**MAE_A ↓**	**R^2_V ↑**	**R^2_A ↑**
ResNet101 [32]	0.02701	0.02246	0.1289	0.1199	0.3644	0.2467
PDANet [16]	0.02589	0.02083	0.1263	0.1159	0.3909	0.3014
ViT [38]	0.03462	0.02705	0.1455	0.1351	0.1852	0.0927
SCEP [17]	0.02539	0.02117	0.1261	0.1160	0.4024	0.2900
ARMNet (ours)	**0.02424**	**0.02012**	**0.1217**	**0.1137**	**0.4294**	**0.3249**

4.4. Ablation Experiments

The proposed ARMNet contains three major components: the multi-level features fusion module, the human eye fixation detection module, and the spatial-channel attention module. To quantitatively show performance improvement, the network structure (the fully connected layers, the holistic feature vector, etc.) is maintained while the major components are removed separately. In addition, the learning rate, weight decay, batch size, and other hyperparameters are the same as in Section 4.1. Finally, MSE is selected as the comparison indicator.

As shown in Table 2, the results reveal the following information:

(1) According to rows 1 and 2 of Table 2, the results of the model with the multi-level features fusion module are better than those without the multi-level features fusion module. The MSE value for valence and arousal of the model with the multi-level features fusion module was reduced by 4.45% and 1.58%, respectively.

(2) According to rows 1, 3, 4, and 5 of Table 2, the eye fixation detection module and the spatial attention mechanism can improve performance. The combination of them performs better than every single module. This proves the necessity of adding a human attention detection module and a spatial attention detection module to the ARMNet. For example, the eye fixation detection module reduced the MSE value for valence and arousal by 1.33% and 1.56%, respectively. Additionally, the spatial attention mechanism module reduces the MSE value by 3.26% in the valence domain, but the MSE value in the arousal domain is almost the same.

(3) According to rows 6 and 7 of Table 2, a comparison shows the performance differences between CAM and SAM. While both combinations (R + M + S + CAM and R + M + S + SAM) yield similar MSE values, R + M + S + SAM slightly outperforms R + M + S + CAM in both the valence and arousal domains. This suggests that although CAM effectively captures channel-wise information, SAM shows more robust performance for spatial attention in the emotional prediction task.

(4) According to rows 6 and 10 of Table 2, the network with the channel attention mechanism module reduces the MSE value for valence and arousal by 2.64% and 1.84% for valence and arousal, respectively, which verifies the validity of the channel attention mechanism module.

(5) When CAM is introduced, as seen in rows 9 and 10 of Table 2, the combination of R + M + S + CBAM + CAM does not outperform R + M + S + SAM + CAM, which delivers better results. This demonstrates that although CBAM has advantages in certain setups, SAM, when combined with CAM, provides more stable and superior performance for emotional prediction.

(6) Furthermore, the spatial-channel attention module is designed based on the CBAM module by adding a gating mechanism, including a spatial attention mechanism module and a channel attention mechanism module. The result is shown in rows 6, 8, 9, and 10 of Table 2. It proves that the CBAM module is effective, while the spatial-channel attention is better. Compared with the CBAM module network, the network with the spatial-channel attention module reduced the MSE value for valence and arousal by 1.03% and 2.49%, respectively.

Table 2. Ablation experiments results (“V”and “A” represent “valence” and “arousal”. “R”, “M”, “S”, “SAM”, “CAM”, and “CBAM” represent “ResNet101”, “Multi-level features fusion module”, “Salicon”, “Spatial Attention Mechanism”, “Channel Attention Mechanism”, and “Convolutional Block Attention Module”, respectively. “↓” indicates that the value should be as smaller as possible, “↑” indicates that the value should be as larger as possible. The bold numbers indicate the best results).

		MSE_V ↓	**MSE_A ↓**	**MAE_V ↓**	**MAE_A ↓**	**R^2_V ↑**	**R^2_A ↑**
1	R	0.02701	0.02246	0.1289	0.1199	0.3644	0.2467
2	R + M	0.02586	0.02211	0.1259	0.1187	0.3912	0.2586
3	R + S	0.02569	0.02050	0.1259	0.1151	0.3956	0.3126
4	R + SAM	0.02521	0.02082	0.1247	0.1153	0.4067	0.3018
5	R + S + SAM	0.02488	0.02050	0.1237	0.1151	0.4145	0.3124
6	R + M + S + SAM	0.02488	0.02049	0.1235	0.1151	0.4145	0.3129
7	R + M + S + CAM	0.02497	0.02100	0.1239	0.1159	0.4123	0.2955
8	R + M + S + CBAM	0.02449	0.02062	0.1225	0.1150	0.4237	0.3084
9	R + M + S + CBAM + CAM	0.02797	0.02433	0.1304	0.1236	0.3417	0.1839
10	R + M + S + SAM + CAM	**0.02424**	**0.02012**	**0.1217**	**0.1137**	**0.4294**	**0.3249**

4.5. Visualization

Attention map reflects the varying importance of input features and is widely used to improve the interpretability of neural networks. The visualizations of the affective regions and the channel attention of the ARMNet are shown in Figure 6.

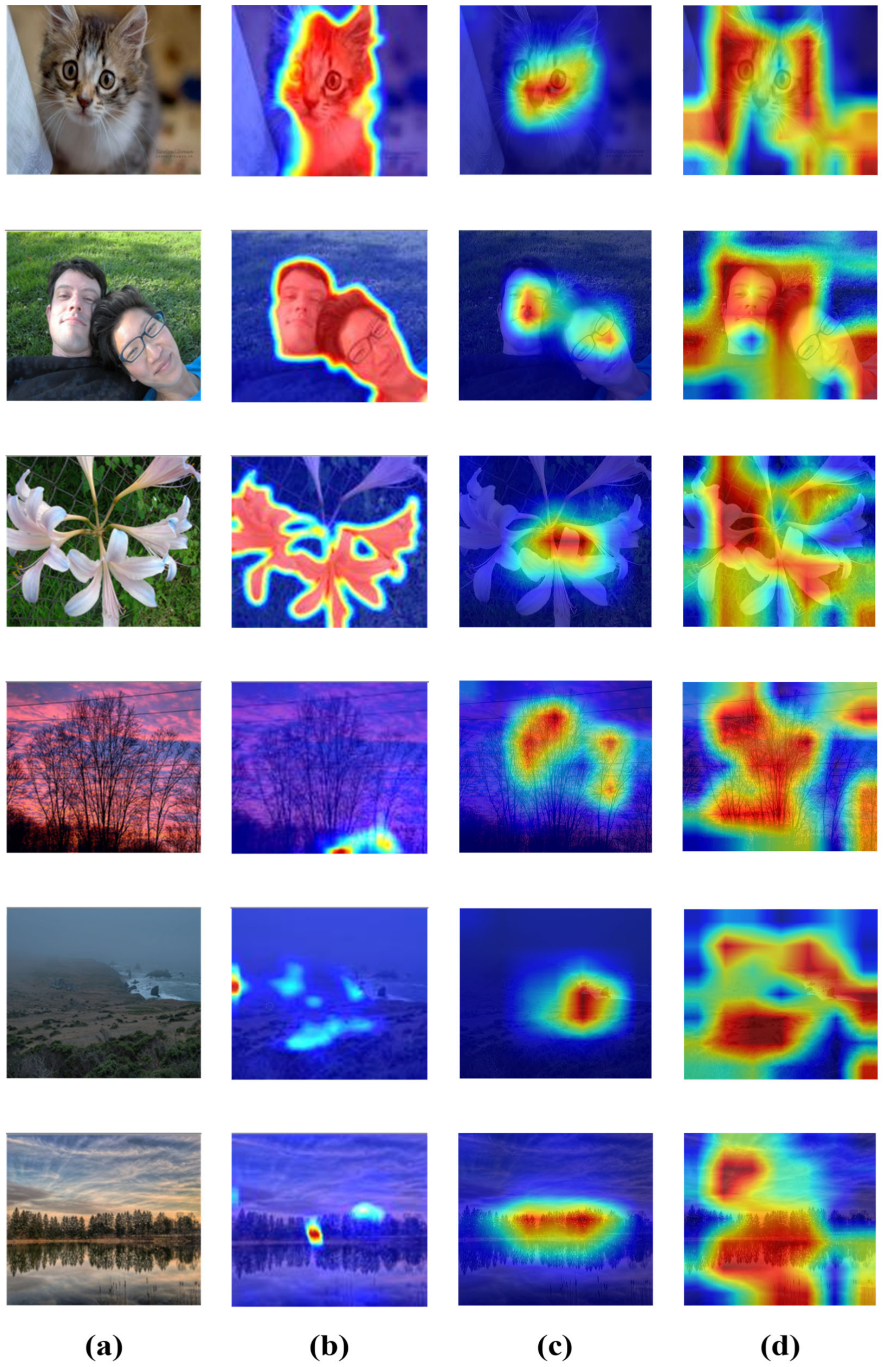

Figure 6. Visualizations of spatial attention: (**a**) the original images; (**b**) the salient object maps; (**c**) the eye fixation maps; (**d**) the union attention region map after integrating the human eye fixation detection module, where the color red represents a higher weight.

4.5.1. Visualization of the Spatial Attention Module

For comparison with existing methods and evaluation of the performance of the union affective region extraction module proposed in this paper, this section conducts a visual analysis. As shown in Figure 6, Figure 6a presents the original picture, while Figure 6b shows one of the affective regions obtained through salient object detection maps using BASNet (Boundary-Aware Salient Object Detection) [39]. Figure 6c,d show the extracted eye fixation map and the union attention region map from ARMNet, respectively.

As shown in Figure 6d, the final attention weights are not only derived from salient objects; this phenomenon is more apparent in images without salient objects. Figure 6c,d show that the eye fixation detection module detects the eye fixation maps, which include the saliency regions, and the spatial attention maps concentrate on the background regions related to emotions. Therefore, the proposed ARMNet can obtain the affective regions with or without the salient objects simultaneously based on the fusion of the human eye fixation detection module and spatial attention mechanism. In other words, the union affective region extraction module enhances the model's ability to capture emotional information.

To further illustrate the effectiveness of the proposed fusion attention mechanism, the images of the CGnA10766 dataset were divided into two categories, as shown in Figure 7. The first category consists of images that contain a specific target object, such as people or animals. The second category includes images without a specific target object, such as landscapes. The first and second categories of images are 6287 and 4479, respectively. Moreover, as shown in Table 3, the valence prediction error of the first category is much smaller than that of the second category, whereas the arousal prediction errors for both categories are almost the same.

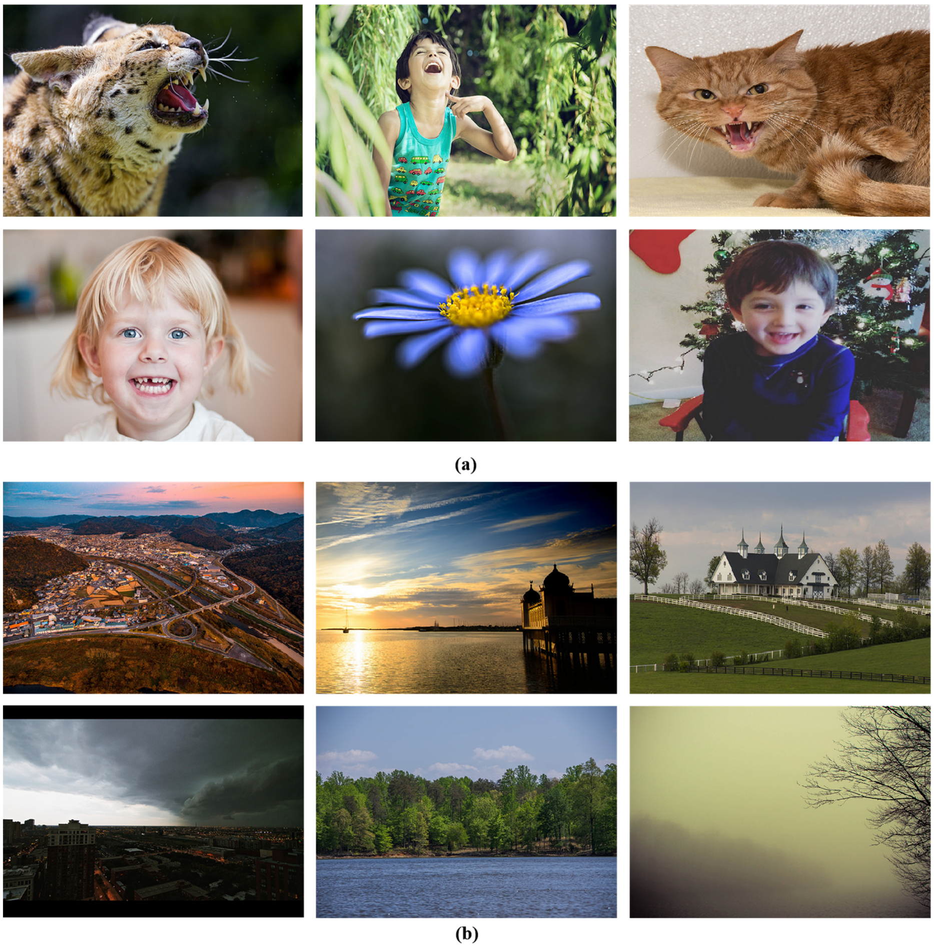

Figure 7. Images from CGnA10766 dataset: (**a**) images with specific target objects (person, animal, etc.); (**b**) images without a specific target subject.

Table 3. Experimental results on two categories of images with or without specific target subjects: (a) images with specific target objects (person, animal, etc.); (b) images without a specific target subject. "↓" indicates that the value should be as smaller as possible, "↑" indicates that the value should be as larger as possible.

	MSE_V ↓	**MSE_A** ↓	**MAE_V** ↓	**MAE_A** ↓	**R^2_V** ↑	**R^2_A** ↑
(a)	0.02528	0.02134	0.1258	0.1167	0.3456	0.2509
(b)	0.01696	0.02167	0.1012	0.1172	0.5229	0.2858

4.5.2. Visualization of the Improved Channel Attention

The three weight parameters of the improved attention channel γ_{c-Avg}, γ_{c-Max}, and γ_{c-l2} are obtained by Equation (11), and the dimension of the channel is 3680. The values of γ_{c-Avg}, γ_{c-Max}, and γ_{c-l2} are shown in Figure 8, Figure 9, and Figure 10, respectively. The visualization results reveal the following information:

(1) The different parameters γ_{c-Avg}, γ_{c-Max}, and γ_{c-l2} indicate that the network assigns different importance to different channel feature descriptors.

(2) The specific values of γ_{c-Avg}, γ_{c-Max}, and γ_{c-l2} are different. The gating weights of the output features from the Conv3 and Conv4 branches are approximately zero, and the gating weights of the output features from the Conv2 branch are small. However, the gating weights of the output features from the Conv5 branch are bigger and fluctuate more sharply, which show a greater influence on the final prediction.

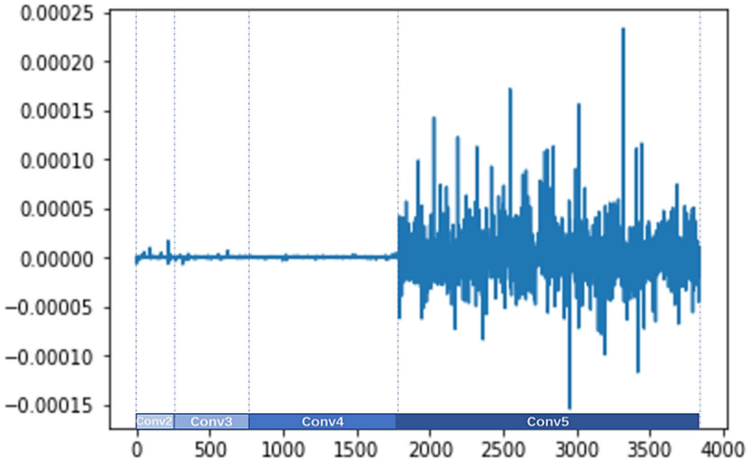

Figure 8. Visualization of γ_{c-Avg}.

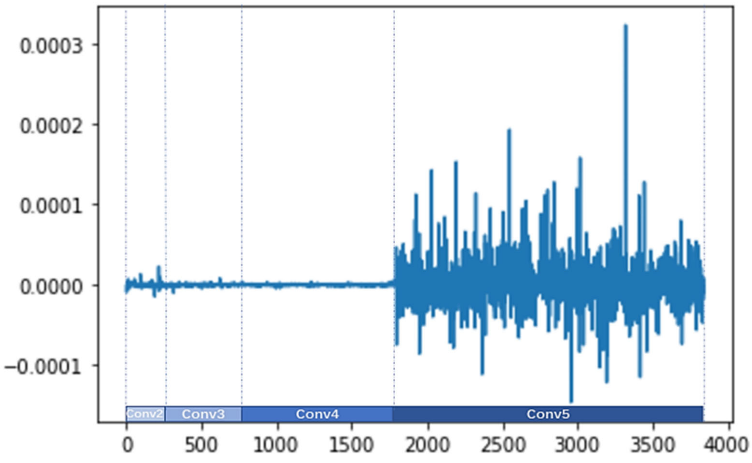

Figure 9. Visualization of γ_{c-Max}.

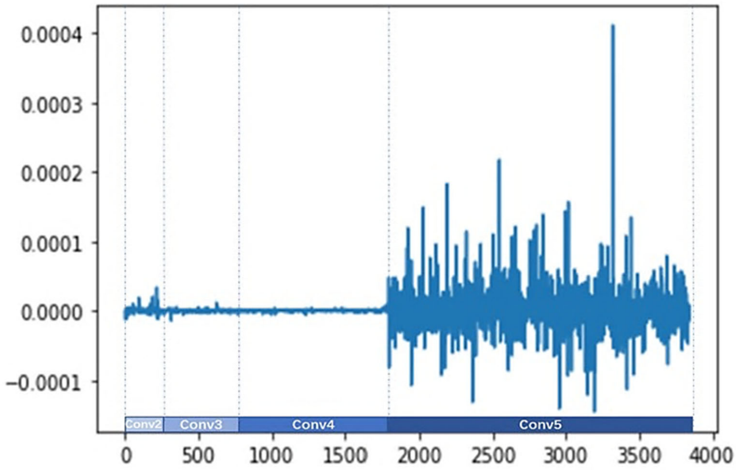

Figure 10. Visualization of γ_{c-l2}.

5. Conclusions

Emotion recognition and processing are essential research directions for developing new intelligent sensing systems in the future. Complex emotions can be described better based on dimensional emotion space, including the subtle emotion representation. This precise representation of emotion necessitates the use of dimensional emotion prediction methods to capture emotional information in images with maximum accuracy and comprehensiveness. However, existing methods mainly extract affective regions with salient objects, neglecting the joint influence of objects and background for the emotion of the image. And even with the same object, there may be differences in emotion due to different backgrounds. Additionally, they fused multi-level features without considering their different contributions to emotion prediction, which makes it challenging to discriminate between valuable and useless features.

This paper proposes an image emotion prediction network named ARMNet. In ARMNet, a union affective region extraction method that combines eye fixation detection and attention detection is proposed to represent the joint influence of objects and background. And the multi-level features are fused with considering their different contributions by an improved channel attention mechanism. We compare the proposed network with other advanced methods on the CGnA10766 dataset. The proposed network performs the best in every indicator. The performance of valence and arousal, as measured by MSE, MAE, and R^2, has improved by 4.74%, 3.53%, 3.62%, 1.93%, 6.29%, and 7.23%, respectively. The effectiveness of each module is also proved by ablation experiments.

Author Contributions: Conceptualization, J.Z., J.S. and C.W.; methodology, J.Z., J.S. and C.W.; software, J.S.; validation, J.Z., J.S. and C.W.; formal analysis, J.Z., J.S., C.W. and F.Z.; investigation, J.Z., J.S. and C.W.; resources, J.Z. and C.W.; data curation, J.S., Z.T. and F.Z.; writing—original draft preparation, J.Z., C.W. and J.S.; writing—review and editing, J.Z., C.W., J.S., Z.T. and F.Z.; visualization, C.W. and J.S.; supervision, J.Z.; project administration, J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62276240) and the Fundamental Research Funds for the Central Universities (No. CUC22JG004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this paper are openly available at: https://figshare.com/articles/CGnA10766_Dataset/5383105 (accessed on 1 September 2020).

Conflicts of Interest: The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

1. Chen, H.; Shao, F.; Mu, B.; Jiang, Q. Image Aesthetics Assessment With Emotion-Aware Multibranch Network. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–15. [CrossRef]
2. Su, Z.; Feng, Y.; Liu, J.; Peng, J.; Jiang, W.; Liu, J. An Audiovisual Correlation Matching Method Based on Fine-Grained Emotion and Feature Fusion. *Sensors* **2024**, *24*, 5681. [CrossRef] [PubMed]
3. Kosti, M.V.; Georgakopoulou, N.; Diplaris, S.; Pistola, T.; Chatzistavros, K.; Xefteris, V.-R.; Tsanousa, A.; Vrochidis, S.; Kompatsiaris, I. Assessing Virtual Reality Spaces for Elders Using Image-Based Sentiment Analysis and Stress Level Detection. *Sensors* **2023**, *23*, 4130. [CrossRef] [PubMed]
4. Horvat, M.; Jović, A.; Burnik, K. Investigation of Relationships between Discrete and Dimensional Emotion Models in Affective Picture Databases Using Unsupervised Machine Learning. *Appl. Sci.* **2022**, *12*, 7864. [CrossRef]
5. Li, H.; Lu, Y.; Zhu, H. Multi-Modal Sentiment Analysis Based on Image and Text Fusion Based on Cross-Attention Mechanism. *Electronics* **2024**, *13*, 2069. [CrossRef]
6. Zhao, S.; Yao, X. An Overview of Image Affective Computing. *Intell. Comput. Appl.* **2017**, *7*, 1–5.
7. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
8. Alarcão, M.; Ribeiro, C.; Garcia, N.; Maruta, C.; Fonseca, M.J. Unfolding hand-crafted features contribution in CNNs for valence and arousal estimation in images. *J. Vis. Commun. Image Represent* **2022**, 13–26. [CrossRef]
9. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Cent. Study Emot. Atten.* **1997**, *1*, 3.
10. Marchewka, A.; Żurawski, Ł.; Jednoróg, K.; Grabowska, A. The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behav. Res. Methods* **2014**, *46*, 596–610. [CrossRef]
11. Dan-Glauser, E.S.; Scherer, K.R. The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behav. Res. Methods* **2011**, *43*, 468–477. [CrossRef] [PubMed]
12. Kurdi, B.; Lozano, S.; Banaji, M.R. Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods* **2017**, *49*, 457–470. [CrossRef] [PubMed]
13. Kim, H.-R.; Kim, Y.-S.; Kim, S.J.; Lee, I.-K. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Trans. Multimed.* **2018**, *20*, 2980–2992. [CrossRef]
14. Yan, M.; Xiong, R.; Wang, Y.; Li, C. Edge Computing Task Offloading Optimization for a UAV-assisted Internet of Vehicles via Deep Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2024**, *73*, 5647–5658. [CrossRef]
15. Yan, M.; Luo, M.; Chan, C.A.; Gygax, A.F.; Li, C.; I, C.-L. Energy-Efficient Content Fetching Strategies in Cache-Enabled D2D Networks via an Actor-Critic Reinforcement Learning Structure. *IEEE Trans. Veh. Technol.* 2024, *early access*. [CrossRef]
16. Zhao, S.; Jia, Z.; Chen, H.; Li, L.; Ding, G.; Keutzer, K. PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 192–201.
17. Li, B.; Ren, H.; Jiang, X.; Miao, F.; Feng, F.; Jin, L. SCEP—A new image dimensional emotion recognition model based on spatial and channel-wise attention mechanisms. *IEEE Access* **2021**, *9*, 25278–25290. [CrossRef]
18. Deng, Z.; Zhu, Q.; He, P.; Zhang, D.; Luo, Y. A Saliency Detection and Gram Matrix Transform-Based Convolutional Neural Network for Image Emotion Classification. *Secur. Commun. Netw.* **2021**, *2021*, 6854586. [CrossRef]
19. Sowmyayani, S.; Rani, P. Salient object-based visual sentiment analysis by combining deep features and handcrafted features. *Multimed. Tools Appl.* **2022**, *81*, 7941–7955. [CrossRef]
20. Rao, T.; Li, X.; Zhang, H.; Xu, M. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing* **2019**, *333*, 429–439. [CrossRef]
21. Zhu, X.; Li, L.; Zhang, W.; Rao, T.; Xu, M.; Huang, Q.; Xu, D. Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; pp. 3595–3601.
22. Rao, T.; Li, X.; Xu, M. Learning multi-level deep representations for image emotion classification. *Neural Process. Lett.* **2020**, *51*, 2043–2061. [CrossRef]
23. She, D.; Yang, J.; Cheng, M.-M.; Lai, Y.-K.; Rosin, P.L.; Wang, L. WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Trans. Multimed.* **2019**, *22*, 1358–1371. [CrossRef]
24. Xiong, H.; Liu, Q.; Song, S.; Cai, Y. Region-based convolutional neural network using group sparse regularization for image sentiment classification. *EURASIP J. Image Video Process* **2019**, *2019*, 30. [CrossRef]
25. Yang, J.; She, D.; Sun, M.; Cheng, M.-M.; Rosin, P.L.; Wang, L. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Trans. Multimed.* **2018**, *20*, 2513–2525. [CrossRef]
26. Yao, X.; She, D.; Zhao, S.; Liang, J.; Lai, Y.-K.; Yang, J. Attention-aware polarity sensitive embedding for affective image retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1140–1150.
27. Wang, W.; Shen, J.; Dong, X.; Borji, A.; Yang, R. Inferring salient objects from human fixations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1913–1927. [CrossRef] [PubMed]

28. Huang, X.; Shen, C.; Boix, X.; Zhao, Q. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 262–270.
29. Zhang, H.; Xu, M. Weakly supervised emotion intensity prediction for recognition of emotions in images. *IEEE Trans. Multimed.* **2020**, *23*, 2033–2044. [CrossRef]
30. Nagappan, S.; Tan, J.Q.; Wong, L.K.; See, J. Context-Aware Multi-Stream Networks for Dimensional Emotion Prediction in Images. In Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 8–11 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 2480–2484.
31. Rapolu, S.; Singh, A.; Dhingra, A. Convolutional Neural Networks for Image Emotion Recognition by Fusing Differential and Supplementary Information. In Proceedings of the 2023 International Conference on Bio Signals, Images, and Instrumentation (ICBSII), Chennai, India, 16–17 March 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
34. Sermanet, P.; Chintala, S.; LeCun, Y. Convolutional neural networks applied to house numbers digit classification. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, 11–15 November 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 3288–3291.
35. Yang, Z.; Zhu, L.; Wu, Y.; Yang, Y. Gated channel transformation for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11794–11803.
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
38. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
39. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. BASNet: Boundary-aware salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7479–7489.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.