S.I. : NCACVIP

Global-local attention for emotion recognition

Nhat Le[1,2] · Khanh Nguyen[1,2] · Anh Nguyen[3] · Bac Le[1,2]

Abstract

Human emotion recognition is an active research area in artificial intelligence and has made substantial progress over the past few years. Many recent works mainly focus on facial regions to infer human affection, while the surrounding context information is not effectively utilized. In this paper, we proposed a new deep network to effectively recognize human emotions using a novel global-local attention mechanism. Our network is designed to extract features from both facial and context regions independently, then learn them together using the attention module. In this way, both the facial and contextual information is used to infer human emotions, therefore enhancing the discrimination of the classifier. The intensive experiments show that our method surpasses the current state-of-the-art methods on recent emotion datasets by a fair margin. Qualitatively, our global-local attention module can extract more meaningful attention maps than previous methods. The source code and trained model of our network are available at https://github.com/minhnhatvt/glamor-net.

1 Introduction

Emotion recognition aims to classify input data into several expressions that convey universal emotions, such as an-gry, disgust, fear, happy, neutral, sad, and surprise. The input data can be one or more of different modalities such as visual information, audio, and text [10, 24, 35]. Due to the availability of a large number of images and videos on the Internet, inferring human emotion from visual content, is considered to be one of the most popular tasks. Recently, automatic emotion recognition has gained a lot of attention in both academia and industry [49]. It enables a wide range of novel applications in different domains, ranging from healthcare [15], surveillance [9] to robotics [42] and human-computer interaction [11].

Traditional methods for emotion recognition combine handcrafted features (e.g, histogram of oriented gradients (HOG) [5], local binary patterns) with classifiers such as SVM [20] or graphical models [27]. With the popularity of deep learning techniques, especially Convolutional Neural Network (CNN) [30], together with the exists of many large-scale datasets, the meaningful features can be extracted using a deep network. However, the majority of previous methods [6, 23, 38, 58] only exploit features from human's face, and use this information to predict human emotions. These works assume that the facial region is the most informative representation of human emotion, therefore they ignore the surrounding context, which is shown to play an important role in the understanding of the perceived emotion, especially when the emotions on the face are expressed weakly or indistinguishable [32].

Recently, researchers have been focusing on incorporating background information such as people's pose, gaits, etc., into the model to improve the performance [39, 43]. In this work, we follow the same direction. However, unlike other works that learn the facial and context information independently [39], we propose to jointly learn both facial and context information using our new Global-Local Attention mechanism. We hypothesize that the local

✉ Anh Nguyen
anh.nguyen@liverpool.ac.uk

Bac Le
lhbac@fit.hcmus.edu.vn

[1] Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam

[2] Vietnam National University, Ho Chi Minh City, Vietnam

[3] Department of Computer Science, University of Liverpool, Liverpool, UK

information (i.e., facial region) and global information (i.e., context background) have a correlative relationship, and by simultaneously learning the attention using both of them, the accuracy of the network can be improved. This is based on the fact that the emotion of one person can be indicated by not only the face's emotion (i.e., local information) but also other context information such as the gesture, pose, or emotion/pose of a nearby person. Figure 1 shows some recognition results of our proposed method.

To verify the effectiveness of our approach, we benchmark on the CAER-S dataset [32], a large-scale dataset for context-aware emotion recognition. We achieved 77.90% top-1 accuracy on the test set, which is an improvement of 4.38% over the recent state-of-the-art method [32]. Furthermore, with the integrated ResNet-18 [25] as the backbone network, we obtained state-of-the-art performance on the CAER-S dataset with 89.88% classification accuracy. We also present a novel way to create a new static-image dataset from videos of the CAER dataset [32]. The experiments on this new dataset also confirm that our proposed method consistently achieves better performance than previous state-of-the-art approaches.

In summary, our contributions are as follows:

- We propose a new deep network, namely, **G**lobal-**L**ocal **A**ttention for E**mo**tion **R**ecognition **Net**work (GLAMOR-Net) that surpasses the state-of-the-art methods in the emotion recognition task.
- In GLAMOR-Net, we proposed the Global-Local Attention module, which successfully encodes both local features from facial regions and global features from surrounding background to improve the human emotion classification accuracy.
- We perform extensive experiments to validate the effectiveness of our proposed method and the contribution of each module on recent challenging datasets.

Fig. 1 Examples of human emotion detection results from our method

The paper is organized as follow: We review the related work in Sect. 2. We then describe our methodology in detail in Sect. 3. In Sect. 4, we present extensive experimental results on challenging datasets and analyze the contribution of each module in GLAMOR-Net. Finally, we conclude the paper and discuss future work in Sect. 5.

2 Related work

2.1 Human emotion

In the late twentieth century, Ekman and Friesen discovered six basic universal emotions including anger, disgust, fear, happiness, sadness, and surprise [18]. Several years later, contempt was added and considered as one of the basic emotions [37]. However, our affective displays in reality are much more complicated and subtle compared to the simplicity of these universal emotions. To represent the complexity of the emotional spectrum, many approaches were proposed such as the Facial Action Coding System [8], where all facial actions are described in terms of Action Units (AUs); or dimensional models [46], where affection is quantified by values chosen over continuous emotional scales like valence and arousal. Nevertheless, those models which use discrete affections are the most popular in automatic emotion recognition task because they are easier to interpret and more intuitive to human.

2.2 Emotion recognition

In automatic human emotion recognition, many approaches mainly focus on analyzing facial expression. Thus, a standard emotion recognition system usually consists of three main stages: face detection, feature extraction, and expression classification [6, 23, 38, 58]). Traditional methods relied on handcrafted features (LBP [51], HOG [5]) to extract meaningful features from input images, and classifiers (such as SVM or random forest) to classify human emotions based on extracted features. With the rise of deep learning, CNN-based methods have made significant progress in the task of emotion recognition [34]. Apart from using input image, other works focus on categorizing emotions by utilizing extra information such as speech [19, 26], human pose [50], body movements and gaits [47, 55]. However, these works have relied on the information coming from a single modality, hence they have limited ability to fully exploit all usable information of human emotions.

To overcome this limitation, many researches have investigated the use of multiple modalities. Primarily, these works tried to fuse multiple channels of information from each modality to predict emotion. Castellano et al. [4] used

extracted features from three different modalities (facial expressions, body gestures and speech expressions), and then fused those modalities in two different levels (i.e. feature level and decision level). Their results showed that the fusion performed at the feature level provided better results than the one performed at the decision level. Sikka et al. [52] extracted different visual features such as SIFT-Bag of Words [53], LPQ-TOP [45], HOG [12], PHOG [3], and GIST [44] and fuse them with audio features by building a kernel from each set of features, then combine them using a SVM classifier. Likewise, the authors in [56] used the same multi-modality approach but using deep learning techniques. In [39], three interpretations of context information are fused together by a deep neural network to classify human emotions in an end-to-end manner.

Recently, many works have focused on exploring context-aware information for emotion recognition. Kosti et al. [29] and Lee et al. [32] proposed two architectures based on deep neural networks for learning context information. Both of them have two separate branches for extracting different kinds of information. One branch focuses on human features (i.e. face for [32] and body for [29]) and the other concentrates on surrounding context. When considering multiple modalities, which have a large amount of information, deep learning-based methods like [16, 29, 32, 39] are more suitable and effective than traditional approaches. These multi-modal approaches often yield better classification performance than uni-modal methods.

2.3 Attention model

Attention was first introduced in machine translation [2], allowing the translation model to search for words in the input sentence that are more relevant to the prediction words. Since then attention models have become an important concept and an essential component of neural network architectures. It has made significant impacts in many application domains, including natural language processing [21], computer vision [57], graph [33], and speech processing [7].

In emotion recognition, attention models were mainly used to discover the attentive areas of the face that need to be focused on [6]. Recently, the work that forced the model to pay attention to the most discriminative regions of the background using attention was proposed in CAER-Net-S [32]. However, previous work only used the background encoding to learn the context saliency map and did not take advantage of the facial representation to assist the process. Therefore, we propose the Global-Local Attention mechanism, which takes both facial and context encoding as inputs, to utilize facial information more efficiently to guide the context saliency map learning procedure.

3 Methodology

3.1 Overview

In this work, we assume that emotions can be recognized by understanding the context components of the scene together with the facial expression. Our method aims to do emotion recognition in the wild by incorporating both facial information of the person's face and contextual information surrounding that person. Our model consists of three components: Encoding Module, Global-Local Attention (GLA) Module, and Fusion Module. Our main contribution is the novel GLA module, which utilizes facial features as the local information to attend better to salient locations in the global context. Figure 2 shows an overview of our method.

3.2 Network architecture

3.2.1 Encoding module

To detect human emotion, many works first process the image by cropping out the human faces from the scene, and then feed them into a convolutional network to extract facially-expressive features [6, 23, 38, 58]. We generally follow this approach in our Encoding Module. In particular, our Encoding Module comprises the Facial Encoding Module to learn the face features, and the Context Encoding Module to learn the context features.

Facial Encoding Module This module aims to learn meaningful features from the facial region of the input image. The facial embedding information can be denoted as $\mathbf{F}_f$:

$$\mathbf{F}_f = \mathfrak{C}(\mathbf{I}_f; \theta_f) \tag{1}$$

where $\mathfrak{C}$ is the convolutional operation parameterized by θ_f, and $\mathbf{I}_f$ is the input facial region. In practice, we use a sub-network (Fig. 3) as the feature extractor for the Facial Encoding Module.

The proposed sub-network has five convolutional layers. Particularly, each convolutional layer has a kernel set of 3×3 filters with strides of 1×1 followed by a Batch Normalization layer and a ReLU activation function. The number of filters starts with 32 in the first layer, increasing by a factor of 2 at each subsequent layer except the last one. Our network ends up with 256 output channels. We also use the padding technique before each convolutional layer to keep the output spatial dimensions the same as the input. The output of each convolutional layer is pooled using a max-pooling layer with strides of 2×2. The encoding module outputs a 256-channel volume feature
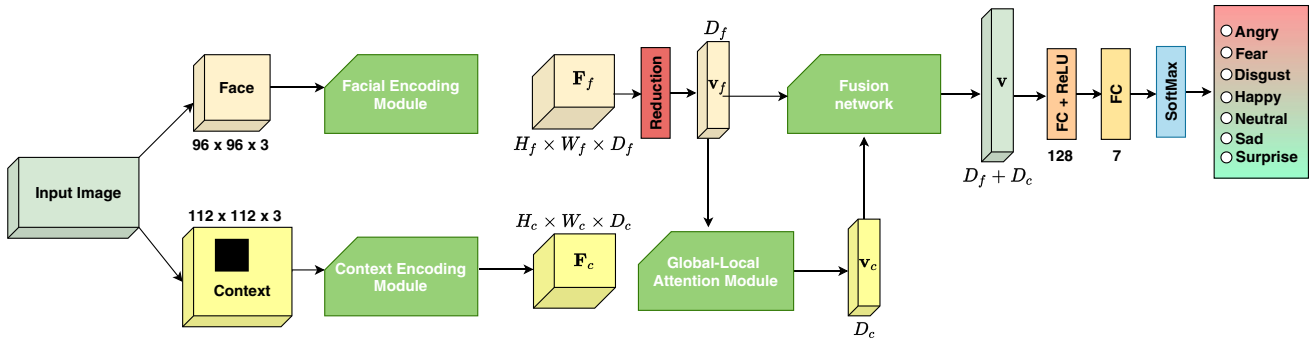
Fig. 2 The architecture of our proposed network. The whole process includes three steps. First, we extract the facial information (local) and context information (global) using two Encoding Modules. Second, we feed the extracted face and context features into the

Global-Local Attention (GLA) module to perform attention inference on the global context. Lastly, we fuse both features from the facial region and output features from GLA into a neural network to make final emotion classification
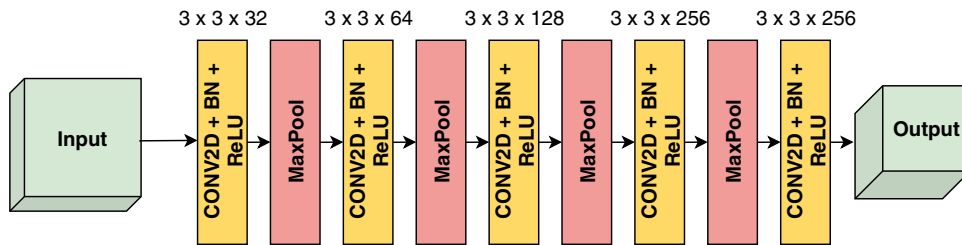

Fig. 3 Our proposed encoder network as the feature extractor for both face and context branches. The network contains five convolutional layers with ReLU non-linearity, each convolution is followed by a max pooling layer except the last one to reduce the spatial dimensions of the input

map, which is the embedded representation with respect to the input image.

Context Encoding Module This module is used to exploit background knowledge to support the emotion predicting process. Similar to the Facial Encoding Module, we follow the same procedure to extract context information contained in the scene with a different set of parameters:

$$\mathbf{F}_c = \mathfrak{C}(\mathbf{I}_c; \theta_c) \tag{2}$$

where $\mathfrak{C}$ is the convolutional operation parameterized by θ_c, and $\mathbf{I}_c$ is the input context. Similar to the Facial Encoding Module, we use the sub-network (Fig. 3) to extract deep features from the background context region in the Context Encoding Module.

After getting these two feature maps, we feed them into the Global-Local Attention Module to calculate the attention scores for regions in the context. However, if we extract the context information in the raw image where the faces apparently exist, the network will also encode the facial information. This problem can make the attention module produce trivial outputs because the network may only focus on the facial region, and omitting the context information in other parts of the image. To address this problem, we first detect the face and then hide it in the raw input by setting all the values in the facial region to zero.

3.2.2 Global-local attention module

Inspired by the attention mechanism [7, 41], to model the associative relationship of the local information (i.e., the facial region in our work) and global information (i.e., the surrounding context background), we propose the Global-Local Attention Module to guide the network focus on meaningful regions (Fig. 4). Specifically, our attention mechanism models the hidden correlation between the face and different regions in the context by capturing their similarity using deep learning techniques. Our attention module takes the extracted face feature map $\mathbf{F}_f$ and the context feature map $\mathbf{F}_c$ from the two encoding modules as input, and then outputs a normalized saliency map that has the same spatial dimension as $\mathbf{F}_c$.

In practice, we first reduce the facial feature map $\mathbf{F}_f$ into vector representation using the Global Pooling operator, denoted as $\mathbf{v}_f$. Note that the context feature map $\mathbf{F}_c$ is a 3D tensor, $\mathbf{F}_c \in \mathbb{R}^{H_c \times W_c \times D_c}$, where H_c, W_c, and D_c are the height, width, and channel dimension respectively. We derive the context feature map $\mathbf{F}_c$ as a set of $W_c * H_c$ vectors with D_c dimensions, each vector in each cell (i, j) represents the embedded features at that location, which can be projected back to the corresponding patch in the input image:

$$\mathbf{F}_c = \{\mathbf{v}_{i,j} \in \mathbb{R}^{D_c} | 1 \leq i \leq H_c, 1 \leq j \leq W_c\} \tag{3}$$

⊙ Element-wise multiplication
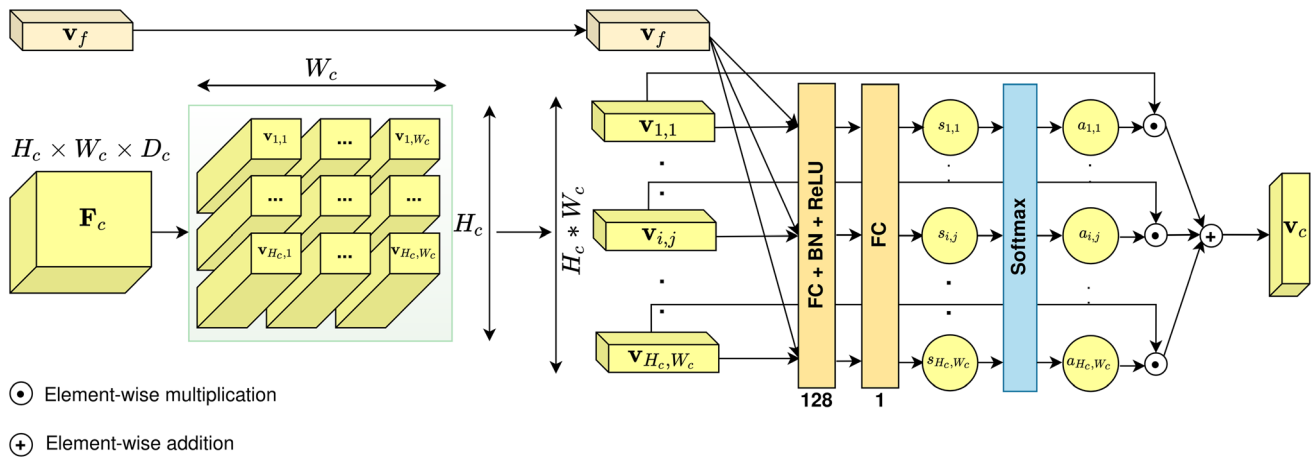
⊕ Element-wise addition

Fig. 4 The proposed Global-Local Attention module takes the extracted face feature vector and the context feature map as the input to perform context attention inference. Each vector $\mathbf{v}_{i,j}$ in the context feature map $\mathbf{F}_c$ is concatenated with the face vector $\mathbf{v}_f$ and then fed into a sub-network to compute the attention weight for the (i, j) position. The final output vector is a linear combination of all regions in the context weighted by the corresponding attention weight. For efficiency, our attention inference network contains a 128-unit Fully Connected layer with the ReLU activation function and a Softmax layer. Weights are shared across all the context regions

At each location (i, j) in the context feature map, we have $\mathbf{F}_c^{(i,j)} = \mathbf{v}_{i,j}$, where $\mathbf{v}_{i,j} \in \mathbb{R}^{D_c}$ and $1 \leq i \leq H_c$, $1 \leq j \leq W_c$.

We concatenate $[\mathbf{v}_f; \mathbf{v}_{i,j}]$ into a big vector $\bar{\mathbf{v}}_{i,j}$, which contains both information about the face and some small regions of the scene. We then employ a feed-forward neural network to compute the score corresponding to that region by feeding $\bar{\mathbf{v}}_{i,j}$ into the network. After repeating the same process for all regions, each region (i, j) will output a raw score value $s_{i,j}$, we spatially apply the Softmax function to produce the attention map:

$$a_{i,j} = \frac{\exp(s_{i,j})}{\Sigma_a \Sigma_b \exp(s_{a,b})} \qquad (4)$$

To obtain the final context representation vector, we squish the feature maps by taking the average over all the regions weighted by $a_{i,j}$ as follow:

$$\mathbf{v}_c = \Sigma_i \Sigma_j (a_{i,j} \odot \mathbf{v}_{i,j}) \qquad (5)$$

where $\mathbf{v}_c \in \mathbb{R}^{D_c}$ is the final single vector encoding the context information, and $\odot$ is the scalar multiplication operation. Additionally, $\mathbf{v}_c$ mainly contains information from regions that have high attention, while other unimportant parts of the context are mostly ignored. With this design, our attention module can guide the network focus on important areas based on both facial information and context information of the image. Note that, in practice, we only need to extract context information once and then using different encoded face representations to make the system look at different regions with respect to that person.

3.2.3 Fusion module

The Fusion Module is used to incorporate the facial and context information more effectively when predicting human emotions. The Fusion Module takes $\mathbf{v}_f$ and $\mathbf{v}_c$ as the input, then the face score and context score are computed independently by two neural networks:

$$s_f = \mathcal{F}(\mathbf{v}_f; \phi_f) \quad s_c = \mathcal{F}(\mathbf{v}_c; \phi_c) \qquad (6)$$

where ϕ_f and ϕ_c are the network parameters of the face branch and context branch, respectively. Next, we normalize those scores by the Softmax function to produce weights for each face and context branch so that these weights sum up to 1.

$$w_f = \frac{\exp(s_f)}{\exp(s_f) + \exp(s_c)} \quad w_c = \frac{\exp(s_c)}{\exp(s_f) + \exp(s_c)} \qquad (7)$$

Notice that the face weight and the context weight are independently computed by their corresponding networks and represent the importance of these branches. We let the two networks competitively determine which branch is more useful than the other. Then we amplify the more useful branch and lower the effect of the other by multiplying the extracted features with the corresponding weight:

$$\mathbf{v}_f \leftarrow \mathbf{v}_f \odot w_f \quad \mathbf{v}_c \leftarrow \mathbf{v}_c \odot w_c \qquad (8)$$

Finally, we use these vectors to estimate the emotion category. Specifically, in our experiments, after multiplying both $\mathbf{v}_f$ and $\mathbf{v}_c$ by their corresponding weights, we concatenate them together as the input for a network to make final predictions. Fig. 5 shows our fusion procedure in detail.
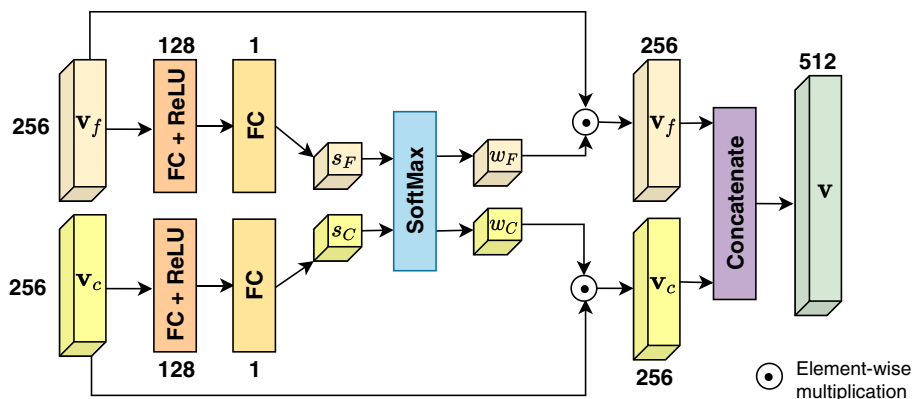
Fig. 5 The Fusion Module consists of two separate sub-networks, each network computes the fusion weights for face branch and context branch. The input vector of each branch is then scaled by its

corresponding weight and combined together into the final representation vector **v**. We use this vector **v** to estimate the emotion category by feeding it into another sub-network (see Fig. 2)

4 Experiments

4.1 Datasets

CAER-S In this work, we only focus on static images with background context as our input. Therefore, we choose the static CAER (CAER-S) dataset [32] to validate our method. The CAER-S dataset contains 70K static images extracted from a total of 13201 video clips of 79 TV shows. Each image is labeled with one of seven universal emotions: anger, fear, disgust, happiness, neutral, sadness and surprise. We follow the standard split proposed by [32] for training, validation and testing, respectively.

Novel CAER-S (NCAER-S) While experimenting with the CAER-S dataset, we observe that there is a correlation between images in the training and test sets, which can make the model less robust to changes in data and may not generalize well on unseen samples. More specifically, many images in the training and the test set of the CAER-S dataset are extracted from the same video, hence making them look very similar to each other. To cope with this issue, we propose a novel way to extract static frames from the CAER video clips to create a new static image dataset called Novel CAER-S (NCAER-S). In particular, frames extracted from the training, validation, and test sets of the CAER dataset are separately put into the corresponding training, validation, and test sets of the new NCAER-S dataset. In particular, for each video in the original CAER dataset, we split the video into multiple parts, each part is approximately 2s long. Then we randomly select one frame of each part to include in the new NCAER-S dataset. Any original video that provides frames for the training set will be removed from the testing set. This process assures the new dataset is novel while the training frames and testing frames are never from one original input video.

With our selection method, we ensure that images in the validation and test sets are independent of those in the training set. We also make sure that the numbers of extracted frames of each emotion category are approximately equal to tackle the imbalance problem of the CAER dataset and prevent bias towards prominent emotions.

The statistics of the original CAER and the new NCAER-S training sets are shown in Fig. 6 and Table 1. The new split NCAER-S dataset can be downloaded at https://bit.ly/NCAERS_dataset.

4.2 Experimental setup

Evaluation Metric Classification accuracy is the standard evaluation metric that is widely used to measure the reliability of automated emotion recognition systems in the literature [13, 32, 34, 36, 40]. To compare our results with previous approaches quantitatively, as in [32, 34] we use the overall classification accuracy as the evaluation metric:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\hat{y}_i = y_i\} \tag{9}$$

where $\mathbb{1}$ is the indicator function, N is the total number of samples in the dataset, $\hat{y}_i$ and y_i is the network prediction and ground-truth category of the i-th example, respectively.

Baselines We compare the results of our proposed **G**lobal-**L**ocal **A**ttention for E**mo**tion **R**ecognition network (GLA-MOR-Net) with the following methods as baselines: AlexNet [31], VGGNet [54], ResNet [25], CAER-Net-S [32]. The results of AlexNet, VGGNet, and ResNet on the CAER-S dataset are reported in two cases: using the ImageNet dataset as the pre-trained model, and fine-tuning these networks on this dataset. Note that, these results are taken from [32] paper. On the CAER-S, we also compare our method to several recent state-of-the-art approaches. GRERN [22] utilized a multi-layer Graph Convolutional
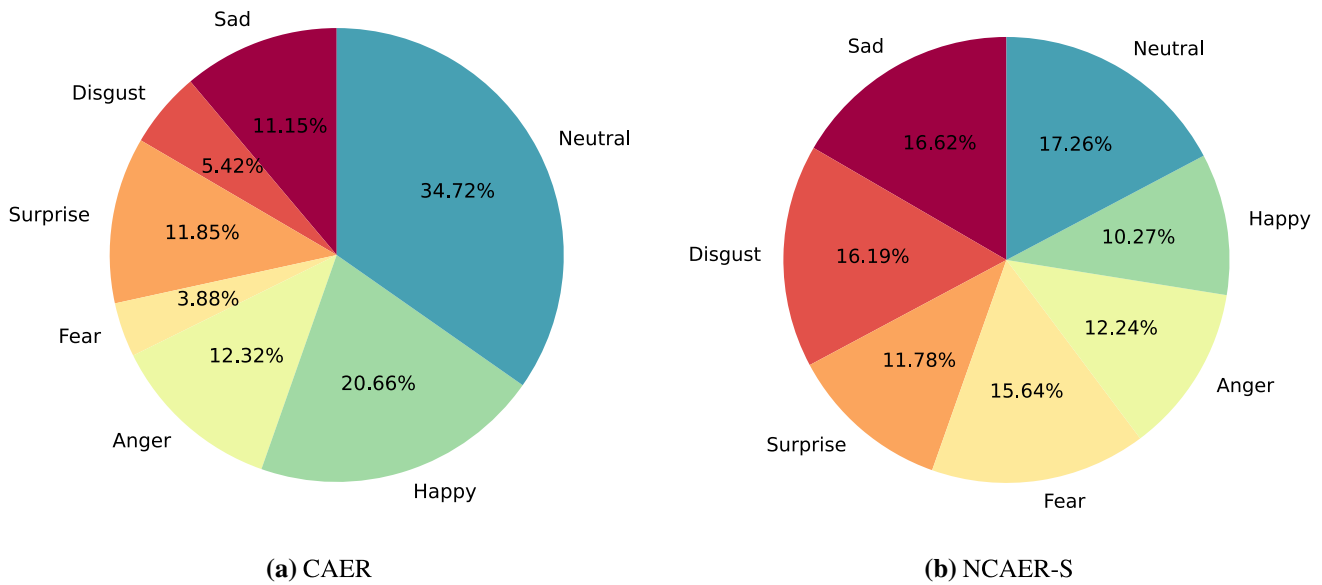
(a) CAER　　　　　　　　　　　　　　　**(b)** NCAER-S

Fig. 6 Percentage of each emotion category in the CAER and the new NCAER-S training sets

Table 1 The number of images in each emotion category in the NCAER-S training set

Emotion	Number of images
Angry	2272
Disgust	3004
Fear	2902
Happy	1905
Neutral	3202
Sad	3084
Surprise	2186
Total	18,555

Network (GCN) to exploit the relationship among different regions in the context. EfficientFace [60] proposed an efficient lightweight network and utilized the label distribution to handle the ambiguity of real-world emotions. MA-Net [59] designed a highly complicated architecture based on ensemble learning of multiple regions to handle the occlusion and pose variation problems. We report the results of our GLAMOR-Net with two different backbones: the original encoding module introduced in Sect. 3.2.1 and ResNet-18 [25].

Implementation Details Our networks are implemented using Tensorflow 2.0 framework [1]. For optimization, we use the SGD optimization algorithm and standard cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log p_i^{(y_i)} \tag{10}$$

where $p_i^{(y_i)}$ is the predicted probability for the true emotion category y_i of the i-th sample and N is the total number of samples in the dataset.

Given an input image, we first use the CNN based face detector in the dlib library [28] to detect the face coordinates. The detected face is then cropped and resized to 96 × 96 and fed to the Facial Encoding Module. To create input for the Context Encoding Module, we mask the facial region in the original image and resize it to 128 × 171, then we apply random crop during the training phase and center crop during the inference phase to the final size of 112 × 112. We use a dropout layer before the final layer with a dropout rate of 0.5 to reduce the effect of overfitting. During training, we observe that the fusion network is very unstable and easily affected by random factors. Specifically, the weights of the face branch or the context branch in the Fusion Module can easily take a value near 0 or 1, which means the model completely ignores information extracted from one of the branches. To tackle this problem, we first train the Facial Encoding Module and the Context Encoding Module separately, then jointly train both modules and the fusion network in an end-to-end manner.

4.3 Results

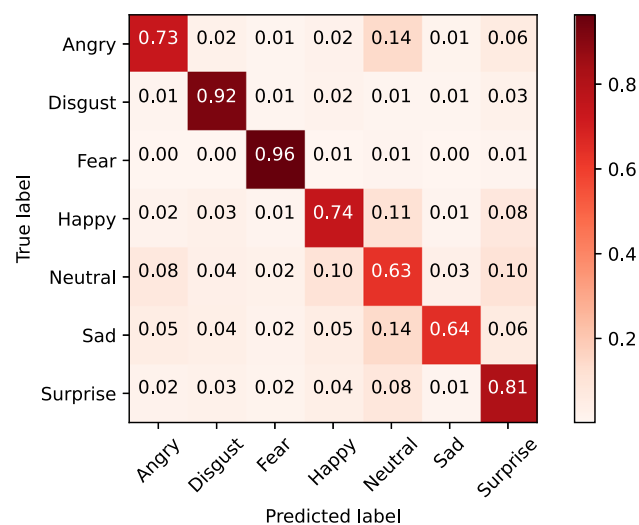#### 4.3.1 Results on the CAER-S dataset

Table 2 summarizes the results of our network and other recent state-of-the-art methods on the CAER-S dataset [32]. This table clearly shows that integrating our GLA module can significantly improve the accuracy performance of the recent CAER-Net. In particular, our

Table 2 Classification accuracy of baseline methods and our GLA-MOR-Net on the CAER-S dataset (bold denotes the best result)

Methods	Year	Accuracy (%)
ImageNet-AlexNet [31]	2012	47.36
ImageNet-VGGNet [54]	2015	49.89
ImageNet-ResNet [25]	2016	57.33
Fine-tuned AlexNet [31]	2012	61.73
Fine-tuned VGGNet [54]	2015	64.85
Fine-tuned ResNet [25]	2016	68.46
CAER-Net-S [32]	2019	73.52
GRERN [22]	2020	81.31
EfficientFace [60]	2021	81.48
MA-Net [59]	2021	88.42
GLAMOR-Net (original)	2021	77.90
GLAMOR-Net (ResNet-18)	2021	**89.88**

GLAMOR-Net (original) achieves 77.90% accuracy, which is a + 4.38% improvement over the CAER-Net-S. When compared with other recent state-of-the-art approaches, the table clearly demonstrates that our GLAMOR-Net (ResNet-18) outperforms all those methods and achieves a new state-of-the-art performance with an accuracy of 89.88%. This result confirms our global-local attention mechanism can effectively encode both facial information and context information to improve the human emotion classification results.

Figure 7 shows the confusion matrix of the GLAMOR-Net (original) on the CAER-S dataset. Overall, the model achieves the highest accuracy on the `fear` class with 0.96 accuracy. The `neutral` class has the lowest accuracy of

Fig. 7 The confusion matrix of our GLAMOR-Net (original) results on the CAER-S test set

0.63 as there are many misclassifications from other classes.

4.3.2 Results on the NCAER-S dataset

On the NCAER-S dataset, we compare our results with three recent methods: VGG16 [54], ResNet50 [25], and CAER-Net-S [32]. The results from the VGG16 and ResNet50 models are reproduced as baseline methods. We finetune the VGG16 and the ResNet50 from the pre-trained models on VGG-Face and ImageNet, respectively. Our GLAMOR-Net (original) and CAER-Net-S are trained from scratch for a fair comparison.

Table 3 reports the comparative results of our GLAMOR-Net and other recent methods. This table shows that the GLAMOR-Net architecture outperforms all other architectures and achieves the highest performance. In particular, our network increases classification accuracy by 2.77% compared to the second-highest model CAER-Net-S. These results also validate the effectiveness of our proposed global-local attention mechanism integrated into the GLAMOR-Net. We note that the result of VGG16 pre-trained on VGG-Face is surprisingly better than the result of ResNet50 pre-trained on ImageNet dataset. This is explainable as the pre-trained weight on VGG-Face carries more meaningful information than the pre-trained weight on ImageNet, which includes many non-face images.

Also from Table 3, we can see that the classification accuracy of the models is much lower than those in Table 2. The reason behind this is the new NCAER-S is more challenging than the original CAER-S dataset. As mentioned earlier, to construct the NCAER-S dataset, we eliminate the correlation between the train and the test samples as much as we can. Specifically, we separately resample image frames from clips of the train and test sets of the CAER dataset to mitigate the train and test dependency. Moreover, note that the size of the new dataset is only less than one-third of the original one, which also limits the amount of information that the models can exploit. However, our GLAMOR-Net still consistently

Table 3 Classification accuracy of baseline methods and our GLA-MOR-Net on the NCAER-S dataset (bold denotes the best result)

Methods	Accuracy (%)
VGG16 [54]	42.85
ResNet50 [25]	41.41
CAER-Net-S [32]	44.14
GLAMOR-Net (original)	**46.91**

outperforms other state-of-the-art methods despite the challenges of the NCAER-S dataset and shows competitive results.

The confusion matrix of our GLAMOR-Net evaluated on the NCAER-S dataset is given in Fig. 8. The two categories with the highest accuracy are `happy` and `neutral` while the `disgust` emotion has the lowest accuracy of 0.28. It can also be inferred from the confusion matrix that our model mostly confuses `neutral` with other emotion categories as most of the misclassified examples of the six categories: `angry`, `disgust`, `fear`, `happy`, `sad` and `surprise` fall into the class `neutral`. It may be because the facial emotion in the NCAER-S dataset is weakly expressed, which makes it more difficult to identify and distinguish other emotions from the `neutral` class.

In summary, we can conclude that our method consistently improves the results on both the original CAER-S and the challenging NCAER-S datasets. Note that although we follow the same procedure as in [32], our proposed Global-Local Attention Module is the key difference that helps enhance the accuracy of the emotion recognition task. The results reported in Tables 2 and 3 verify that with the assistance of our attention strategy, the classification accuracy is significantly improved. We believe that if a more sophisticated neural architecture is adopted, the performance will be further boosted.

4.3.3 Analysis

To further analyze the contribution of each component in our proposed method, we experiment with 4 different input settings on the NCAER-S dataset: (*i*) face only, (*ii*) context only with the facial region being masked, (*iii*) context only
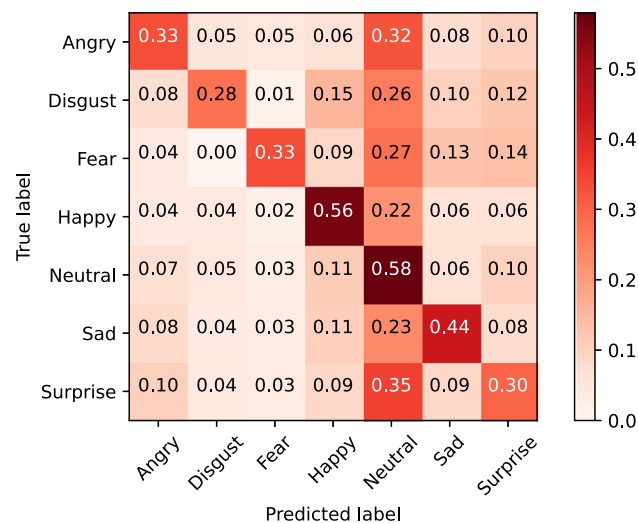
with the facial region visible, and (*iv*) both face and context (with masked face). When the context information is used, we compare the performance of the model with different context attention approaches (no attention, standard attention module in CAER-Net-S and our GLA module). Note that to compute the saliency map with the proposed GLA in the (*ii*) and (*iii*) setting, we extract facial features using the Facial Encoding Module, however, these features are only used as the input of the GLA module to guide the context attention map learning process and *not* as the input of the Fusion Network to predict the emotion category. The performances of these settings are summarized in Table 4.

The results clearly show that our GLA consistently helps improve performance in all settings. Specifically, in setting (*ii*), using our GLA achieves an improvement of 1.06% over method without attention, 0.97% over standard attention module in CAER-Net-S [32]. It is also noteworthy that when the context with visible faces is utilized as in setting (*iii*), using the attention module in the CAER-Net-S achieves 41.94% accuracy, lower than the one using only the cropped face in setting (*i*) by 0.64%, while using our GLA module achieves higher accuracy (42.66% vs. 42.58%). Our GLA also improves the performance of the model when both facial and context information is used to predict emotion. Specifically, our model with GLA achieves the best result with an accuracy of 46.91%, which is higher than the method with no attention 3.72% and standard attention module in [32] 2.77%. The results from Table 4 show the effectiveness of our Global-Local Attention module for the task of emotion recognition. They also verify that the use of both the local face region and

Fig. 8 The confusion matrix of our GLAMOR-Net (original) results on the NCAER-S test set

Table 4 Ablation study of our proposed method on the NCAER-S dataset (bold denotes the best result)

Settings	w/F	w/mC	w/fC	w/CA	w/GLA	Accuracy (%)
(i)	✓					42.58
(ii)		✓				41.18
		✓		✓		41.27
		✓			✓	42.24
(iii)			✓	✓		41.94
			✓		✓	42.66
(iv)	✓	✓				43.19
	✓	✓		✓		44.14
	✓	✓			✓	**46.91**

'w/F', 'w/mC', 'w/fC', 'w/CA', 'w/GLA' denote using the output of the Facial Encoding Module, the Context Encoding Module with masked faces as input, the Context Encoding Module with visible faces as input, the standard Context Attention in CAER-Net-S [32] and our Global-Local Attention Module, respectively, as input to the Fusion Network

global context information is essential for improving emotion recognition accuracy.

In order to emphasize the contribution of the Attention module to the final results, we conduct Stuart-Maxwell test for each pair of methods that are used in the setting (iv) of Table 4.The Stuart-Maxwell test is the generalized version of McNemar test [14] which is generally used for testing the significant difference of multi-class classification models. The resulted p-values of the tests are shown in Table 5. Note that the lower p value indicates stronger statistical disagreement between the two compared methods. Overall, we can see that all of the models have significant different error rates. Furthermore, the higher value on the main diagonal would imply stronger agreement between the model prediction and the observed data, which means the performance is better. In conjunction with the results in Table 4, we can statistically confirm that our GLA module performs better than other attention mechanisms.

4.3.4 Fusion methods comparison

To study the effectiveness of the information obtained from multiple modalities via different fusion strategies, we conduct experiment by alternatively changing the Fusion Module with multiple Fusion operators while keeping other components of the system unchanged. Specifically, the Element-wise Addition (*Fusion Add*), Element-wise Maximum (*Fusion Max*)) and our *Fusion Net* are studied in our experiment. Furthermore, we also compare our method with recent work by Dubey et al. [17]. Table 6 summarizes the results from our experiment. As shown in this table, the performance of our network using *Fusion Net* is superior to other fusion strategies. However, we notice that the results from other fusion techniques are also very competitive. This shows that the fusion strategy is also an important module in the emotion recognition task, however the final result is also affected by the extracted features from the feature extraction and attention modules.

Table 5 p value of the Stuart-Maxwell test for each pair of methods that are used in the setting (iv) of Table 4

Methods	w/GLA	w/CA	w/o Attention
w/GLA	3.0×10^{-2}	1.69×10^{-14}	1.38×10^{-53}
w/CA		1.33×10^{-10}	3.33×10^{-14}
w/o Attention			2.81×10^{-38}

Each element on the main diagonal is the test result of the agreement between the model prediction and the observed data (ground-truth label)

Table 6 Results of different fusion strategies on the NCAER-S dataset (bold denotes the best result)

Methods	Accuracy (%)
Dubey et al. [17]	44.33
GLAMOR-Net + *Fusion Add*	45.62
GLAMOR-Net + *Fusion Max*	46.26
GLAMOR-Net + *Fusion Net*	**46.91**

4.3.5 Backbone architectures

We further study the effect of different Encoding network architectures. Specifically, the MobileNetV2 [48] and ResNet-18 [25] are adopted as the backbone network to extract features for both face and context branches in our study. We use the output of the last convolutional layer as the represented feature maps. These feature maps are then fed into the GLA module and processed as in Sect. 3. We summarize the total amount of network parameters and the classification results on CAER-S and NCAER-S in Table 7. We observe that the ResNet-18 significantly outperforms other shallower architectures (Original and MobileNetV2) and yields the best performance with 89.88% and 48.40% accuracy on CAER-S and NCAER-S. However, using such complex model resulted in more memory footprint as well as computational cost. Additionally, the MobileNetV2 can balance the trade-off between accuracy and the speed of the model, which is a considerable option for deploying in environments with limited resources such as mobile devices.

4.3.6 Visualization

Figure 9 shows the qualitative visualization with learned attention maps obtained by our method GLAMOR-Net in comparison with CAER-Net-S. It can be seen that our Global-Local attention mechanism produces better saliency maps and helps the model attend to the right discriminative regions in the surrounding background than the attention map produced by CAER-Net-S [32]. As we can see, our model is able to focus on the gesture of the person (Fig. 9f) and also the face of surrounding people (Fig. 9c, d) to infer the emotion accurately.

Figure 10 shows some emotion recognition results of different approaches on the NCAER-S dataset. More specifically, the first two rows (i) and (ii) contain predictions of the CAER-Net-S while the last two rows (iii) and (iv) show the results of our GLAMOR-Net. In some cases, our model was able to exploit the context effectively to perform inference accurately. For instance, with the same

Table 7 Accuracy of different encoding network architectures (bold denotes the best result)

Method	Backbone	#Params	CAER-S	NCAER-S
GLAMOR-Net	Original	2.23M	77.90	46.91
GLAMOR-Net	MobileNetV2 [48]	5.83M	85.44	47.52
GLAMOR-Net	ResNet-18 [25]	22.90M	**89.88**	**48.40**

(a) Anger   (b) Disgust   (c) Fear   (d) Happy   (e) Sad   (f) Surprise

Fig. 9 Visualization of the attention maps. From top to bottom: original image in the NCAER-S dataset, image with masked face, attention map of the CAER-Net-S, and attention map of our GLAMOR-Net

sad image input (shown on the (*i*) and (*iii*) rows), the CAER-Net-S misclassified it as neutral while the GLAMOR-Net correctly recognized the true emotion category. It might be because our model was able to identify that the man was hugging and appeasing the woman and inferred that they were sad. Another example is shown on the (*i*) and (*iii*) rows of the fear column. Our model classified the input accurately, while the CAER-Net-S might be confused between the facial expression and the wedding surrounding, thus incorrectly predicted the emotion as happy.

On the other hand, we can also see on the (*iv*) rows of Fig. 10, the GLAMOR-Net misclassified the disgust and the surprise images as happy and the neutral

image as sad. The reason might be that these images look quite confusing even to humans. Our model also failed to recognize emotions in the anger, fear, happy and sad images on the (*iv*) rows and predicted them as neutral instead. It can be because the facial expression in these images does not manifest clearly enough, which makes it difficult to distinguish between the neutral class and these emotion categories. This uncertainty was previously shown in the confusion matrix in Fig. 8.

4.3.7 Emotion recognition in the wild

As both the CAER-S dataset and its new split NCAER-S dataset contain only images from movie settings, they have

Fig. 10 Predictions on the NCAER-S test set. The first two rows (*i*) and (*ii*) show the results of the CAER-Net-S while the last two rows (*iii*) and (*iv*) demonstrate predictions of our GLAMOR-Net. The columns' names from (a) to (g) denote the ground-truth emotion of the images

Fig. 11 Human emotion detection results in the wild setting

a very limited number of people in a constrained environment. Therefore, the model trained using these datasets potentially do not work well on real-world image setting. Despite this challenge, Fig. 11 shows that our GLAMOR-Net can successfully detect and recognize human emotion in these challenging settings. Note that, the input images in this setup do not share any overlap with the movie settings as in the training set. This again confirms the generalization ability of our proposed method.

5 Conclusions and future work

In this work, we presented a novel method to exploit context information more efficiently by using the proposed global-local attention model. We have shown that our approach can considerably improve the emotion classification accuracy compared to the current state-of-the-art result in the context-aware emotion recognition task. The results on the CAER-S and the NCAER-S dataset consistently demonstrate the effectiveness and robustness of our method.

Our approach currently only takes static images as input, which limits the amount of knowledge that can be exploited. We are planning to utilize temporal information in dynamic videos and other modalities such as audio in order to further improve the performance. We also consider releasing a more challenging emotion recognition dataset that contains rich background contexts with multiple faces in the same frame and take advantage of our attention model to extract the context saliency map for each face in a more effective manner. We hope that our work will pave the way for future work in which predicting the emotions of different people simultaneously is tackled.

Availability of data and material The NCAER-S dataset can be downloaded at https://bit.ly/NCAERS_dataset.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Code availability The source code and trained model of our network are available at https://github.com/minhnhatvt/glamor-net.

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X (2016) Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on operating systems design and implementation, OSDI'16, p. 265–283

2. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y (eds.) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings. http://arxiv.org/abs/1409.0473

3. Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM international conference on image and video retrieval, CIVR '07. Association for computing machinery, New York, NY, USA, p. 401–408. https://doi.org/10.1145/1282280.1282340

4. Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter C, Beale R (eds) Affect and emotion in human-computer interaction, from theory to applications, lecture notes in computer science. Springer, New York, pp 92–103

5. Chen J, Chen Z, Chi Z, Fu H et al (2014) Facial expression recognition based on facial components detection and hog features. In: International workshops on electrical and computer engineering subfields, pp 884–888

6. Chen Y, Wang J, Chen S, Shi Z, Cai J (2019) Facial motion prior networks for facial expression recognition. In: 2019 IEEE visual communications and image processing, VCIP 2019, Sydney, Australia, December 1–4, 2019 IEEE, pp. 1–4. https://doi.org/10.1109/VCIP47243.2019.8965826

7. Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds.) Advances in neural information processing systems 28: annual conference on neural information processing systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp. 577–585. http://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition

8. Clark EA, Kessinger J, Duncan SE, Bell MA, Lahne J, Gallagher DL, O'Keefe SF (2020) The facial action coding system for characterization of human affective response to consumer product-based stimuli: asystematic review. Front Psychol 11:920

9. Clavel C, Vasilescu I, Devillers L, Richard G, Ehrette T (2008) Fear-type emotion recognition for future audio-based surveillance systems. Speech Commun 50:487–503. https://doi.org/10.1016/j.specom.2008.03.012

10. Corneanu CA, Simón MO, Cohn JF, Guerrero SE (2016) Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. IEEE Trans Pattern Anal Mach Intell 38(8):1548–1568

11. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor J (2001) Emotion recognition in human-computer interaction. Signal Process Mag IEEE 18:32–80. https://doi.org/10.1109/79.911197

12. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition

(CVPR'05) - Volume 1 - Volume 01, CVPR '05. IEEE Computer Society, USA, p. 886-893. https://doi.org/10.1109/CVPR.2005.177

13. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE MultiMedia 19(3):34–41. https://doi.org/10.1109/MMUL.2012.26

14. Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10(7):1895–1923. https://doi.org/10.1162/089976698300017197

15. Do T, Nguyen BX, Tjiputra E, Tran M, Tran QD, Nguyen A (2021) Multiple meta-model quantifying for medical visual question answering. arXiv preprint arXiv:2105.08913

16. Do TT, Nguyen A, Reid I (2018) Affordancenet: an end-to-end deep learning approach for object affordance detection. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE. pp. 5882–5889

17. Dubey SR, Roy SK, Chakraborty S, Mukherjee S, Chaudhuri BB (2020) Local bit-plane decoded convolutional neural network features for biomedical image retrieval. Neural Comput Appl 32(11):7539–7551

18. Ekman P, Friesen W (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124–129

19. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 44(3):572–587

20. Evgeniou T, Pontil M (2001) Support vector machines: theory and applications. Machine learning and its applications. Springer, Berlin Heidelberg, pp. 249–257

21. Galassi A, Lippi M, Torroni P (2020) Attention in natural language processing. IEEE Trans Neural Netw Learn Syst

22. Gao Q, Zeng H, Li G, Tong T (2021) Graph reasoning-based emotion recognition network. IEEE Access 9:6488–6497. https://doi.org/10.1109/ACCESS.2020.3048693

23. Georgescu M, Ionescu RT, Popescu M (2019) Local learning with deep and handcrafted features for facial expression recognition. IEEE Access 7:64827–64836. https://doi.org/10.1109/ACCESS.2019.2917266

24. Han K, Yu D, Tashev I (2014) Speech emotion recognition using deep neural network and extreme learning machine. In: Interspeech 2014

25. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

26. Hyun KH, Kim EH, Kwak YK (2007) Emotion recognition using voice based on emotion-sensitive frequency ranges. Springer, Berlin, Heidelberg, pp 217–223

27. Jordan MI (2004) Graphical models. Stat Sci 19(1):140–155

28. King D (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10:1755–1758

29. Kosti R, Alvarez JM, Recasens A, Lapedriza A (2019) Context based emotion recognition using emotic dataset. IEEE Trans Pattern Anal Mach Intell 42(11):2755–2766

30. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems, vol 25. Curran Associates Inc, Red Hook, pp 1097–1105

31. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems - Volume 1, NIPS'12. Curran Associates Inc., Red Hook, NY, USA, pp. 1097–1105

32. Lee J, Kim S, Kim S, Park J, Sohn K (2019) Context-aware emotion recognition networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10142–10151. https://doi.org/10.1109/ICCV.2019.01024

33. Lee JB, Rossi RA, Kim S, Ahmed NK, Koh E (2019) Attention models in graphs: a survey. ACM Trans Knowl Discov Data 13(6):1–25

34. Li S, Deng W (2020) Deep facial expression recognition: asurvey. IEEE transactions on affective computing p. 1–1. http://dx.doi.org/10.1109/TAFFC.2020.2981446

35. Liu X, Kumar BVKV, You J, Jia P (2017) Adaptive deep metric learning for identity-aware facial expression recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 522–531

36. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition -workshops, pp. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

37. Matsumoto D (1992) More evidence for the universality of a contempt expression. Motiv Emot 16:363–368

38. Meng D, Peng X, Wang K, Qiao Y (2019) Frame attention networks for facial expression recognition in videos. In: 2019 IEEE international conference on image processing (ICIP), pp. 3866–3870. https://doi.org/10.1109/ICIP.2019.8803603

39. Mittal T, Guhan P, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emoticon: context-aware multimodal emotion recognition using frege's principle. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14222–14231. https://doi.org/10.1109/CVPR42600.2020.01424

40. Mollahosseini A, Hasani B, Mahoor MH (2019) Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 10(1):18–31. https://doi.org/10.1109/TAFFC.2017.2740923

41. Nguyen A, Do TT, Reid I, Caldwell DG, Tsagarakis NG (2019) V2cnet: a deep learning framework to translate videos to commands for robotic manipulation. arXiv preprint arXiv:1903.10869

42. Nguyen A, Nguyen N, Tran K, Tjiputra E, Tran QD (2020) Autonomous navigation in complex environments with deep multimodal fusion network. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE. pp. 5824–5830

43. Nguyen BX, Nguyen BD, Do T, Tjiputra E, Tran QD, Nguyen A (2020) Graph-based person signature for person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshop, pp. 3492–3501

44. Oliva A, Torralba A (2006) Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 155:23–36

45. Päivärinta J, Rahtu E, Heikkilä J (2011) Volume local phase quantization for blur-insensitive dynamic texture classification. In: Heyden A, Kahl F (eds) Image analysis. Springer, Berlin, Heidelberg, pp 360–369

46. Paulmann S, Bleichner M, Kotz SA (2013) Valence, arousal, and task effects in emotional prosody processing. Front Psychol 4:345

47. Randhavane T, Bhattacharya U, Kapsaskis K, Gray K, Bera A, Manocha D (2020) Identifying emotions from walking using affective and deep features. arXiv preprint arXiv:1906.11884

48. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520

49. Sariyanidi E, Gunes H, Cavallaro A (2015) Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 37(6):1113–1133

50. Schindler K, Van Gool L, de Gelder B (2008) Recognizing emotions expressed by body pose: a biologically inspired neural model. Neural Netw 21(9):1238–1246

51. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816

52. Sikka K, Dykstra K, Sathyanarayana S, Littlewort G (2013) Multiple kernel learning for emotion recognition in the wild. In: ICMI 2013 - Proceedings of the 2013 ACM international conference on multimodal interaction. https://doi.org/10.1145/2522848.2531741

53. Sikka K, Wu T, Susskind J, Bartlett M (2012) Exploring bag of words architectures in the facial expression domain. In: Proceedings of the 12th international conference on computer vision - Volume 2, ECCV'12. Springer-Verlag, Berlin, Heidelberg, p. 250–259https://doi.org/10.1007/978-3-642-33868-7_25

54. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds.) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. http://arxiv.org/abs/1409.1556

55. Stathopoulou IO, Tsihrintzis GA (2011) Emotion recognition from body movements and gestures. In: Tsihrintzis GA, Virvou M, Jain LC, Howlett RJ (eds) Intelligent interactive multimedia systems and services. Springer, Berlin, Heidelberg, pp 295–303

56. Sun B, Li L, Zhou G, Wu X, He J, Yu L, Li D, Wei Q (2015) Combining multimodal features within a fusion network for emotion recognition in the wild. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, ICMI '15. Association for Computing Machinery, New York, NY, USA, p. 497–502 https://doi.org/10.1145/2818346.2830586

57. Wang F, Tax DMJ (2016) Survey on the attention based RNN model and its applications in computer vision. CoRR. http://arxiv.org/abs/1601.06823

58. Wang K, Peng X, Yang J, Meng D, Qiao Y (2019) Region attention networks for pose and occlusion robust facial expression recognition. CoRR. http://arxiv.org/abs/1905.04075

59. Zhao Z, Liu Q, Wang S (2021) Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Trans Image Process 30:6544–6556

60. Zhao Z, Liu Q, Zhou F (2021) Robust lightweight facial expression recognition network with label distribution training. In: Proceedings of the AAAI conference on artificial intelligence, 35:3510–3519

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.