

# Facial Expression Recognition with Face Mask using Attention Mechanism<sup>\*</sup>

Shunsuke HAYASHI<sup>1</sup>, Yuga ONO<sup>1</sup>, Ryuto ISHIBASHI<sup>1</sup>, Qi LI<sup>1</sup> and Lin MENG<sup>2,†</sup>

<sup>1</sup>College of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan 525-8577

<sup>2</sup>Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan 525-8577

## Abstract

Facial expression recognition (FER) has been studied widely up to the present because FER is highly applicable. However, the COVID-19 pandemic forces people to wear face masks daily. A face mask removes information necessary for FER from a face and contains unnecessary information such as color and material. This paper aims to prevent the influence of meaningless regions for recognition, such as masks and backgrounds, due to the use of the attention mechanism. To obtain the optimal attention map, this paper has prepared two-stage networks. In stage 1, the network learns a simple task like masked/unmasked classification to generate an attention map. In stage 2, we train the network and use an attention map to guide the network to focus on the meaningful region to FER. As a result, the accuracy of our proposal achieved 86.8% and is 2.5% higher than VGG. The experimental results have proved the effectiveness of the proposed method.

## 1. Introduction

Facial expression recognition (FER) is one of the major recognition themes because this is expected to be widely used in many fields, such as healthcare and smartphone applications. Various deep-learning models have been proposed to achieve these goals until now[1]. Since 2019, COVID-19 has spread worldwide and forced people to wear face masks outside[2]. The percentage of people wearing a mask in Japan is high in 2023. The demand for FER on masked faces has increased. Wearing a mask covers most of the face with meaningful information for FER, such as the mouth and nose. Additionally, masks have redundant features like color and material. These have a negative impact on facial expression recognition. The attention mechanism is one method that solves this problem. The attention mechanism dynamically decides where to focus attention on input data and guides the classification network to focus on the area around the eyes, and these issues are alleviated.

---

*The 5th International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2023), August 28-29, 2023, Kusatsu, Japan*


<sup>\*</sup>These authors contributed equally.

<sup>†</sup>Corresponding author.

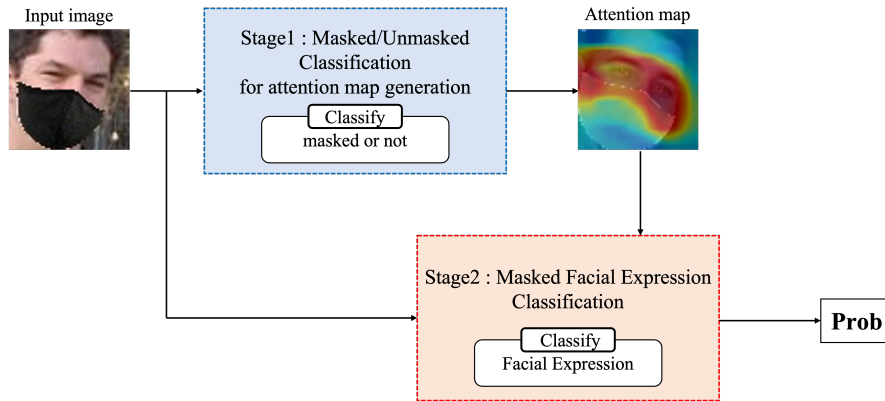
✉ ri0104ss@ed.ritsumei.ac.jp (S. HAYASHI); ri0098ss@ed.ritsumei.ac.jp (Y. ONO); ri0097fx@ed.ritsumei.ac.jp (R. ISHIBASHI); gr0517rs@ed.ritsumei.ac.jp (Q. LI); menglin@fc.ritsumei.ac.jp (L. MENG)

🌐 <http://http://www.ihpc.se.ritsumei.ac.jp/> (L. MENG)

🆔 0000-0003-4351-6923 (L. MENG)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Our FER system is divided into two stages. In stage 1, the network is trained to classify input images into masked/unmasked binary classes and to generate an attention map. In stage 2, the attention map guides the network to pay more attention to the region that is essential to Masked FER. The network predicts which facial expression class the input image is classified into.

To resolve this problem, this paper proposes the FER system based on CNN that introduces an attention mechanism divided into two stages. In stage 1, the network classifies whether the input image contains a mask and generates an attention map from this learning. This attention map has information about where to focus attention in the spatial direction of the input image or feature map. Moreover, this stage also aims to stabilize the generation of the attention map by learning a simple classification. In stage 2, the network performs classification by synthesizing the feature map and the attention map obtained in the previous stage. Due to synthesizing them, the region necessary for recognition is given attention, and the rest is ignored. In short, the major contribution of the paper has two aspects:

- By generating and using the attention mechanism, our proposal prevents influence from extraneous information such as masks and backgrounds. Our proposal also has improved accuracy compared to other CNNs.
- Our method has dynamically achieved stable attention map generation without human power like annotation and image processing such as landmark detection.

The remaining parts of this paper are organized as follows. Section 2 introduces the related works. In Section 3, we propose our FER system. Section 4 shows the experimental method and dataset. Section 5 reports experimental results. Finally, this paper is concluded in Section 6.

## 2. Related work

### 2.1. Deep Learning

Deep learning is one of the popular techniques of machine learning[3, 4, 5]. The reason deep learning has been able to evolve so far is due to advances in hardware and the ease of acquiring big data. In particular, CNN uses convolution layers and pooling layers to imitate the human visual mechanism called the local receptive field. And CNN achieves high recognition accuracy.

Lightweight models [6][7] and improvement models[8][9] have been proposed. These have been applied in many fields up to now. These have been widely used as a base model for facial expression recognition.

## **2.2. Facial Expression Recognition (FER)**

FER is approximately divided into a static image-based FER network and dynamic-based FER networks[1]. Especially, static image-based studies tend to be popular due to the ease of statistical analysis and the availability of relevant training and test data. Static image-based FER networks are categorized as ensembles[10], vision transformer-based networks[11], attention mechanisms, and so on.

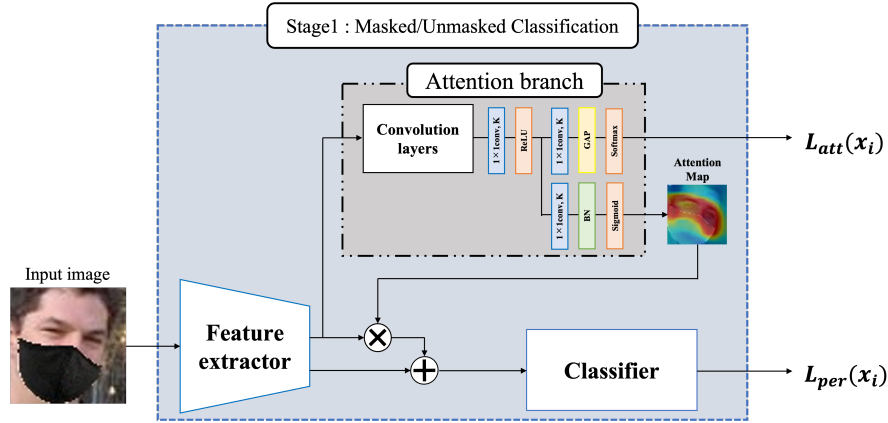
## **2.3. Attention Mechanism**

One of the major problems in facial expression recognition is the occlusion problem. This problem is that the face is partially hidden by obstacles such as glasses or hands. Wearing a face mask falls into the category of this issue. To solve this, various attention mechanisms have been proposed. Attention mechanisms' approaches to occlusion problems are broadly divided into two categories, such as patch-base-CNN[12][13][14] and attention-map-base-CNN. Patch-base-CNN divides the face image into patch images based on facial organ points and independently extracts feature vectors. After that, classification is performed by concatenating each feature vector and inputting it into the fully connected layers. By weighting patches necessary for facial expression recognition, the network pays attention to specific regions. Most of these studies use landmarks as reference points for each facial organ. They are detected even if facial images are partially hidden by obstructions. However, face masks cover most of the facial image, and the detection rate of landmarks is not stable. Attention-map-base-CNN generates an attention map that expresses where to pay attention in the spatial direction during training[15][16][17]. These networks combine feature maps with attention maps into weighted feature maps. Their networks focus on the valid area for FER by using attention maps. FERatt[18] incorporates the segmentation branch to generate an attention map. FERatt generates a mask image of the face from the input image and uses it to reconstruct the image to remove the wasted region. However, this method requires the mask image to be annotated by hand. The problem with this approach is that it takes a lot of time and manpower. Attention Branch Network(ABN)[16] is an end-to-end network that consistently performs feature extraction, attention map generation, and classification. The method of generating attention maps is based on CAM[19]. ABN visualizes the basis for classification decisions as well as achieves improved recognition accuracy in various tasks by introducing an attention mechanism.

## **3. Proposed Method**

As shown in Figure 1, our FER system consists of two-stages such as Masked/Unmasked Classification and Facial Expression Classification.

### 3.1. Stage 1: Masked/Unmasked Classification Network (MCN)



**Figure 2:** Masked/Unmasked Classification Network (MCN)

In stage 1, the Masked/Unmasked Classification Network (MCN) is intended to dynamically generate the appropriate attention map from simple classification such as masked/unmasked binary classification. MCN, designed based on ABN as shown in Figure 2, incorporates the traditional recognition network structure and the Attention branch. The attention branch compresses the feature map into a 1 channel by inputting the  $1 \times 1$  convolution layer and outputs it as attention maps. To ensure that the attention branch learns optimally, This branch also makes a class prediction and calculates the loss. Accordingly, the overall loss function of this network is determined by Eq1.

$$L(x_i) = L_{att}(x_i) + L_{per}(x_i) \quad (1)$$

The output attention map is normalized from 0 to 1 by the sigmoid function before the final output. The attention map output from the attention branch has different features depending on the prediction class of the attention branch because of the different points of the GPA and  $1 \times 1$  convolution layer response. If the network containing the Attention branch directly learns to classify facial expressions, the appearance of the output attention map varies in accordance with expression classes. To avoid this, MCN classifies whether the input image is masked or not. When masked facial images are inputs, return a common attention map regardless of facial expression. Moreover, expression recognition is a difficult task because it requires looking at every detail of the face. It makes the generation of attention maps tricky. Thus, this network replaces this with a simple classification like the masked/unmasked binary classification to stabilize. When learning of this classification is complete, unnecessary nodes to output the attention map are disconnected from the network which fixes the weight parameters of this network. Therefore, MCN becomes a generator that returns an appropriate attention map for the input images as shown in Figure 2 (b).

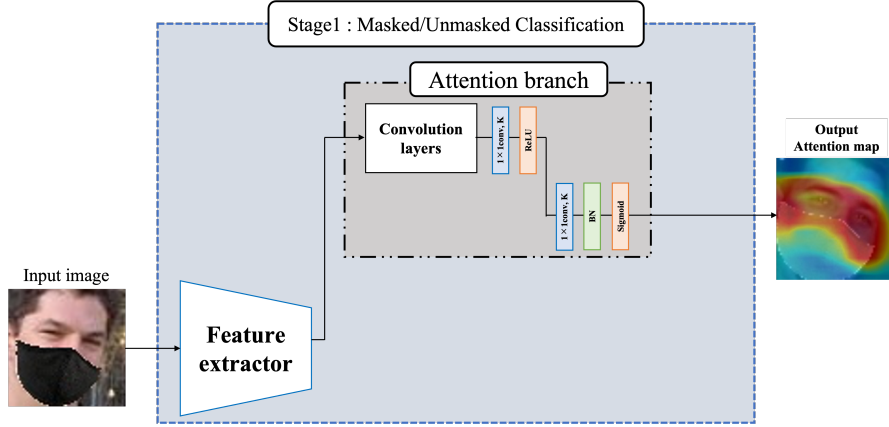


Figure 3: Delete unnecessary nodes after learning

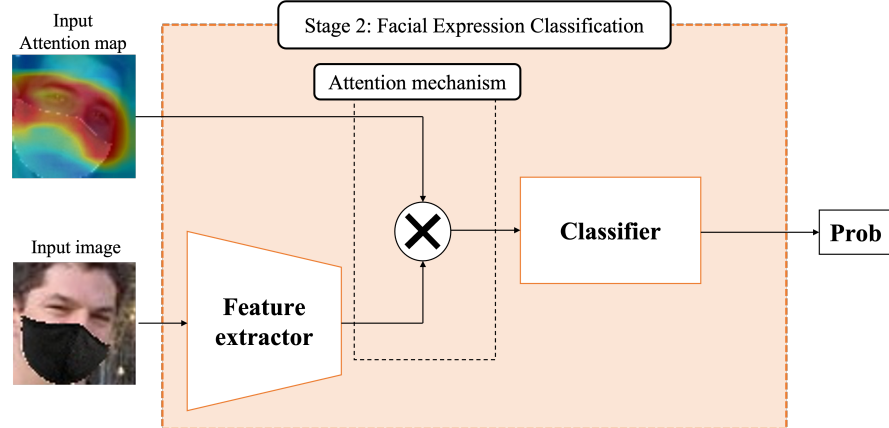


Figure 4: Facial Expression Classification Network (FECN)

### 3.2. Stage 2: Facial Expression Classification Network (FECN)

In Stage 2, Facial Expression Classification Network (FECN) uses the attention map to guide the network to pay more attention to the region that is essential to FER when this network classifies facial expressions. FECN combines feature maps generated from input images with attention maps by hadamard product and outputs the final probability of each class as shown in Figure 4. The attention map obtained from MFC is  $M(x_i)$ , and a feature map is  $F_c(x_i)$ . Weighted feature map is defined  $F'_c(x_i)$ , i.e., Eq.2.

$$F'_c(x_i) = M(x_i) \odot F_c(x_i) \tag{2}$$

Due to using this weighted feature map, this network is trained while ignoring the unnecessary regions, including face mask and background.

## 4. EXPERIMENTS

### 4.1. Dataset

Large amounts of data are required to train deep learning adequately. Unfortunately, it is difficult to collect facial expression data, including masks. Moreover, these data are not widely available. Therefore, in this paper, we create a new dataset by pasting a mask image on the facial expression dataset called RAF-DB[20]. RAF-DB is a large-scale facial expression database with around 30K great-diverse facial images downloaded from the Internet and includes subjects of various ethnicities, ages, and genders. By 40 experts, these data were categorized into 7 classes such as angry, fearful, disgusted, happy, sad, surprised, and normal. We paste the face mask image into this dataset following the method[21]. It detects facial contour points (landmarks) from a face image and synthesizes a mask image according to the coordinates. In reality, people wear face masks of various textures and colors, so in this experiment, white non-woven cloth, black cloth, and gray cloth are equally included in the dataset as shown in Figure 5. We removed data that cannot be visually classified as facial expressions and failed to

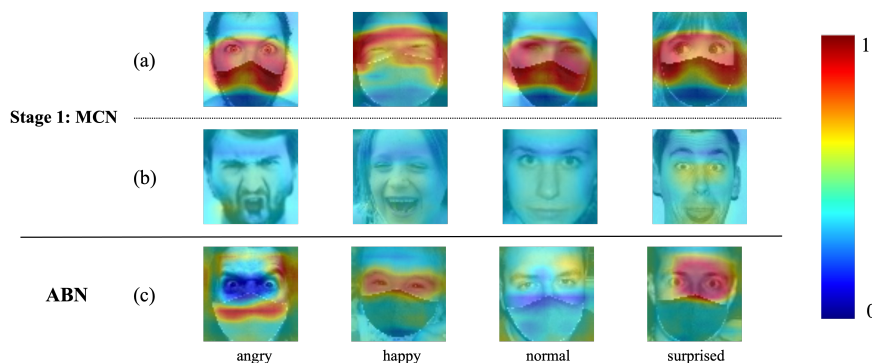


**Figure 5:** Example of mask pattern

paste a face mask. This experiment has prepared two datasets as shown in Table 1. The first dataset is the masked/unmasked dataset (MU-dataset), including unprocessed face images and mask-wearing images to train the masked/unmasked binary classifier. MU-dataset has about 500 images for training and 100 images for testing and includes multiple emotions. The second dataset is the Masked FER dataset (M-FER-dataset) containing expression classes. However, the mouth has meaningful information in helping to distinguish between the emotions of fear and surprise[22]. By wearing a face mask, it is difficult for both humans and computers to classify these expressions. Therefore, this FER focuses on four target classes such as angry, surprised, normal, and happy. M-FER-dataset has about 4000 images for training and 400 images for testing. These datasets are performed parallel shifts and random horizontal flips as data augmentation.

**Table 1**  
Datasets and classes

Dataset	Classes	Train	Test
<b>MU-dataset</b>	Masked face /Unmasked face	500	100
<b>M-FER-dataset</b>	Angry/happy/normal/surprised	4000	400



**Figure 6:** Visualizing attention maps for each class. (a) are the attention maps that MCN outputs when being input masked face images and (b) are the attention maps in case input is unmasked facial face. (c) are the attention map outputs from ABN

## 4.2. Experimental Method

This paper compares the accuracy of our proposal with the VGGNet and simple ABN to elucidate our superiority. Table 2 indicates the dataset used by each network as well as whether each network outputs the attention map. ABN and VGGNet are trained on M-FER-dataset to classify facial expressions. In our proposal, the MCN is firstly trained on MU-dataset to generate the attention map, and FECN is nextly trained on M-FER-dataset. This experiment used a GPU device called GeForce RTX 3080 and Pytorch as the Python framework.

**Table 2**  
Dataset used to train, Whether each network outputs an attention map

	Network	Dataset	Attention Map
	<b>VGG</b>	M-FER-dataset	-
	<b>ABN</b>	M-FER-dataset	output
<b>Ours</b>	<b>Stage1:MCN</b>	MU-dataset	output
	<b>Stage2:FECN</b>	M-FER-dataset	-

## 4.3. Experimental results of Attention Map Output

Figure 6 visualizes the Attention Map. Both (a) and (b) are the attention map output from MCN, which is trained on MU-dataset for masked/unmasked binary classification, while Figure 6 (c) is the output from ABN, which is trained on M-FER-dataset for facial expression classification. The color of the attention map refers to the strength of attention, and red means the highest.

Figure 6 (a) is the attention map for each expression when masked facial images are input into MCN, and Figure 6 (b) is the attention map in case unmasked facial images are input. As shown in Figure 6 (a), MCN outputs a common attention map, which pays attention to the area around the regardless of expression class when a mask image is an input. On the other hand, Figure 6 (b) means no attention is paid anywhere. Figure 6 (c) shows the attention map that ABN outputs represent different appearances for each facial class. Attention maps of happiness and surprise turn red around the eyes, but the others are irregular.

#### 4.4. Experimental results of Facial Expression Classification

**Table 3**

Evaluation of Facial Expression Classifier

Network	Angry	Happy	Normal	Surprised	Total(%)
VGG11	<b>71.0</b>	94.0	84.0	89.0	84.4
ABN	63.0	93.0	89.0	<b>92.0</b>	84.2
FECN(ours)	69.0	<b>98.0</b>	<b>90.0</b>	90.0	<b>86.7</b>

Table 3 shows that FECN has the highest total accuracy of all the models as well as FECN achieves higher accuracy in happiness, anger, and surprise compared to VGG11. ABN has a significantly lower accuracy class like angry. Thus, ABN had lower total recognition accuracy than VGG11. Moreover, The anger class is less accurate in recognition for all facial expression classifiers as shown in Figure 7.

## 5. Discussion and Future Work

### 5.1. Discussion

This experiment prepares ABN which directly is trained in FER to compare with our proposal, which is divided into two-stage. When a masked face image is input to ABN, the output is the unique attention map that depends on the input facial expression, as shown in Figure 6 (c). In addition, difficult classifications such as FER make it hard to generate attention maps, and Figure 6 (c) shows that the angry class is one example. When these unstable attention maps are used as the attention mechanism, Table 3 indicates that recognition performance decreases on the contrary. To avoid this situation, we replaced the learning to generate attention maps with a simple classification. As a result, an attention map was generated focused on the area around the eyes for all classes, as shown in Figure 6 (a). These attention maps make FECN pay attention to the region that is meaningful to FER, such as the area around the eyes, and remove extraneous information, including background and mask patterns. Table 3 shows our proposal has achieved improved recognition accuracy on Masked FER.

### 5.2. Future Work

FER requires real-time performance. The network must also be lightweight so that it runs on small devices. In this paper, our FER system is divided into two stages and trains different



True Label	angry	0.69	0.2	0.08	0.03
	happy	0	0.98	0.01	0.01
	normal	0.01	0.05	0.9	0.04
	surprised	0	0.02	0.081	0.9
		angry	happy	normal	surprised

(a) Confusion Matrix of VGG

True Label	angry	0.63	0.17	0.13	0.07
	happy	0.01	0.93	0.05	0.01
	normal	0.01	0.06	0.89	0.04
	surprised	0.01	0.03	0.04	0.92
		angry	happy	normal	surprised

(b) Confusion Matrix of ABN

True Label	angry	0.71	0.16	0.1	0.03
	happy	0.01	0.94	0.04	0.01
	normal	0.04	0.07	0.84	0.05
	surprised	0.03	0.01	0.071	0.89
		angry	happy	normal	surprised

(c) Confusion Matrix of Ours

**Figure 7:** Confusion matrix of each facial expression classifier

networks at each stage. Thus, the number of parameters for the entire network considerably has increased. Both MCN and FECN have their own feature extractors. Therefore, we aim to solve this problem by having these feature extractors share weights.

## 6. Conclusion

As COVID-19 is spreading worldwide, people are forced to wear masks when going out. Therefore, the demand for FER while wearing a face mask has increased. However, masks hide most of the face as well as contain extraneous information such as the pattern and color of the mask. To prevent these negative effects, our FER system has incorporated an Attention mechanism and consists of two stages such as classification and facial expression classification. By dividing our FER system into two stages, our proposal has succeeded in generating an attention map that is closer to the ideal for this task than other attention models. As a result, our proposal

achieved 86.7% recognition accuracy. In the future, the challenge is to reduce the size of the network to decrease the number of parameters and amount of calculations.

## References

- [1] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, O. Krejcar, Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey, *IEEE Transactions on Instrumentation and Measurement* 72 (2023) 1–31.
- [2] M. Wen, K. Yokoo, X. Yue, L. Meng, An ai-based mask-wearing status recognition and person identification system, in: *2021 International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2021)*, August. 2021.(In JAPAN), 2021.
- [3] X. Yue, H. Li, Y. Fujikawa, L. Meng, Dynamic Dataset Augmentation for Deep Learning-Based Oracle Bone Inscriptions Recognition, *J. Comput. Cult. Herit.* 15 (2022). URL: <https://doi.org/10.1145/3532868>. doi:10.1145/3532868.
- [4] X. Yue, H. Li, L. Meng, An ultralightweight object detection network for empty-dish recycling robots, *IEEE Transactions on Instrumentation and Measurement* 72 (2023) 1–12. doi:10.1109/TIM.2023.3241078.
- [5] Y. Fujikawa, H. Li, X. Yue, C. V. Aravinda, G. A. Prabhu, L. Meng, Recognition of Oracle Bone Inscriptions by using Two Deep Learning Models, *International Journal of Digital Humanities* (2022).
- [6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861* (2017).
- [7] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 2584–2593.
- [8] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
- [9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [10] K. Mohan, A. Seal, O. Krejcar, A. Yazidi, Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks, *IEEE Transactions on Instrumentation and Measurement* 70 (2021) 1–12. doi:10.1109/TIM.2020.3031835.
- [11] F. Ma, B. Sun, S. Li, Facial expression recognition with visual transformers and attentional selective fusion, *IEEE Transactions on Affective Computing* (2021) 1–1. URL: <https://doi.org/10.1109/TAFFC.2021.3122146>. doi:10.1109/taffc.2021.3122146.
- [12] G. Chen, J. Peng, W. Zhang, K. Huang, F. Cheng, H. Yuan, Y. Huang, A region group adaptive attention model for subtle expression recognition, *IEEE Transactions on Affective Computing* (2021) 1–1. doi:10.1109/TAFFC.2021.3133429.
- [13] Y. Li, J. Zeng, S. Shan, X. Chen, Patch-gated cnn for occlusion-aware facial expression recognition, in: *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2209–2214. doi:10.1109/ICPR.2018.8545853.

- [14] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using cnn with attention mechanism, *IEEE Transactions on Image Processing* 28 (2019) 2439–2450. doi:10.1109/TIP.2018.2886767.
- [15] S. Minaee, M. Minaei, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, *Sensors* 21 (2021) 3046.
- [16] H. Fukui, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Attention branch network: Learning of attention mechanism for visual explanation, 2019. arXiv:1812.10025.
- [17] Y. Zhang, C. Wang, X. Ling, W. Deng, Learn from all: Erasing attention consistency for noisy label facial expression recognition, 2022. arXiv:2207.10299.
- [18] P. D. M. Fernandez, F. A. G. Peña, T. I. Ren, A. Cunha, Feratt: Facial expression recognition with attention net, 2019. arXiv:1902.03284.
- [19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, 2015. arXiv:1512.04150.
- [20] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2584–2593.
- [21] A. Anwar, A. Raychowdhury, Masked face recognition for secure authentication, 2020. arXiv:2008.11104.
- [22] M. W. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Y. Chiao, S. L. Franconeri, Eye movements during emotion recognition in faces, *Journal of Vision* 14 (2014) 14–14. URL: <https://doi.org/10.1167/14.13.14>. doi:10.1167/14.13.14. arXiv:[https://arvojournals.org/arvo/content\\$public/journal/jov/933685/i1534](https://arvojournals.org/arvo/content$public/journal/jov/933685/i1534) – 7362 – 14 – 13 – 14.pdf.