Spatial Attention Network with High Frequency Component for Facial Expression Recognition*

Seongmin Kim and Kanghyun Jo[0000−0001−8317−6092]

Dept. of Electrical, Electronic and Computer Engineering,
University of Ulsan, Ulsan, Korea
asdfhdsa1234@mail.ulsan.ac.kr
acejo@ulsan.ac.kr

Abstract. Indeed, object classification is one of the most advanced fields in computer vision today, and there are ongoing efforts to classify datasets used in real-world industries, beyond just public experimental data. Facial expression recognition is indeed one of the most prominent examples of such tasks, closely related to the Human-Computer Interaction (HCI) industry. Unfortunately, facial expression classification tasks are often more challenging compared to classifying public benchmark datasets. This paper aimed to address these challenges by mimicking human facial expression recognition processes and proposed an attention network that leverages high-frequency components to recognize expressions, inspired by how humans perceive emotions. The presented attention module vectorizes the singular value matrices of the query (the high-frequency component of the 1-channel input tensor) and the key (the 1-channel input tensor) and prepares a pairwise cross-correlation matrix by performing an outer product between them to create the attention scores. The correlation matrix is transformed into an attention score by passing through a convolution layer and sigmoid function. After that, it is used for element-wise multiplication with the value (input tensor) to perform attention. This paper conducted experiments using the ResNet18 and MobileNetV2 models along with the FER2013, JAFFE, and CK+ datasets to demonstrate the significant impact of the proposed attention module. The experimental results in this study have demonstrated the effectiveness of the proposed attention network and suggest its potential significance in real-time facial expression recognition tasks.

Keywords: Facial Expression Recognition · Spatial Attention Network · Spatial Frequency-domain Filtering

1 Introduction

Image recognition (or classification) task is the most basic and important research theme of computer vision. Efforts to employ artificial intelligence for object recognition in images commenced in the 20th century and, as of 2024, have

* This result was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

demonstrated superior accuracy compared to human performance [2, 3, 6, 8, 11, 23, 26]. In recent times, endeavors have been undertaken to classify data from diverse industries beyond publicly available experimental datasets, such as CIFAR-100 [12]. The representative example is facial expression recognition [4,5,7,17,22]. As the demand for Human-Computer Interaction (HCI) continues to rise, the comprehension and recognition of facial expressions have emerged as pivotal tasks for facilitating more natural interactions. However, upon examination of the graph depicted in Fig. 1 below, it becomes evident that the task of recognizing facial expressions using Convolutional Neural Network (CNN) is notably challenging when contrasted with public experimental images. On the contrary,

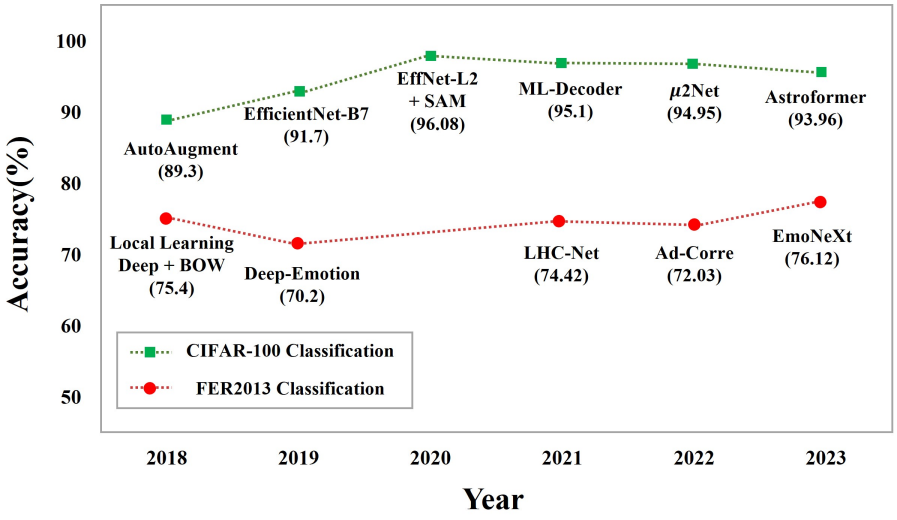

Fig. 1: The graphs of illustration depict the comparison of difficulty between CIFAR-100 classification and FER2013 classification.

humans can discern facial expressions significantly more effortlessly than neural networks. Because the human can recognize the tiny differences in facial expressions. This distinction signifies the variance between the orbicularis oculi muscle, orbicularis oris muscle [18], and alterations in the regions of the eyes, nose, and mouth. Hence, using just the difference of pixel values is insufficient to describe the distinction. According to this paper [20], humans rely meaningfully on the high-frequency components of facial images when classifying facial expressions. The high frequency component of the face, Fig. 2(b), contains detailed information that is effective in recognizing human facial expressions, such as eyes, nose, mouth, and wrinkles. This image processing result can support the hypothesis that humans use not only spatial domain information but also frequency domain information to classify facial expressions. This paper conducted an extensive exploration of various research approaches aimed at adapting the human facial

(a) Original Image (b) High Frequency Component (c) Low Frequency Component

Fig. 2: This illustration represents the spatial frequency domain filtering result. The face image is from CelebA dataset [13] and converted RGB to grayscale.

expression recognition process to CNN. Although there were many approaches to induce the model to utilize specific information intensively, the most effective one was the attention mechanism [4, 19, 22, 27]. The attention mechanism encourages the CNN to enhance the elements in its feature map tensor on the part highly relevant to the query suggested by the network designer. This study presents a simple attention network to emphasize the high-frequency components in the feature map, aiming to enhance the performance of CNN in facial expression recognition tasks. It also demonstrates the effectiveness of the proposed network by evaluating its contributions through FER2013 [9], JAFFE [15, 16], and CK+ [14] datasets.

2 Related Work

The most crucial aspect of an Attention network lies in comprehending the relationship between the query and the key. Various methodologies have been studied to grasp this relationship. They can be categorized into the following two groups.

- **Explicit Method:** The method for calculating the relationship through direct mathematical operations (e.g. Dot Product) after converting query and key data into a more manageable form such as Vector Embedding.
- **Implicit Method:** The method for calculating the relationship through an additional neural network after concatenating (or Stacking) query and key data.

Multi-head Attention [27], the most effective attention method in machine learning, is a representative example of the explicit method. Multi-head Attention [27] generates attention scores by performing a dot product operation between specific query vectors and all preprocessed key data, following the transformation of the vector-level embedded data into query, key, and value vectors. The explicit method can represent the reason for the relation between the query and keys.

Because it uses a mathematical approach to measure the correspondence. However, during vector-level embedding, the purity of the original data is decreased. The human-made mathematical operation, such as dot product, has a limit to represent the complex relationship. So explicit method has low representation capacity. The well-known implicit methods are Seq2Seq+Attention [1] and Bottleneck Attention Module (BAM) [19]. Seq2Seq+Attention [1] is designed for natural language processing. This method treats the decoder state of the previous step to query and the encoder hidden state of all steps to key. It measures the attention score with additional neural networks. BAM [19] also uses additional neural networks to calculate attention scores about the channel axis and spatial axis. This attention network is designed to perform channel attention and spatial attention within a deep convolutional neural network. Because of using nonlinear neural networks, the implicit method has a high representation capacity for relationships. However, the relationship-finding process is only dependent on many hidden layers of the neural network. So, the attention network cannot produce any evidence of a relationship. Therefore, the results of the implicit method cannot be fully trusted. This paper has designed an attention network by appropriately combining both explicit and implicit methods to selectively leverage the advantages of each.

3 Proposed Method

In this chapter, a detailed description of the proposed attention network designed to accentuate the high frequency components of the face within the feature map is provided. In this paper, the feature map inputted into the attention network before creating query and key was compressed into a single-channel through pointwise convolution. This is because, as can be observed in Fig. 3, the high-frequency components are quite similar across each channel, and thus, there is no need to handle a multi-channel tensor while increasing computational complexity. Therefore, before performing attention, this paper undergoes pre-

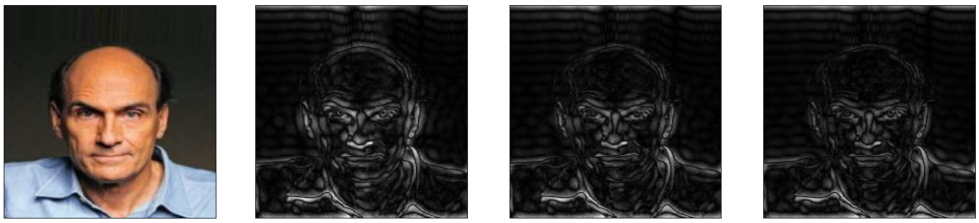

Fig. 3: The high pass filtering results of the 3-channel color image. HFC denotes the high frequency component.

processing of the input tensor ($\mathbf{x} \in \mathbb{R}^{c \times n \times n}$) as shown in Eq. 1, where it is

compressed through pointwise convolution.

$$\tilde{\mathbf{x}} = \mathcal{C}_{1\times1}(\mathbf{x}) \tag{1}$$

Where $\mathcal{C}_{1\times1}(\cdot)$ denotes pointwise convolution. For the query used in performing attention, the high frequency component ($\mathbf{h}_{\tilde{\mathbf{x}}}$) of a 1-channel tensor ($\tilde{\mathbf{x}} \in \mathbb{R}^{n\times n}$) was selected. To obtain $\mathbf{h}_{\tilde{\mathbf{x}}}$, $\tilde{\mathbf{x}}$ is first mapped into the frequency domain through Fourier transformation ($\mathfrak{F}$) as described in Eq. 2, followed by the execution of a high pass filtering process.

$$\tilde{\mathbf{x}}_h = \mathfrak{F}\{\tilde{\mathbf{x}}\}\mathcal{H}_{HPF} \tag{2}$$

Where $\mathcal{H}_{HPF}$ is high pass filter on frequency domain. Subsequently, by employing the inverse Fourier transformation ($\mathfrak{F}^{-1}$) as in Eq. 3, $\mathbf{h}_{\tilde{\mathbf{x}}}$ is obtained by remapping it back into the spatial domain.

$$\mathbf{h}_{\tilde{\mathbf{x}}} = \mathfrak{F}^{-1}\{\tilde{\mathbf{x}}_h\} \tag{3}$$

It was decided to use $\tilde{\mathbf{x}}$ as the key data. This paper attempted to obtain evidence for the relationship between query and key through mathematical operations, similar to explicit methods. In this case, the Kronecker product between two tensors(e.g. $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$) is computed as shown in Eq. 4.

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & \cdots & a_{11}b_{1q} & \cdots & a_{1n}b_{11} & a_{1n}b_{12} & \cdots & a_{1n}b_{1q} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{11}b_{p1} & a_{11}b_{p2} & \cdots & a_{11}b_{pq} & \cdots & a_{1n}b_{p1} & a_{1n}b_{p2} & \cdots & a_{1n}b_{pq} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{11} & a_{m1}b_{12} & \cdots & a_{m1}b_{1q} & \cdots & a_{mn}b_{11} & a_{mn}b_{12} & \cdots & a_{mn}b_{1q} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{p1} & a_{m1}b_{p2} & \cdots & a_{m1}b_{pq} & \cdots & a_{mn}b_{p1} & a_{mn}b_{p2} & \cdots & a_{mn}b_{pq} \end{bmatrix} \tag{4}$$

However, the Kronecker product between tensors required a significant amount of memory. To perform the Kronecker product between tensors with a batch size of 32, 1-channel, and a resolution of 112×112, a memory requirement of 600.25GB was needed. Therefore, this paper decided to approximate the process of creating the pairwise cross-correlation matrix. To achieve a reliable approximation, an investigation was conducted to determine which characteristics could effectively represent the unique spatial information of the images. Various characteristics were explored, and among them, this paper chose to utilize the singular values of the images. When performing Singular Value Decomposition (SVD) on an image, it can be represented as a combination of rank-1 matrices ($\mathbf{u}\mathbf{v}^{\mathrm{T}}$), obtained through the outer product of the left singular vector ($\mathbf{u}$) and the right singular vector ($\mathbf{v}$), as shown in Eq. 5.

$$\mathbf{I} = \mathbf{U}\boldsymbol{\Sigma}_{\boldsymbol{I}}\mathbf{V}^{\mathrm{T}} = \sigma_1^I\mathbf{u}_1\mathbf{v}_1^{\mathrm{T}} + \sigma_2^I\mathbf{u}_2\mathbf{v}_2^{\mathrm{T}} + \cdots\sigma_n^I\mathbf{u}_n\mathbf{v}_n^{\mathrm{T}} \tag{5}$$

In this context, the vectors are considered as column vectors, and σ_1 represents the largest singular value. The singular values ($\sigma_1\cdots\sigma_n$) of an image ($\mathbf{I} \in \mathbb{R}^{n\times n}$)

represent the contributions of each of the constituent elements ($\mathbf{u}\mathbf{v}^{\mathrm{T}}$) that make up the image. Therefore, the singular values of an image can be interpreted as important intrinsic information for representing the image in the spatial domain. The Fig. 4 below has been included to aid in understanding the Singular Value Decomposition (SVD) of images. In this paper, the characteristics of the singular

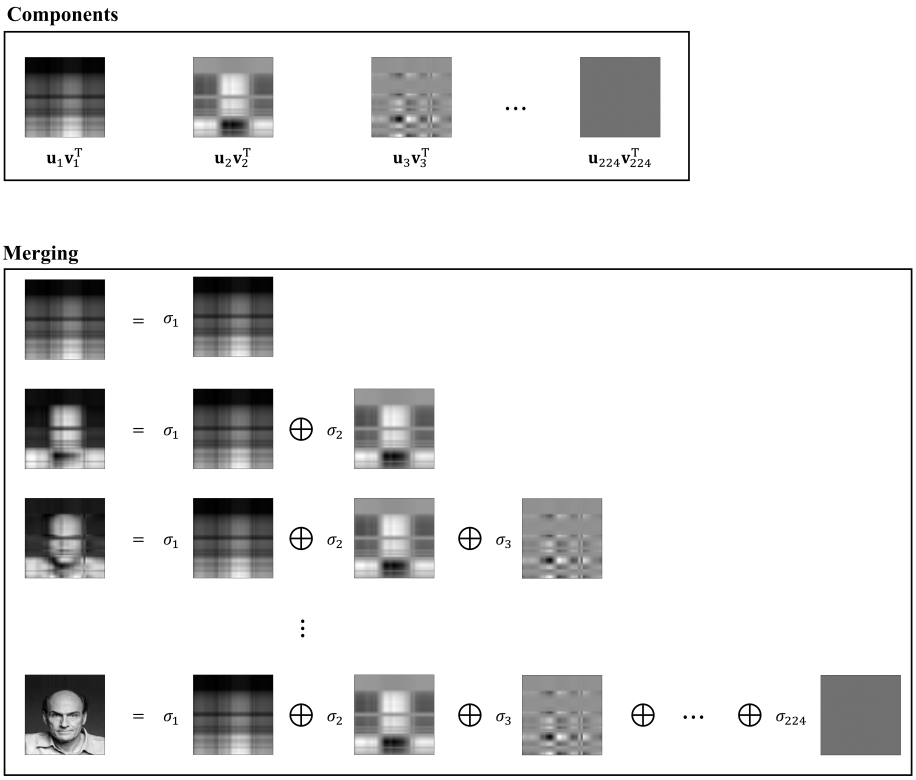

Fig. 4: The example process of singular value decomposition on grayscale image.

values of images are utilized to create a pairwise cross-correlation matrix by first flattening singular value matrices of the high frequency component ($\boldsymbol{\Sigma}_h$) and the compressed tensor ($\boldsymbol{\Sigma}_x$) , as shown in Eq. 6 and 7, and then applying vector outer products as depicted in Eq. 8. Where vec$(\cdot)$ is vectorization operation.

$$\boldsymbol{\Sigma}_h^{'} = \mathrm{vec}(\boldsymbol{\Sigma}_h) \tag{6}$$

$$\boldsymbol{\Sigma}_x^{'} = \mathrm{vec}(\boldsymbol{\Sigma}_x) \tag{7}$$

$$\boldsymbol{\Sigma}_h^{'} \otimes \boldsymbol{\Sigma}_x^{'} = \boldsymbol{\Sigma}_h^{'} \boldsymbol{\Sigma}_x^{'\mathrm{T}} = \begin{bmatrix} \sigma_1^h \sigma_1^x & \sigma_1^h \sigma_2^x & \cdots & \sigma_1^h \sigma_n^x \\ \sigma_2^h \sigma_1^x & \sigma_2^h \sigma_2^x & \cdots & \sigma_2^h \sigma_n^x \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_n^h \sigma_1^x & \sigma_n^h \sigma_2^x & \cdots & \sigma_n^h \sigma_n^x \end{bmatrix} \tag{8}$$

The generated cross-correlation matrix is not used directly but rather passes through a convolution layer and sigmoid function ($f_\sigma(\cdot)$)as shown in Eq. 9.

$$\mathbf{a} = f_\sigma(\mathcal{C}_{7\times7}(\boldsymbol{\Sigma}_h^{'} \otimes \boldsymbol{\Sigma}_x^{'}) \tag{9}$$

The reason for passing through the convolution layer is to leverage the implicit method's approach, which involves generating an attention score matrix from the cross-correlation matrix. In contrast to the traditional implicit approach, which seeks to establish the relationship between query and key without any evidence, this paper generates attention scores through the cross-correlation matrix, providing concrete evidence for the relationship. The attention score matrix, generated in this manner, undergoes broadcasting and is then subjected to an elementwise product ($\odot$) with the value $\mathbf{x}$, as represented in Eq. 10.

$$\mathbf{x}^{\text{refined}} = \mathbf{a} \odot \mathbf{x} \tag{10}$$

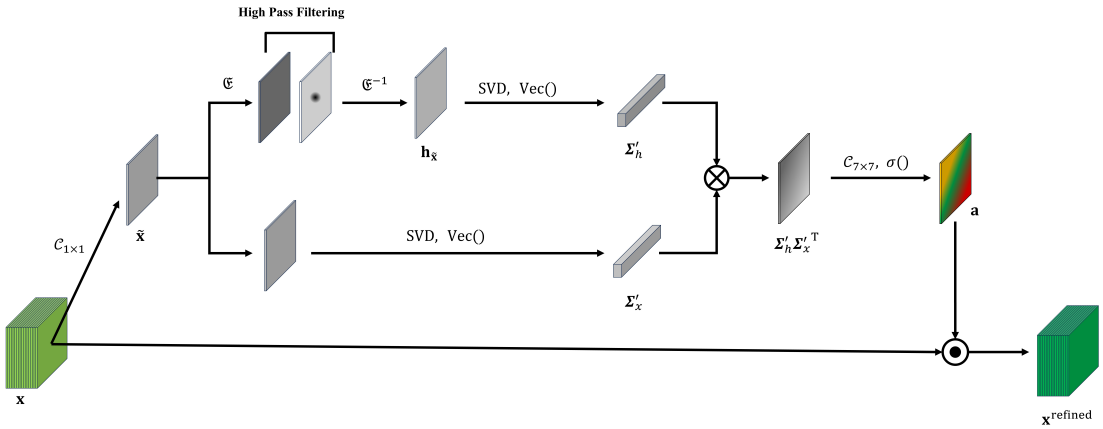

Fig. 5: This illustration depicts entire process of proposed attention network.

4 Experiment

4.1 Dataset

FER2013: FER (Facial Expression Recognition) 2013 [9] is a dataset that comprises grayscale images representing seven basic emotions: angry, disgust,

fear, happy, sad, surprise, and neutral. Each image in the dataset has a resolution of 48×48 pixels, and it consists of a substantial collection of 35,887 images in total. The dataset contains facial expression data representing various races, ages, and genders. Fig. 6 below illustrates examples from the FER2013 dataset. In this experiment, 28,707 images were used as the training dataset, and 7,180 images were allocated for the test dataset. For all training using this dataset, a batch size of 32 was configured.

Fig. 6: The example images of FER2013 dataset.

JAFFE: JAFFE (Japanese Female Facial Expression) [15,16] is a dataset consisting of a total of 213 grayscale images. The dataset uses the following 7 basic emotions as class labels: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. In this experiment, 158 images were used as the training dataset, and 55 images were designated for the test dataset. For all training utilizing this dataset, a batch size of 8 was configured. The Fig. 7 below represents examples from the JAFFE dataset.

Fig. 7: The example images of JAFFE dataset.

CK+: The CK+ (Extended Cohn-Kanade) dataset [14] comprises 593 sequences and 123 subjects. In this experiment, the last 3 frames of each sequence were used, resulting in a total of 981 images being utilized. This dataset employs 7 emotions (anger, contempt, disgust, fear, happy, sadness, and surprise) as class labels. In this paper, 735 images were allocated for the training dataset, while 246 images were designated for the test dataset out of the 981 total images. A batch size of 16 was utilized for all training using this dataset. Fig. 8 below provides examples from the CK+ dataset.

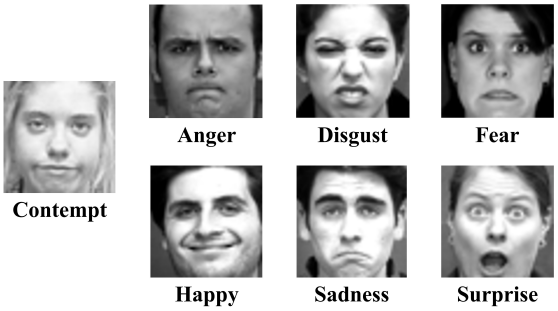

Fig. 8: The example images of CK+ dataset.

4.2 Experimental Setup

Experimental Equipment: The experiment was conducted using the equipment listed below.

- **CPU:** Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz (1EA)
- **GPU:** NVIDIA Geforce RTX 3090 24GB VRAM (4EA)
- **RAM:** Samsung DDR4 32GB (6EA)

Train Setup: The models used in this experiment are ResNet18 [10] and MobileNetV2 [24]. In this experiment, all training images were upscaled to a resolution of 224x224 pixels before use. The loss function employed in this experiment was cross-entropy, and the number of training epochs was fixed at 100 for all experiments. The initial learning rate was assigned differently for each model. ResNet18 was set to an initial learning rate of 1e-4 for all datasets, including FER2013, JAFFE, and CK+. On the other hand, MobileNetV2 was configured with a learning rate of 1e-4 for FER2013 and 1e-3 for both JAFFE and CK+. The learning rate scheduler employed the "ReduceLROnPlateau" provided by PyTorch [21]. It updated the learning rate to 0.3 times its previous value when the validation loss did not decrease continuously for 5 epochs. ResNet18 has five big convolution block (64-channel, 64-channel, 128-channel, 256-channel, and

512-channel). The attention modules are attached between each big convolution block. It means that 4 modules are used for ResNet18. MobileNetV2 has 8 big convolution block (32-channel, 16-channel, 24-channel, 32-channel, 64-channel, 96-channel, 160-channel, and 320-channel) and one pointwise convolution (1280-channel). The attention modules are attached between each big convolution block. In other words, the 7 attention modules are used for MobileNetV2.

4.3 Ablation Study

In this section, the results of the verification of whether high frequency components can indeed provide meaningful assistance in performing attention are presented. The control groups were prepared as follows: (1) Vanilla ResNet18, (2) when the query of the attention module was provided with the high frequency component (HFC), (3) when the query of the attention module was provided with the low frequency component (LFC), (4) when both HFC and LFC were used as separate queries, and (5) when the additional cross-correlation matrix between LFC and HFC was used. In the case of control group (4), both the cross-correlation matrices obtained by using HFC and LFC as queries were concatenated, and the resulting matrix passed through a convolutional layer. Control group (5) further concatenated the cross-correlation matrix obtained by providing HFC as the query and LFC as the key to the result of control group (4). Subsequently, the combined matrix was passed through a convolutional layer. The following Table 1 presents the results of the ablation study. Looking at Table 1, it is evident that providing LFC as the query results in a

Table 1: Ablation study results with ResNet18 and CK+.

Model	**Best Epoch**	**Accuracy (%) @ Best Epoch**
ResNet18	100	86.9
ResNet18 + HFC	41	**92.31** (+5.41)
ResNet18 + LFC	17	86.88 (-0.02)
ResNet18 + (HFC, LFC)	38	91.7 (+4.8)
ResNet18 + (HFC, LFC, HL)	19	87.73 (+0.83)

decrease in performance compared to the vanilla model. When both LFC and HFC were used simultaneously, the accuracy increased compared to the Vanilla model. However, it showed lower performance compared to when HFC alone was used. These research results support the argument presented in this paper that providing HFC as the query improves classification performance. Furthermore, it can be observed that providing LFC as the query actually has a negative impact on performance.

4.4 Comparison

In this section, we verify whether the attention network proposed in this paper has a beneficial impact on ResNet18 and MobileNetV2 using the FER2013, JAFFE, and CK+ datasets. Furthermore, we compare its performance with the existing implicit method, BAM (Bottleneck Attention Module) [19]. All performance comparisons are conducted based on the models that achieved the highest validation accuracy among the entire 100 epochs. Through Table 2, it can be

Table 2: Comparison results about ResNet18, MobileNetV2, BAM, and ours with FER2013, JAFFE, and CK+ datasets.

Dataset	Model	Param.	Best Epoch	Accuracy (%) @ Best Epoch							
	Classes			**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**	**Average**
FER2013	ResNet18	11.17M	100	56.98	47.45	43.36	82.19	61.48	46.19	78.94	59.56
	ResNet18 + BAM	11.198M (+23,808)	30	52.71	54.05	45.7	81.29	61.15	51.16	76.41	60.35 (+0.79)
	ResNet18 + Ours	11.17M (+708)	90	54.18	54.05	41.6	81.62	59.53	54.05	77.98	**60.43** (+0.87)
	MobileNetV2	2.23M	100	41.75	54.05	38.77	80.61	58.16	46.11	73.16	56.09
	MobilNetV2 + BAM	2.24M (+11,192)	38	46.76	54.05	42.29	81.12	56.61	41.38	73.77	56.56 (+0.47)
	MobileNetV2 + Ours	2.23M (+767)	57	50.87	53.15	40.82	80.05	48.74	52.13	73.41	**57.02** (+0.93)
	Classes			**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**	**Average**
JAFFE	ResNet18	11.17M	100	100	85.71	100	87.5	100	100	100	96.17
	ResNet18 + BAM	11.198M (+23,808)	30	87.5	100	100	87.5	100	100	100	**96.43** (+0.26)
	ResNet18 + Ours	11.17M (+708)	31	75	100	100	87.5	100	100	100	94.64 (-1.53)
	MobileNetV2	2.23M	100	50	71.43	62.5	87.5	87.5	62.5	62.5	69.13
	MobilNetV2 + BAM	2.24M (+11,192)	67	75	71.43	75	75	87.5	75	75	**76.28** (+7.15)
	MobileNetV2 + Ours	2.23M (+767)	32	87.5	42.86	50	75	87.5	87.5	87.5	73.98 (+4.85)
	Classes			**Anger**	**Contempt**	**Disgust**	**Fear**	**Happy**	**Sadness**	**Surprise**	**Average**
CK+	ResNet18	11.17M	100	76.47	100	100	84.21	100	47.62	100	86.9
	ResNet18 + BAM	11.198M (+23,808)	45	82.35	100	93.18	73.68	98.08	85.71	100	90.43 (+3.53)
	ResNet18 + Ours	11.17M (+708)	41	79.41	100	93.18	89.47	100	85.71	100	**92.31** (+5.41)
	MobileNetV2	2.23M	100	58.52	100	90.91	63.16	98.08	52.38	98.39	80.25
	MobilNetV2 + BAM	2.24M (+11,192)	87	76.47	100	100	94.74	94.23	57.14	95.16	88.25 (+8)
	MobileNetV2 + Ours	2.23M (+11,192)	50	73.53	100	93.18	84.21	100	85.71	88.71	**89.34** (+9.09)

observed that the attention network proposed in this paper generally enhances the facial expression classification performance of ResNet18 and MobileNetV2. Especially in the case of FER2013 and CK+, it outperforms the traditional implicit method, BAM, in terms of performance. In the JAFFE dataset, while MobileNetV2 exhibited significant performance improvement, it was observed that ResNet18, on the contrary, experienced a decrease in performance. This could be interpreted as occurring due to overfitting caused by the small size of the dataset. However, in situations where the dataset is abundant, it can be observed that the performance of the Vanilla model is significantly improved, and it outperforms BAM, demonstrating superior performance. Furthermore, compared to BAM, the increase in parameters is minimal, which raises expectations for its effective use in real-time inference tasks for facial expression recognition. The following Fig. 9 shows the Grad-CAM (Gradient-weighted Class Activation Mapping) [25]extracted for Vanilla MobileNetV2 and MobileNetV2 with the attention network proposed in this paper, both trained on the CK+ dataset. By examining the figure, it can be observed that the model with the added attention network extracts more generalized features from the dataset compared to the conventional MobileNetV2. Using Contempt as an example, CK+ Contempt images commonly exhibit wrinkles around the mouth area. Vanilla MobileNetV2 tends to focus more on the eyes rather than these common features. However, in

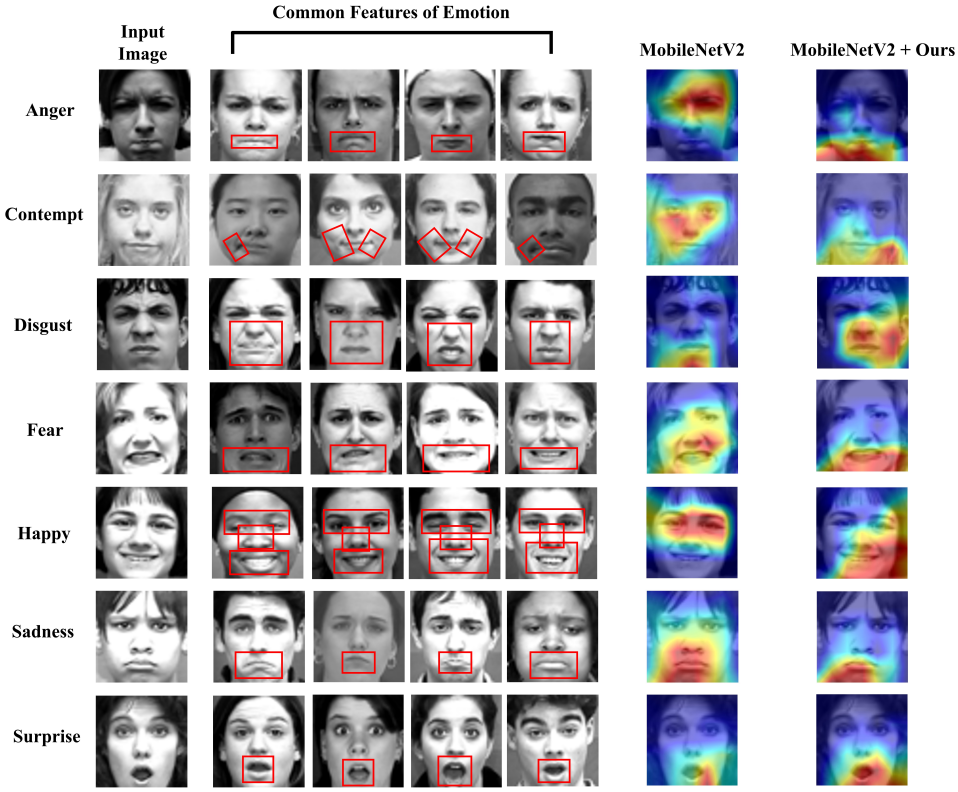

Fig. 9: This illustration represents the Grad-CAM of vanilla MobileNetV2 and ours. All the facial expression images are sourced from the CK+ dataset.

the model with the added attention network, it is evident that it concentrates more on the wrinkles around the mouth area. This paper attributed this phenomenon to the attention network's ability to highlight information inherent in HFC, such as wrinkles around the mouth and changes in lip shapes, which are essential for recognizing facial expressions within the feature map. In addition, it can be interpreted that the performance of the model has also improved due to the utilization of a feature map in which information necessary for emotion recognition is highlighted.

5 Conclusion

This paper introduced an attention network that encourages the utilization of high-frequency components to emulate the human process of recognizing facial expressions, aiming to enhance the facial expression recognition rate of CNN. The presented attention network combined the explicit method of mathematically

calculating the relationship between query and key with the implicit method of calculation through a neural network. Rather than simply inputting query and key into a neural network, the paper first calculated the mathematical relationship between query and key, using this as evidence to guide the neural network in computing the attention score. To efficiently operate the attention network, this paper compressed the input feature map into a 1-channel tensor. Furthermore, instead of computing pixel-wise pairwise cross-correlation matrices, this paper vectorized the singular value matrices of query and key, performing outer product operations to create pairwise cross-correlation matrices. The attention network proposed in this study significantly improved the performance of ResNet18 and MobileNetV2 models trained on the FER2013, JAFFE, and CK+ datasets. Furthermore, compared to the existing implicit method, BAM, it generally demonstrated superior performance in most scenarios. And, the increase in the number of parameters was much smaller compared to BAM. The attention network proposed in this paper is expected to provide meaningful assistance in tasks that require real-time facial expression recognition.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
3. Dagli, R.: Astroformer: More data might not be all you need for classification. arXiv preprint arXiv:2304.05350 (2023)
4. El Boudouri, Y., Bohi, A.: Emonext: an adapted convnext for facial emotion recognition. In: 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6 (2023). https://doi.org/10.1109/MMSP59012.2023.10337732
5. Fard, A.P., Mahoor, M.H.: Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. IEEE Access **10**, 26756–26768 (2022). https://doi.org/10.1109/ACCESS.2022.3156598
6. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)
7. Georgescu, M.I., Ionescu, R.T., Popescu, M.: Local learning with deep and handcrafted features for facial expression recognition. arXiv preprint arXiv:1804.10892 (2018)
8. Gesmundo, A., Dean, J.: An evolutionary approach to dynamic introduction of tasks in large-scale multitask learning systems. arXiv preprint arXiv:2205.12755 (2022)
9. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20. pp. 117–124. Springer (2013)
10. 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. pp. 770–778 (2016)

11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
12. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009), https://api.semanticscholar.org/CorpusID:18268744
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
14. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: 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 (2010). https://doi.org/10.1109/CVPRW.2010.5543262
15. Lyons, M.J.: " excavating ai" re-excavated: Debunking a fallacious account of the jaffe dataset. arXiv preprint arXiv:2107.13998 (2021)
16. Lyons, M.J., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets (ivc special issue). arXiv preprint arXiv:2009.05938 (2020)
17. Minaee, S., Abdolrashidi, A.: Deep-emotion: Facial expression recognition using attentional convolutional network. arXiv preprint arXiv:1902.01019 (2019)
18. Minji Park, Jung Nyun Lee, J.C.Y.J.K.J.Y., Whang, M.: Facial vibration analysis for emotion recognition (2016), https://api.semanticscholar.org/CorpusID:137695591
19. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. arXiv preprint arXiv:1807.06514 (2018)
20. Park, S., Jung, W.: The effect of spatial frequency filtering on facial expression recognition and age perception. Korean Journal of Cognitive and Biological Psychology **18**(4), 311–324 (2006)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
22. Pecoraro, R., Basile, V., Bono, V., Gallo, S.: Local multi-head channel self-attention for facial expression recognition. arXiv preprint arXiv:2111.07224 (2021)
23. Ridnik, T., Sharir, G., Ben-Cohen, A., Ben-Baruch, E., Noy, A.: Ml-decoder: Scalable and versatile classification head. arXiv preprint arXiv:2111.12933 (2021)
24. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)