Facial Expression Recognition: A Lite Deep Learning Based Approach

Vo Hoang Chuong[1][0000-0002-9910-7494] and Vo Hung Cuong[2][0000-0003-3989-4921]

[1] National Taiwan University of Science and Technology, Taipei, Taiwan
[2] Vietnam-Korea University of Information Technology and Communication - Danang University
vhchuong1997@gmail.com
vhcuong@vku.udn.vn

Abstract. Human expression is one of the most powerful and natural signals for human to convey their emotional states and intentions. With the recent success of deep learning techniques in many fields, along with the appearance of many datasets, the Facial expression recognition has made a transition from with-in laboratory problems to real-world applications. In fact, the deep learning has played an integral part in enriching the facial features being exploited from plenty of faces that further improve the recognizing accuracy. Recent methods have been witnessing two major problems: the first one is lacking sufficient training data, and the second one is that many methods cannot overcome hardship conditions such as: illumination, head poses, and complex background, etc. In this paper, we propose a CNN based network to deal with FER in videos. The input frames will be fed to a deep neural network (lite-XceptionFCNet) to extract the spatial features. Then, these features go through a classifier in the later component of the architecture. We used the confusion matrix and the parameters inferred from it such as accuracy, misclassification, precision, recall (sensitivity) and specificity to evaluate our fire detection module and model. The system achieved high accuracy in the FER dataset, respectively. The proposed system behaved robustly and showed the potential of being applied to real-time facial expression recognition.

Keywords: expression recognition, deep learning, human, facial expression, lite model.

1 Introduction

Facial expression is critical for the development of machines' and robots' cognitive systems. It focuses on how to utilize the arrangement and strength of facial muscle movements to estimate an expression label in a static picture or video. For its practicality in interactive robots, medical treatment, driver drowsiness detection, and many other human-computer interface systems, automatic facial expression analysis has been the subject of several studies.

Because of its pioneering research and the straightforward and intuitive characterization of facial expressions, the categorical model, which explains emotions in terms of distinct fundamental emotions, is still the most common perspective for FER. We will focus on FER producing high-quality features in this paper.

According to the feature extraction methods, FER systems can be classified into two categories: traditional features designing and learnt feature presentation. In some works, other modalities, including as audio and physiological channels, have been utilized in multimodal systems to aid in the recognition of expression based on these two vision-based techniques, however most of the systems support the images collected from cameras as it's only input.

Most of the traditional approaches preferred handcrafted features or shallow learning such as local binary pattern (LBP) [6], LBP on three orthogonal planes (LBP-TOP) [7]. In recent years, a set of standard algorithmic pipelines for FER has been created. J. Liu et al. [4] tackled the wide range recognition problem including various subjects, lightning conditions, and obstructions. In order to improve the contrast of the raw input data, the data augmentation approach is proposed. In 2020, Y. Liu et al. [5] wanted to eliminate unnecessary regions of input facial images and focus on the parts that are truly crucial in determining human's emotion.

The handcrafted features work well in many cases, however, there still exist issues when it comes to FER. Due to differences in personal characteristics such as age, gender, ethnic origin, and level of expressiveness, there are considerable subjective variances. Variations in posture, lighting, and obstructions are common in unconstrained facial expression settings, in addition to subject identification bias. Recently, deep learning's tremendous feature learning capabilities have proved its effectiveness. There are methods that uses Recurrent neural network in its architecture, however, if one wants to use this kind of deep learning component, the input signal must be processed carefully. The main reason is that the facial expression is sometimes very ambiguous; therefore, the changes of facial muscle are not obviously clear, or the changes are very local and small that need to be amplified. J. Guo et al. [3] produced a method call a hybrid facial expression recognition in which the features are formed by combining both spatial and temporal characteristics while the RGB images are not themselves being fed directly into the RNN module but being transformed into landmarks and considering the other signals to enhance the accuracy. Several other works often focus on exploiting a well-designed backbone model to produce high-quality encoded features as in [12], [13], [14] all performed very well in this task of recognition, however, only few of those can meet the requirements of being able to be at runtime processing with relatively small hardware.

In this paper, we propose a simple but effective architecture for the facial expression features recognition which can exploit the rich spatial features and light enough to be run on small devices. This paper is organized as follows. Section 2 presents the proposed method. Section 3 shows the experimental results, Section 4 shows discussions.Corporate social responsibility definition and theory.

Corporate social responsibility is the business responsibilities on the basis of respecting the law and commitment to stakeholders, being able to link business activities with solving social problems, ethically, protecting the environment, human rights and responding to customer concerns, with the aim of maximizing the benefits of business owners, stakeholders and society as a whole, prevent and minimize possible negative impacts from business activities of enterprises to ensure sustainable development goals (Clark et al., 2016; Agudelo, 2019).

Corporate social responsibility disclosure is a way to implement the transparency of the business to ensure that shareholders and management agencies can timely and fairly access information. Adequate and accurate information disclosure play an important role to investors, shareholders, and state management agencies. Previous studies on CSRD mentioned the theories such as agency theory, stakeholder theory, proprietary cost theory, political cost theory, signal theory, legitimacy theory (Hackston and Milne, 1996).

2 Method

In this section, we introduce common stages often being found in automatic FER systems: preprocessing, feature extracting. For each of these stages, we will briefly introduce the used algorithms, and mention the existing methods that practice the implementation with their referenced papers.

2.1 Preprocessing

Environmental variations have a huge impact on how well the FER systems behave. Every single factor such as: complex backgrounds, illuminations, head-poses all affect the performance of the algorithms. Thus, before training it is necessary to preprocess the data. This stage aims to ease the training and equip the system with an ability to deal with complex external changes. In fact, this step consists of several algorithms which help normalize and align the meaningful facial information.

Face alignment.

The Viola-Jones (V&J) face detector [2] is a well-known and extensively used implementation for near-frontal face identification that is both robust and computationally simple. The Viola-Jones Object Detection Framework combines the concepts of Haar-like Features, integral images, the AdaBoost Algorithm, and the Cascade Classifier to create a system for object detection that is fast and accurate. Firstly, V&J face detector detects the face from the input image. Then, the region of interest which is defined as the face region is cropped out for the next step of pre-processing.

Recently, many methods [3], [4], [5] has applied landmark features point in their preprocessing steps. Originally, landmark points can be used to find out the importance of the main components of the face and the other positions which has less information. However, those points also provide accurate information of the important points' location. For instance, as in [3] the method exploited the 12 points that surrounding the eyes. From 12 points found, the centroids of two eyes are calculated by taking the average of six points at each side. After that, the rotation angle and image-scale are produced which help to rotate the face along with the horizontal axis and zoom the face out or in to best eliminate the background. This kind of alignment is very effective in case of complex background and in case of that the actors are not in formal posing.

Spatial Features Extraction

Our model is inspired by the Xception [8]. This architecture combines the use of residual modules and depth-wise separable convolutions. Residual modules change the intended mapping between two successive layers so that the learned features are the difference between the original and desired feature maps. This architecture has proven itself to be efficient and lite weight in dealing with problems that require models with as least parameter as possible.

Because most of the fully connected layers were removed, we were able to cut the number of parameters even more by eradicating them from the convolutional layers. This was accomplished by employing depth-wise parable convolutions. Convolutions that are depth-wise separable are made up of two layers: depth-wise convolutions and pointwise convolutions. The primary goal of these layers is to distinguish between spatial and channel cross correlations [8]. They accomplish this by first applying a D×D filter to each of the M input channels, and then combining the M input channels into N output channels using N 1×1×M convolution filters. Using 1×1×M convolutions, each value in the feature map is combined without regard for their spatial relationship within the channel.

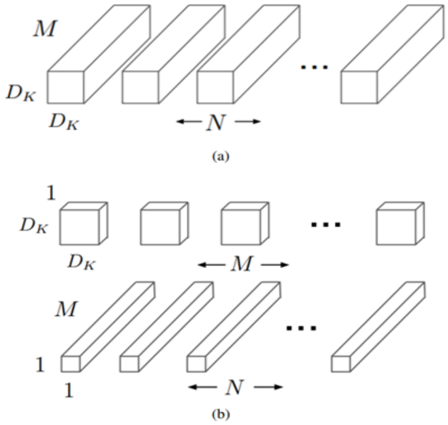


Fig. 1. Difference between (a) standard convolutions and (b) depth-wise separable convolutions. [9]

Depth-wise separable convolutions take less time to calculate than ordinary convolutions. **Fig. 1** shows a visual representation of the difference between a standard Convolution layer and a depth-wise separable convolution.

Table 1. LITE- XCEPTION BODY ARCHITECTURE

Stage i	Operator $\widehat{F}_i$	Resolution $\widehat{H}_i \times \widehat{W}_i$	Channels. $\widehat{C}_i$
1	Conv3 × 3	46 × 46	8
2	Conv3 × 3	44 × 44	8
3	XceptionRes1	22 × 22	16
4	XceptionRes2	11 × 11	32
5	XceptionRes3	6 × 6	64
6	XceptionRes4	3 × 3	128
7	Fully Connected	1 × 1	7

Our final design is a fully convolutional neural network with four residual depth-wise separable convolutions and a batch normalization and ReLU activation function after each convolution. To make a prediction, the final layer uses global average pooling and a fully connected layer followed by soft-max activation function. This architecture contains around 488,700 parameters, which is several times size reduction over original Xception Net. Our final architecture, which we call lite-Xception, is illustrated in Figure 3 and listed in **Table 1**.

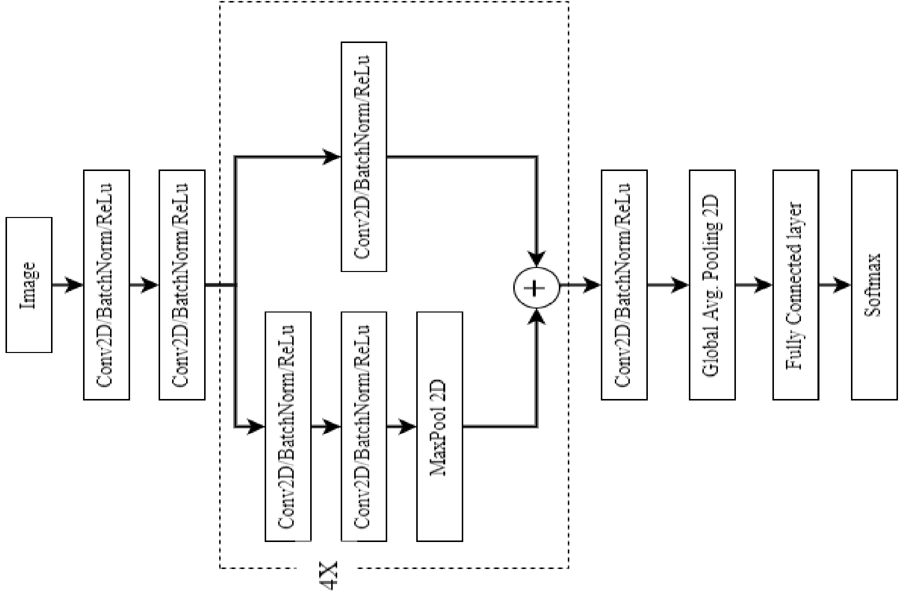


Fig. 2. The proposed pipeline.

2.2 Facial expression recognition system

We propose a real-time framework to detect human face and infer the facial expression from the face obtained. Firstly, human's face is detected using Viola-Jones

(V&J) face detector [2]. The facial region will then be cropped out of the frame and go through preprocessing function. After acquiring the desired format, the input facial image will be fed into the model to yield the classification prediction. The result and the predicted percentage are shown onto the interface. The pipeline is illustrated in **Fig. 2**.

2.3 Dataset

For the creation of a deep emotion recognition system, having enough labeled training data that includes as many variants of people and settings as feasible is critical. The primary reference, number of participants, number of pictures or video samples, collecting environment, expression distribution, and other information are all listed in **Table 2.** Facial expression dataset. In this work, we use 1 dataset as described. We analyzed the dataset and checked for the are noisy instances.

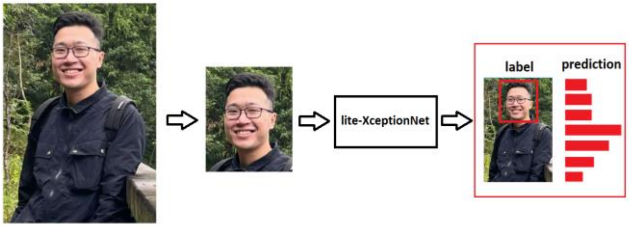


Fig. 3. Real-time facial expression recognition system.

Table 2. Facial expression dataset

Database	Samples	Subject	Condit.	Expression distribution
FER-2013 [1]	35,887 images	N/A	Web	7 basic expressions

FER2013 [1] The Google image search API automatically collects FER2013, which is a large-scale, unrestricted database. After eliminating improperly identified frames and fine-tuning the cropped region, all pictures were registered and shrunk to 48*48 pixels. With seven expression labels, FER2013 has 28,709 training photos, 3,589 validation images, and 3,589 test images. **Fig. 4**. shows some instances from FER2013 dataset.

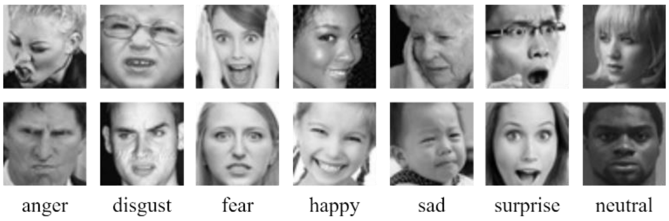


Fig. 4. Some samples of FER2013 database.

3 EXPERIMENTAL RESULTS AND EVALUATION

This paper developed a system with the ability of detecting face, recognize the facial expression from the detected face and show the prediction percentage onto display. In order to execute these tasks, Viola-Jones (V&J) face detector was employed, the lite-XceptionFCNet is proposed with XceptionRes blocks. In this experiments, lite-XceptionFCNet is tested through confusion matrices and the metrics inferred from them including accuracy, misclassification, precision, recall (sensitivity) and F1-Score. These metrics are calculated as follow [10] [11]:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2}$$

$$\text{Misclassification} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{F1} - \text{score} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall+Precision}} \tag{6}$$

with TP: True positive, TN: True negative, FP: False positive, FN: False negative.

Overall, the results obtained from the experiments show that the proposed system has reasonable accuracy and high computational efficiency.

Result of Facial expression recognition lite-XceptionFCNet

The qualitative results of the fire detection module are shown in **Fig. 5**

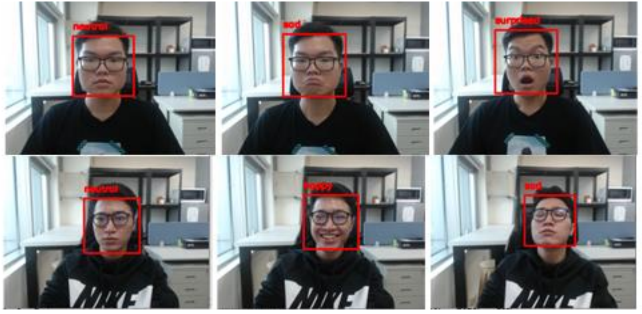
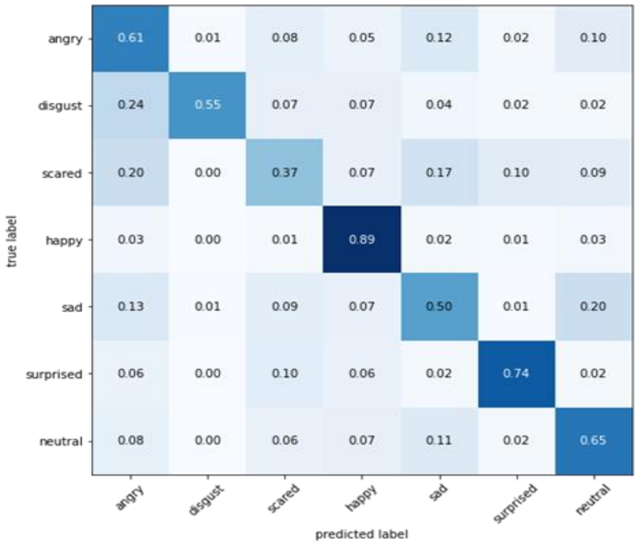


Fig. 5. The qualitative results of the lite-XceptionFCNet

Evaluation of Facial expression recognition lite-XceptionFCNet

The lite-XceptionFCNet is tested with 3,589 images divided equally into 7 classes: "angry", "disgust", "scared", "happy", "sad", "surprised", "neutral". In this testing set, the number of images in each class are 958, 111, 1024, 1774, 1233, 1247, 831 respectively. The evaluation metrics of the lite-XceptionFCNet model is shown in Table III. The confusion matrix is shown in **Fig. 6**.



	angry	disgust	scared	happy	sad	surprised	neutral
angry	0.61	0.01	0.08	0.05	0.12	0.02	0.10
disgust	0.24	0.55	0.07	0.07	0.04	0.02	0.02
scared	0.20	0.00	0.37	0.07	0.17	0.10	0.09
happy	0.03	0.00	0.01	0.89	0.02	0.01	0.03
sad	0.13	0.01	0.09	0.07	0.50	0.01	0.20
surprised	0.06	0.00	0.10	0.06	0.02	0.74	0.02
neutral	0.08	0.00	0.06	0.07	0.11	0.02	0.65

Fig. 6. The confusion matrix of testing the lite-XceptionFCNet on FER2013.

It can be observed from **Table 3** that the lite-XceptionFCNet model has relatively decent performance with average misclassification rate. The metrics prove that the proposed model after end-to-end learning process has obtained high accuracy, precision, recall and specificity throughout 3 classes e.g., happy, surprised, and neutral.

Table 3. Evaluation of lite-XceptionFCNet model

Class	Accuracy	Misclassification	Precision	Recall	F1-score
Angry	0.61	0.49	0.51	0.61	0.56
Disgust	0.55	0.35	0.65	0.55	0.59
Scared	0.37	0.49	0.51	0.37	0.42
Happy	**0.89**	**0.19**	**0.81**	**0.89**	**0.85**
Sad	0.5	0.46	0.54	0.5	0.52
Surprised	0.74	0.23	0.77	0.74	0.75
neutral	0.65	0.38	0.62	0.65	0.63

Misclassification cases. There appear some cases of misclassification. The lite-XceptionFCNet mistook the expressions of anger and disgust to neural and sad. These cases are shown in **Fig. 7**. These are apparently resulted from the similarity among the expressions that are easily confused between each other.

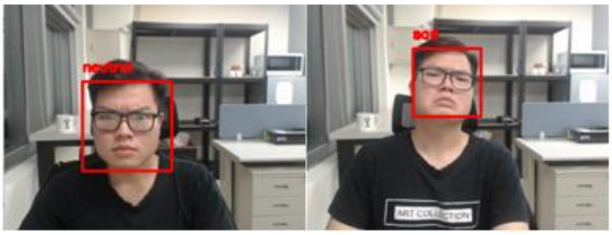


Fig. 7. The misclassification cases.

Baseline. The comparison between the proposed lite-XceptionFCNet performance with several well-known architectures applied in in FER 2013 dataset. The results claimed from the papers will be used for subjective comparison. The detailed comparison is shown in **Table 4**

Table 4. Performance evaluation of well-known classification networks and our method on FER2013

Technique	**Parameters**	**Accuracy**
VGG19 [12]	139.5M	70.80
ResAttNet56 [13]	29.0M	72.63
Densenet121 [14]	6.9M	73.16
Resnet152 [15]	58.1M	73.22
Cbam resnet50 [16]	28.5M	**73.39**
Our lite-XceptionFCNet	**488.7K**	64.53

Result of facial expression recognition system

The results of the facial expression recognition system involved processing the input video frame and yielding the output of face detection, emotion recognition and display prediction label correspondingly. In Fig. 5. the system has successfully detected the person’s face, recognized their emotion and given out correct label prediction according to the above-mentioned elements. The results obtained suggest that the proposed system is feasible and can be applied to real-life condition e.g., indoor situation.

4 DISCUSSION

In this paper, we have proposed a lite weight but efficient architecture for recognizing human facial expression. The algorithm first captures the facial region to eliminate the background. After that, the patch is preprocessed and being fed into the lite-XceptionFCNet feature extractor to extract learnt feature. Consequently, the vector is then classified to predict the expression label. Compared to other previous works, the proposal is slightly lower in terms of accuracy (64.53%). However, the number of parameters is superiorly smaller, with only 488K parameters which allows it to be used ideally in at-run-time required applications. In addition, the proposal is implemented using Keras framework, and a demonstration program is built along with it. In conclusion, the proposal is very suitable for applications which needs to balance both the accuracy and the hardware resource. However, to satisfy ones who ask for a higher accuracy but still desire this smallness of model size, the feature extractor should be kept on refined to be more well-produced highly qualified features. Finally, the model is suggested for various kinds of applications especially the ones with medium requirements on correctness and the ones with limitations of hardware.

References

1. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee et al.: Challenges in representation learning: A report on three machine learning contests, in International Conference on Neural Information Processing. Springer, 2013, pp. 117–124.
2. P. Viola and M. Jones: Rapid object detection using a boosted cascade of simple features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, pp. I-I, doi: 10.1109/CVPR.2001.990517.
3. J. Guo, P. Huang and L. Chang: A Hybrid Facial Expression Recognition System Based on Recurrent Neural Network, 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2019, pp. 1-8, doi: 10.1109/AVSS.2019.8909888.
4. J. Liu, H. Wang and Y. Feng: An End-to-End Deep Model With Discriminative Facial Features for Facial Expression Recognition, in IEEE Access, vol. 9, pp. 12158-12166, 2021, doi: 10.1109/ACCESS.2021.3051403.
5. Y. Liu, X. Zhang, Y. Lin and H. Wang: Facial Expression Recognition via Deep Action Units Graph Network Based on Psychological Mechanism, in IEEE Transactions on Cognitive and Developmental Systems, vol. 12, no. 2, pp. 311-322, June 2020, doi: 10.1109/TCDS.2019.2917711.
6. C. Shan, S. Gong, and P. W. McOwan: Facial expression recognition based on local binary patterns: A comprehensive study, Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.
7. G. Zhao and M. Pietikainen: Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 6, pp. 915–928, 2007.
8. F. Chollet: Xception: Deep Learning with Depthwise Separable Convolutions, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.
9. Howard, Andrew G., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017).
10. Ting K.M: Confusion Matrix, In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA, 2017, pp. 260, https://doi.org/10.1007/978-1-4899-7687-1_50
11. Powers, David M. W. (2011): Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. In: Journal of Machine Learning Technologies. 2 (1): 37–63.
12. Simonyan, Karen, and Andrew Zisserman: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

13. F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang: Residual attention network for image classification, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164, 2017.
14. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger: Densely connected convolutional networks, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708, 2017.
15. K. He, X. Zhang, S. Ren, and J. Sun: Deep residual learning for image recognition, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.
16. S. Woo, J. Park, J.-Y. Lee, and I. So Kweon: Cbam: Convolutional block attention module, in Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, 2018