RARN: Lightweight Deep Residual Learning with Attention for Human Emotions Recognition

Zhenyuan Zhu[a]
Master of Professional Study in Data Science, University of Auckland, Auckland, New Zealand

Keywords: Emotion Recognition, Rotation-Aware Residual Network, Facial Expressions, Convolutional Neural Networks.

Abstract: Human emotion identification represents a formidable challenge within computer vision research. This study endeavours to classify human emotions across seven discrete categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. To address this challenge, this paper introduces the Rotation-Aware Residual Network (RARN), a novel framework leveraging convolutional neural networks (CNNs) and spatial attention mechanisms. Notably, this approach is designed to excel in accurately discerning facial emotions amidst complex real-world contexts. Experimental validation conducted on the FER-2013 Dataset underscores the efficacy of our proposed model, demonstrating notable improvements in emotion recognition accuracy. Crucially, the Rotation-Aware Residual Network's innovative integration of multi-scale fusion and angle-sensitive spatial attention modules underscores its unique capacity to capture nuanced facial expressions. This breakthrough has significant implications for diverse applications, including human-computer interaction, psychological health assessment, and social signal processing. Moving forward, future research endeavours will focus on further refining the network architecture and expanding the diversity of datasets to enhance the model's performance across various scenarios.

1 INTRODUCTION

Automatic facial expression analysis is widely used in computer vision for many applications, such as emotion prediction, expression retrieval, and image album summarization, and has been extensively studied (He, 2016)(Szegedy, 2017)(Zhou, 2023). The generalised classification model categorises emotion detection into happiness, sadness, fear, fury, disgust, and surprise. These categories are established to streamline the identification and description process through common terminology. The universality hypothesis of emotion (Ekman, 1969) is widely used in emotional computing research due to its simplicity and universality, making it the preferred theory. The implementation of mobile and embedded computing requires not just stronger hardware, more datasets, and more complex models for autonomous facial expression analysis but also network topologies that are efficient in terms of power consumption and memory utilisation (Szegedy, 2017).

The most straightforward strategy to enhance the effectiveness of deep neural networks is to increase their depth and breadth (Arora, 2014). Nevertheless, this would unavoidably lead to a significant rise in the network's parameter count, leading to overfitting (He, 2016). Szegedy et al. (Szegedy, 2017) from Google introduced the Inception module of deep convolutional networks as a solution to the aforementioned issues. This module's core principle is the parallel integration of several convolutional layers; concatenating the output matrices from each layer in the depth dimension produces a more complex matrix. By repeatedly stacking the Inception module, a more extensive network can be created, effectively increasing the network's depth and breadth. This, in turn, enhances the accuracy of the deep learning network and prevents overfitting. One benefit of utilising the Inception module is its capacity to merge visual data of varying dimensions while reducing the size of matrices containing numerous entries. This aggregation technique facilitates extracting characteristics from images of various sizes.

In addition to the Inception module, this study delves into the residual connections proposed by He

[a] https://orcid.org/0009-0009-7527-5395

Zhu, Z.
RARN: Lightweight Deep Residual Learning with Attention for Human Emotions Recognition.
DOI: 10.5220/0012911100004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 127-134
ISBN: 978-989-758-713-9

et al. (He, 2016). They contend that residual connections are inherently vital for efficiently training deep networks. Their ResNet addresses the issue of model degradation caused by increased network depth through the incorporation of a deep residual learning module. Specifically, this module employs a stacking mechanism that combines the input and output of each layer without introducing additional parameters or computations, thereby enhancing the convergence speed of model training. Another study (Wang, 2017) has demonstrated that incorporating a spatial attention mechanism into ResNet ensures that an image retains its original features even after undergoing operations like cropping, translation, or rotation. This enhancement significantly improves the accuracy of model predictions.

This study encapsulates the foundational concepts of the previously mentioned network modules and introduces a novel and versatile facial expression detection model, termed the Rotation-Aware Residual Network (RARN). RARN is designed to balance both network performance and efficiency, addressing critical aspects overlooked by existing architectures. Through rigorous experimentation on the FER-2013 dataset, the effectiveness and practicality of RARN are thoroughly evaluated. Comparative analyses against conventional ResNet and InceptionNet architectures highlight the unique contributions of RARN in achieving superior performance metrics while maintaining computational efficiency. This research underscores the significance of incorporating rotation-aware mechanisms in facial expression detection, offering valuable insights into improving accuracy and robustness.

2 LITERATURE REVIEW

2.1 Deep Residual Learning

He et al. (He, 2016) proposed including a residual framework in the network architecture to address the issue of training deep networks. The residual network is based on the concept of a highway network, which has shortcut connections in its construction. This allows the input to be immediately sent to the output. Specifically, the fundamental concept underlying ResNet is the presumption that an optimal solution exists for the model's network architecture. In other words, ResNet holds that numerous network layers are frequently redundant in the actual deep network architecture. To achieve the completion of the identity mapping in these redundant levels and verify that the input and output of the identity layer are identical, as seen in Figure 1, ResNet modifies the input of the residual module from $F(X)$ to $H(X) = F(X) + x$. If the network layers are redundant, then just let $F(X)$ equal zero to achieve the identity mapping. Through the incorporation of this residual learning module, the network can substantially augment the network layer's depth throughout the design phase.

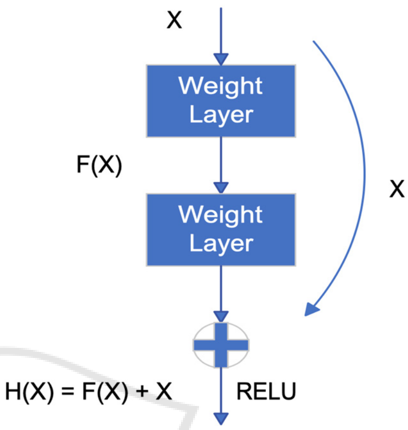

Figure 1: Residual Module (Photo/Picture credit: Original).

The ResNet architecture is often used for object recognition because of its efficient design. The architecture has a single convolutional layer, multiple convolutional blocks in the intermediate section, and an output layer. The ResNet architecture is classified as ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152, according to the quantity of convolutional blocks included in the centre. As more blocks are included, the network increases in depth, enabling the detection of increasingly complex feature patterns.

2.2 Inception Module

The primary concept behind the Inception module (Szegedy, 2015) is to simultaneously apply multiple convolution operations or pooling operations to the input image. This enables the retrieval of different dimensions of feature data from the input picture. The convolution output results are then merged and concatenated to create a more comprehensive feature map, resulting in an enhanced image representation. This not only significantly expands the breadth of the network, but it may also serve as a substitute for manually picking the filter type in a convolutional layer or deciding whether to set convolutional and pooling layers.

The Size-Aware Parallel Residual network architecture suggested in this paper is a deep and intricate structure composed of linked modules inspired by the concept of the Inception module (Szegedy, 2016). Each module has several convolutional and pooling layers specifically designed to extract distinct characteristics from the input picture. The Inception module consists of two main components: decomposed convolution and batch normalisation. Decomposed convolutions use a blend of convolutional filters with varying kernel sizes to extract characteristics from the input picture. 1x1 convolutional filters are used to decrease the input dimensionality, while high-latitude convolutional filters are utilised to extract more intricate features from the input image. Batch normalisation is a technique that helps to stabilise the training process and mitigate the problem of internal covariate shift. Internal covariate shift refers to the changes in the distribution of network inputs that occur during training.

2.3 Attention

While convolutional neural networks possess a robust capacity for nonlinear expression, in cases when the information is exceedingly complicated, it becomes imperative to construct a more intricate network model in order to get a more potent expression ability. To simplify the model, the attention mechanism may enhance the neural network's capacity to process information by mimicking the way the human brain handles excessive data. During face expression identification tasks, the gathered photographs are often categorised into distinct classification outcomes as a consequence of varying shooting angles. An effective approach is to enhance the network's sensitivity to angles by including a spatial attention mechanism in the network design. A spatial transformation neural network (Zhang, 2023) has the ability to convert different types of deformed data in space and automatically identify the features of significant areas. This module guarantees that the resulting picture after performing cropping, translation, or rotation operations will be identical to the original image before the operations were applied.

3 ROTATION-AWARE RESIDUAL NETWORK (RARN)

3.1 Overview

The RARN builds upon the ResNet framework for facial expression classification, leveraging both high-level and low-level image features while integrating an angle-sensitive attention mechanism. Illustrated in Figure 2, the network architecture comprises a multi-scale fusion module and an angle-sensitive spatial attention module. The former extracts features from input images using a combination of down-sampling, up-sampling, and lateral connections, facilitating the fusion of low-level and high-level information for comprehensive feature representation. Meanwhile, the angle-sensitive spatial attention module enhances feature extraction by incorporating angle information into the feature map, allowing for adaptive feature weighting tailored to different facial expressions. This innovative approach enables RARN to capture both global and local characteristics more effectively, thereby enhancing its ability to classify facial expressions accurately.

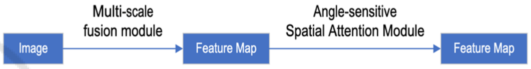

Figure 2: Rotation-Aware Residual Network (Photo/Picture credit: Original).

3.2 Unit for Fusion on Multiple Scales

The unit for fusion on multiple scales integrates micro-expression features into the generalized facial expression recognition process. Specifically, it incorporates low-level characteristics to detect subtle changes in expression, as relying solely on high-level features may overlook them. Typically, low-level characteristics offer limited semantic information but provide precise physical position details. On the other hand, high-level characteristics contain rich semantic details but lack complete spatial position information. The module aims to optimise the utilisation of global features by integrating low-level and high-level features.

To mitigate the potential issue of gradient vanishing in the deep feature extraction network, this study employs Resnet as the feature extraction network for the detection network. After a thorough experimental comparison, it is determined that Resnet152 has a superior detection effect, and the time loss is also within an acceptable range. Therefore, Resnet152 is chosen as the backbone network. Comprising three steps - downsampling, upsampling, and lateral connection - as depicted in Figure 3, the multi-scale fusion module orchestrates the integration process.

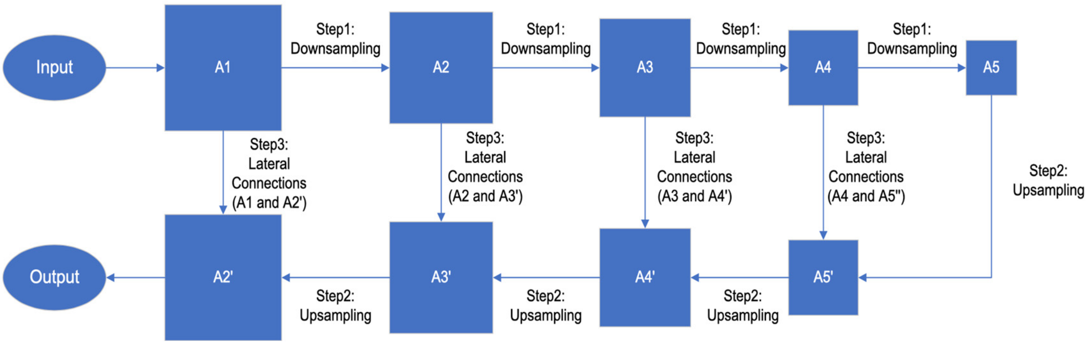

Figure 3: Multi-scale fusion module (Photo/Picture credit: Original).

The downsampling process involves utilizing the Residual Module architecture to extract feature maps from five layers of varying depths, starting from the bottom and progressing upwards, with a scaling factor of two. This phase establishes a feature hierarchy comprising feature maps of diverse sizes. Considering that the bottommost layer of each stage has the most durable traits, the output from the last layer of each stage is chosen for further operations. Downsampling allows the network to retrieve feature maps, transitioning from high-level to low-level. This enables the network to capture both the overall characteristics of the expression as well as the subtle variations in micro-expressions for each face.

To fully use these feature maps, this study uses upsampling and lateral connection techniques to effectively merge the features at each layer. Specifically, starting with the fifth layer feature map A5, the feature map A5 is upsampled to match the size of the fourth layer feature map A4. Subsequently, the newly generated feature map A5' and the feature map from the fourth layer A4 are joined in a lateral manner to produce a novel feature map. Similarly, up sampling this new feature map to match the dimensions of the feature map A3, which outputs a new feature map A4'. And then laterally connect A4' and A3. This iterative procedure continues until all feature maps have been computed. By combining high-level and low-level features, the network is able to effectively capture both global and local features without experiencing overfitting or underfitting. Furthermore, this strategy does not adversely impact the performance of the model.

3.3 Angle-Sensitive Spatial Attention Module

The angle-sensitive spatial attention module comprises six distinct phases, as seen in Figure 4.

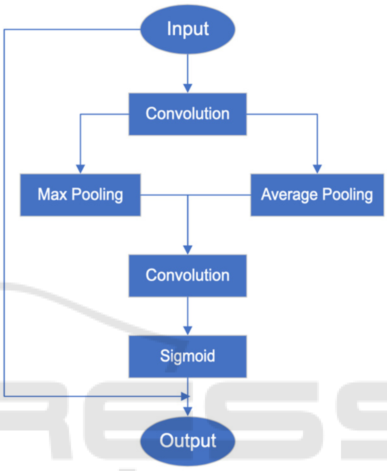

Figure 4: Angle-sensitive Spatial Attention Module (Photo/Picture credit: Original).

First, the input feature map is convolutionally transformed into the Angle feature map, which is subsequently subjected to average pooling and max-pooling across column channels. In the spatial dimension, global maximum pooling and global average pooling reduce the size of the feature maps. Pooling at various levels results in more intricate high-level characteristics being recovered. More precisely, the process of global average pooling allows for the inclusion of each individual pixel on the feature map, while global max pooling allows for the identification of the location in the feature map with the highest response during gradient backpropagation.

Next, the two graphs generated in the previous phase are joined along the channel axis. Then, a convolutional network is used to enhance the capacity for nonlinear expression. Subsequently, the feature maps undergo normalisation via a Sigmoid function to derive weights for the channel characteristics, resulting in the angle-sensitive spatial attention weight map. Ultimately, the acquired angle-sensitive spatial attention weight map is multiplied by the input feature map to allocate distinct attention weights to the feature map based on the spatial angle.

4 EXPERIMENTS

4.1 FER-2013 Dataset

The FER-2013 Dataset was first shown at the Facial Expression Recognition Challenge at the ICML 2013 session (Goodfellow, 2013). The dataset comprises 35,887 grayscale photographs of faces with dimensions of 48x48 pixels, as shown in Figure 5.

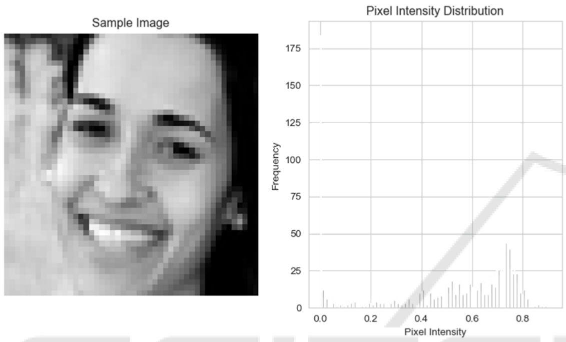

Figure 5 Sample Image & Pixel Intensity Distribution (Photo/Picture credit: Original).

The majority of these photos have been mechanically aligned to ensure that the faces are roughly centred and occupy a similar amount of space in each image. A comparison of the training and test datasets' label distributions is shown in Figure 6.

The training sample consists of two columns, namely "emotion" and "pixel". The Emotion column categorises each face into one of seven distinct categories depending on the specific emotion shown in the facial expression. These categories are as follows: anger, disgusted, fearful, happy, neutral, sad, and surprised. The "pixels" column comprises a string of characters surrounded by quotation marks for each picture. The test sample only consists of the "pixel" column.

4.2 Image Pre-Processing

This study employs many popular data augmentation techniques on the training dataset to generate new and different pictures from the original photos and enhance the performance and robustness of Size-Aware Parallel Residual. These include various modifications such as rotation, mirroring, and cropping, along with brightness, contrast, and colour adjustments. Engaging in this strategy mitigates overfitting and enhances the generalisation capacity of models trained on image data. Table 1 presents the data augmentation settings used for the FER-2013 Dataset in this study.

Table 1: Image Augmentation Parameters.

Parameter	Value
Horizontal Flip	1
Vertical Flip	1
Random Grayscale	0.2
Height Shift Range	0.28
Width Shift Range	0.75
Rotation Range	90
Colour Jitter	brightness=0.2, contrast=0.2, hue=0.2
Normalisation	1

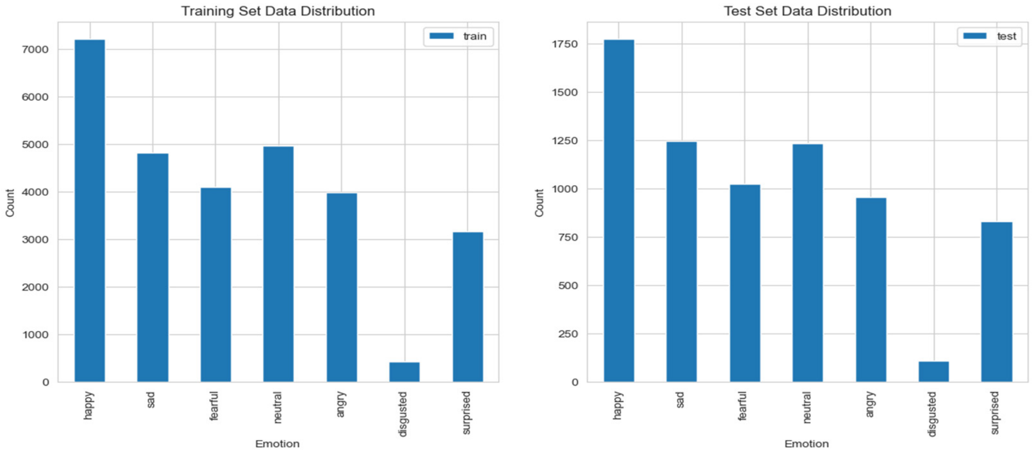

Figure 6: Training and Test Set Data Distribution (Photo/Picture credit: Original).

4.3 Training

The network is trained using the PyTorch and Torchvision libraries in Python for 33 epochs. The batch size is 64. The optimisation uses the SGD optimiser with an initial learning rate of 0.01 and a decay rate of 0.9 after three epochs of repetition. The loss function employed is categorical cross-entropy, calculated using the formula $\frac{-1}{N}\sum_{i=1}^{N} logp_{model}\left(y_i \in c_{yi}\right)$. In this calculation, $p_{model}\left(y_i \in c_{yi}\right)$ represents the chance that a picture y_i belongs to category C_{yi}. All tests are conducted under the same running environment, as shown in Table 2.

Table 2: Running Environment.

System	Ubuntu 22.04
CPU	AMD Ryzen 9 5900HS
Memory	32GB
GPU	NVIDIA RTX 3060

4.4 Evaluation

Figure 7 illustrates the precision of the model across both the training and test datasets. It vividly portrays the model's progressive convergence, culminating instability on the test set, ultimately achieving a final accuracy of 57.51%. By observing the changes in the test curves of the two graphs, it becomes evident that the model has reached a state of stability by the 25th epoch of training. Meanwhile, as the model is further trained, the evaluation accuracy and loss are correspondingly enhanced, although the training accuracy is constantly improving. This indicates that the proposed model has not entered an overfitting state and can stimulate the best performance of the model.

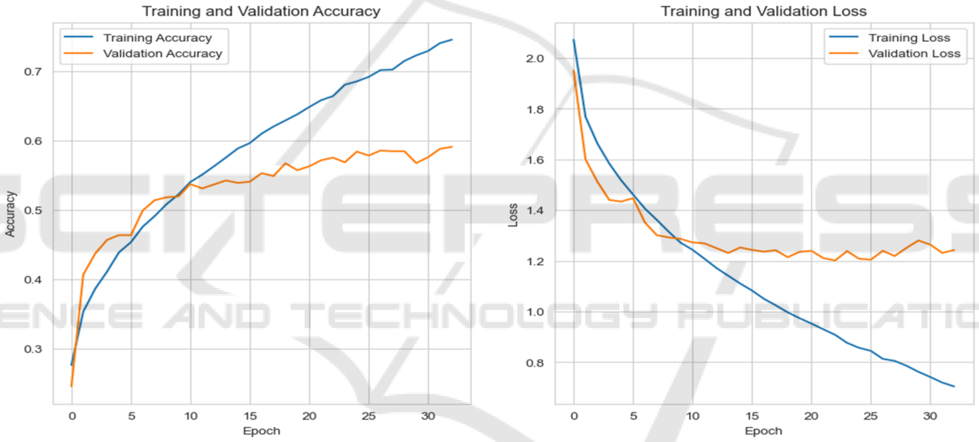

Figure 7: Accuracy & Loss Distribution (Photo/Picture credit: Original).

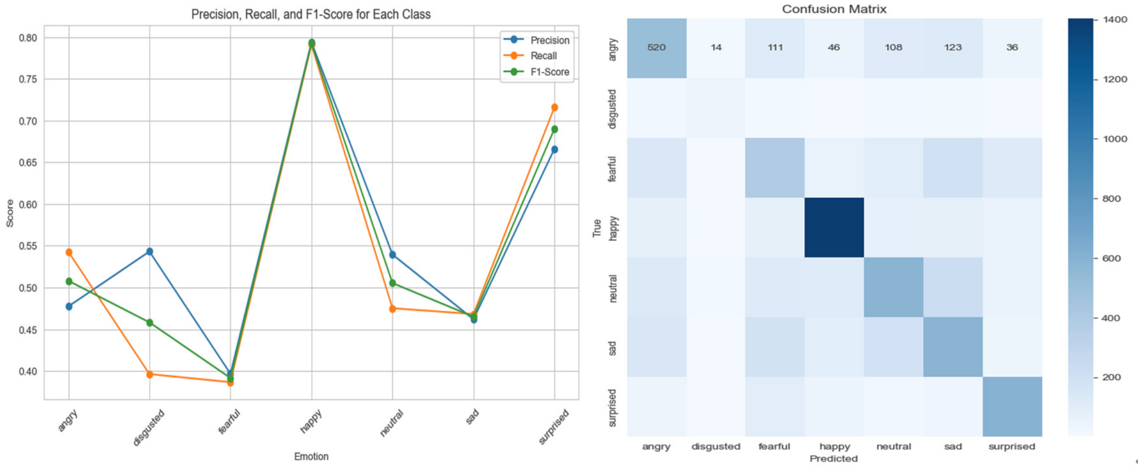

Figure 8: Classification Report (Photo/Picture credit: Original).

The detailed statistics presented in Figure 8 meticulously evaluate the accuracy of the forecast outcomes. It's noteworthy that the training efficacy of the network model is perceptibly impacted by the quantity of training datasets. The category of disgust has the smallest dataset, which explains the reason why it has the most notable disparity in terms of Precision, Recall, and F1-Score. The two classes that showed the most significant discrepancies in the effect of model predictions on the test set are the fearful and happy classes. Based on this outcome, it can be deduced that some expressions exhibit notable variations in prediction outcomes as a consequence of their intricacy. Hence, investigating the network structure with the explicit goal of identifying a certain kind of expression might be regarded as a prospective area of study.

To validate the improved classification performance of the proposed network, this research also conducts a comparative analysis of many popular classification neural networks, including InceptionNet (Szegedy, 2017) and MobileNet (Howard, 2017). Table 3 demonstrates that when using the same training settings and environment, RARN outperforms other models in terms of obtaining convergence and producing a final model with greater accuracy. RARN enhances the accuracy of the model's recognition rate while only requiring a minimal amount of parameters. This demonstrates that RARN guarantees the performance of the network while also assuring the benefits of operational efficiency. RARN achieved an Accuracy of 57.51%, with gains of 0.54% and 21.17% compared to InceptionNet and MobieNet

Table 3: Comparison of Accuracy.

Network	Accuracy
RARN	57.51%
InceptionNet	56.47%
MobielNet	36.34%

5 CONCLUSIONS

This study presents a comprehensive facial expression categorization technique that harnesses attention mechanisms and deep learning. The approach integrates a multi-scale fusion module and an angle-sensitive spatial attention module to drive the classification function. While the multi-scale fusion module captures both global and specific characteristics of the input image, the angle-sensitive spatial attention module enhances feature mapping by incorporating angle information. Experimental results showcase the method's superior recognition rate and substantial improvement in facial expression categorization. Future research endeavours will delve into refining network structures, exploring parameters like convolution core size and step size, and further defining network levels. Additionally, the inclusion of more extensive datasets will enhance the evaluation of the network's performance.

REFERENCES

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).

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on artificial intelligence (Vol. 31, No. 1).

Zhou, J., Xiong, Y., Chiu, C., Liu, F., & Gong, X. (2023). Sat: Size-aware transformer for 3d point cloud semantic segmentation. arXiv preprint arXiv:2301.06869.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. Science, 164(3875), 86-88.

Arora, S., Bhaskara, A., Ge, R., & Ma, T. (2014, January). Provable bounds for learning some deep representations. In International conference on machine learning (pp. 584-592). PMLR.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., ... & Tang, X. (2017). Residual attention network for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conf. on Computer Vision and Pattern Recognition (pp. 1-9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conf. on computer vision and pattern recognition (pp. 2818-2826).

Zhang, X., Liu, C., Yang, D., Song, T., Ye, Y., Li, K., & Song, Y. (2023). Rfaconv: Innovating spatial attention and standard convolutional operation. arXiv preprint arXiv:2304.03198.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Bengio, Y. (2013). 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 berlin heidelberg.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.