

## Image Pre-processing Effects on Attention Modules in Facial Emotion Recognition

H. Arabian\*, V. Wagner-Hartl\*, \*\* K. Moeller\*

\* *Furtwangen University, Institute of Technical Medicine (ITeM), VS-Schwenningen 78054, Germany*  
(e-mail: [h.arabian@hs-furtwangen.de](mailto:h.arabian@hs-furtwangen.de)).

\*\* *Furtwangen University, Campus Tuttlingen, 78532 Tuttlingen, Germany*}

**Abstract:** Neural Networks have achieved a reputation for their ability to outperform traditional machine learning algorithms. The robustness of network models to unseen data is important. In this study the implementation of an attention module to an existing network architecture is studied for robustness improvement with extensive image pre-processing for facial emotion recognition, a component in a closed loop feedback system being developed for the emotional training of people with autism spectrum disorder. The squeeze and excitation (SE) attention module was selected and combined with the network architecture of VGG16. The performance of the model, trained with the OULU-CASIA dataset, was based on statistical data generated from validation accuracies with a 10-fold partition, while the robustness analysis was based on data generated from cross database prediction accuracies. The visualization technique of Grad-CAM was used to observe the regions of impact on classification. Validation accuracies of up to 98.39% were achieved and generalization performance was 52.43% on the FACES, and 26.10% on the JAFFE datasets. The improvements of the SE over the base model were observed in the visualizations of the validation datasets. Data analyzed showed that extensive image pre-processing diminished effects of SE in relation to model robustness.

Copyright © 2024 The Authors. This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

**Keywords:** Attention modules, Autism spectrum disorder, Deep learning, Model robustness, Prediction visualizations.

### 1. INTRODUCTION

Artificial Intelligence (AI) has been a hot topic over recent years. The popularity of AI has grown, and its use incorporated into daily life, from the simplest applications in leisure of photo capturing to advanced financial and security applications. Neural Networks have shown they outperform traditional machine learning algorithms, however, they are still recognized as “Black Boxes” due to their behind the scenes decision making processes (Samek et al., 2017). The robustness of network models to unseen data is important to provide accurate results.

One application in the field of medicine is emotion identification for the use in therapeutic intervention e.g. emotion training for children with autism to help cope better in social interactions (Golan et al., 2010; Yuan & Ip, 2018). The use of emotion recognition is currently being studied for the incorporation in a closed loop feedback system to help in the treatment of people suffering from autism spectrum disorder (ASD). ASD is a neurological developmental disorder that affects the social relations, communications, and behaviors of individuals (Committee on Children With Disabilities, 2001). It is estimated that one out of fifty-nine people are affected by ASD (Rylaarsdam & Guemez-Gamboa, 2019). The challenges of dealing with ASD increases on the individual when there is a deviation from a state of routine e.g. stay at home instructions and returning from lockdown measures, this also adds stress on the caregivers (Lugo-Marin et al., 2021).

Facial emotion recognition (FER) was chosen as the identification method for emotions as it was observed that 55% of a person’s emotions are recognized through facial expressions (Mehravian, n.d.). A virtual world tailored to the individual combined with a novel reward system that rewards emotion expressions, provides a suitable atmosphere for treatment. The system could also be beneficial in improving mental health given the stresses caused by the pandemic (Wind et al., 2020). In (Voss et al., 2019) a clinical trial performed showed that there was improvement in the socialization of children with ASD. (Ravindran et al., 2019) conducted a pilot study with a closed loop virtual environment achieving encouraging results. These studies highlight the benefits of such a system in supporting treatment of patients suffering from ASD.

There have been many studies highlighting the importance of model robustness and methods created to improving network representations (Carlini & Wagner, 2017; Hendrycks & Dietterich, 2019; Hu et al., 2018; Raj et al., 2020; Woo et al., 2018). One of the most recent methods has been the introduction of attention modules. The concept behind these modules is to improve the decision-making process of a neural network during the training phase.

In (Hu et al., 2018) tests conducted on ImageNet 2012 (Russakovsky et al., 2015) dataset using architectures of VGGNet (Simonyan & Zisserman, 2015) and different residual networks showed that attention incorporated networks outperformed baseline models. In (Woo et al., 2018) an

attention module extending the concept of squeeze and excitation (SE) module was developed and tests conducted on the ImageNet 1K (Deng et al., 2009) database using ResNet50 (He et al., 2016) as the base architecture showed better performance over the base and SE incorporated models. In (Ling et al., 2019) two attention modules were developed and tests using residual networks revealed better performance, on the respected datasets, against the base model. In (Fukui et al., 2019) the attention branch network (ABN) was introduced, results showed that ABN combined architectures improved performance over base models. In this study the implementation of the squeeze and excitation (Hu et al., 2018) module has been analyzed for the robustness improvement of a network model.

The aim of this study is to show that extensive image pre-processing diminishes the effects of the attention modules in improving robustness of a network model.

## 2. METHODS

### 2.1 Methodology

Input images from three different datasets of OULU-CASIA (Zhao et al., 2011), FACES (Ebner et al., 2010) and Japanese Female Facial Expressions (JAFFE) (Lyons et al., 2020) were first processed by a segmentation algorithm developed in (Arabian et al., 2021), where the image is cropped to allow the appearance of the face of the subject without background noise. The base convolutional neural network (CNN) model of VGG16 (Simonyan & Zisserman, 2015) architecture was used in combination with two initial weight settings. Two network architectures were used in this study, the first was the base model and the second a combination of VGG16 architecture with the SE attention module. Both models were trained on the two initial weight settings. The first setting utilized the weights of the VGG16 architecture trained on the ImageNet (Deng et al., 2009) dataset i.e. transfer Learning, while the second was a random initialization of weights i.e. training from scratch. The models were trained with the OULU-CASIA (Zhao et al., 2011) dataset partitioned according to a 10-Fold cross validation scheme so that each image is included during training and testing phases.

The datasets of FACES (Ebner et al., 2010) and JAFFE (Lyons et al., 2020) were used for cross dataset evaluation, to evaluate the robustness of the models. The mean of the predicted validation set of the OULU-CASIA (Zhao et al., 2011) across the validation scheme was chosen as the model performance. As a supplementary class existed in the FACES (Ebner et al., 2010) and JAFFE (Lyons et al., 2020) datasets, this data was removed from further processing to have a fair comparison of the model robustness. To interpret the trained model's focus area for classification, the visualization technique of gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) was used.

A separate study was performed on the original images without pre-processing to have a fair comparison, with the same methodology and model architectures.

### 2.2 Attention modules placement and training parameters

Attentions modules are a relatively new concept. They were designed to improve the performance of network models by making them focus on informative features more and diminishing the effects of less informative ones (Hu et al., 2018). The squeeze and excitation (SE) module was developed to enhance a CNNs representation power by calculating the interdependencies between the different channels at different stages in the architecture during training (Hu et al., 2018). Attention modules have displayed improvement from base models without significantly adding to the computational demand.

The SE modules are simple in structure and implementation. They are designed to enhance the representation power of the model by refining channel wise features without significantly affecting the computational performance. The input features are first squeezed by applying global average pooling (GAP) and then passed through gating mechanisms by means of two fully connected (FC) layers. A reduction coefficient (R) is applied for the first FC layer, for this study the R value was set to 16, a Rectified Linear Unit (ReLU) activation followed. In the second FC layer the features are scaled up to the dimensions of the GAP, this is referred to as the excitation phase. The features then pass through a Sigmoid activation function, after which they are multiplied element wise with the input features to produce refined features as output (Hu et al., 2018). Figure 1 represents the SE module architecture.

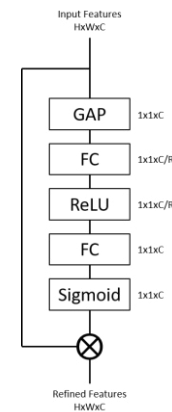


Figure 1. SE model architecture. H, W and C represent the Height, Width and Channel dimensions respectively.

The SE module was adopted into the VGG16 architecture at three different stages of the architecture. The placements were made following the ReLU activation function of each of the convolutional blocks of 2, 3 and 4 before the pooling operation with the same layers and settings from (Hu et al., 2018). The last 3 layers of both models were altered to fit the task at hand such that the Fully Connected layer (FC) was set to 6 classes followed by a Softmax activation function then a classification layer.

### 2.3 Database Description

The Oulu-CASIA database is composed of image sequences from 80 subjects, with varying age, gender, and ethnicity. The subjects are expressing 6 basic emotions of Anger, Disgust, Fear, Happiness, Sadness and Surprise (Zhao et al., 2011). The

images were captured under three illumination settings and two capturing techniques of visible light and near infra-red. For this study the image sequences of original RGB, with visible light and strong illumination lighting were selected. The dataset selected consisted of 10,379 images in total.

The Japanese female facial expressions (JAFPE) (Lyons et al., 2020) database is made up of 213 facial portrait images of 10 different Japanese female students. The images are represented in grey scale color format. The subjects express 7 emotions (six basic emotions plus neutral). The FACES (Ebner et al., 2010) dataset is a collection of facial portraits from subjects of varying age and gender. It contains a total of 2,052 images expressing 6 emotion classes of anger, disgust, fear, happiness, sadness and neutral.

#### 2.4 Image pre-processing

To highlight the face of the subject in the image and reduce background noise, an image pre-processing algorithm was developed. The pre-processing model utilizes the object detection algorithm of (Viola & Jones, 2001) in combination with the approach from (Ojala et al., 2002) to first segment the face region. After which the model was designed such that the left eye and mouth regions are located by implementing the approach of (Castrillón et al., 2007), then these locations were used as guidelines to further crop the face image. The model was designed such that if any region failed to be detected the image was re-moved from further analysis.

#### 2.5 Performance criteria

Training parameters of the model were selected according to the details in Table 1. The data was split according to the 10-fold cross validation scheme, i.e. the data is distributed into 10 different combinations of training and validation set. The 10-fold provides 90% of the data for training and 10% for the validation at each distribution i.e. fold. The training and validation sets are different at every fold so that all the data is used for the analysis of the network model. The mean of true positive accuracies predicted from the validation set were used. The statistical data of mean, standard deviation (SD), Median (Med), Median Absolute Deviation (MAD), Inter Quartile Range (IQR), Upper 95% (UCI) and Lower 95% Confidence Intervals (LCI) from the 10-fold scheme was used as the performance criteria of the model and study approach.

A threshold value was also set for the models, such that if any model validation set accuracy did not achieve this threshold, then the model was excluded from further evaluation of robustness and generalization ability. A threshold of 35% was selected, taking into consideration the uniform distribution of the six classes, with each class making up around 17% of the total data.

The Grad-CAM (Selvaraju et al., 2017) visualization technique was used to analyze the regions of impact for the classification. The visualizations were performed on the ReLU of the last convolution block in the architecture the “relu5”. The Grad-CAM feature maps were extracted for each image of the validation set, then the mean of the maps across the class was used for the analysis. This was done to see the region of impact focus for each class as the image pre-processing

restricts the images to a certain form that is common across all the classes.

**Table 1. Model training options.**

Parameter	Method / Value
Solver	Stochastic gradient descent with momentum (SGDM)
Mini Batch Size	50
# of Epochs	150
Shuffle	Every-Epoch
Initial Learning Rate	0.0001
Momentum	0.9

### 3. EXPERIMENTAL RESULTS

#### 3.1 Data selection and distribution

Table 2 shows the different distribution of the images into their respective classes for each of the three databases of OULU-CASIA, FACES and JAFPE. After performing the extensive image pre-processing on the original images, the algorithm excluded 10.41%, 8.77% and 1.41% the images from further processing for each of the three datasets of OULU-CASIA, FACES and JAFPE respectively, due to the inability of the method to correctly segment the regions prescribed. The images of the neutral class were removed from the analysis procedure. The distribution of the images into the relative classes is near uniform with the range remaining in less than a 3% margin.

**Table 2. Database image distribution per class before and after pre-processing.**

Class	OULU-CASIA		FACES		JAFPE	
	Original	Pre-Processed	Original	Pre-Processed	Original	Pre-Processed
Ang	1790	1538	342	292	30	30
Dis	1633	1425	342	292	29	29
Fear	1796	1734	342	303	32	32
Hap	1791	1725	342	338	31	31
Sad	1668	1445	342	317	31	31
Sur	1701	1432	0	0	30	27
Total	10379	9299	1710	1542	183	180

#### 3.2 Model performance

Four different models were trained and analyzed. The base network architecture was the same for all the four models with the change in the initial weight settings between 2 of the models and the incorporation of the attention modules in the other 2 models. The results from the original image experiments, without image pre-processing, did not reach the threshold mark. The findings from those experiments are not mentioned in this report and limited to the previous statement.

The detailed results shown hereafter are from modelling using extensive image pre-processing on the input images. Table 3 shows the different statistical results of the 4 models on the OULU-CASIA validation dataset. As can be observed from the results, the base models performed slightly better than the attention integrated models in both initialization methods. All

models showed good performance, reaching accuracy results above a 90% threshold in the 10-fold cross validation scheme. The mean of the data was observed to be near the median. SE integrated models showed a higher variance than the base model.

**Table 3. Model accuracy performance results (%) on OULU-CASIA validation dataset.**

	Transfer Learning		Training from Scratch	
	VGG16	VGG16-SE	VGG16	VGG16-SE
Mean $\pm$ SD	99.49 $\pm$ 0.17	99.11 $\pm$ 0.42	98.96 $\pm$ 0.24	98.39 $\pm$ 0.27
Median $\pm$ MAD	99.52 $\pm$ 0.16	99.19 $\pm$ 0.27	98.98 $\pm$ 0.16	98.39 $\pm$ 0.27
IQR	0.32	0.65	0.32	0.54
LCI $\pm$ SD	99.37 $\pm$ 0.12	98.81 $\pm$ 0.29	98.78 $\pm$ 0.17	98.19 $\pm$ 0.19
UCI $\pm$ SD	99.62 $\pm$ 0.31	99.41 $\pm$ 0.77	99.13 $\pm$ 0.44	98.58 $\pm$ 0.50

**Table 4. Model accuracy robustness results (%) using FACES dataset.**

	Transfer Learning		Training from Scratch	
	VGG16	VGG16-SE	VGG16	VGG16-SE
Mean $\pm$ SD	43.93 $\pm$ 5.47	42.84 $\pm$ 6.40	54.75 $\pm$ 4.43	52.43 $\pm$ 3.87
Median $\pm$ MAD	44.29 $\pm$ 3.08	40.73 $\pm$ 5.61	54.86 $\pm$ 3.24	51.88 $\pm$ 1.95
IQR	6.16	11.74	5.64	5.51
LCI $\pm$ SD	40.01 $\pm$ 3.76	38.26 $\pm$ 4.40	51.59 $\pm$ 3.05	49.66 $\pm$ 2.66
UCI $\pm$ SD	47.85 $\pm$ 9.99	47.42 $\pm$ 11.68	57.92 $\pm$ 8.08	55.20 $\pm$ 7.06

**Table 5. Model accuracy robustness results (%) using JAFFE dataset.**

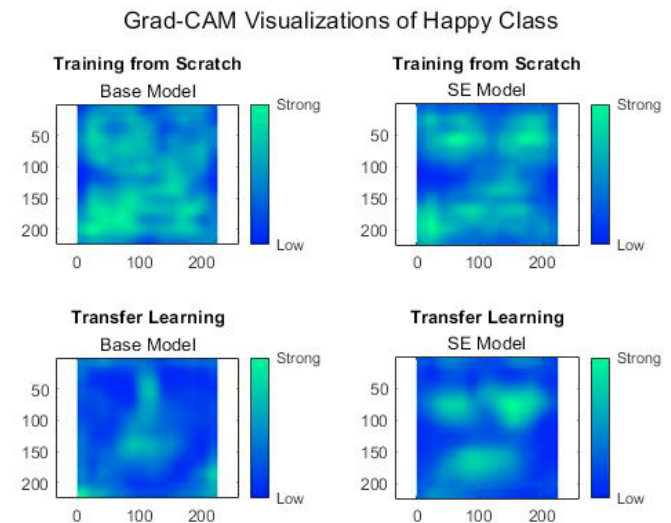
	Transfer Learning		Training from Scratch	
	VGG16	VGG16-SE	VGG16	VGG16-SE
Mean $\pm$ SD	32.94 $\pm$ 4.82	26.33 $\pm$ 5.59	29.72 $\pm$ 6.84	30.44 $\pm$ 1.41
Median $\pm$ MAD	32.50 $\pm$ 2.78	25.83 $\pm$ 7.14	30.56 $\pm$ 6.39	30.28 $\pm$ 0.83
IQR	6.11	10.29	12.22	2.22
LCI $\pm$ SD	29.50 $\pm$ 3.31	22.34 $\pm$ 3.84	24.83 $\pm$ 4.70	29.44 $\pm$ 0.97
UCI $\pm$ SD	36.39 $\pm$ 8.79	30.33 $\pm$ 10.20	34.61 $\pm$ 12.48	31.45 $\pm$ 2.57

Table 4 showcases the results from the cross-dataset validation using the FACES dataset on the four different models. The models were not able to generalize well on the FACES dataset, achieving at best a 54.75% mean accuracy at the base model architecture with scratch training. The variations of the base models were also lower than that with the attention module

integrated. The range of the data distribution also varied between the attention and base models suggesting that the models were not generalizing well to unseen data.

Table 5 highlights the results obtained from the cross-dataset validation using the JAFFE dataset. The results show that the base models performed slightly better and in one case lower than the attention integrated model. The confusion matrices were also investigated to view the misclassifications. The models misclassified most of the categories of the JAFFE dataset into the Surprise class which holds around 15% of the total number of images.

In Fig. 2 the mean Grad-CAM visualizations of the four different models on the happy class are represented. As observed from the figure the regions that have a high influence on the classification process are those of a light blue color.



**Figure 2. Mean Grad-CAM visualizations of the different models on the happy class.**

#### 4. DISCUSSION

The results obtained showed that the attention modules did not have a significant difference from the base model mean accuracies, when tested on pre-processed images. The variations of the SE models were also higher than that of the base models, highlighting the increase in the range of distribution of the model accuracies.

To better highlight the generalization ability or robustness of the different models, they were tested with the images of the FACES, and JAFFE datasets. The data also showed the further lack in robustness of the models when dealing with different image sources and of different color schemes. The models were able to achieve accuracies greater than a threshold of 35% for the FACES dataset suggesting that it was able to distinguish between the different classes but lacked in robustness. This also signified that the models were not extracting the relevant feature patterns that are important for FER, but rather was dependent on the color combination of the images.

To get a better understanding of the areas of focus for the network models the prediction visualizations were calculated

and reviewed. From the visualization it is noticed that the base model VGG16 trained from scratch performed the worst as it was not able to identify the key areas that are of importance for FER, which are the eyebrows, eye, and mouth regions. The attention integrated model VGG16-SE was able to identify and focus on those regions more, thereby improving classification and in term robustness. This highlights the ability and importance of using attention modules to focus on more informative features and diminish non useful ones. Even though the statistical data suggested otherwise, attention module implementation is advised for network architectures so that the areas of focus are more related to the features of relevance.

The transfer learning models did perform slightly better in a visual comparison perspective. The VGG16 base model had less disturbances but lacked the essential components of the identification. The VGG16-SE was able to focus on the regions of importance more than the base model as noticed from Fig. 2. However, a more detailed look into the filters revealed the necessity for further training of the scratch models since no concrete pattern was observable.

When analyzing the data from the original image study, without image pre-processing, the threshold value set was not achieved and therefore robustness evaluation was not performed, and results not mentioned in this report. The study showed the weakness of the models to background noise, as it classified a significant portion of the data into two classes only as noticed by the confusion matrices. The prediction visualizations also revealed that they were biased to background noise rather than focusing on facial features for correct classification which explains the low model performance and bias to the two classes.

#### 4.1 Study limitations

A few limitations were accounted for during this study, such as the removal of the extra class from the databases thereby reducing the image count. The use of parameters, suggested by the SE reference paper, without performing an ablation study that is specific to the task of FER. The bias of having an extensive image pre-processing step.

### 5. CONCLUSION

In this study the effects of image pre-processing on attention integrated network models were analyzed. The results showed that the addition of the SE module did not influence the mean accuracy of the models, where only a small variation of up to 2% in the accuracies was noticed when using pre-processed images. The attention module however was able to concentrate on more informative areas in the image than that of the base models, this how-ever failed to improve the generalization ability or robustness of the models to un-seen images as was noticed from the cross-database evaluations. Therefore, the integration of attention modules into network architectures showed better representation ability which is important in classification tasks.

Further research is required to enhance and improve the robustness of the models using attention integrated networks. The influence of the weight initialization schemes on the

performance as well as different combinations of the three datasets for training purposes with data augmentation are also currently being studied to obtain a model that can generalize well to unseen data which is critical in high-risk applications.

### ACKNOWLEDGEMENT

This research was partially funded by the German Federal Ministry of Research and Education (BMBF) under grant LESSON FKZ: 3FH5E10IA, a grant from KOMPASS funded by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) of Baden-Wuerttemberg Germany, a grant from the ERAPERMED2022-276—ETAP BMG FKZ 2523FSB110, and a DAAD grant AIDE-ASD FKZ 57656657.

### AUTHOR'S STATEMENT

The authors declare no conflict of interest.

### REFERENCES

- Arabian, H., Wagner-Hartl, V., Chase, J. G., & Möller, K. (2021). Facial Emotion Recognition Focused on Descriptive Region Segmentation. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3415–3418.
- Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. <https://doi.org/10.1109/SP.2017.49>
- Castrillón, M., Déniz, O., Guerra, C., & Hernández, M. (2007). ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, 18(2), 130–140.
- Committee on Children With Disabilities. (2001). The Pediatrician's Role in the Diagnosis and Management of Autistic Spectrum Disorder in Children. *Pediatrics*, 107(5), 1221–1226. <https://doi.org/10.1542/peds.107.5.1221>
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42, 351–362.
- Fukui, H., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10705–10714.
- Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V., & Baron-Cohen, S. (2010). Enhancing Emotion Recognition in Children with Autism Spectrum Conditions: An Intervention Using Animated Vehicles

- with Real Emotional Faces. *Journal of Autism and Developmental Disorders*, 40(3), 269–279. <https://doi.org/10.1007/s10803-009-0862-9>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv Preprint arXiv:1903.12261*.
- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.
- Ling, H., Wu, J., Wu, L., Huang, J., Chen, J., & Li, P. (2019). Self residual attention network for deep face recognition. *IEEE Access*, 7, 55159–55168.
- Lugo-Marín, J., Gisbert-Gustemps, L., Setien-Ramos, I., Español-Martín, G., Ibañez-Jimenez, P., Forner-Puntonet, M., Arteaga-Henríquez, G., Soriano-Día, A., Duque-Yemail, J. D., & Ramos-Quiroga, J. A. (2021). COVID-19 pandemic effects in people with Autism Spectrum Disorder and their caregivers: Evaluation of social distancing and lockdown impact on mental health and general status. *Research in Autism Spectrum Disorders*, 83, 101757. <https://doi.org/10.1016/j.rasd.2021.101757>
- Lyons, M. J., Kamachi, M., & Gyoba, J. (2020). Coding facial expressions with Gabor wavelets (IVC special issue). *arXiv Preprint arXiv:2009.05938*.
- Mehrabian, A. (n.d.). Communication Without Words | 15 | v2 | *Communication Theory* | Albert. Retrieved January 17, 2024
- Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971–987.
- Raj, A., Bresler, Y., & Li, B. (2020). Improving robustness of deep-learning-based image reconstruction. *International Conference on Machine Learning*, 7932–7942.
- Ravindran, V., Osgood, M., Sazawal, V., Solorzano, R., & Turnacioglu, S. (2019). Virtual Reality Support for Joint Attention Using the Floreo Joint Attention Module: Usability and Feasibility Pilot Study. *JMIR Pediatrics and Parenting*, 2(2), e14429. <https://doi.org/10.2196/14429>
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211–252.
- Rylaarsdam, L., & Guemez-Gamboa, A. (2019). Genetic Causes and Modifiers of Autism Spectrum Disorder. *Frontiers in Cellular Neuroscience*, 13. <https://www.frontiersin.org/articles/10.3389/fncel.2019.00385>
- Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models (arXiv:1708.08296). *arXiv*. <https://doi.org/10.48550/arXiv.1708.08296>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
- Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition (arXiv:1409.1556). *arXiv*. <https://doi.org/10.48550/arXiv.1409.1556>
- Viola, P., & Jones, M. (2001). 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*, 1, I–I. <https://doi.org/10.1109/CVPR.2001.990517>
- Voss, C., Schwartz, J., Daniels, J., Kline, A., Haber, N., Washington, P., Tariq, Q., Robinson, T. N., Desai, M., Phillips, J. M., Feinstein, C., Winograd, T., & Wall, D. P. (2019). Effect of Wearable Digital Intervention for Improving Socialization in Children With Autism Spectrum Disorder: A Randomized Clinical Trial. *JAMA Pediatrics*, 173(5), 446–454. <https://doi.org/10.1001/jamapediatrics.2019.0285>
- Wind, T. R., Rijkeboer, M., Andersson, G., & Riper, H. (2020). The COVID-19 pandemic: The ‘black swan’ for mental health care and a turning point for e-health. *Internet Interventions*, 20, 100317. <https://doi.org/10.1016/j.invent.2020.100317>
- Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.
- Yuan, S. N. V., & Ip, H. H. S. (2018). Using virtual reality to train emotional and social skills in children with autism spectrum disorder. *London Journal of Primary Care*, 10(4), 110–112. <https://doi.org/10.1080/17571472.2018.1483000>
- Zhao, G., Huang, X., Taini, M., Li, S. Z., & Pietikäinen, M. (2011). Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9), 607–619.