Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 227 (2023) 372–380

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023)

Attention is Everything You Need: Case on Face Mask Classification

Nanda Pratama[a*], Dody Harianto[a], Stefan Filbert[a], Harco Leslie Hendric Spits Warnars[b], Maybin K. Muyeba[c]

[a]*School of Computer Science, Bina Nusantara University, West Jakarta, Jakarta and 11480, Indonesia*
[b]*Computer Science Department, BINUS Graduate Program, Bina Nusantara University, West Jakarta, Jakarta and 11480, Indonesia*
[c]*MIST Consulting, Manchester and M410SA, United Kingdom*

Abstract

Automated face mask classification has surfaced recently following the COVID-19 mask wearing regulations. The current State-of-The-Art of this problem uses CNN-based methods such as ResNet. However, attention-based models such as Transformers emerged as one of the alternatives to the status quo. We explored the Transformer-based model on the face mask classification task using three models: Vision Transformer (ViT), Swin Transformer, and MobileViT. Each model is evaluated with a top-1 accuracy score of 0.9996, 0.9983, and 0.9969, respectively. We concluded that the Transformer-based model has the potential to be explored further. We recommended that the research community and industry explore its integration implementation with CCTV.

Keywords: Face Mask Classification; Convolutional Neural Network; Attention; Transformer; Deep Learning

1. Introduction

CoronaVirus Disease 2019 (COVID-19) is a disease caused by the SARS-Cov-2 virus (severe acute respiratory syndrome coronavirus) [1]. Around 6.82 million people perished from the virus, which infected nearly 670 million people. Particularly in elderly individuals and those with underlying medical issues, this virus can produce severe

* Corresponding author.
E-mail address: nanda.pratama@binus.ac.id

10.1016/j.procs.2023.10.536

symptoms and occasionally result in death. Mild to severe symptoms can include fever, coughing, and breathing problems [2].

Wearing a mask is one of the best strategies to stop the COVID-19 virus from spreading (combined with vaccination) [3]. People who may have the virus but are asymptomatic can avoid spreading it to others by wearing masks. Additionally, they guard against virus-containing droplets being inhaled by the individual wearing the mask. According to [4], wearing a mask effectively prevents virus transmission. As a result, numerous governments and health organizations worldwide have encouraged or even mandated the use of masks in public.

However, human resource shortage makes it challenging to enforce mask usage [5]. Many companies and organizations lack the resources necessary to police the wearing of masks by customers or employees [6]. Additionally, it may be difficult to enforce compliance because some people can refuse to wear masks for various reasons. Because those who do not wear masks can spread the infection to others, it is challenging to contain the spread of the virus. To stop the spread of this virus, it is crucial to inform individuals about the significance of wearing masks and to collaborate [7].

As a result, automation is required to handle mask-wearing in crowded places [8]. When a person is not wearing a mask, AI-related applications will classify and alert the person that is not wearing a mask. This will aid in the enforcement of mask-wearing regulations and the control of COVID-19 spread. It will also alleviate the strain on human resources and make it easier for businesses, organizations, and governments to enforce mask-wearing policies.

Machine learning and, intense learning, have emerged as a viable solution for automating mask-wearing enforcement [9, 10]. These sophisticated techniques can detect faces and masks with high accuracy and can even be trained to recognize different types of masks and variations in facial expressions. This technology can be used in various settings, including public places, office buildings, and even public transportation, to ensure that mask-wearing regulations are followed and to help control the spread of COVID-19 [8].

Researchers are investigating the use of automation for face mask classification. These studies aim to improve mask-detection systems, making them more reliable and widely available [11]. Automated face mask classification can enforce mask-wearing regulations in various settings, including public places, office buildings, and public transportation [9]. The development and deployment of these systems will be critical in the fight against COVID-19.

Deep learning models have received significant attention to improve the automated face mask classification systems. Researchers have been experimenting with different techniques in their research and consequently creating various deep-learning models that may influence these systems' performance.

2. Related Works

Most face mask classification tasks used convolutional neural networks (CNN), including its optimized methods, to solve the problem. CNNs are currently state-of-the-art models that outperform traditional models. [13]. The further development of CNN-based models has resulted in using models with trained layers, known as the pre-trained model since it is more challenging to train deeper neural networks due to high computational cost. One method is to use residual learning, where the layers are treated as residual functions to optimize the neural network more efficiently [17]. ResNet is one of the models based on CNN, which achieves the highest accuracies when tested using the ImageNet [18] database. Other well-known CNN-based models are VGGNet or commonly known as VGG [19] which examine the effect of the depth of the neural network using the most miniature convolutional filters (3x3) that leads to a significant improvement in model performance compared to the usual plain CNN models [16]. To reduce the risk of encountering overfitting, L_2 regularization is used to lower the variance while keeping the bias low between the predictions and the ground-truth data to obtain the best fit.

AlexNet utilizes non-saturating neurons and parallel processing implementation of the convolutional operation. AlexNet is trained for classifying images and can achieve great model accuracy using a smaller number of convolutional layers [20]. The dropout regularization technique reduces overfitting due to high train and test data variance.

Pre-trained CNN models have their expansion as deeper architectures by increasing the number of layers or more comprehensive architectures by increasing the number of trainable parameters to obtain better performance, such as Inception [21], Xception [22], and DenseNet [23]. These models are heavier in complexity and architecture, so a higher computational cost is required, which leads to the innovation for recent researchers to build a smaller and more

efficient CNN model and blend them with other types of models and/or mechanisms such as Attention, first mentioned in [14].

Attention solves text translation problems. It can be defined as knowing one's position concerning others. This mechanism proves to be better than the conventional approaches to the problem. Because of this, the researcher started to dive deep into this mechanism, but the most successful is the Transformer model.

Transformer [30] is the further expansion of recurrent neural networks (RNN) originally proposed for solving NLP problems and text analysis. Transformer consisted of two smaller architectures, an encoder and a decoder. It expands the attention mechanism further by introducing scaled dot-product attention that can be defined by equation (1),

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where matrix Q is the query of the input, K is the key of the input, d_k is the dimension of the matrix K. Matrix V is the values matrix.

Gradually, further developments enable transformers to solve tasks related to computer vision, one of which is image classification. Vision transformers (ViT) [27] use additional large-scale data sources than the standard ImageNet dataset, which allows it to perform better than its counterpart [24].

In recent research, there are models in the form of hybrid architectures which combine both CNN and Transformers to yield a lightweight and general-purpose model that can also be utilized on mobile devices [25]. The model with such configurations as MobileViT can be an interesting comparison with other models, either transformers-based or the de facto CNN pre-trained models.

However, CNN and Transformer have opposing viewpoints on image classification. This is evident in both the CNN model and the Transformer itself. The CNN model categorizes the similarity pattern into stages. The pattern similarity component is arguably a disadvantage of CNNs because it introduces differences in how different layers capture and encode information, potentially making it difficult to interpret and analyze the relationships between features across the network. On the other hand, Transformer has a more uniform similarity structure throughout their layers. The obvious grid-like structure indicates that Transformer's self-attention mechanism can capture pairwise relationships between all patches equally well [13].

3. Proposed Model

Fig. 1. Sample data from dataset [11]

This paper proposes using three different models: Vision Transformer (ViT) [27], Swin Transformer [28], and MobileViT [25], aiming to provide effective solutions. The three models will then be compared [11] to evaluate their performance and capabilities. All models are already pre-trained with ImageNet before. Reference [11] contains photos of masked and unmasked people. The dataset consists of 95,000 images which have two classes, masked (5,000 images) and unmasked (90,000 images). The dataset will be divided with a proportion of 80% data for training, 10% data for validation, and 10% unseen images for final model evaluation. Sample from the dataset can be seen in figure 1.

3.1. Vision Transformer (ViT)

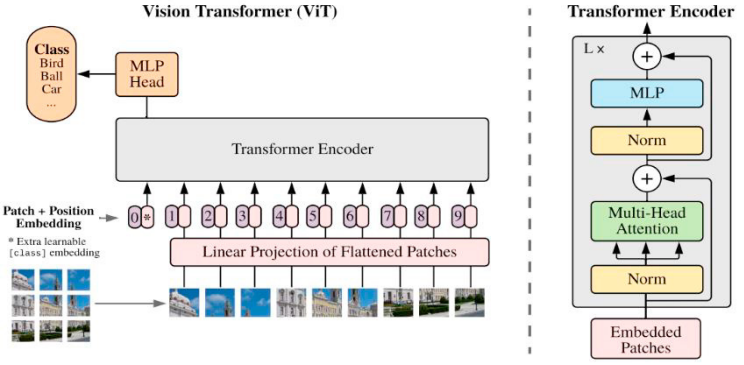

Fig. 2. Left: Architecture of Vision Transformer (ViT) [27]; Right: Encoder architecture in ViT [27]

In Vision Transformer (ViT), the main idea refers to the technique used to weigh the value of different regions in an input image. When producing predictions, it uses a self-attention mechanism to focus on the most relevant portions of the image dynamically. The input image is separated into non-overlapping patches in ViT before turning into high-dimensional embeddings [12]. The embeddings are subsequently processed by multi-head self-attention layers, which allow the model to weigh the value of various visual regions.

ViT is a deep learning architecture based on transformers built explicitly for computer vision tasks. The primary idea behind ViT is to treat the entire image as a succession of tokenized image patches rather than processing each patch individually, as is done with typical Convolutional Neural Networks (CNN). Referring to figure 2, the vision transformer divides the picture input into size patches, embeds each patch size with linear embeddings, positions the embeddings, and feeds the result sequence of vectors to an encoder. It adds a classification token to the sequence for the model to classify.

ViT is available in three models, that is base, large, and giant [27]. Each configuration is represented by 16, 32, and 64, which refers to the size of the image patches. The entire image is separated into patches in the base setup, and each patch is regarded as a token. In contrast, the compact configuration employs down-sampled patches and fewer tokens, resulting in reduced model size and faster calculation. However, this comes at the expense of lower performance than the base setup.

In this case, we are using configuration ViT-S16 from [26]. The ViT-S16 is a miniature version of the Vision Transformer (ViT) model. It is intended for resource-limited applications since it has a smaller model size and faster calculation than the default configuration. Down-sampled patches and fewer tokens are used in the ViT-S16 configuration, resulting in a more compact model. This makes it appropriate for deployment in real-world applications with restricted computational resources. ViT-S16 can also cut training periods and memory needs, giving it a cost-effective option for face mask classification jobs.

3.2. Swin Transformer

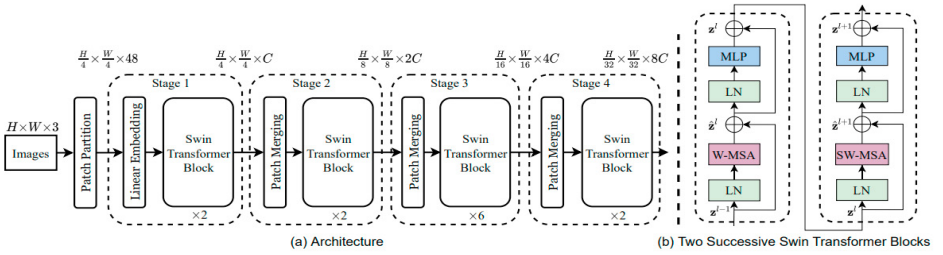

Fig. 3. (a) Architecture of Swin Transformer [28]; (b) Two successive Swin Transformer Blocks [28]

The Swin Transformer [28] attention idea refers to the technique through which the model focuses on distinct sections of the input image to create predictions. The Swin Transformer employs self-attention layers, allowing it to weigh the relevance of various picture regions flexibly and adaptively. The model estimates the attention scores between each patch in the image and all other patches in each self-attention layer. The attention ratings reflect how meaningful each patch is concerning the others, and they are used to generate a weighted sum of all patches, representing each patch's context. The Swin Transformer may then learn and adapt to the task at hand, making predictions based on the most relevant sections of the input image.

Swin Transformer architecture is based on the Transformer architecture. Swin Transformer simulates the CNN process with the transformer paradigm. The usage of a "Shifted Windows" block in place of the multi-head self-attention mechanism distinguishes Swin Transformer from other Transformers.

Swin Transformer comes in two varieties, that is, Swin-T and Swin-S. To minimize computational cost, the Swin-T configuration employs a "Window" mechanism in the self-attention layer, whereas the Swin-S configuration replaces the layer with a "Split-Attention" mechanism that divides attention into numerous sub-spaces. In this case, we use configuration Swin-T for the model Swin Transformer. According to figure 3, Swin-T consists of input, Stage 1, Stage 2, Stage 3 & Stage 4, and Swin Transformer Block. Also, Swin-T can achieve good performance even with fewer parameters than other models, making it computationally more efficient.

3.3. MobileViT

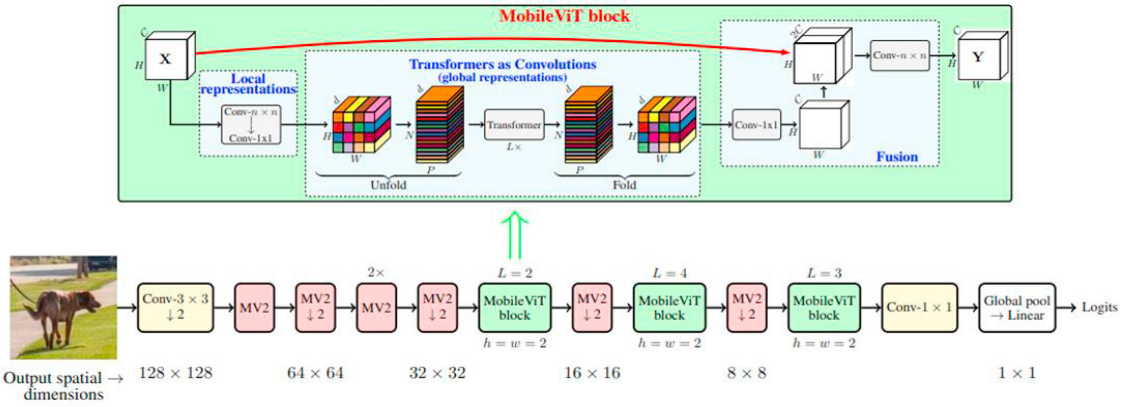

Fig. 4. Top: MobileViT block [29]; Bottom: Architecture of MobileViT [29]

MobileViT [29] is a small vision transformer architecture that employs attention to detect links between visual elements. Its self-attention layer employs a depth-wise separable convolution method to reduce computing costs compared to typical vision transformers. In MobileViT, the attention mechanism oversees recording the dependencies between picture patches, which are then turned into feature vectors via a linear projection. The attention technique is intended to be computationally efficient while capturing long-term dependencies in the input data.

The Vision Transformer (ViT) architecture is condensed into the MobileViT architecture. It substitutes the original ViT's global average pooling with a lightweight pointwise convolution; the architecture can be seen in figure 4. The depth-wise separable convolution layer extracts spatial features, whereas the pointwise convolution layer processes channel-wise features. The skip connection directs the gradient from the input to the output while maintaining the information from the input. A deep neural network for diverse computer vision applications may be formed by stacking numerous MobileViT blocks.

4. Experiments and Results

All models were trained with NVIDIA Tesla T4 GPU that uses the Turing (TU104) architecture. The environment used was Kaggle Kernel. The RAM used had a capacity of 13 GB, and the used GPU VRAM had 14 GB. In the training process, all models were trained using the hyperparameters in Table 1. The total parameters for each model can be seen in Table 2. The models have been pretrained previously on ImageNet dataset [18].

Table 1. Model Hyperparameters

Hyperparameter	Value
Optimizer	Adam [15]
Learning Rate	0.001
Loss Function	Cross-entropy
Batch Size	64
Epoch	3

The hyperparameters were chosen because the computational resources were scarce. Accuracy (Top-1) used to evaluate the performance of the models along with Recall, to know the model's ability to recognize the actual members from its class.

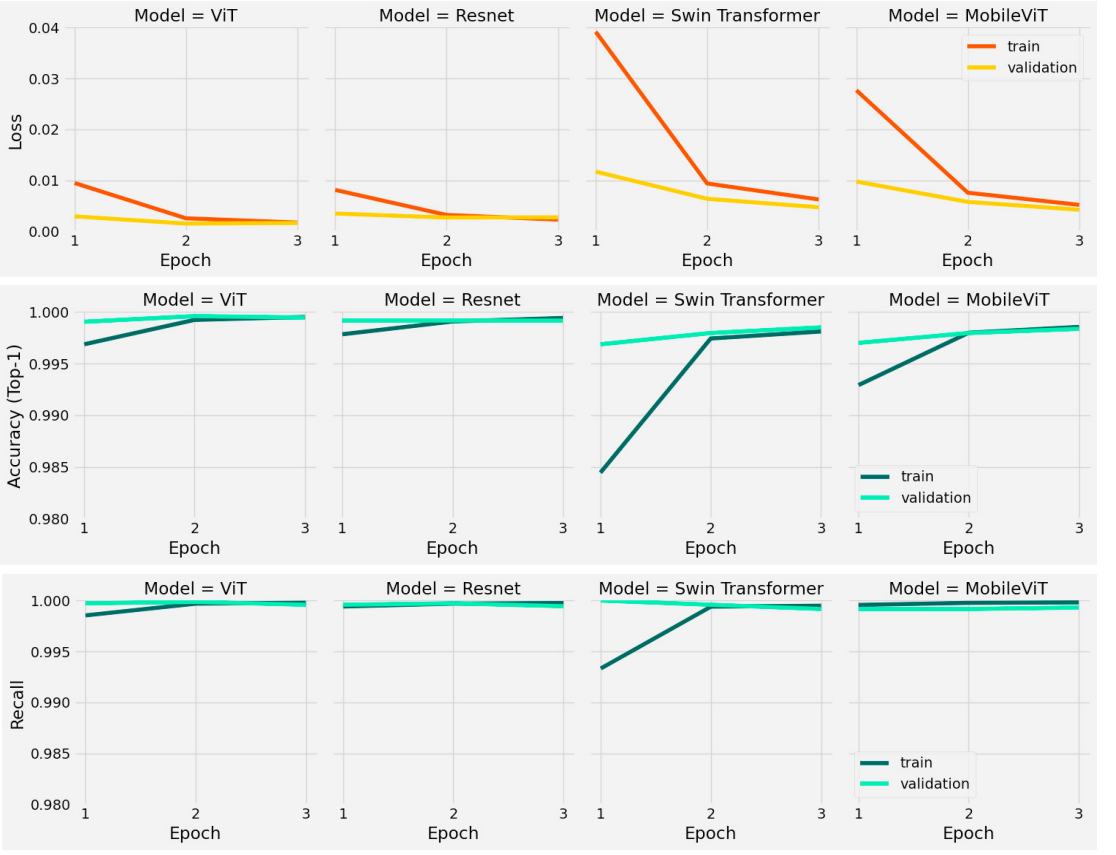

Fig. 5. Top: Loss Plot of Train and Validation dataset; Middle: Accuracy Plot of Train and Validation dataset; Bottom: Recall Plot of Train and Validation dataset.

Table 2. Model Type and Total Parameters.

Model	Type	# Parameters
ViT-S16	Transformer	21.7M
ResNet50 V2	CNN	23.6M
Swin-T	Transformer	27.5M
MobileViT-S	CNN + Transformer	5M

Several conclusions can be drawn from Figure 5. When the losses of the various models are compared, the ViT-S16, Swin Transformer, and MobileViT-S all show consistent decreases in both training and validation losses across the epochs, indicating effective learning and generalization. The losses of these models are relatively low and similar, as are the model's accuracy and recall in the validation dataset, indicating good performance and no overfitting. ResNet50 V2, on the other hand, shows a different pattern, with the training loss consistently decreasing but little progress on the validation loss, as well as the same thing happening with training and validation accuracy and recall. This highlights the importance of further research to improve its generalization capabilities.

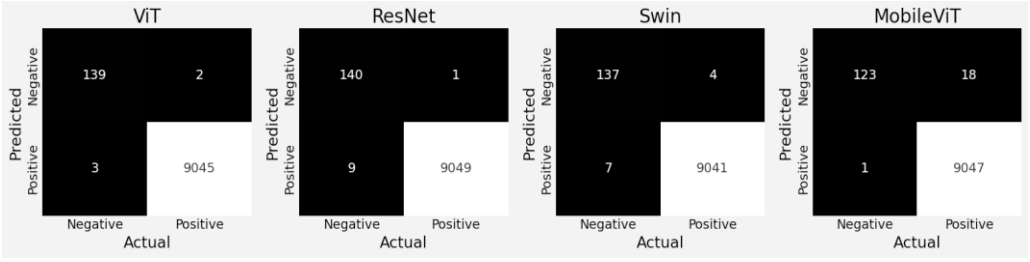

Fig. 6. Confusion Matrix of Test data from each Model

Figure 6 shows that the ViT-S16 model performs well, with high true positives and true negatives, implying accurate classification for both classes. As shown in figure 6, it has a low rate of misclassification, as evidenced by the low number of false positives and false negatives, yielding an accuracy of around 0.9996 with parameter 21.7M. ResNet50 V2 outperforms ViT-S16 in terms of true positives and true negatives, with more accurate negative classification and a slightly higher false negative rate. The precision of ResNet50 V2 with parameter 23.6M is approximately 0.9986, as shown in figure 6. Swin-T achieves accurate classification with high true positives and true negatives, but its slightly higher false negative rate indicates that some positive misclassifications occur. As shown in figure 6, Swin-T has a parameter of 27.5M and an accuracy of about 0.9983. Finally, with parameter 5M as shown in figure 6, MobileViT-S exhibits accurate classification with a moderate false positive rate and a low false negative rate, yielding an accuracy of around 0.9969. Finally, all four models (ViT-S16, ResNet50 V2, Swin-T, and MobileViT-S) perform well on the test data, with the ViT-S16 achieving highest overall accuracy and lower misclassification rates. It can also be inferred that the usage of transfer learning is effective in face mask classification task.

For future works, more extensive hyperparameter tuning and more data gathering can be done to boost the model's performance and validity to obtain more accurate predictions. Integrating model into IoT (Internet of Things) devices, specifically public CCTV. By integrating the model into IoT devices, the model can be continuously updated and perfected, and the model will remain relevant and effective. It is hoped that by performing hyperparameter tuning, data collection, and model integration into IoT devices, will be a breakthrough in various fields, particularly in the healthcare sector, will be achieved. By constantly refining and improving the model, it can provide more accurate and reliable predictions that can inform policies and interventions that can have a positive impact on society.

5. Conclusion

All models are trained using a small number of epochs and the same learning rate, which are just adjusted such that the model can be trained in a shorter amount of time and using less GPU and RAM capacity. If there is sufficient

computational power and time, these models can be optimized using a learning rate scheduler and more complex deep learning optimizers, which can cut the time needed for the algorithm to converge while the peak performance has not been achieved yet. Most tried models in this research are small in architecture, i.e., the number of layers, neurons, and parameters. Models with immense complexity can be tried, either transformer-based or hybrid, to achieve better accuracy.

Although CNN is commonly used in classifying images, transformers-based models can achieve the same result (or better) than CNN in its current state. The current hypothesis is that there is a model that can be applied to many communities without bottlenecks.

References

[1] Chavez S, Long B, Koyfman A, Liang SY. Coronavirus Disease (COVID-19): A primer for emergency physicians. Am J Emerg Med. 2021; 44:220–9. Available from: http://dx.doi.org/10.1016/j.ajem.2020.03.036

[2] Cao B, Jing X, Liu Y, Wen R, Wang C. Comparison of laboratory parameters in mild vs. severe cases and died vs. survived patients with COVID-19: systematic review and meta-analysis. J Thorac Dis. 2022;14(5):1478–87. Available from: http://dx.doi.org/10.21037/jtd-22-345

[3] Brüssow H, Zuber S. Can a combination of vaccination and face mask wearing contain the COVID-19 pandemic? Microb Biotechnol [Internet]. 2022;15(3):721–37. Available from: http://dx.doi.org/10.1111/1751-7915.13997

[4] Feng S, Shen C, Xia N, Song W, Fan M, Cowling BJ. Rational use of face masks in the COVID-19 pandemic. Lancet Respir Med [Internet]. 2020;8(5):434–6. Available from: http://dx.doi.org/10.1016/S2213-2600(20)30134-X

[5] Crupi A, Liu S, Liu W. The top-down pattern of social innovation and social entrepreneurship. Bricolage and agility in response to COVID-19: cases from China. R D Manag. 2022;52(2):313−30. Available from: http://dx.doi.org/10.1111/radm.12499

[6] Tiesman H, Marsh S, Konda S, Tomasi S, Wiegand D, Hales T, et al. Workplace violence during the COVID-19 pandemic: March-October, 2020, United States. J Safety Res [Internet]. 2022;82:376–84. Available from: http://dx.doi.org/10.1016/j.jsr.2022.07.004

[7] Rab S, Javaid M, Haleem A, Vaishya R. Face masks are new normal after COVID-19 pandemic. Diabetes Metab Syndr [Internet]. 2020;14(6):1617–9. Available from: http://dx.doi.org/10.1016/j.dsx.2020.08.021

[8] Kumar TA, Rajmohan R, Pavithra M, Ajagbe SA, Hodhod R, Gaber T. Automatic face mask detection system in public transportation in smart cities using IoT and deep learning. Electronics (Basel) [Internet]. 2022;11(6):904. Available from: http://dx.doi.org/10.3390/electronics11060904

[9] Alzu'bi A, Albalas F, AL-Hadhrami T, Younis LB, Bashayreh A. Masked face recognition using deep learning: A review. Electronics (Basel) [Internet]. 2021;10(21):2666. Available from: http://dx.doi.org/10.3390/electronics10212666

[10] Mbunge E, Simelane S, Fashoto SG, Akinnuwesi B, Metfula AS. Application of deep learning and machine learning models to detect COVID-19 face masks - A review. Sustainable Operations and Computers [Internet]. 2021;2:235–45. Available from: http://dx.doi.org/10.1016/j.susoc.2021.08.001

[11] Wang Z, Wang G, Huang B, Xiong Z, Hong Q, Wu H, et al. Masked Face Recognition Dataset and Application. 2020; Available from: http://dx.doi.org/10.48550/ARXIV.2003.09093

[12] Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell [Internet]. 2023;45(1):87–110. Available from: http://dx.doi.org/10.1109/TPAMI.2022.3152247

[13] Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do Vision Transformers see like convolutional neural networks? [Internet]. arXiv [cs.CV]. 2021. Available from: http://arxiv.org/abs/2108.08810

[14] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [Internet]. arXiv [cs.CL]. 2014. Available from: http://arxiv.org/abs/1409.0473

[15] Kingma DP, Ba J. Adam: A method for stochastic optimization [Internet]. arXiv [cs.LG]. 2014. Available from: http://arxiv.org/abs/1412.6980

[16] Qin B, Li D. Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19. Sensors (Basel) [Internet]. 2020;20(18):5236. Available from: http://dx.doi.org/10.3390/s20185236

[17] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2015; Available from: http://dx.doi.org/10.48550/ARXIV.1512.03385

[18] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009.

[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014; Available from: http://dx.doi.org/10.48550/ARXIV.1409.1556

[20] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. 2016; Available from: http://dx.doi.org/10.48550/ARXIV.1602.07360

[21] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. Proc Conf AAAI Artif Intell [Internet]. 2017;31(1). Available from: http://dx.doi.org/10.1609/aaai.v31i1.11231

[22] Chollet F. Xception: Deep learning with depthwise separable convolutions. 2016; Available from: http://dx.doi.org/10.48550/ARXIV.1610.02357

[23] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2016; Available from: http://dx.doi.org/10.48550/ARXIV.1608.06993

[24] Liu Y, Zhang Y, Wang Y, Hou F, Yuan J, Tian J, et al. A survey of visual transformers. 2021; Available from: http://dx.doi.org/10.48550/ARXIV.2111.06091

[25] Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. 2021; Available from: http://dx.doi.org/10.48550/ARXIV.2110.02178

[26] Steiner A, Kolesnikov A, Zhai X, Wightman R, Uszkoreit J, Beyer L. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. 2021; Available from: http://dx.doi.org/10.48550/ARXIV.2106.10270

[27] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020; Available from: http://dx.doi.org/10.48550/ARXIV.2010.11929

[28] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows. 2021; Available from: http://dx.doi.org/10.48550/ARXIV.2103.14030

[29] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017; Available from: http://dx.doi.org/10.48550/ARXIV.1706.03762