Lightweight Face Mask Detection Utilizing Supervised Contrastive Learning

Philipp Geppner*
University of Applied Sciences Upper Austria

David Aigner†
University of Applied Sciences Upper Austria

Figure 1: Grad-CAM visualization of face mask classification predictions made by the hybrid model. The heatmaps highlight the regions of interest used by the model to distinguish between 'Mask' and 'No Mask' classifications. Correctly predicted cases are labeled accordingly.

ABSTRACT

The COVID-19 pandemic heightened global awareness of disease prevention measures, particularly the use of face masks to limit respiratory illness transmission. Although the pandemic is officially over, mask-wearing remains essential in contexts such as healthcare, public transportation, and crowded events. To ensure timely and efficient face mask detection in these scenarios and directly on edge devices, lightweight models with minimal resource consumption are crucial.

This study presents a lightweight and efficient approach that combines Mobile Inverted Bottleneck Convolution (MBConv2D) layers, inspired by MobileNet and EfficientNetV2, with a Fully Convolutional Neural Network (FCNN) trained using a supervised contrastive learning objective. The FCNN, appended with a projection head, generates 128-dimensional feature embeddings, which serve as input to a gradient boosting classifier (XGBoost) for binary classification.

While EfficientNetV2-B0—with approximately six million parameters—serves as the high-quality lightweight baseline, our proposed solution uses only 5% of these parameters (around 300,000) and requires just 1.1 MB of memory. The hybrid approach, combining the feature extraction capabilities of the FCNN and the robust classification performance of XGBoost, achieves comparable accuracy while significantly reducing the memory footprint. The study evaluates the system's performance on diverse datasets to ensure robustness and reliability. These findings underscore the practicality of our method for on-device applications, enabling real-time deployment and supporting ongoing public health efforts in the post-pandemic era.

Index Terms: Image Classification—Face-Mask Detection—Synthetic Data—Lightweight Models

1 INTRODUCTION

The COVID-19 pandemic has fundamentally reshaped global attitudes toward public health and disease prevention. Although the pandemic has officially ended, increased awareness of the importance of mitigating the spread of infectious diseases remains. Preventive measures such as wearing face masks continue to play a critical role in reducing transmission, particularly in high-risk environments such as healthcare facilities, public transportation, and densely populated areas. However, monitoring face mask compliance on a large scale presents logistical challenges, particularly in real-time scenarios where manual oversight is impractical.

Automated face mask detection offers an efficient and scalable solution to address these challenges. By leveraging advancements in computer vision and deep learning, such systems can detect mask usage accurately and in real-time, making them suitable for deployment in diverse settings. For practical implementation, especially on mobile and edge devices, the efficiency of these systems is as important as their accuracy. Lightweight models, which combine high performance with minimal computational and memory requirements, are particularly well-suited for these applications.

*e-mail: S2310595005@fhooe.at
†e-mail: S2310595002@fhooe.at

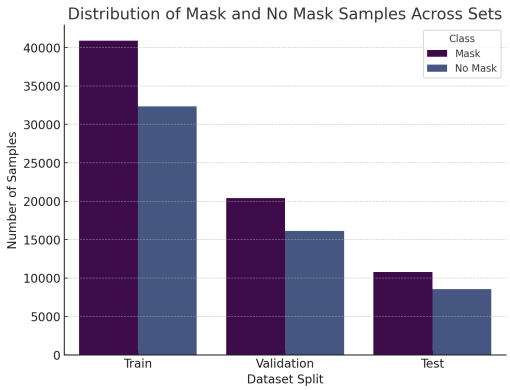


Figure 2: Distribution of Mask and No Mask samples across the three datasets.

1.1 Research Questions

1.2 Related Work

Mohammed Ali and Al-Tamimi [15] published a review on real-time face mask detection methods and techniques, summarizing 18 different publications and highlighting their advantages and disadvantages. Detecting a face mask in an image involves the classification of a face mask in the input picture, which is the part of the pipeline our work focuses on. The different approaches involve Multi-Stage Detectors [18], Convolutional Neural Networks (CNN) [1, 23, 21], Region-based Convolutional Neural Networks (R-CNN) [14, 28], Fast Region-based Convolutional Neural Networks (FAST R-CNN) [25, 17, 26, 5], Faster Region-based Convolutional Neural Networks (FASTER R-CNN) [19, 4, 27, 20] and Mask Region-based Convolutional Neural Networks (MASK FAST R-CNN) [2, 10, 11].

2 Dataset Oveview

This study uses images from three distinct datasets (see Figure 2). The intution behind combining three different sets lies in their quality and diversity tradeoffs. While the MaskFaceNet dataset provides a strong baseline with high diversity of subjects and scenes, the masks themselves are always blue surgical masks. Additionally, the masks themselves are only *stitched* onto the images in an augmentation step. To mitigate this, another dataset from a Kaggle competition is included, which contains a mixture of real and augmented masks on subjects. However, we noticed that a substantial amount of images come from a small subset of subjects (people). A third synthetically generated dataset is therefore included to further increase diversity of people, scenes and types of masks. Additionally, the synthetic dataset allows for easy inclusion of edge cases and difficult adversarial examples, such as faces obscured by scarfs, or clothing that contains additional faces on it.

The images are categorized into two classes: individuals wearing masks and individuals not wearing masks. Some images were augmented to include masks artificially. To ensure data integrity and prevent potential data leakage, duplicate images were excluded. Again, the datasets were considered to represent a diverse range of ethnicities, color schemes, mask types, and age groups.

2.1 Kaggle Dataset

The first featured dataset has been published as a Kaggle competition[1]. It includes 7,553 images of people wearing masks, wearing mask-similar clothing such as bandanas or scarfs, or not wearing masks. Some pictures have been augmented, but most are photo-realistic. The dataset consists of 1,776 images taken from another github page[2] and 5,777 images taken from the Google search engine. See Figure 3 for example images of this dataset.

(a) Person wearing a mask (b) Person wearing a bandana

(c) Person without a mask (d) Person without a mask

Figure 3: Four example pictures from the Kaggle dataset resized to resolution 224x224.

2.2 MaskedFace-Net Dataset

The second featured dataset has been published in a GitHub repository titled MaskedFace-Net[3]. It includes 67,049 images of people without masks and the same amount of pictures with augmented masks. Additionally, the dataset also includes 66,734 images of incorrectly worn masks, but these are not relevant for this publication. The unaugmented images have been taken from the Flickr-Faces-HQ (FFHQ)[4] and augmented with a simple mask overlay. See Figure 4 for example images of this dataset.

2.3 Diffusion Model generated Dataset

A third dataset consisting of 2,400 synthetically generated images was created using the Stable Diffusion XL [16] model. According to the authors' assessment, these images exhibit characteristics resembling photo-realistic imagery. The dataset includes individuals of diverse ethnicities, with a wide variety of clothing colors and styles, as well as different types of face masks. See Figure 5 for example images of this dataset.

2.4 Preprocessing

The python library Open-CV was utilized to develop a script for resizing pictures to a specified resolution. The resizing process does not preserve the original aspect ratio, resulting in image distortion. However, we argue that this distortion is preferable to the alternative of cropping, as it ensures that no image content is omitted. This preprocessing script has subsequently been used to resize the previously described datasets to a target resolution of 224 x 224 pixels. We decided to use a target resolution of 224 x 224 pixels because this image format is the expected input size for EfficientNet models and also widely adopted in other publications.

[1] https://www.kaggle.com/datasets/omkargurav/face-mask-dataset

[2] https://github.com/prajnasb/observations

[3] https://github.com/cabani/MaskedFace-Net

[4] https://github.com/NVlabs/ffhq-dataset

(a) Person with augmented mask

(b) Person with augmented mask

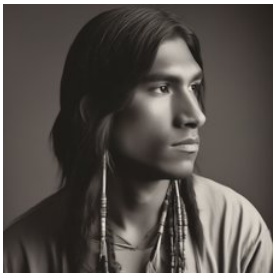
(c) Person without a mask

(d) Person without a mask

Figure 4: Four example pictures from the MaskedFace-Net dataset resized to resolution 224x224.

(a) Person with augmented mask

(b) Person with augmented mask



(c) Person without a mask

(d) Person without a mask

Figure 5: Four example pictures generated by the diffusion model in resolution 224x224.

3 METHODOLOGY

Modern deep learning architectures on the upper end of performance rankings tend to be large in size. This comes with obvious hardware requirements and constraints. For real-world face-mask detection on mobile edge devices, the small memory footprint of light-weight models is a key advantage. Two of the best light-weight models, when considering size-to-performance, are the MobileNet and EfficientNet architectures. We therefore based our approach on the ideas presented in the MobileNet [3] and EfficientNetv2 [24] publications, in particular the use of *Mobile Inverted Bottleneck Convolution Layers (MBConv2D)*. Our custom approach to light-weight face mask detection uses a Fully Convolutional Neural Network (FCNN) [12]. The FCNN is trained on a supervised contrastive learning objective, as formulated by Khosla et al. [7], to arrive at disentangled class representations. The output is a 128-dimensional feature vector, that can be used to train any modern machine learning model as classifier. We demonstrate the efficacy of our approach with a gradient boosting algorithm, XGBoost.

3.1 EfficientNetV2

EfficientNetV2-B0 - the smallest variant of the model family - is used as a baseline model for all the performance comparisons on our facemask detection dataset. It combines a fairly lightweight model architecture (around 6 million parameters) with state-of-the-art performance on various benchmarks. Its architecture employs compound scaling, depthwise separable convolutions, and an optimized multi-stage structure, making it suitable for resource-limited scenarios like real-time facemask detection. For our purpose, the model-weights were pretrained on ImageNet [9]. Transfer Learning was done with both frozen model weights and with a full finetuning on our dataset.

3.2 Hybrid Approach: SupCon-FCN and XGBoost

As previously mentioned, we implemented a custom solution that combines the frozen weights of the backbone fully convolutional network (FCN) and a strong feature-based gradient boosting method, XGBoost. To arrive at feature representations that are meaningful and disentangled, we first use supervised contrastive learning - *SCL* - as described by Khosla et al. [8] in 2020, on the FCN. We then use Global Average pooling and a linear projection head to reduce the feature dimensionality. The supervised contrastive learning objective tries to find class represetnations, so that images with the same class are close together in the feature space, while simultaneously pushing images of differing classes further apart.

Formally, supervised contrastive loss is defined as:

$$\mathscr{L}_{\text{SupCon}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}$$

where I represents the set of all image indices in a batch, $P(i)$ denotes the set of positive samples corresponding to the anchor i, and $A(i)$ includes all instances except the anchor itself. The feature vectors z are normalized, and τ represents a temperature scaling parameter that controls the separation in feature space.

This formulation encourages positive pairs (same-class samples) to have high similarity while ensuring that negative pairs (different-class samples) remain distinct.

The original SupCon Paper recommends very large batch sizes - over 1000 samples per batch - and a temperature setting of 0.1. While we keep the temperature for our approach at 0.1, initial experiments showed that larger batch sizes do not yield any real upsides four our limited, binary classification task. Instead, a batchsize of 1000 samples, increased training times from 2 minutes to upwards of 5 minutes per epoch. Therefore, we kept the batchsize at 32.

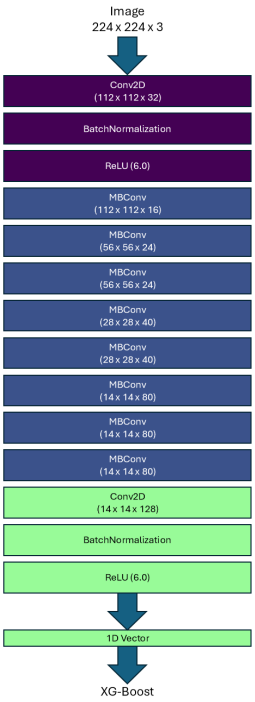


Figure 6: Architecture of the hybrid model approach.

After training on the SupCon objective, the disambiguated FCN outputs are used as features for our binary classification task with XGBoost. Compared to the baseline Efficientnet-V2-B0 model, our FCN approach needs only 5% of the parameters (around 3e5) and memory (1.1 Mb). The memory footprint of the final model is therefore, also influenced by the choice of classifier. Our lightweight XGBoost model adds only 0.17 Mb to the model size. Figure 6 contains a schematic overview of our approach.

3.3 Data Augmentation

Data Augmentation is vital for diverifying data during training and assure that models learn more robust representations. We employ the following set of image augmentation steps on the train and test set:

- **Random Flipping:** Images are flipped horizontally to introduce spatial invariance.
- **Random Zoom:** The image is randomly scaled in or out, up to 20%, to simulate different distances and perspectives.
- **Random Shear:** A shear transformation is applied to distort the image along the x and y axis,again up to 20%, mimicking viewpoint variations.
- **Random Rotation:** The image is rotated by a random angle to improve rotational invariance of the model.
- **Random Translation:** The image is shifted horizontally or vertically to simulate changes in positioning.
- **Random Brightness:** The brightness of the image is adjusted randomly to account for varying lighting conditions.
- **Random Contrast:** Contrast levels are randomly altered to improve robustness to lighting variations.
- **Gaussian Noise:** Random Gaussian noise is added to the image to simulate low quality images.

Table 1: Model performances measured by different scores

Model	Accuracy %	MCC	F1 Score
EfficientnetV2B0 (Frozen)	99.94	0.998	0.9995
EfficientnetV2B0	99.95	0.9991	0.9996
MBConvolutional FCN and XGBoost	99.61	0.9921	0.9965
MBConvolutional FCN and MLP	96.65	0.9324	0.9665
Convolutional FCN and XGBoost	98.6	0.973	0.988

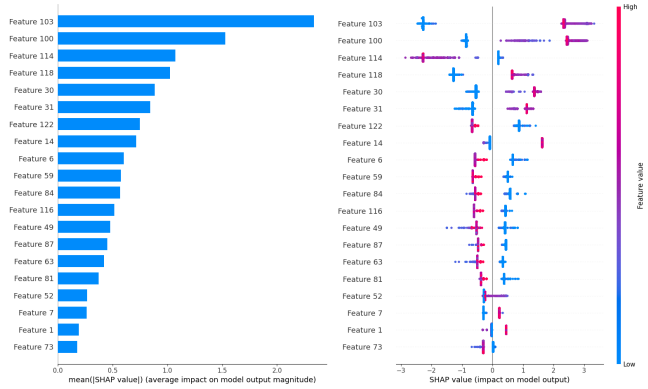


Figure 7: Barchart visualizing feature importances and violin plot visualizing how the features impacted the classification.

4 Implementation

All deep learning model implementations for this study are done using keras 3.1, with tensorflow 2.18 as the backend framework. The Scikit-Learn and XGBoost packages are used for metrics and the gradient boosting algorithm, respectively. Preprocessing and data augmentation is done with keras, as well as OpenCV and numpy.

For the synthetic data generation with StableDiffusion-XL, the huggingface diffusers library, which builds on pytorch, is used.

For our prototypical integration into a real-time detection pipeline, we use YOLOv11, provided by the ultralytics package.

5 Results

5.1 Classification

We evaluate and compare the classification results between the baseline model and different formulations of the hybrid models. Accuracy is used, as the datasets are balanced, as previously shown in figure 7. For basic ablation purposes, we used XGBoost and a simple Multi-Layer-Perceptron to test the influence of the classifier used on top of the extracted features. To show the influence of the MBConvolution architecture, we also trained a simple FCN network using normal convolutions and max pooling on the contrastive objective. All the classification results are detailed in table 1.

5.2 Training

As shown in Figure 8, the training loss for the baseline EfficientNet-V2-B0 model flattens after epoch 2 and remains consistent with no real further improvement. This could be in part because the dataset is quite diverse and large, but might also hint at the ability of the model to adapt its already well pretrained weights.

For our Hybrid Model, the supervised contrastive loss is also evaluated over the training history. Results can also be seen in figure 8. Compared to the pretrained EfficientNet-V2-B0's cross-entropy loss, the contrastive loss keeps getting lower for almost the entirety of the training run.

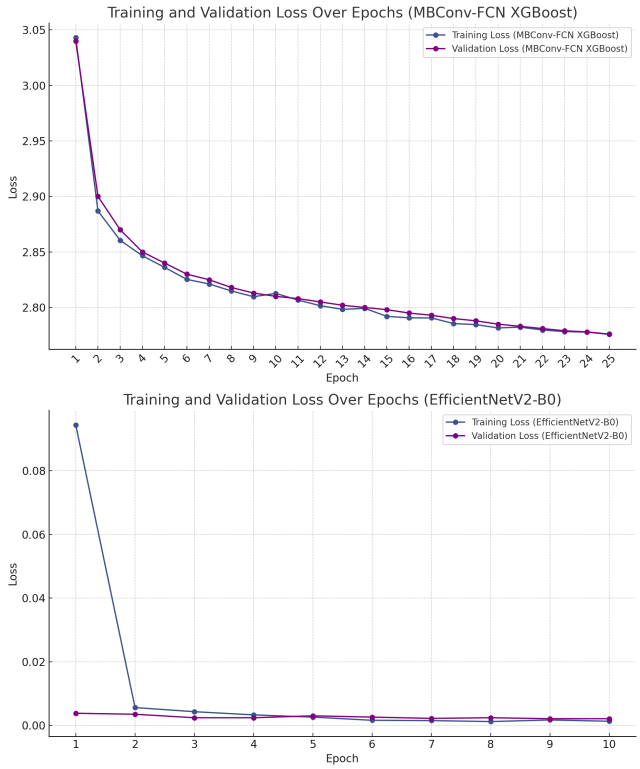


Figure 8: Visualisation of the training and validation loss over 25 epochs on the SupCon loss of the hybrid model (top) and 10 epochs on the Cross-Entropy loss of EfficientNetV2-B0 (bottom).

5.3 Feature Importance with GradCAM

Different from classical image feature engineering, the filters of convolutional layers are automatically learned by the network. For this reason, it is important to evaluate whether a trained network has learned fitting representations of our class instances. If the dataset has inconsistencies across the different classes, be it other co-occuring objects, graphical artifacts, or biases, it might generalise to a wrong concept. In the case of face mask detection, we expect the model to look at the mouth and nose area for shape and structural information.

One way to evaluate this is to look at the feature maps with Grad-CAM [22], which highlights the most important regions in the image that influence the model's predictions by using gradient-based localization. Grad-CAM generates a heatmap on top of the input image, allowing us to visually inspect if the model is focusing on relevant areas, such as the mouth and nose region for face mask detection. If the heatmap reveals attention on unrelated areas, such as background, hair, or clothing, it could indicate that the model has learned spurious correlations rather than meaningful facial features. Figure 1 and 9 show the GradCAM results for our SupCon-FCN and the EfficientNetV2B0 model, respectively. As we can see, both models have learned to look at the facial area, which is a good indication for their generalisation capabilities.

5.4 Feature Importance with SHAP

Using XGBoost as the classifier allows us to also investigate the contrastive learning approach and its effect on separation of class representations. We use the SHAP library, which is based on Shapley values and was initially introduced by Lundberg and Lee [13]. It allows for not only global feature importance analysis but also the local influence for each sample. This way, we can evaluate which

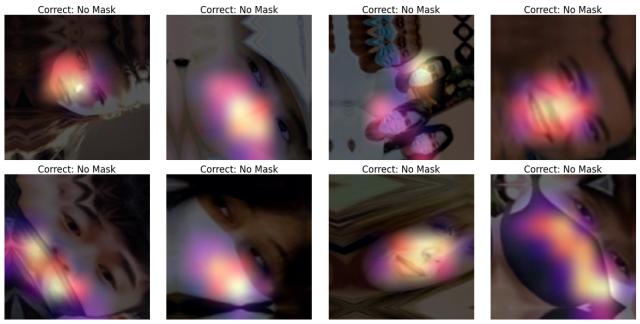


Figure 9: Grad-CAM visualization of face mask classification predictions made by the Efficient-Net model.

Figure 10: Four frames of a classified video featuring a girl being equipped with a mask.

extracted features vote for and against facial masks in the sample image. Figure 7 shows the results for our hybrid approach, with cumulated absolute shapely values on the bar chart, and directional influences on the violin graph. A negative value in the violin plot means that the feature voted for a face mask in the image and a positive one against it.

For the first 20 features, the gradient boosting model seemed to be able to find well separated interpretations. This could indicate a good separation between the class embeddings.

5.5 Prototypical Integration with YOLOv11

To explore potential extensions of our approach, we integrated YOLOv11 [6], a state-of-the-art object detection model, with DeepSort, a robust tracking algorithm, to enable real-time detection and tracking of individuals wearing or not wearing face masks (see Figure 10). However, a comprehensive evaluation of this integration is beyond the scope of this publication. Further investigation into the performance and real-world feasibility of such a system is left for future work.

6 Discussion and Conclusion

The findings of this study highlight the potential of lightweight models in achieving efficient and accurate face mask detection for real-time applications on edge devices. By using supervised contrastive learning combined with FCNN and XGBoost, the proposed hybrid model demonstrates comparable accuracy to baseline models like EfficientNetV2-B0 while significantly reducing memory and computational requirements. This efficiency is critical for deployment in resource-constrained environments such as mobile devices and embedded systems.

Through the integration of diverse datasets, including synthetic data, the model effectively handles variations in mask types, subject characteristics, scenes and image quality. We include difficult synthetic, adversarial examples, such as faces covered by scarfs or clothing that contains additional faces.

Grad-CAM visualizations confirm the model's focus on relevant facial regions, for the custom hybrid approach and the baseline models. Additionally, SHAP analyses validate the effectiveness of

supervised contrastive learning to disentangle class representations in feature space.

As part of this work, a prototypical integration of a realtime face mask detection pipeline was developed. It uses YOLOv11 and DeepSort to detect and track faces in a video. The obtained regions of interest are then classified using our model. Apart from functioning, further performance analysis of this prototype is left for future work.

In conclusion, this study presents a lightweight and efficient approach to face mask classification and detection, achieving minimal performance loss on a diverse dataset when compared to EfficientNetV2-B0, while having only 5% as many parameters.

Despite the good accuracy on the datasets featured in this publication, further testing should be done on different datasets to assess the overall performance of the approach. Alternate datasets for testing could include data with varying resolution or in real world assessment as part of a face detection pipeline.

Future work can also involve hyperparameter optimization and improvements in XG-Boost. Additionally to hyper parameter optimization, the hybrid approach also allows for the integration of additional statistical parameters. Further research could also evaluate the influence of such a feature engineering. Another promising approach would revolve around the usage of fusion MB-Convolution2D layers in earlier parts of the model, as proposed by EfficientNet V2, as they are even more efficient than the normal MB-Convolutional layers.

Acknowledgments

This work has been part of the Computer Vision lecture in the department of Data Science and Engineering at the University of Applied Sciences Upper Austria.

References

[1] P Deval et al. "CNN based face mask detection integrated with digital hospital facilities". In: *Int. J. Adv. Res. Sci., Commun. Tech* 4.2 (2021), pp. 492–497.

[2] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[3] Andrew G Howard. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017).

[4] Huaizu Jiang and Erik Learned-Miller. "Face detection with the faster R-CNN". In: *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE. 2017, pp. 650–657.

[5] Lin Jiang et al. "Application of a fast RCNN based on upper and lower layers in face recognition". In: *Computational intelligence and neuroscience* 2021.1 (2021), p. 9945934.

[6] Rahima Khanam and Muhammad Hussain. *YOLOv11: An Overview of the Key Architectural Enhancements*. 2024. arXiv: 2410.17725 [cs.CV]. URL: https://arxiv.org/abs/2410.17725.

[7] Prannay Khosla et al. "Supervised Contrastive Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 18661–18673. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

[8] Prannay Khosla et al. *Supervised Contrastive Learning*. 2021. arXiv: 2004.11362 [cs.LG]. URL: https://arxiv.org/abs/2004.11362.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Commun. ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

[10] Kaihan Lin et al. "Face Detection and Segmentation Based on Improved Mask R-CNN". In: *Discrete dynamics in nature and society* 2020.1 (2020), p. 9242917.

[11] Kaihan Lin et al. "Face detection and segmentation with generalized intersection over union based on mask R-CNN". In: *Advances in Brain Inspired Cognitive Systems: 10th International Conference, BICS 2019, Guangzhou, China, July 13–14, 2019, Proceedings 10*. Springer. 2020, pp. 106–116.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[13] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.

[14] NA Mohamed and MSH Al-Tamimi. "Image fusion using a convolutional neural network". In: *Solid State Technol* 63.6 (2020), pp. 13149–13162.

[15] Firas Mohammed Ali and Mohammed Al-Tamimi. "Face mask detection methods and techniques: A review". In: *International Journal of Nonlinear Analysis and Applications* 13.1 (2022), pp. 3811–3823. ISSN: 2008-6822. DOI: 10.22075/ijnaa.2022.6166. eprint: https://ijnaa.semnan.ac.ir/article_6166_2c38029239e424401afbfb4a1aa1e196.pdf. URL: https://ijnaa.semnan.ac.ir/article_6166.html.

[16] Dustin Podell et al. *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. 2023. arXiv: 2307.01952 [cs.CV]. URL: https://arxiv.org/abs/2307.01952.

[17] Rongqiang Qian et al. "Road surface traffic sign detection with hybrid region proposal and fast R-CNN". In: *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE. 2016, pp. 555–559.

[18] J Redmon. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[19] Shaoqing Ren et al. "Object detection networks on convolutional feature maps". In: *IEEE transactions on pattern analysis and machine intelligence* 39.7 (2016), pp. 1476–1481.

[20] Mosab Rezaei et al. "Assessing the effect of image quality on SSD and faster R-CNN networks for face detection". In: *2019 27th Iranian Conference on Electrical Engineering (ICEE)*. IEEE. 2019, pp. 1589–1594.

[21] Jansi Rani Sella Veluswami, Sai Prakash, Niel Parekh, et al. "Face mask detection using SSDNET and lightweight custom CNN". In: *Proceedings of the International Conference on IoT Based Control Networks & Intelligent Systems-ICICNIS*. 2021.

[22] Ramprasaath R Selvaraju et al. *Grad-CAM: Why did you say that?* 2017. arXiv: 1611.07450 [stat.ML]. URL: https://arxiv.org/abs/1611.07450.

[23] S Shivaprasad et al. “Real time CNN based detection of face mask using mobilenetv2 to prevent Covid-19”. In: *Annals of the Romanian Society for Cell Biology* 25.6 (2021), pp. 12958–12969.

[24] Mingxing Tan and Quoc Le. “EfficientNetV2: Smaller Models and Faster Training”. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 10096–10106. URL: https://proceedings.mlr.press/v139/tan21a.html.

[25] Kai Wang et al. “Use fast R-CNN and cascade structure for face detection”. In: *2016 Visual Communications and Image Processing (VCIP)*. IEEE. 2016, pp. 1–4.

[26] Qihang Wang and Junbao Zheng. “Research on face detection based on fast R-CNN”. In: *Recent Developments in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2017*. Springer. 2019, pp. 79–85.

[27] Wenqi Wu et al. “Face detection with different scales based on faster R-CNN”. In: *IEEE transactions on cybernetics* 49.11 (2018), pp. 4017–4028.

[28] Chenchen Zhu et al. “Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection”. In: *Deep learning for biometrics* (2017), pp. 57–79.