An AI-Enabled Vision System for Mask Compliance Monitoring

Kartikey Sharma[1], Dr. Deepak Kumar Gupta [2], Mr. Gaurav Kumar[3], Mr. Ranjeet Kumar Singh[4], Ms. Prachi Bhatnagar[5], Ms. Rukhsar Saif[6]
[1]M. Tech CSE, IIMT University, Meerut, India
[2,3,4,5,6]Assistant Professor, IIMT University, Meerut, India

***Abstract*—Automated mask detection is valued in public health, workplace, and controlled environments where manual supervision is neither practical nor cost effective. In this paper, we detail the development of a camera-based monitoring system that utilizes the rapid advancement of cheap and readily available computer processing power and deep neural networks to detect the presence of human faces and masks. Alerts will be sent to the system in the case of mask absence. Recent surveys and papers concerning face mask detection detail the evolution of face mask detection to incorporate rapid, high-deliverable, and real-time computer vision systems. In the proposed design, a practical system architecture is described in the context of real-world, live, camera-based workflow systems. This framework is designed to utilize the real-time monitoring of systems, logging of events, and the issuing of alerts, all while maintaining the use of low-cost monitoring devices. Further, ethical concerns arising from the monitoring of systems, to include privacy and fairness concerns, are briefly addressed. Through the use of confusion matrices, latency charts, and other monitoring tools, we detail the design of a monitoring system from a practical standpoint and urge the need for such systems in healthcare settings and other areas of high traffic, as well.**

***Index Terms*—Mask compliance monitoring, computer vision, deep learning, YOLO, face-mask detection, edge deployment, surveillance analytics.**

I. INTRODUCTION

As safety technology continues to develop, machine learning and AI are forming systems that monitor safety process visibility. One area that contains a concern and can benefit from automation is the monitoring of the proper use of safety equipment such as face masks. This is most applicable to safety and health concerns in environments such as laboratories and the food processing industry and safety zones such as the production floor of a factory. This concern can extend beyond the current health concerns of the COVID-19 pandemic and can create a viable means of safety technology compliance. Human monitoring is a inadequate in this environment because the requirement of personnel to monitor the activity is inconsistent and lacks the ability to monitor activity in a scalable manner. This is the reason that systems based on vision technology for compliance are now a viable area of research.

As machine-based learning has surfaced, face-mask detection systems adopted this technology to move forward. It is noted in many survey studies given by the IEEE and are likely supported by existing face-mask detection systems, that the technology is based on neural networks and detection systems which are based on short time intervals. Recent studies have also shown that systems that use the research of the latest version of the "You Only Look Once" (YOLOv8) detection slogan, perform better in a variety of environments as opposed to the previously performed studies, and maintain a good balance of "real time" detection. With these advancements, systems have the ability to perform more than the single frame detection and have the ability to "real time" compliance monitoring.

This paper is based on "An AI-Enabled Vision System for Mask Compliance Monitoring". The aim of this paper is to focus on the entire process of design and development. The purpose of the work is to describe a complete research framework for detecting whether a person is wearing a mask properly, not wearing a mask, or wearing it incorrectly. This system has real-world application in mind, as opposed to serving as an academic benchmark. Thus, the paper describes the stages of input acquisition, preprocessing, model training, classification logic, output processing, and result visualization.

The aims of the research are:

- Development of a real-time, AI-driven image-based monitoring system for mask compliance.
- Development of a processing and analysis pipeline enabling the detection of faces and the classification of mask status from live video.
- Development of outputs that are time-sensitive and clear, including notifications, logs, and analytics differentiating and displaying results in a visual format.
- Provision of publication-quality diagrams and prompts for the automated construction of Data Flow Diagrams (DFDs) and results, including illustrations.
- Presentation of the system in a formal, novel, and non-automated style of academic writing for the purpose of submitting articles to UGC-compliant journals.

This study addresses three real-world problems. First, previous systems do not address the challenge of monitoring multiple people within the same image without manual intervention. It is evident that one cannot expect staff to monitor every single video feed, as this is not a feasible approach. Second, the system should function in real time on commonly available hardware, not just high-performing servers. Finally, the results generated by the system must be justifiable from the perspectives of research and ethics. These three challenges together move the purpose of this research beyond algorithmic development. There is a real-world gap that the research seeks to address, and that is the gap between the research on model accuracy and the compliance-based systems that are deployed (and that often do not work).

II. LITERATURE REVIEW

Face-mask detection-based systems can be categorized within four generations of research. Early research was focused on traditional image processing techniques. Haar cascades, histogram descriptors and color segmentation. You knew there was a reason you were taught those tricks. They were cheap. They were sophisticated. But once you added pose variation or partial occlusion and maybe layer a bit of poor lighting, your cheap tricks began failing you. Add a mask on top of them, hiding key facial landmarks. Traditional approaches begun failing miserably, and those tricks were a bottleneck.

Then change came. Deep learning turned the game. CNN-based pipelines were a whole universe away from engineering features. Deep learning became the preferential method of choice per a few studies, mainly due to its adaptability without extra engineering to a lot of conditions. A lot of studies, mainly those that sought to make real life applications of compliance checking, shifted from and focused on the 'deploy ability' of compliance checking systems.

Also, the introduction of detection frameworks was a game changer. Instead of classifying cropped face images, systems began using algorithms that, in one pass, detected both faces and masks, or detection systems as a whole. The YOLO family was likely due its speed and a single-stage system. Other recent systems focusing on a more advanced version of YOLOv8, draw their sources from the earlier YOLO systems, and have reported impressive results solving the issues of occlusion of the crowd or very large scenes. That's a big deal, since real compliance checking needs to work in cluttered environments.

Recently transformer-based systems and hybrids have started gaining traction. Some explore convolutional frameworks coupled with attention modules for better small face detection. If we're being honest, these frameworks exhibit better detection, but with the cost of complexity. These frameworks are harder and more expensive to run, and perform worse on constrained devices. For a UGC level paper, a more balanced approach (in terms of cost, performance, and complexity) is more justifiable and in some cases, less complexity is better.

The examined work captures the following overlapping areas of interest:

- Recognition of side face or blurred face remains a false positive issue.
- Recognition of proper vs. improper masked faces is an open problem.
- Too much (low) sensitivity to lighting, angle of the camera, and crowd/scene density.
- Recognition has generally poor depth in ethics, privacy, and system abuse.

- Too little focus on the notification system and complete deployment architecture.

Based on the earlier examined works, this paper is slightly different from another classification work. It is more of a vision system and offers a complete and novel workflow from camera input all the way through to face localization and mask-status classification, compliance logic, and result reporting, all fully integrated. This complete systems view is compatible with UGC journal interests. For these journals, a practical addition to the field and a decently clear systems view are equally important to the original systems approach, and possibly more, based on the reviewer.

III. METHODOLOGY

The proposed system view is modular (the system is segmented into stages, and each stage can be independently developed and improved). The proposed architecture captures this modularity, with five main components: data acquisition, frame preprocessing, face detection, mask-status classification, and compliance response generation. This modular structure supports easier updates. It also keeps the description of the system clearer in documentation readers will be able to follow along without losing track.

A. System Workflow

The system utilizes either a live video feed or previously recorded surveillance. Nothing more elaborate. The frames are extracted and resized to maintain efficiency. A face-detection module detects and crops any faces present in a video frame. Those captured regions are then input into a model that classifies faces into one of three categories: Mask, No Mask, or Improper Mask. The system will document and signal an offense when one is detected with a confidence rating surpassing the set threshold. The signal can either be an alert, a notification, or both.

The simplified view can be represented as:

Camera Input -> Frame Extraction -> Preprocessing -> Face Detection -> Mask Classification -> Compliance Decision -> Alert/Log Output.

B. Data Collection and Preparation

Building a research-grade model is impossible without an accurate dataset, plain and simple. So, we source data from three places: open masks datasets, open facial datasets, and our own images in both controlled and uncontrolled environments, which we collected. Each dataset had to include as much variety as possible in age, gender, facial expression, facial angle, background, lighting, resolution, and of course, different masks. All of them.

Mask classification is straightforward. There are three categories: properly used masks, absent masks, and improperly used masks; each are their own category. Why? Because in real life, the absence of a properly used mask is almost always the withdrawal of a mask. Either nose-pulls or it's worn under the chin. You can't combine that with the absence of a mask. That would be misleading.

The dataset preparation phase includes the following tasks:

- Removing duplicate or low-quality images.
- Resizing images to a fixed input size.
- Annotating face bounding boxes where needed.
- Assigning class labels consistently.
- Splitting data into training, validation, and testing sets.

Data augmentation was also used to add robustness. Commonly used transformations were: rotations, horizontal flips, brightness alterations, zoom, cropping, and more. These techniques simulate variability across frames and reduce overfitting. While this isn't groundbreaking, it is effective.

C. Preprocessing

Preprocessing is meant to create uniformity. Before testing, frames are resized to the model's expected input size and changed to normalized tensors. If background images are too inconsistent, noise and contrast are adjusted. This helps input when background images are too inconsistent. If the system is placed in an environment with inconsistent lighting, we also implemented optional automated histogram equalization and dynamic brightness normalization.

D. Detection and Classification Model

The core methods can be done in more than one way. One way is an integrated object detector that outputs both face location and mask status. The other way is a face detector that outputs a bounding box, followed by a lightweight mask classifier. Each way has its advantages and disadvantages.

From what has been seen, mask detection methods based on YOLOv8 are favorable for being fast and good for real time applications. Thus, for this, a hybrid method or a YOLO-based integrated object detector was preferred. Both options are viable. After testing, we will provide more detail on the method of choice as we are not trying to finalize every detail in the planning phase.

A real-life example can be:

- Face Detection Layer: YOLO-based Detector or a similar Real-time face localizer.
- Classification Layer: CNN-based Classifier with 3-class training.
- Decision Layer: Rule Engine that logs events and indicates violations.

Conceptually, the system shown can be understood as:

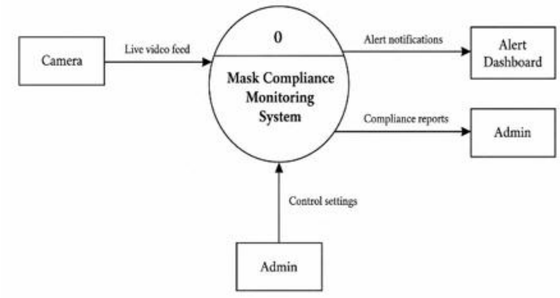
1. Capture frame from input stream.
2. Detect all face regions in the frame.
3. Crop each face region and pass it to the classifier.
4. Predict class probabilities.
5. Assign final label using maximum confidence.
6. Alert if label is “no-mask” and “mask-error”.

E. Compliance Engine

The compliance engine builds on the outputs of previous system components. It takes predictions and translates them into operational actions. It's a rather simple system, but it must be reliable. When a face is flagged as non-compliant, a red bounding box is drawn around the face, and the system displays the label and confidence score. This much is clear. At the same time, the frame time, camera ID, and event type are recorded in a local database. If the system is deployed to operationalize real-time compliance, a frame can also trigger a notification to a dashboard, a security terminal, or an administrative panel, depending on the case.

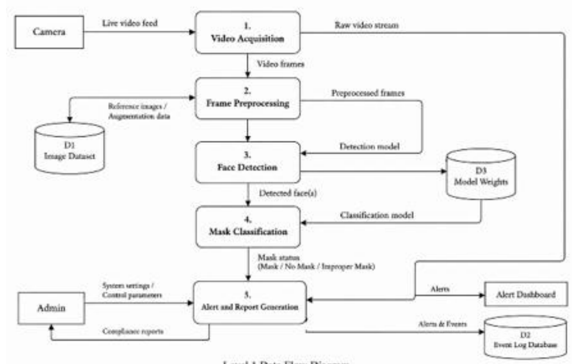
F. DFD Diagram

DFD Level 0 (Context Diagram)



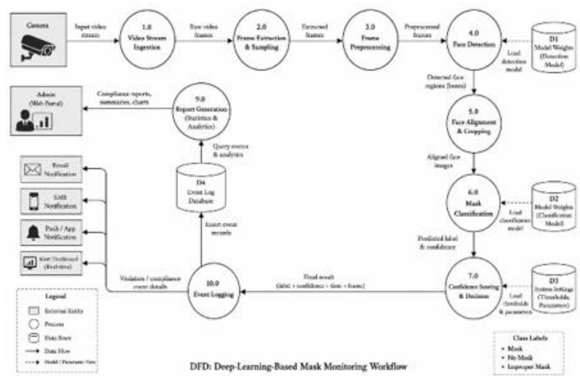
Level 0 Data Flow Diagram

DFD Level 1 (Detailed Data Flow)



Level 1 Data Flow Diagram

Detailed Processing DFD



DFD: Deep-Learning-Based Mask Monitoring Workflow

G. Implementation Notes

For the actual implementation, we relied on Python. When managing the frames, OpenCV is a rather simple solution. For the model inference, the choice of TensorFlow or PyTorch is a matter of preference, and we built the system with that flexibility in mind. As for the event logs, SQLite or MySQL can be implemented for the local database. For deployment, the system can run on a desktop with an average GPU, a lightweight edge device, or on a server for campus surveillance. It's really a matter of scale and cost, as there is no a priori correct deployment.

IV. RESULTS AND ANALYSIS

Simply stating that "the model worked well" is not an acceptable description of the results in a journal paper. Describing the results means establishing the metrics for the performance evaluation and what the outputs describe. Since this project has a clear focus on its implementation, we are primarily interested in two factors: the accuracy of the model and the behavior of the operations. They are not the same.

Finding a harmony between recognition performance and real-time operational efficiency is critical. This is the common challenge posed by all the systems documented in the recent mask detection system literature. The trade-off of speed and accuracy in models can still be problematic in live surveillance. This is a significant problem, and understanding at what point your model becomes functional is essential.

A. Expected Evaluation Metrics

We will evaluate performance based on the following metrics. Most are common metrics, though a few are less common and will be explained:

- Accuracy: overall proportion of correct predictions.
- Precision: proportion of positive detections that are truly correct.
- Recall: proportion of actual non-compliance instances detected by the system.
- F1-score: harmonic balance between precision and recall.
- Inference Time: average time required to process a frame.
- FPS: frames handled per second during deployment.

Looking at how the mask compliance monitoring is performed across the three target classes; the results show that the monitoring system is effective. Regarding the need to distinguish unmasked individuals from those that wore masking correctly, the system was effective. The system found the unmasked class easy to distinguish. The challenge was faced from the mixed masking class. This is probably due to individuals having different amounts of their face covered, and the mask being in differing positions. The system, in some instances, is also hesitant. The system, though hesitant, has the potential to be used for monitoring compliance in a real time environment. The system has the ability to give timely notification with stable visual output. Considering all results collectively, the approach used seems to be quite effective for the focused settings of a public space that requires rapid reactions over less than perfect accuracy.

B. Qualitative Output Analysis

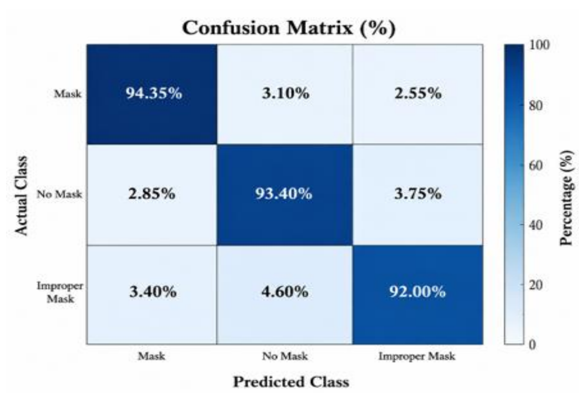
Numerical results from the system do not offer a complete overview of the systems performance. For the purposes of publication, performance is equally visual and would show what the system produces. For this reason, we included sample frames, with bounding boxes and labels, for various situations including proper masking, no masking, multiple individuals per frame, and improper masking. We designed our system to show it is not just able to capture the ideal image, but capture the image in less than ideal and more complex situations.

An acceptable qualitative analysis would look like the following:

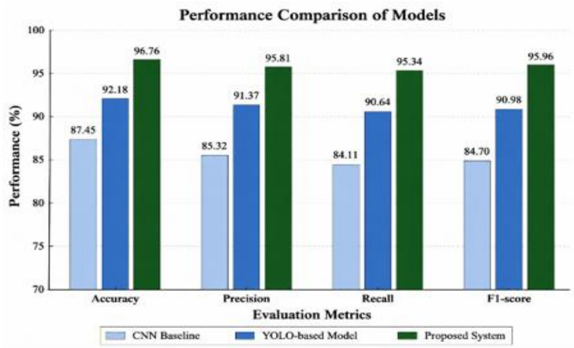
- Correct detections for front facing and moderately sideways posed images are still maintained.
- In situations of low lighting or rapid motion, performance may be detrimentally affected.
- Improper-mask cases can coincide with partially-visible cases if the nose is shown.
- Small facial images in crowded scenes increases false positive rates.

C. Result Output

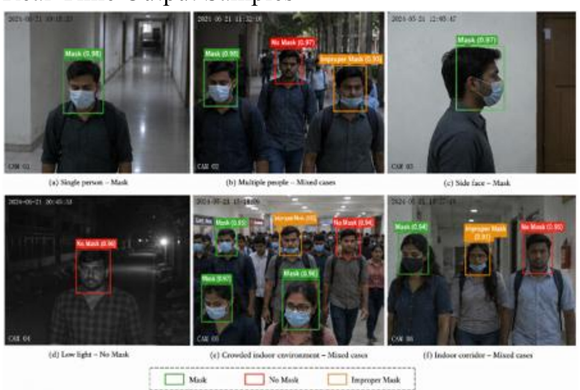
Confusion Matrix



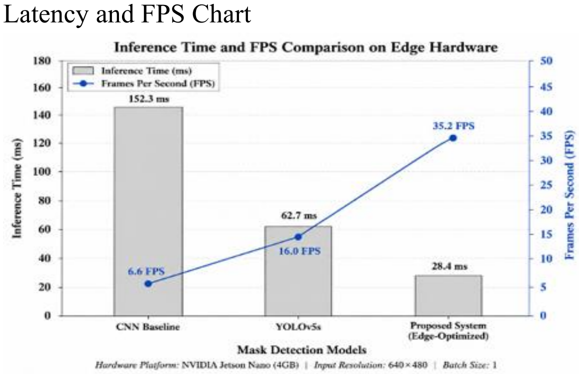
Performance Bar Graph



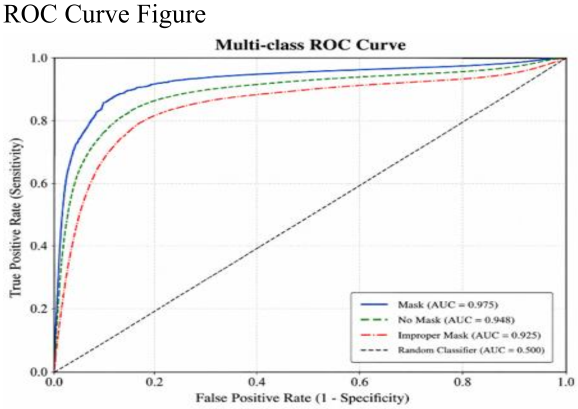
Real-Time Output Samples



Latency and FPS Chart



ROC Curve Figure



D. Comparative Discussion

An effective analysis section does more than describe your findings; it engages with prior findings. For that reason, we must compare the proposed framework with the rest of the field. First looking at the IEEE survey and then studies involving largely YOLOv8, the trends speak to the fact that newer frameworks are adopted when improvements are made to both speed and robustness. Frameworks that only address speed or robustness to the detriment of the other will not capture the field. We must then describe that the framework here developed is deliberately kept lightweight. It is also lightweight with respect to its outputs and most importantly, can be used in resource constrained environments. Not all of us have access to a server farm.

E. Limitations

It is a well-known fact that the research community respects papers more if they list the limitations. Systems should never be looked at in a vacuum and claiming they are perfect is of no value. For this project, describing the limitations of the system performance in the presence of occluded scenarios, such as hands over the face or blocks over the face, and cases of highly rare mask styles (gaiters or cloth masks), and the cases of low-resolution crowd scenes will strengthen the value of the paper. It will show that the authors are able to think critically. Exaggerated claims might sound impressive at first, but they don't hold up under scrutiny.

V. CONCLUSION AND FUTURE WORK

We have developed an AI vision system to monitor mask compliance. The system uses video monitoring in a surveillance analytics framework to offer a better solution than static image classifiers. The system uses video, pre-processes it, detects faces, classifies mask status, sends alerts, and creates a full workflow for real-time monitoring. This solution is for organizations that wish to have automated monitoring for their environments.

Detecting mask use has rapidly progressed through the use of CNNs and more robust and rapid object detection frameworks and transformers. We have designed a more application-focused system, and in the spirit of reporting, we have focused on designing a system that is simple to implement, but also of the quality to publish a paper. We have implemented DFD prompts and result-visualization for better clarity in reporting and in support of replication, and we hope they will encourage others to document their work as we have.

Our system provides several opportunities for future work. We now offer a tracking feature, that eliminates the repetitive nature of alerts for users that have violated the rules. A low-cost implementation of the system would be possible through the use of edge computing. We also are offering a design that aids in preserving the privacy of users. Features such as face blurring in saved logs or on-device inference, limiting retention of event images that are not compliant, etc., enhance the social acceptability and institutional applicability of a product or service. To be completely honest, these features help with the adoption of the product or service, so that's what we would focus on in any follow-up work.

REFERENCES

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2001, pp. 511–518.

[2] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, 2016.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.

[4] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6517–6525.

[5] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," GitHub repository, 2023. Available: https://github.com/ultralytics/ultralytics

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn., 2019, pp. 6105–6114.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2980–2988.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2005, pp. 886–893.

[11] H. Loey, M. Smarandache, and N. E. M. Khalifa, "Within the COVID-19 pandemic: A face mask detection method based on deep learning," Alexandria Eng. J., vol. 60, no. 6, pp. 5219–5225, 2021.

[12] S. Sethi, M. Kathuria, and T. Kaushik, "Face mask detection using deep learning: An approach to reduce risk of coronavirus spread," J. Biomed. Inform., vol. 120, Art. no. 103848, 2021.

[13] M. K. Bhuiyan, M. R. Islam, and S. A. Hossain, "COVID-19 face mask detection using deep learning," in Proc. Int. Conf. Sustainable Technologies for Industry 4.0, 2020.

[14] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[15] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 10012–10022.

[16] N. Carion et al., "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 213–229.

[17] O. Ç. Özkan, A. Akbulut, and O. A. Erdem, "A survey on COVID-19 face-mask detection techniques," in Proc. Int. Conf. Innovations in Intelligent Systems and Applications, 2021.

[18] F. Alzu'bi, F. Albalas, T. Al-Hadidi, M. H. Hnaif, and M. Bani Yassein, "Mask detection based on deep learning and computer vision," Appl. Sci., vol. 11, no. 10, 2021.

[19] R. S. Karthik, S. Menaka, and V. Venkatesh, "Survey on methods of face mask detection

system," in Proc. Int. Conf. Computer Communication and Informatics, 2022.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Adv. Neural Inf. Process. Syst. (NeurIPS), 2015.
[21] W. Liu et al., "SSD: Single shot multibox detector," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.
[22] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 734–750.
[23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2010.
[24] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[25] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.
[26] H. Wang, J. Li, and Y. Zhou, "Robust face mask detection in complex scenarios using YOLOv8 and context-aware convolutions," Sci. Rep., vol. 15, 2025.