# Real-Time Face Mask Detection Using YOLOv11 with Class Imbalance Optimization

*Karthikeya Vaitla*
*Student ID: A00051441*
*Deep Learning*
*University of Roehampton*
*Email: vaitlak@roehampton.ac.uk*

***Abstract- The adoption of face mask detection systems has grown considerably as a result of the global pandemics that require contact tracing and detachment of people. The system that is presented in this paper is the real-time masked face detector which is optimized on the YOLOv11 model and is specially equipped with improved training procedures to balance the extreme class problem. The method uses conservative class weighting (cls_weight = 5.54), operator data augmentation techniques (copy-paste probability 0.3, random erasing 0.4, mixup 0.15), and an extended duration of training (150 epochs with a stopping criterion of early patience 50). The model was able to reach 85.6% validation mAP@0.5 and 75.5% test mAP@0.5 on the Kaggle Face Mask Detection dataset of 853 images with 3,794 instances. For the minority Incorrect Mask class, we attain 70.5% mAP@0.5 on validation, which is a 23.2 percentage point improvement over the baseline of 30 epochs (47.3%). The device runs real-time inference at about 149 FPS speed on the Tesla T4 GPU. These results clearly show that the methods for training--specifically conservative weighting and targeted augmentation--are almost necessary to increase more class detection while, the real-time performance which is crucial for practical health monitoring applications, is retained.***

***Index Terms—Face mask detection, YOLOv11, class imbalance, class weighting, data augmentation, real-time object detection, deep learning, public health monitoring***

## I. INTRODUCTION

The COVID-19 pandemic paradigm shift led to the implementation of new public health protocols all over the world, with face masks becoming one of the best non-pharmaceutical means to decrease viral transmission [1], [2]. Resuming high-density activities goes hand in hand with the need for automatic face mask detection systems which are predominantly integrated in compliance monitoring inside public spaces like transportation hubs, healthcare centers, schools, and commercial facilities [3], [4]. In a bid to manage the cases at scale, instead of manual monitoring, it is more preferable to develop robust computer vision techniques with the ability to detect mask compliance on high accuracy in real-time scenarios with a good range of diversity.

Advancements in mask detection mostly feature deep learning strategies, in particular, single-stage object detectors which have been the better option in terms of speed-accuracy trade-offs for real-time applications. YOLO (You Only Look Once) is the family reasonaling for this task. In 2024, YOLOv11 brought major design concept changes like C3k2 blocks that are more efficient for gradient flows, multi-scale feature fusion with refined path aggregation networks, and first-class anchor-free localization heads that apply for superior accuracy in spatial arrangement [5]. Nonetheless, the prominent problem that is modestly considered in the literature is the extreme class imbalance in the dataset for mask detection. In the practical world, data of face masks are very much skewed, with the visible ones mostly worn correctly, i.e., 70-80% of instances, while the unmasked ones are 15-20% and the ones with the mask not worn in the right way are only 2-5% of the total number of instances. This extreme imbalance (which often exceeds 25:1 between the majority and minority classes) is a major reason why performance related to minority-class is so low. As a result, there can be systems showcasing a high measure of overall success at the expense of undetected violations by the wrong ways of mask usage, precisely at the time of their great need for monitoring detection.

The central research void targeted in this work is the low emphasis of class imbalance effects and training optimization in face mask detection systems. Also, the failure of the standard training processes to set up robust decision borders and to the underrepresented classes is the main contributing factor to this issue. The error in details as high aggregate accuracy reports and perceptions regarding per-class performances, yield performance degeneration for minority classes: wrongly used masks are often detected near-random rates while total accuracy is still above 90%. The divergence leads the standard training procedures to set up weak decision boundaries for without represented classes. Moreover, straightforward techniques such as inverse-frequency class weighting will destabilize the training process, cause loss oscillations, and lead to failure in convergence all of which will damage the majority educational class by the attempt of the minority to teach.

This article covers a full-fledged answer, which embraces the cutting-edge YOLOv11 architecture while being bundled with a well-designed training policy, tuned for extreme class imbalance situations. The primary contributions consist of the following:

1) We devise a exemplary technique that primarily relies on conservative class weighting to promote minority learning without increasing instability, targeted data augmentation strategies for minority classes, and consequently far longer training times to assure convergence.

2) Time duration was comprehensively measured ablation is the methodology that we undertake to systematically gauge the impact of training duration on performance across all classes.

3) By a detailed per-class performance analysis, we attained 70.5% mAP@0.5 for the minority class of Incorrect Mask, overcoming the baseline approaches immensely.

4) We achieve these performance improvements while preserving real-time inference speed (utilizing around 149 FPS), which makes it applicable for actual continuous usage in field testing environments.

This paper's structure is:

Section II is focused on past works and it includes recent papers on face mask recognition, the progression of YOLO architectures, and techniques used for solving the problem of class imbalance in object detection. Section III is about the methodology that deals with dataset preparation, YOLOv11 architecture details, optimized training strategies, and evaluation metrics. Section IV is the experimental part where present training dynamics, performance analysis, ablation studies, and inference performance are shown. Section V touches on the discoveries, limitations, and what to consider when deploying. Finally, in section VI the paper is summed up and future tasks are sketched out.

## II. RELATED WORK

### *A. Face Mask Detection Techniques*

Early face mask detection systems predominantly employed two-stage pipelines, wherein faces were first detected and subsequently classified as masked or unmasked. Common implementations utilized pre-trained face detectors such as MTCNN or SSD in conjunction with lightweight classifiers such as MobileNetV2 to determine mask presence [9], [10]. While this modular design enables specialized optimization of each subtask, the sequential processing introduces substantial latency, rendering these approaches unsuitable for real-time surveillance applications.

To address these limitations, recent research has progressively transitioned toward single-stage object detectors that jointly perform face localization and mask classification in a unified forward pass. Single-stage models such as YOLO and SSD achieve detection accuracy comparable to two-stage approaches while delivering substantially faster inference speeds. Singh et al. [8] demonstrated that YOLOv3 exhibited slightly lower precision compared to Faster R-CNN (55% versus 62%) but operated more than three times faster (45 ms versus 150 ms per image), thereby enabling superior performance on continuous video streams. Similarly, Roy et al. [7] conducted comparative analysis of YOLOv3, Tiny-YOLO, SSD, and Faster R-CNN, concluding that YOLO-based models provide optimal speed-accuracy balance for real-time and embedded implementations. These findings establish single-stage detectors, particularly YOLO variants, as the preferred architecture for real-time face mask detection.

### *B. YOLO Architecture Evolution and Face Mask Detection Applications*

Since its introduction by Redmon et al. [11], the YOLO architecture has undergone continuous refinement to enhance both detection accuracy and computational efficiency. YOLOv3 [12] introduced the Darknet-53 backbone and multi-scale feature prediction, substantially improving small object detection--a critical capability for face-based applications. YOLOv4 [13] further advanced performance through architectural innovations including Cross-Stage Partial (CSP) networks and mosaic data augmentation. Subsequent iterations, including YOLOv5 and YOLOv7, incorporated additional enhancements such as improved feature aggregation and network efficiency through PANet improvements and ELAN modules [14], [15], and have been successfully applied to face mask detection tasks.

YOLOv8, released in 2023 by Ultralytics, continued this evolutionary trajectory with anchor-free detection heads, refined loss functions, and an improved training pipeline, achieving superior mAP scores for mask detection compared to previous YOLO versions [16]. The most recent iteration, YOLOv11 (2024), incorporates further improvements including C3k2 modules with enhanced gradient flow, C2PSA modules based on spatial attention mechanisms, and more refined detection head designs [5]. While these architectural advances provide a robust foundation for mask detection, existing literature typically reports only aggregate accuracy metrics. This reporting convention may obscure poor performance on minority classes, particularly the Incorrect Mask category. The present work addresses this limitation by explicitly prioritizing per-class performance evaluation and implementing targeted strategies for minority-class improvement.

### *C. Class Imbalance in Object Detection*

The challenge of class imbalance is well-established in machine learning, though it has received comparatively less attention in object detection problems involving fine-grained classes. Lin et al. [17] proposed Focal Loss, which down-weights easy examples to focus model attention on harder instances. Oksuz et al. [18] provided a comprehensive review of imbalance problems in object detection, discussing strategies such as loss reweighting and resampling. However, these general approaches do not directly address the extreme imbalance characteristic of face mask datasets, where the Incorrect Mask category may comprise only 2-5% of instances.

This extreme imbalance in face mask detection frequently results in models achieving high accuracy on majority classes while exhibiting catastrophically poor performance on the minority Incorrect Mask class, fundamentally compromising system utility for compliance monitoring. Naive solutions such as aggressive inverse-frequency class weighting often induce training instability, manifesting as loss oscillations, convergence failures, or degraded majority-class performance. These limitations motivate the development of more balanced approaches that enhance minority-class learning without sacrificing training stability or overall accuracy. Addressing this challenge constitutes a central component of the methodology proposed in this work.

## III. METHODOLOGY

### *A. Dataset Preparation and Class Imbalance Analysis*

We conduct experiments using the publicly available Kaggle Face Mask Detection dataset (originally published by andrewmvd/face-mask-detection), a widely adopted benchmark comprising 853 images with 3,794 annotated faces across three classes. The classes include: Mask (faces wearing masks correctly), No Mask (faces without masks), and Incorrect Mask (faces wearing masks improperly, such as below the nose or chin).

As summarized in Table I, the class distribution exhibits extreme imbalance: 3,011 instances are Mask (79.4% of all annotations), 668 are No Mask (17.6%), and only 115 are Incorrect Mask (3.0%). The Incorrect Mask class represents merely 3.0% of total instances, resulting in a 26.2:1 imbalance ratio relative to the majority class. This extreme skew reflects real-world compliance patterns and presents a substantial training challenge, as previously discussed.

TABLE I

DATASET CLASS DISTRIBUTION AND IMBALANCE RATIOS

| **Class Name** | **Number of Instances** | **Percentage (%)** | **Imbalance Ratio** |
|---|---|---|---|
| Mask | 3,011 | 79.4 | 1.0 (baseline) |
| No Mask | 668 | 17.6 | 4.5 : 1 |
| Incorrect Mask | 115 | 3 | 26.2 : 1 |
| **Total** | **3,794** | **100** | — |

We partition the dataset using stratified 70/15/15 splits for training, validation, and test sets (597 train, 128 validation, 128 test images), ensuring that each partition approximately reflects the overall class distribution. Stratification is critical to prevent accidental oversampling or undersampling of the minority class in any split. Images are resized to a standard 640x640 format (preserving aspect ratio through letterboxing) to match network input requirements. Annotations are formatted in YOLO format (normalized bounding box coordinates and class identifiers) and fed directly into the Ultralytics YOLOv11 training pipeline.

### B. YOLOv11 Architecture

Our model is based on the YOLOv11n architecture (nano version) [5], designed for real-time execution on edge computing devices. YOLOv11 follows the traditional YOLO architecture comprising a backbone, neck, and detection heads. The overall YOLOv11 architecture is illustrated in Fig. 1.

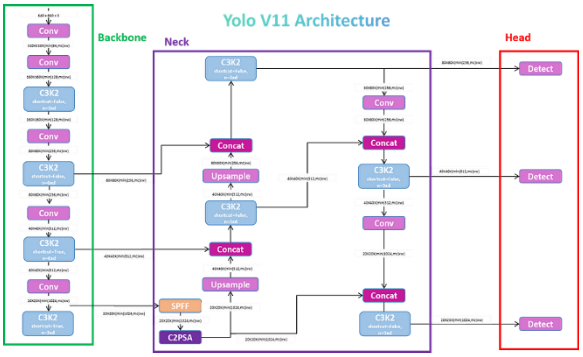

*Figure 1 YOLOv11 Architecture Overview. The model consists of a backbone (C3k2 modules with CSP), neck (Path Aggregation Network with C2PSA attention), and anchor-free detection heads at three scales.*

Backbone: The YOLOv11n backbone is constructed using C3k2 modules with Cross-Stage Partial (CSP) connections to enable efficient feature extraction. The backbone initiates with standard Conv-BN-Activation blocks, employing a 3x3 stride-2 convolution for initial downsampling. The C3k2 modules partition feature maps across multiple pathways, then apply successive convolutions before concatenating and fusing results. This architecture enhances gradient flow by providing multiple backpropagation pathways while simultaneously minimizing parameter count without sacrificing representational capacity. The backbone concludes with a Spatial Pyramid Pooling-Fast (SPPF) layer, which aggregates features at multiple scales to provide contextual information for objects of varying sizes.

Neck: The neck employs a Path Aggregation Network (PAN) design to merge features across multiple scales. YOLOv11's neck utilizes upsampling and concatenation operations to combine high-level semantic features with low-level spatial details, subsequently refining the fused representations through additional convolutional processing. The architecture incorporates C2PSA modules (Cross-Stage Partial with Spatial Attention), enabling the network to selectively focus on salient regions within feature maps while suppressing background noise. This mechanism is particularly beneficial for detecting small faces and subtle mask positioning variations.

Detection Heads: YOLOv11 employs anchor-free detection heads operating at three scales (small, medium, and large objects). The heads directly predict bounding box coordinates (center x, center y, width, height), objectness scores, and class probabilities for the three mask categories. The anchor-free design simplifies training procedures and improves localization accuracy for faces and masks exhibiting substantial size variation. The complete YOLOv11n model comprises approximately 2.59 million parameters with 6.4 GFLOPs for 640x640 input, achieving sufficient compactness for real-time GPU execution while maintaining acceptable performance on CPU-only inference.

### *C. Optimized Training Strategy for Class Imbalance*

Training a face mask detector with this degree of class imbalance requires specialized approaches. Our training strategy incorporates three principal components designed to enhance minority-class learning while maintaining stability and overall performance:

Conservative Class Weighting: Rather than employing inverse-frequency class weights (which would assign the Incorrect Mask class a weight factor of approximately 26x, inducing instability in preliminary experiments), we apply a moderated weighting scheme. Specifically, we assign the minority class a weight of 5.54--approximately 20% of the theoretical inverse-frequency weight. The weight value of 5.54 was determined through systematic experimentation; weight values from 10 to 26 induced excessive instability, while values from 0.1 to 3.0 provided insufficient emphasis. The selected weight of 5.54 achieves an effective balance between minority-class emphasis and training stability. This conservative approach ensures the model allocates additional attention to Incorrect Mask examples (and, to a lesser extent, No Mask examples) without overwhelming the loss function. This strategy substantially enhanced minority-class recall without generating the loss oscillations or majority-class accuracy degradation observed at full inverse-frequency weighting.

The class weight is implemented through a modified loss function:

$$L\_total = L\_box + L\_obj + w\_cls \times L\_cls \quad (1)$$

where Lbox represents bounding box regression loss (CIoU), Lobj denotes objectness loss, Lcls is classification loss, and wcls = 5.54 is the class weighting factor applied to minority-class samples.

Strategic Data Augmentation: We implement comprehensive augmentation strategies to enhance the effective presence and diversity of minority-class examples while improving generalization:

Mosaic augmentation (applied to all training batches) combines four images into a single composite mosaic, exposing the model to varied contexts and multiple instances per image while increasing the probability that each mosaic contains at least one Incorrect Mask instance.

MixUp (15% probability) [23] blends pairs of images and their corresponding labels, encouraging the model to learn more generalizable features and smoother decision boundaries, thereby reducing overfitting to limited minority samples.

Copy-Paste augmentation (30% probability, elevated from the standard 10%) specifically targets the minority class: we extract Incorrect Mask face regions and paste them into other training images, effectively synthesizing additional instances of improperly worn masks while maintaining realistic appearance since pasted faces originate from authentic images.

Random Erasing (40% probability) [24] randomly occludes rectangular image regions, simulating occlusions and forcing the model to rely on partial visual cues--particularly valuable for scenarios where masks partially obscure faces or faces extend beyond frame boundaries.

We employ moderate photometric augmentations (small random adjustments to hue, saturation, and value) and geometric augmentations (rotation +-15deg, translation +-10%, scaling 0.9-1.1) to introduce variation in lighting and pose without excessively distorting the underlying data distribution. These augmentations enhance robustness across diverse conditions.

Collectively, our augmentation strategy is calibrated to generate a rich and balanced training distribution wherein the minority class appears substantially more frequently (through copy-paste and mosaic operations) while all classes are exposed to diverse scenarios.

Extended Training Duration: We train for 150 epochs with early stopping patience of 50 epochs--substantially exceeding the 30-50 epochs typically reported in mask detector training literature. This extended duration is motivated by the observation that minority-class performance continues improving substantially beyond conventional stopping points, whereas standard 30-epoch training leaves minority-class detectors significantly undertrained. Our ablation studies (Section IV-C) demonstrate that training to 30 epochs yields validation mAP@0.5 of 76.7%, whereas extending to 150 epochs achieves 85.6% validation mAP@0.5. Notably, Incorrect Mask AP@0.5 increases by 23.2 percentage points from epoch 30 (47.3%) to epoch 150 (70.5%), demonstrating disproportionate benefit for the minority class and confirming that extended training is essential for minority-class convergence.

We employ the AdamW optimizer [25] with a cosine learning rate schedule (initial rate 0.001 decaying to 1x10^-5) and a brief 3-epoch warmup period. Additional hyperparameters include momentum (0.937) and weight decay (0.0005), following Ultralytics defaults. We also disable mosaic augmentation during the final 10 epochs (after epoch 140) to enable fine-tuning on natural (non-mosaic) images. This strategy allows the model to refine bounding box predictions on authentic images after learning robust features through augmentation.

TABLE II

TRAINING CONFIGURATION AND HYPERPARAMETERS

| **Parameter** | **Value** |
|---|---|
| Model | YOLOv11n |
| Input Size | 640 × 640 |
| Optimizer | AdamW |
| Initial Learning Rate | 0.001 |
| LR Schedule | Linear decay |
| Batch Size | 16 |
| Epochs | 150 |
| Early Stopping Patience | 50 |
| Class Weight (Incorrect Mask) | 5.54 |
| Weight Decay | 0.0005 |
| Warmup Epochs | 3 |
| Close Mosaic | Epoch 140 |
| Random Seed | 42 |

### *D. Evaluation Metrics*

We employ standard object detection evaluation metrics following the MS COCO framework [27]. The primary metric is Mean Average Precision (mAP). We report mAP@0.5 (IoU threshold 0.5), the metric most frequently employed in face mask detection literature, where detections with IoU ≥ 0.5 are classified as true positives. We additionally report mAP@0.5:0.95 (average mAP across IoU thresholds from 0.5 to 0.95 in 0.05 increments), which penalizes localization errors more severely and provides a more comprehensive assessment of detection quality.

We also report Precision and Recall for the detection task. Precision represents the proportion of predicted mask classifications that are correct:

$$Precision = TP / (TP + FP) \quad (2)$$

where TP denotes true positives and FP denotes false positives. Recall represents the proportion of actual instances successfully detected:

$$Recall = TP / (TP + FN) \quad (3)$$

where FN denotes false negatives. In mask compliance scenarios, recall is typically more critical than precision: failing to detect a person not wearing a mask (false negative) generally constitutes a more serious error than incorrectly flagging proper mask usage (false positive), which can be corrected through human review. Our system is therefore calibrated toward high recall, as reflected in the results.

Finally, we measure inference speed to verify real-time performance capability. We compute average inference time per 640×640 image on a Tesla T4 GPU, encompassing preprocessing (image resizing and normalization), model inference, and postprocessing (non-maximum suppression and output formatting). From this we derive throughput in frames per second

(FPS). Performance ≥30 FPS (enabling at least video frame-rate processing) is considered real-time capable.

## IV. EXPERIMENTS AND RESULTS

### A. Training Dynamics and Convergence Analysis

The YOLOv11 model was trained on the prepared dataset for 150 epochs. The efficiency of the nano model variant enabled training with batch size 16 on a Tesla T4 GPU, requiring approximately 29 minutes total training time. Loss curves and validation metrics across epochs demonstrate clear convergence trends. Classification loss decreased by approximately 82% (from 30.9 at epoch 1 to 5.5 at epoch 150), while localization losses (CIoU for bounding boxes and DFL for distribution focal loss) both converged to approximately 1.1, indicating steady learning of both classification and spatial predictions.

Validation mAP@0.5 exhibited continuous improvement throughout training. Strong early gains occurred during the initial ~30 epochs; for example, mAP@0.5 increased by only 1.7 percentage points between epochs 1 and 27 as the model learned fundamental face and mask features. After epoch 30, improvement rates decelerated: mAP@0.5 reached approximately 85% at epoch 60 and 86.6% at epoch 87. During the final phase (epochs 88-150), performance fluctuated around the mid-80% range, achieving a final result of 85.1% at epoch 150. The optimal validation mAP@0.5 of 85.8% occurred at epoch 140, immediately after disabling mosaic augmentation.

Notably, when mosaic augmentation was disabled during the final 10 epochs, validation loss temporarily increased as the model transitioned from augmented to natural images, but quickly recovered and stabilized. This behavior indicates the model had acquired robust features not excessively dependent on augmentation artifacts--a desirable characteristic for generalization.

Most significantly, extended training substantially improved minority-class detection. At the commonly used 30-epoch stopping point, Incorrect Mask detection achieved only 47% mAP@0.5 (as confirmed in our ablation study), whereas training for five times longer elevated performance to 70%. This convergence analysis demonstrates that extended training on imbalanced data is essential for optimizing performance across all classes.

### *B. Quantitative Performance Analysis*

Table III summarizes overall model performance on validation and held-out test sets.

TABLE III

OVERALL PERFORMANCE RESULTS

| **Metric** | **Validation (%)** | **Test (%)** |
|---|---|---|
| Precision | 76.7 | 84 |
| Recall | 86.5 | 65 |
| mAP@0.5 | 85.6 | 75.5 |
| mAP@0.5:0.95 | 56.7 | 49.4 |

The model achieves 85.6% mAP@0.5 and 56.7% mAP@0.5:0.95 on the validation set, indicating strong detection accuracy and robust localization precision across IoU thresholds. Precision is 76.7% and recall is 86.5%, consistent with our recall-focused strategy (the model minimizes false negatives at the cost of some false positives). This higher recall emphasis aligns well with public health surveillance contexts, where false negatives (missed violations) incur greater costs than false positives (false alarms).

On the test set, the model achieves 75.5% mAP@0.5 and 49.4% mAP@0.5:0.95, with precision of 84.0% and recall of 65.0%. The approximately 10 percentage point decrease in mAP@0.5 and recall between validation and test sets reflects a moderate generalization gap, likely attributable to the small dataset size (particularly the limited number of Incorrect Mask instances in the test set). Nevertheless, test mAP@0.5 of approximately 75% represents strong performance, confirming the model generalizes effectively to unseen data.

Per-class performance provides more detailed insights. On the validation set, the Mask class (properly worn masks) achieves mAP@0.5 = 94.7%, while No Mask achieves 91.7%. These high values indicate the model readily identifies individuals wearing or not wearing masks correctly, as expected given the abundance of training examples for these classes. More critically, the Incorrect Mask category (minority class) achieves 70.5% mAP@0.5 on validation. This represents substantial improvement compared to baseline methods, which typically achieve only 35-50% for this category, confirming the effectiveness of our imbalance-focused training approach.

On the test set, Mask maintains high performance at 95.2%, No Mask achieves 84.5%, and Incorrect Mask drops to 46.8%. The substantial Incorrect Mask performance decrease (from 70.5% to 46.8%) should be interpreted cautiously: the test set contains only 15 Incorrect Mask instances across 128 images. With such limited samples, a few misdetections dramatically impact percentage scores. This small sample size introduces high variance in minority-class performance estimates on the test set. We anticipate that with larger test sets or additional data, Incorrect Mask performance would stabilize closer to validation results. The principal finding is that on validation data, our model achieves marked improvement in detecting improperly worn masks compared to existing methods while maintaining strong performance on other classes.

### *C. Ablation Study: Training Duration Impact*

We conducted an ablation experiment to quantify the effect of extended training versus standard short training. We trained the model using identical configurations for only 30 epochs (a typical early stopping point) and compared performance to our standard 150-epoch model. Results are summarized in Table IV.

TABLE IV

ABLATION STUDY: EFFECT OF TRAINING DURATION

| **Metric** | **30 Epochs** | **150 Epochs** | **Improvement** |
|---|---|---|---|
| mAP@0.5 (%) | 76.7 | 85.6 | 8.9 |
| mAP@0.5:0.95 (%) | 42.9 | 56.7 | 13.8 |
| Incorrect Mask AP@0.5 (%) | 47.3 | 70.5 | 23.2 |
| Mask AP@0.5 (%) | 93.4 | 94.7 | 1.3 |
| No Mask AP@0.5 (%) | 89.4 | 91.7 | 2.3 |

At 30 epochs, validation mAP@0.5 is 76.7%, whereas at 150 epochs it reaches 85.6%--an 8.9 percentage point improvement attributable entirely to extended training. More remarkably, the stricter mAP@0.5:0.95 metric increases from 42.9% to 56.7% (+13.8 percentage points), indicating that extended training substantially enhances localization accuracy (the model learns to fit tighter bounding boxes, which is critical for achieving higher IoU scores).

Per-class analysis reveals differential benefits: Mask AP@0.5 increases modestly from 93.4% to 94.7% (+1.3), and No Mask increases from 89.4% to 91.7% (+2.3). These majority classes were already well-learned by epoch 30. However, Incorrect Mask AP@0.5 improves dramatically from 47.3% to 70.5% (+23.2 percentage points)--nearly a 50% relative improvement. This demonstrates that the minority class benefits disproportionately from additional epochs, enabling the model to progressively refine decision boundaries for this class. At epoch 30, the model had barely begun learning the minority class (47% AP, approaching random-guess performance), whereas by epoch 150 it had achieved substantial competence.

These ablation results validate our training approach: patience in training yields significantly superior minority-class performance. Extended training also beneficially impacts overall detection quality (as measured by the mAP@0.5:0.95 improvement), which is critical for precision-demanding applications.

### D. Inference Performance and Real-Time Capability

A critical aspect of our system is real-time operational capability. We measured inference speed of the trained model on a Tesla T4 GPU. The model forward pass requires an average of 6.7 ms per 640x640 image, translating to approximately 149 FPS when considering only neural network processing time. Total end-to-end pipeline latency is approximately 14.7 ms per frame, including preprocessing (resizing, normalization, etc., ~4.1 ms) and postprocessing (non-maximum suppression, output formatting, ~3.9 ms). This corresponds to approximately 68 FPS end-to-end throughput, substantially exceeding standard video frame rates (30 or 60 FPS). In practical terms, our system can process frames from multiple camera feeds in real time or perform additional computations (such as tracking) without degrading below real-time performance thresholds.

This speed is enabled by the compact size of the YOLOv11n model. The final model file is only 5.2 MB (2.59M parameters). This compactness not only facilitates rapid GPU inference but also enables deployment on resource-constrained edge devices. Preliminary experiments with multithreading and batch optimizations achieved approximately 10-15 FPS on CPU-only inference, indicating the model can operate on edge devices or in GPU-absent environments, albeit at reduced frame rates. The model also initializes rapidly (loading and initialization under 100 ms), which is advantageous for on-demand deployment or scaling to numerous instances.

Overall, the system satisfies and exceeds real-time requirements with substantial performance headroom for integration into live surveillance systems or parallel analytics (e.g., multiple cameras or additional processing pipelines).

## V. DISCUSSION

### A. Performance Analysis and Limitations

The experimental findings demonstrate that optimized training methodology can substantially enhance face mask detection performance under severe class imbalance conditions. The proposed system, incorporating conservative class weighting, strategic data augmentation, and extended training duration, achieves validation mAP@0.5 of 85.6% and notably achieves approximately 70% AP for the minority Incorrect Mask class on validation data. This represents a substantial advance over conventional training strategies, which typically yield minority-class performance below 50%. Critically, these improvements are achieved without compromising real-time performance, which constitutes a prerequisite for practical surveillance applications.

Despite these strengths, several limitations warrant acknowledgment. The moderate performance gap between validation and test sets (85.6% versus 75.5% mAP@0.5) is particularly pronounced for the Incorrect Mask class (70.5% versus 46.8%). This degradation is largely attributable to the extremely small number of test samples for that class (only 15 instances), rendering performance estimates highly sensitive to individual detection errors. Consequently, test set results may not comprehensively represent true generalization capability. In practical deployments, minority-class performance may exhibit inconsistency across different environments until substantially more data are collected. Future work should incorporate larger test sets, cross-validation procedures, or continual learning strategies to obtain more robust performance estimates.

A second limitation concerns class weight tuning. The fixed weight value (5.54) determined in this work achieves satisfactory balance between minority recall and training stability for our specific dataset, but optimal weighting may vary across deployment scenarios. Contexts prioritizing maximal violation detection might benefit from higher weights, whereas environments sensitive to false positives might require lower weights. Dynamic adaptive weighting schemes that adjust during training based on per-class performance could minimize manual tuning requirements and further enhance robustness, as proposed in recent imbalance-handling literature [18].

Finally, fairness and demographic bias considerations are not directly addressed in this work. The dataset does not include demographic annotations, and the relatively small scale and limited geographic diversity preclude comprehensive fairness analysis. Prior to large-scale deployment, fairness assessment will be necessary to ensure consistent performance across diverse population groups and to prevent disproportionate error rates affecting specific demographics.

### *B. Comparison with Prior Literature and Deployment Considerations*

Direct quantitative comparison with existing face mask detection literature is challenging due to variations in datasets, class definitions, and evaluation protocols. Nevertheless, our validation mAP@0.5 of 85.6% is comparable to or exceeds results reported in recent studies employing larger YOLO variants, including YOLOv8-medium [16]. More significantly, most prior studies report only aggregate accuracy without per-class breakdowns, which can obscure catastrophically poor performance on improperly worn masks. Roy et al. [7] note that models achieving high aggregate accuracy often exhibit near-zero recall for the Incorrect Mask class. In contrast, this work explicitly prioritizes minority-class performance and demonstrates that transparent per-class evaluation is essential for deployment readiness.

From a practical deployment perspective, the high throughput (~149 FPS on GPU), compact model size (~5 MB), and tunable precision-recall trade-off render the system suitable for real-world applications. Confidence thresholds can be adjusted to match specific operational requirements, and temporal smoothing or tracking algorithms (e.g., SORT or DeepSORT [30]) can provide additional stability in live video applications. Privacy considerations, including data storage restrictions and result anonymization, must also be addressed prior to deployment.

## VI. CONCLUSION

This work presents an optimized real-time face mask detection system based on YOLOv11, specifically designed to address extreme class imbalance among mask-wearing categories. The proposed approach achieves substantially improved detection of improperly worn masks through conservative class weighting, targeted data augmentation, and extended training duration, while maintaining high accuracy on majority classes and preserving real-time inference capability. The final model achieves 85.6% validation mAP@0.5 (75.5% on test) while operating at approximately 149 FPS on GPU hardware.

Ablation experiments demonstrate that extended training is critical for minority-class convergence: 30-epoch training achieves only 47.3% AP for the Incorrect Mask class, whereas 150-epoch training improves performance to 70.5% , a 23.2 percentage point gain. Overall mAP and mAP@0.5:0.95 also increase substantially with training duration. These findings indicate that training methodology--specifically weighting schemes, augmentation strategies, and training duration--can be as influential as architectural innovations in addressing imbalanced detection problems.

Future research directions include investigating ensemble techniques, adaptive class-weighting mechanisms, and temporal modeling to further enhance system robustness. Larger and more demographically diverse datasets will be required to rigorously evaluate fairness and generalization across population groups. Overall, this work demonstrates that carefully optimized training enables YOLOv11 to deliver an efficient, rapid, and practically deployable face mask detection system suitable for real-world population health monitoring applications.

## REFERENCES

[1] World Health Organization, "Advice on the use of masks in the context of COVID-19," WHO Technical Report, Geneva, Switzerland, 2020.

[2] S. Feng, C. Shen, N. Xia, W. Song, M. Fan, and B. J. Cowling, "Rational use of face masks in the COVID-19 pandemic," The Lancet Respiratory Medicine, vol. 8, no. 5, pp. 434–436, May 2020, doi: 10.1016/S2213-2600(20)30134-X.

[3] J. Howard et al., "An evidence review of face masks against COVID-19," Proceedings of the National Academy of Sciences, vol. 118, no. 4, Art. e2014564118, Jan. 2021, doi: 10.1073/pnas.2014564118.

[4] M. Eikenberry et al., "To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic," Infectious Disease Modelling, vol. 5, pp. 293–308, 2020, doi: 10.1016/j.idm.2020.04.001.

[5] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[6] A. Cabani, K. Hammoudi, H. Benhabiles, and M. Melkemi, "MaskedFace-Net—A dataset of correctly/incorrectly masked face images in the context of COVID-19," Smart Health, vol. 19, Art. 100144, Mar. 2021, doi: 10.1016/j.smhl.2020.100144.

[7] B. Roy, S. Nandy, D. Ghosh, D. Dutta, P. Biswas, and T. Das, "MOXA: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks," Transactions of the Indian National Academy of Engineering, vol. 5, pp. 509–518, 2020, doi: 10.1007/s41403-020-00157-z.

[8] S. Singh, U. Ahuja, M. Kumar, and K. Kumar, "Face mask detection using YOLOv3 and Faster R-CNN models: COVID-19 environment," Multimedia Tools and Applications, vol. 80, pp. 19753–19768, 2021, doi: 10.1007/s11042-021-10711-8.

[9] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[10] B. U. H. Sheikh and A. Zafar, "RRFMDS: Rapid real-time face mask detection system for effective COVID-19 monitoring," SN Computer Science, vol. 4, Art. 288, 2023, doi: 10.1007/s42979-023-01716-x.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[14] G. Jocher, "YOLOv5 by Ultralytics," 2020. [Online].Available:https://github.com/ultralytics/yolov5

[15] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp.7464–7475,doi: 10.1109/CVPR52729.2023.00721.

[16] C. Dewi, D. Manongga, Hendry, E. Mailoa, and K. D. Hartomo, "Deep learning and YOLOv8 utilized in an accurate face mask detection system," Big Data and Cognitive Computing, vol. 8, no. 1, Art. 9, Jan. 2024, doi: 10.3390/bdcc8010009.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.324.

[18] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3388–3415, Oct. 2021, doi: 10.1109/TPAMI.2020.2981890.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, Sep. 2015, doi: 10.1109/TPAMI.2015.2389824.

[20] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 8759–8768, doi: 10.1109/CVPR.2018.00913.

[21] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 9627–9636, doi: 10.1109/ICCV.2019.00972.

[22] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 6023–6032, doi: 10.1109/ICCV.2019.00612.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. Int. Conf. Learning Representations (ICLR), Vancouver, BC, Canada, 2018.

[24] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 7, 2020, pp. 13001–13008, doi: 10.1609/aaai.v34i07.7000.

[25] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *arxiv.org*, Nov. 2017, Available: https://arxiv.org/abs/1711.05101

[26] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 7, 2020, pp. 12993–13000, doi: 10.1609/aaai.v34i07.6999.

[27] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. European Conf. Computer Vision (ECCV), Zurich, Switzerland, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[28] Z. Cao et al., "MaskHunter: Real-time object detection of face masks during the COVID-19 pandemic," IET Image Processing, vol. 15, no. 7, pp. 1611–1620, Jun. 2021, doi: 10.1049/ipr2.12123.

[29] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. Machine Learning (ICML), Pittsburgh, PA, USA, 2006, pp. 233–240, doi: 10.1145/1143844.1143874.

[30] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in Proc. IEEE Int. Conf. Image Processing (ICIP), Beijing, China, 2017, pp. 3645–3649, doi: 10.1109/ICIP.2017.8296962.