Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Pipeline Optimisation for Real-World Masked Face Recognition: YOLOv8 Detection, Landmark-Guided Cropping, and Multi-Model Embedding Benchmarking

Hamza Zidoum

zidoum@squ.edu.om

Sultan Qaboos University

Shihab Al Busaidi

Sultan Qaboos University

Article

Keywords: masked face recognition, YOLOv8, FaceNet, ArcFace, landmark-guided cropping, RMFRD, occlusion-robust biometrics

Posted Date: June 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9435637/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Face masks, now commonplace in post-pandemic environments, expose critical weaknesses in face recognition systems trained exclusively on unoccluded faces. This paper presents a systematic end-to-end evaluation pipeline that combines a fine-tuned YOLOv8 masked face detector with landmark-guided upper-face cropping and a benchmarking study across three embedding models — FaceNet, ArcFace, and VGG-Face — on the Real-World Masked Face Recognition Dataset (RMFRD). A balanced 10,000-image subset (5,000 masked / 5,000 unmasked) was partitioned 80:20 by identity under a closed-set protocol. The fine-tuned detector achieved an image-level IoU ≥ 0.5 success rate (ISR@0.5) of 95.76% overall and 91.9% on masked faces, statistically validated with 95% confidence intervals and McNemar's test against MTCNN, RetinaFace, and Dlib. At the recognition stage, FaceNet consistently outperformed ArcFace and VGG-Face in baseline masked conditions (Rank-1: 46.5%, 45.8%, 34.2% respectively). Cosine similarity outperformed Euclidean and Manhattan distance across all models and conditions, while photometric augmentation produced changes within ± 2 percentage points. Landmark-guided cropping with MediaPipe — retaining only eyes, forehead, and upper cheeks — raised FaceNet's Rank-1 accuracy from 46.5% to 70.1% (+ 23.6 pp), with comparable gains for ArcFace (+ 21.6 pp) and VGG-Face (+ 8.1 pp). These results establish that multi-stage pipeline optimization, rather than single-model tuning, is the practical path to robust masked face recognition.

1 Introduction

Face recognition has become a standard component of security cameras, airport gates, and building-access systems [1]. The COVID-19 pandemic exposed a critical vulnerability: face masks cover the nose, mouth, and chin — regions that both detection and recognition models rely on heavily — causing systems trained on fully visible faces to misidentify or miss masked individuals altogether [2, 3]. This gap has created an urgent demand for recognition solutions that operate reliably under real-world occlusion without requiring system replacement.

A standard recognition pipeline comprises two linked stages: face detection, followed by identity matching via embedding comparison. Errors in detection cascade directly into recognition inaccuracies, underscoring the necessity for robust detectors capable of handling masked occlusions [4, 5]. Classical methods such as Haar cascades and Histogram of Oriented Gradients (HOG) perform adequately in controlled settings but fail under partial occlusion, poor lighting, and pose variation [6]. CNN-based detectors such as Multi-task Cascaded Convolutional Networks (MTCNN) and RetinaFace improved general-purpose performance but remain underexplored for the specific challenge of masked faces [7].

On the recognition side, models such as FaceNet [8], ArcFace [9], and VGG-Face [10] achieve above 99% accuracy on unmasked benchmarks such as Labeled Faces in the Wild (LFW). Under mask occlusion, however, accuracy drops substantially [11]. Most existing studies aggravate this problem by relying on synthetically overlaid masks [12, 13, 14], which fail to capture real-world variance in fabric, fit, and wearing style. Ritchie et al. [15] demonstrated empirically that both humans and algorithms perform

markedly worse on genuinely masked faces than on digitally masked ones, underscoring the need for real-world evaluation datasets.

A review of the literature reveals five persistent gaps: (i) reliance on synthetic mask datasets that overestimate real-world performance; (ii) limited benchmarking of modern one-stage detectors such as YOLOv8 against masked faces; (iii) narrow recognition benchmarking, typically comparing only one or two models without systematic cross-model analysis; (iv) incomplete evaluation of similarity metrics beyond cosine similarity; and (v) absence of quantified assessment of landmark-guided cropping within an integrated end-to-end pipeline. This paper directly addresses all five gaps through the following contributions:

(1) A fine-tuned YOLOv8 detector achieving ISR@0.5 of 95.76% overall and 91.9% on masked faces, statistically validated against MTCNN, RetinaFace, and Dlib.

(2) A systematic before/after benchmark of FaceNet, ArcFace, and VGG-Face under real mask occlusion using three similarity metrics (cosine, Euclidean, Manhattan) and photometric augmentation.

(3) Landmark-guided cropping via MediaPipe that removes occluded lower-face regions and raises FaceNet Rank-1 accuracy from 46.5% to 70.1%, with comparable gains across all three recognizers.

(4) An integrated end-to-end pipeline (YOLOv8 → MediaPipe cropping → FaceNet → cosine similarity) evaluated on the RMFRD real-world masked dataset.

2 Related Work

2.1 Deep Learning Face Recognition Under Occlusion

FaceNet [8], introduced by Schroff et al. (2015), uses triplet-loss training to produce compact 128-dimensional embeddings that maximize inter-class separation. Its strong performance on LFW (99.63%) established it as a de-facto benchmark. Deng et al. introduced ArcFace [9] with an additive angular margin loss, achieving 99.83% on LFW via 512-dimensional embeddings and enhanced intra-class compactness. VGG-Face [10], based on the VGG-16 architecture, was among the first deep models trained for face recognition at scale, achieving 98.95% on LFW, though its softmax-based training is less suited to angular similarity metrics. Despite these gains, Jeevan et al. [11] demonstrated that all three architectures degrade substantially under mask occlusion: ArcFace dropped from ~ 99% to 71.9% on real masked faces. Huang et al. [16] further quantified this on the WebFace-OCC benchmark, finding ArcFace accuracy reduced specifically when lower-face regions (mouth and chin) were occluded.

2.2 Synthetic vs. Real-World Masked Evaluation

Most prior work on masked face recognition uses synthetically overlaid masks [12, 13, 14]. Mishra and Reza [13] achieved 96% accuracy using Inception ResNet V1 on synthetically masked CASIA-WebFace, while Cheng et al. [14] used a genetic algorithm with FaceNet on VGGFace2 to improve from 55% to 65%

under synthetic occlusion. Ritchie et al. [15] provided the critical empirical correction: both human observers and algorithms perform significantly worse on genuine masks than on digitally superimposed ones, invalidating synthetic results as proxies for real-world deployment. This finding directly motivates the use of RMFRD in the present study.

2.3 Masked Face Detection

MTCNN [17] and RetinaFace [18] remain the standard baselines for face detection. MTCNN's three-stage cascade provides five-point landmark alignment but struggles under heavy lower-face occlusion. RetinaFace offers single-pass multi-scale detection using a Feature Pyramid Network (FPN), demonstrating strong performance on the WIDER dataset but degrading under partial occlusion. YOLOv8, validated by Li et al. [19] for helmet detection and Dewi et al. [20] for mask-wearing compliance, has demonstrated accuracy up to 99.1% on specific detection tasks when fine-tuned on domain-appropriate data. Its decoupled head and C2f backbone enable effective localization of partially visible facial regions. Systematic comparison of YOLOv8 against MTCNN and RetinaFace specifically on masked faces from a real-world dataset has not been previously published.

2.4 Similarity Metrics and Preprocessing Strategies

Huzaimi and Mazalan [21] demonstrated that the choice of similarity metric significantly impacts masked face matching, with cosine similarity generally outperforming Euclidean distance. However, Shi and Jain [22] cautioned that cosine similarity can yield misleadingly high scores under severe occlusion, and Pham et al. [23] showed that learned similarity functions such as Siamese networks with contrastive loss can outperform fixed metrics. Arif et al. [24] employed MTCNN and VGG-Face with cosine similarity and achieved approximately 87% on a mixed dataset, though no comparison of similarity metrics was conducted. Salim and Surantha [25] demonstrated that zeroing the masked lower-facial region improved recognition without model retraining. The present work extends this by using dynamic landmark-guided cropping (rather than fixed masking) and evaluating across all three major recognizers.

3 Methodology

The experimental pipeline follows the sequence: (i) dataset preparation and annotation; (ii) YOLOv8 fine-tuning and benchmarking; (iii) face recognition baseline across three embedding models; (iv) similarity metric and augmentation analysis; and (v) landmark-guided cropping (Fig. 1). All stages are evaluated on the RMFRD real-world masked dataset to ensure ecological validity.

3.1 Dataset

Real-World Masked Face Recognition Dataset (RMFRD) [26] contains approximately 95,000 images of 525 identities, with ~ 5,000 genuinely masked and ~ 90,000 unmasked images (Fig. 2). A balanced subset of 10,000 images (5,000 masked / 5,000 unmasked) was constructed and partitioned 80:20 by identity using stratified grouping to prevent identity leakage between training and testing sets. All images were standardized to 640×640 pixels in RGB format. Duplicates, corrupt files, and low-quality images (blurred,

poorly lit, or with occluded landmarks beyond the mask) were removed via automated Python scripts followed by manual review.

RMFRD was chosen specifically because it contains genuine masks in varied styles, fabrics, and wearing conditions — unlike MLFW and MFR2 which use synthetic overlays.

Table 1
Dataset partition structure (RMFRD balanced subset).

Subset	Masked	Unmasked	Total
Train (80%)	4,000	4,000	8,000
Test (20%)	1,000	1,000	2,000
Total	5,000	5,000	10,000

Bounding-box annotations for the YOLO detection task were created with LabelImg in YOLO format using two classes: Class 0 (unmasked) and Class 1 (masked). Each annotation was manually verified for bounding-box accuracy.

3.2 Face Detection

3.2.1 YOLOv8 Architecture and Fine-tuning

YOLOv8 (Ultralytics, v8.1) was selected due to its anchor-free decoupled prediction heads, C2f cross-stage feature fusion module, and Spatial Pyramid Pooling Fast (SPPF) layer — all of which support effective localization of partially visible facial regions. Transfer learning from pre-trained yolov8n.pt weights (COCO) was used, with the detection head retrained on the RMFRD training split. Table 2 summarizes the hyperparameters.

Table 2
YOLOv8 training hyperparameters.

Hyperparameter	Value
Pre-trained weights	yolov8n.pt (COCO)
Optimizer	Adam
Learning rate	0.01 (constant)
Batch size	32
Max epochs	1,000 (early stopping enabled)
Input size	640 × 640 px
Deployed confidence threshold	0.524
Hardware	Apple M1 Pro, 16 GB RAM (MPS backend)

3.2.2 Baseline Detectors

Three detectors were benchmarked against YOLOv8 using the 1,000-image masked test subset, with no retraining applied to any baseline: (a) Dlib HOG + SVM — a classical gradient-based detector providing a non-CNN baseline; (b) MTCNN [17] — a three-stage cascaded CNN providing five-point landmarks, representing the most widely used benchmark in the masked face literature; and (c) RetinaFace [18] — a single-stage FPN-based detector trained on WIDER. A detection was considered correct if the predicted bounding box achieved IoU ≥ 0.5 with the ground-truth face annotation.

3.3 Face Recognition Models

FaceNet [8] — 128-dimensional triplet-loss embeddings (Inception-ResNet backbone); **ArcFace** [9] — 512-dimensional additive angular margin embeddings (ResNet backbone); and **VGG-Face** [10] — 512-dimensional softmax-trained embeddings (VGG-16 backbone). All three models were used as pre-trained feature extractors without further fine-tuning on RMFRD, following a zero-shot embedding evaluation protocol. Each test face embedding was compared against a reference gallery containing one embedding per identity extracted from the training set. This constitutes a closed-set identification protocol.

3.4 Similarity Metrics

Cosine similarity (Eq. 1) — angular distance, scale-invariant; **Euclidean distance** (L2 norm, Eq. 2); and **Manhattan distance** (L1 norm, Eq. 3). A test image is assigned the gallery identity yielding the highest cosine similarity (or lowest distance for L2/L1) above a fixed threshold.

3.5 Data Augmentation

To assess whether photometric augmentation improves robustness, the following randomized transforms were applied during training-gallery embedding extraction: rotation (± 15°), brightness/contrast jitter (± 20%), Gaussian blur ($\sigma \in [0.5, 1.5]$), translation (± 5%), and scaling (± 10%). Augmentation was not applied during test embedding extraction to avoid artificially inflating results.

3.6 Landmark-Guided Cropping

MediaPipe FaceMesh was used to extract 468 facial landmarks per image. Only the subset of landmarks corresponding to the upper facial region — eyes, eyebrows, forehead, and upper cheeks — was retained, deliberately excluding mouth, chin, and nose bridge coordinates. A tight bounding crop was then applied to this upper-face region before embedding extraction. This strategy converts the full masked face into an occlusion-free upper-face image, providing cleaner input to the embedding model. The deployed confidence threshold of MediaPipe was set to 0.5; images where landmark detection failed were excluded.

3.7 Evaluation Metrics

ISR@0.5 (Image-level Success Rate at IoU ≥ 0.5) as the primary metric, plus precision, recall, and F1 at the deployed confidence threshold; 95% confidence intervals (Wilson method); and McNemar's test (χ^2 statistic, 1 d.f.) for pairwise significance. For recognition: **Rank-1 accuracy** — the proportion of test images where the top-1 gallery match corresponds to the true identity.

4 Results

4.1 YOLOv8 Detection Performance

Evaluated on the full 2,000-image test set (masked + unmasked), YOLOv8 at the deployed threshold (conf = 0.524) achieved ISR@0.5 = 95.76%. Class-level analysis shows 98.66% recall on masked faces and 92.86% on unmasked faces, with a 7.14% rate of unmasked faces misclassified as masked (Table 3). The F1-confidence curve peaks at F1 = 0.99 at conf = 0.90, confirming the model can trade recall for very high precision at stricter thresholds. Performance degraded appreciably only above conf = 0.91, indicating the deployed threshold is well-placed.

Table 3
YOLOv8 classification metrics at deployed threshold (conf = 0.524, n = 2,000).

Metric	Value (%)	Note
Precision	93.2	Low false-positive rate
Recall	98.66	High sensitivity to masked faces
F1-Score	95.9	Strong precision–recall balance
ISR@0.5 (overall)	95.76	n = 2,000 (masked + unmasked)
ISR@0.5 (masked only)	91.9	n = 1,000 (benchmarking subset)

Note: ISR@0.5 = 95.76% is the overall metric on the 2,000-image test set; ISR@0.5 = 91.9% is the masked-only ISR used for baseline comparisons in Table 4 (n = 1,000). These two values are not contradictory and reflect different evaluation subsets.

4.2 Detection Benchmarking and Statistical Validation

Table 4 reports ISR@0.5 for all four detectors on the 1,000-image masked test subset with 95% confidence intervals and McNemar's test results (YOLO vs. each baseline).

Table 4
Detection benchmarking results on 1,000 masked face images.

Model	ISR@0.5 (%)	95% CI Lower	95% CI Upper	McNemar vs. YOLOv8
YOLOv8 (proposed)	91.9	90.1%	93.5%	—
MTCNN	47.3	44.2%	50.5%	χ^2=444, $p < 0.001$
RetinaFace	14.7	12.4%	17.5%	$p < 0.001$
Dlib (HOG + SVM)	4.7	3.1%	6.9%	$p < 0.001$

YOLOv8's 91.9% ISR@0.5 is separated from the next-best detector (MTCNN, 47.3%) by a non-overlapping confidence interval gap of approximately 41 percentage points. The McNemar contingency table for YOLO vs. MTCNN yielded A = 446 (YOLO correct, MTCNN incorrect) and B = 0 (MTCNN correct, YOLO incorrect), giving $\chi^2 = 444.00$ ($p < 0.001$). RetinaFace and Dlib, both trained on full-face datasets, perform near chance levels on masked faces, confirming that domain-specific fine-tuning is essential.

4.3 Baseline Face Recognition

Table 5 reports Rank-1 accuracy for all three recognizers under full masked faces using cosine similarity. FaceNet achieves the highest accuracy at 46.5%, closely followed by ArcFace at 45.8%. VGG-Face performs considerably worse at 34.2%. All three values represent substantial drops from the models' reported LFW accuracies (> 99%), quantifying the severity of mask occlusion on current state-of-the-art embeddings.

4.4 Similarity Metrics Comparison

Table 5
Rank-1 accuracy (%) by model and similarity metric on 1,000 masked test faces.

Model	Cosine (%)	Euclidean (%)	Manhattan (%)
FaceNet	46.5	35.2	24.8
ArcFace	45.8	34.7	23.9
VGG-Face	34.2	13.5	22.1

Cosine similarity is the dominant metric across all models and conditions, typically outperforming Euclidean by approximately 11 percentage points for FaceNet and ArcFace. Manhattan distance generally performs worst, with the notable exception of VGG-Face where it outperforms Euclidean — an anomaly attributed to VGG-Face's unnormalized embedding space, where L1 norms are more stable than L2 under high-dimensional redundancy.

4.5 Effect of Photometric Augmentation

Augmentation produced minimal effects across all models: FaceNet + 0.3 pp (46.5% → 46.8%), ArcFace + 0.1 pp (45.8% → 45.9%), VGG-Face − 2.1 pp (34.2% → 32.1%). All changes fall within ± 2 percentage points of baseline, consistent with the thesis prediction. The negative effect on VGG-Face suggests that photometric transforms introduce noise that disrupts its already-fragile embedding consistency under occlusion.

4.6 Landmark-Guided Cropping

Table 6
Rank-1 accuracy (%) before and after landmark-guided upper-face cropping (cosine similarity).

Model	Baseline (%)	Cropped (%)	Gain (pp)
FaceNet	46.5	70.1	**+ 23.6**
ArcFace	45.8	67.4	+ 21.6
VGG-Face	34.2	42.3	+ 8.1

Landmark-guided cropping yields the largest accuracy improvement of any technique tested. FaceNet's 23.6 pp gain is the most pronounced, reflecting the benefit of its compact 128-dimensional embedding space when upper-face features are cleanly isolated. ArcFace's 21.6 pp gain confirms the technique generalizes beyond a single architecture. VGG-Face's smaller gain (+ 8.1 pp) is consistent with its architectural limitations under partial occlusion; its softmax-trained embeddings are less suited to the angular structure that upper-face cropping preserves. The FaceNet-cropped result of 70.1% is

comparable to results reported by Wang et al. [26] on RMFRD (~ 65%), achieved here without any model retraining — a practically significant finding for deployment scenarios where fine-tuning is not feasible.

4.7 Consolidated Pipeline Summary

Table 7
Summary of all experiments and key results.

Exp.	Focus	Key Result	Takeaway
E1	YOLOv8 fine-tuning	ISR@0.5 = 95.76% (overall)	Domain-specific fine-tuning critical
E2	Detector benchmarking (masked)	YOLOv8: 91.9% vs. MTCNN: 47.3%	YOLOv8 dominates; $p < 0.001$
E3	Recognition baseline (masked)	FaceNet: 46.5%, ArcFace: 45.8%, VGG: 34.2%	All models degrade severely under occlusion
E4	Similarity metric comparison	Cosine best across all models	Angular metrics preferred for these embeddings
E5	Photometric augmentation	Changes within ± 2 pp	Negligible benefit; VGG-Face degraded
E6	Landmark-guided cropping	FaceNet: 46.5% → 70.1% (+ 23.6 pp)	Largest single improvement in the pipeline

5 Discussion

5.1 Detection

The YOLOv8 result of 91.9% ISR@0.5 on masked faces, validated with confidence intervals and McNemar's test, directly confirms the hypothesis that fine-tuning a modern one-stage detector on real-world masked data yields significant gains over models trained on full-face corpora. MTCNN's 47.3% mirrors the finding of Yurtkan and Hamdi [27], who reported MTCNN performance dropping under real-world masked inputs, while RetinaFace's collapse to 14.7% is consistent with its training objective on the WIDER dataset which emphasizes full-face scenarios. The 91.9% figure also aligns with Li et al. [19] (94.7% precision on occluded industrial workers) and Dewi et al. [20] (99.1% on synthetic masked compliance detection), lending external validity to the fine-tuning approach.

YOLOv8's 7.14% rate of unmasked faces misclassified as masked reflects a precision–recall trade-off at the deployed threshold. In security contexts where false negatives (missed masked faces) are more costly than false positives (unmasked faces flagged as masked), this asymmetric error profile is acceptable. Future work should evaluate whether raising the confidence threshold or integrating image quality assessment (IQA) via BRISQUE [28] can reduce this without sacrificing recall.

5.2 Face Recognition

FaceNet's superiority over ArcFace under masked conditions is counterintuitive given ArcFace's superior LFW benchmark. We attribute this to the curse of dimensionality: ArcFace's 512-dimensional embeddings encode fine-grained feature distributions that rely on lower-face completeness. When 30–40% of the face is occluded, these dimensions become noisy, reducing cosine separability. FaceNet's 128-dimensional triplet-loss embeddings are more compact and regularized, maintaining discriminability even with incomplete inputs. This explanation is supported by Huang et al. [16], who found ArcFace accuracy specifically sensitive to mouth and chin occlusion. VGG-Face's poor performance (34.2%) is consistent with Li et al. [29], who reported its accuracy dropping by over 50% under mask-like conditions, attributable to its softmax-based training objective which is misaligned with cosine similarity at inference time.

5.3 Similarity Metrics and Augmentation

Cosine similarity's consistent superiority aligns with Huzaimi and Mazalan [21]. Its angular invariance makes it robust to the magnitude distortions that occlusion introduces into embedding vectors. However, the 70.1% ceiling for FaceNet with cropping suggests a residual limitation under severe or multi-type occlusion, consistent with the caution raised by Shi and Jain [22]. The minimal augmentation effect reinforces findings by Yang et al. [30]: RMFRD's inherent real-world diversity (varied lighting, pose, mask styles) acts as implicit augmentation, reducing marginal value from additional synthetic transforms. Generic photometric transforms may also introduce asymmetries that confuse models — particularly for VGG-Face, whose − 2.1 pp drop under augmentation highlights the risk of applying indiscriminate augmentation to pre-trained models under occlusion.

5.4 Landmark-Guided Cropping

The landmark-guided cropping result (+ 23.6 pp for FaceNet) is the most practically significant finding of this study. It demonstrates that removing occluded image regions — without any model retraining — is more effective than any other technique evaluated, including augmentation and metric selection. The principle echoes findings from adjacent domains: Yao et al. [31] improved flood mapping accuracy by isolating flood-prone regions in satellite imagery, and Pham and Ghimire [32] improved water segmentation by cropping out urban noise. In each case, spatial exclusion of irrelevant or obstructed content improved downstream model performance. Unlike the static lower-face masking approach of Salim and Surantha [25], this study uses MediaPipe's 468-landmark dense mesh for adaptive, per-image cropping, and evaluates it across three recognizers — establishing broader generalizability.

5.5 Limitations

This study has several limitations that constrain generalizability. First, all experiments are conducted on a single dataset (RMFRD), which is demographically narrow (predominantly East Asian subjects); cross-dataset validation on more diverse corpora such as LFW, MFR2, or AgeDB is needed. Second, the evaluation does not include compound occlusions (masks combined with eyewear, hats, or scarves), which are common in practice. Third, head pose variation and image resolution effects on the cropping strategy were not systematically studied — MediaPipe's landmark detection degrades under extreme

angles. Fourth, recognition significance testing was not conducted for all pairwise comparisons, which should be addressed in follow-on work. Finally, the pipeline was implemented and benchmarked on a consumer-grade Apple M1 Pro; inference speed and latency at scale were not measured.

6 Conclusion

This paper presented a systematic end-to-end evaluation of masked face recognition, yielding four key findings. (1) A fine-tuned YOLOv8 detector achieves 95.76% ISR@0.5 overall and 91.9% on masked faces — a statistically significant ($p < 0.001$) improvement over MTCNN (47.3%), RetinaFace (14.7%), and Dlib (4.7%), confirming that domain-specific fine-tuning is essential. (2) Among three recognizers, FaceNet's compact 128-dimensional triplet-loss embeddings prove more resilient to occlusion than ArcFace's 512-dimensional angular embeddings or VGG-Face's softmax-trained features. (3) Cosine similarity is the preferred matching metric for all three recognizers under masked conditions; photometric augmentation yields minimal benefit when training data already reflects real-world variation. (4) Landmark-guided cropping via MediaPipe — retaining only the unoccluded upper face — is the single most effective enhancement, raising FaceNet Rank-1 accuracy from 46.5% to 70.1% without model retraining.

The practical implication is clear: reliable masked face recognition is achievable through multi-stage pipeline optimization (detector fine-tuning + preprocessing + metric selection) rather than through end-to-end model replacement. The proposed pipeline has direct applications in access control, healthcare screening, and surveillance systems where mask occlusion is persistent. Future work should validate the pipeline on demographically diverse cross-dataset benchmarks, extend evaluation to compound occlusions, explore learned similarity functions (Siamese networks, contrastive loss), and integrate image quality assessment to handle degraded inputs.

Declarations

Author Contributions: H.Z. and **S.A.** conceptualized the study and developed the methodology. **S.A.** was responsible for software implementation, formal analysis, and writing the original draft. Validation was performed jointly by both authors. **H.Z.** provided supervision and contributed to the review and editing of the manuscript. All authors have read and approved the final version.

Funding: This research received no external funding.

Data Availability: The RMFRD dataset used in this study is publicly available at https://doi.org/10.48550/arXiv.2003.09093.

Conflicts of Interest: The authors declare no conflicts of interest.

AI-Assisted Writing Disclosure: During preparation of this manuscript, the authors used an OpenAI GPT-series language model for copy-editing and language refinement of draft text. No AI tool was used for

analysis, data interpretation, or methodological decisions. All AI-assisted content was reviewed and edited by the authors, who take full responsibility for the final manuscript.

References

1. Li, L., Mu, X., Li, S. & Peng, H. *Rev. Face Recognit. Technol. IEEE Access.*, **8**, 139110–139120, (2020).
2. Wang, B., Zheng, J. & Chen, C. L. P. A Survey on Masked Facial Detection Methods and Datasets for Fighting Against COVID-19. *IEEE Trans. Artif. Intell.* **3** (3), 323–343 (2022).
3. Mamieva, D., Abdusalomov, A. B., Mukhiddinov, M. & Whangbo, T. K. Improved Face Detection Method via Learning Small Faces on Hard Images Based on a Deep Learning Approach, Sensors, vol. 23, no. 1, p. 502, (2023).
4. Patel, V. et al. Robust Face Detection: Challenges and Trends. *IEEE Trans. PAMI*, (2019).
5. Naser, O. A. et al. Comparative Analysis of MTCNN and Haar Cascades for Face Detection. *J. Commun. Softw. Syst.* **21** (1), 109–119 (2025).
6. Bansal, A., Agarwal, N. & Gupta, R. Real-time face detection and recognition: Current challenges and deep learning-based solutions. *J. Real-Time Image Process.* **18** (2), 345–358 (2021).
7. Kumar, A., Kaur, A. & Kumar, M. Face detection techniques: A review. *Artif. Intell. Rev.* **52** (2), 927–948 (2019).
8. Schroff, F., Kalenichenko, D. & Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering, Proc. CVPR, pp. 815–823, (2015).
9. Deng, J., Guo, J., Xue, N. & Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition, Proc. CVPR, pp. 4690–4699, (2019).
10. Parkhi, O. M., Vedaldi, A. & Zisserman, A. Deep Face Recognition, Proc. BMVC, (2015).
11. Jeevan, G., Zacharias, G. C., Nair, M. S. & Rajan, J. An empirical study of the impact of masks on face recognition, Pattern Recognit., vol. 122, p. 108308, (2022).
12. Mahmoud, M. M., Kasem, M. & Kang, H. S. Artificial intelligence-based masked face detection: A survey. *Appl. Sci.* **14** (19), 8781 (2024).
13. Mishra, A. & Reza, M. S. Robust face recognition under mask occlusion using synthetic masked datasets. *Multimed Tools Appl.* **81**, 35853–35872 (2022).
14. Cheng, W. C. et al. Masked face recognition using FaceNet and genetic algorithm. *IEEE Access.* **11**, 17345–17356 (2023).
15. Ritchie, K. L., Papesh, M. H., Dunn, J. D. & Davis, J. P. Face masks and fake masks: The effect of real and superimposed masks on face matching. *Cogn. Res. Princ Implic*. **9** (1), 5 (2024).
16. Huang, B. et al. When face recognition meets occlusion: A new benchmark. *Multimed Tools Appl.* **80**, 26219–26243 (2021).
17. Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal. Process. Lett.* **23** (10), 1499–1503 (2016).

18. Deng, J. et al. RetinaFace: Single-stage dense face localisation in the wild, Proc. CVPR, pp. 5203–5212, (2020).
19. Li, Y., Chen, Y., Zhang, Y. & Liu, J. A novel approach for real-time helmet-wearing detection based on improved YOLOv8, Proc. ICISCAE, pp. 1–6, (2023).
20. Dewi, C., Nugroho, H. A. & Santoso, A. YOLOv8-based mask-wearing compliance detection using real-time video streams. *Int. J. Comput. Vis. Intell. Syst.* **11** (4), 201–212 (2023).
21. Huzaimi, M. H. & Mazalan, L. Similarity measures for person identification with face mask, Proc. IEEE ISIEA, pp. 76–82, (2023).
22. Shi, Y. & Jain, A. K. Masked face recognition: Human vs. machine. *IEEE Trans. Biom Behav. Identity Sci.* **4** (1), 89–101 (2022).
23. Pham, Q. H., Nguyen, K., Tran, T. & Phung, D. Rethinking similarity metrics for occluded face recognition, Proc. WACV, pp. 1128–1137, (2023).
24. Arif, H. N. et al. Comparative analysis of masked and unmasked face recognition using VGG Face and MTCNN. *Procedia Comput. Sci.* **216**, 300–307 (2023).
25. Salim, R. J. & Surantha, N. Masked face recognition by zeroing the masked region without model retraining. *Int. J. Innov. Comput. Inf. Control*. **19** (4), 1087–1101 (2023).
26. Wang, Z. et al. Masked Face Recognition Dataset and Application, arXiv preprint arXiv:2003.09093, (2020).
27. Yurtkan, K. & Hamdi, H. E. Masked face recognition based on FaceNet pre-trained model. *J. Image Graph*. **8** (2), 35–42 (2020).
28. Mittal, A., Moorthy, A. K. & Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **21** (12), 4695–4708 (2012).
29. Li, J., He, L., Guo, Y. & Liu, Y. Benchmarking masked face recognition: A comparative study. *J. Vis. Commun. Image Represent*. **78**, 103123 (2021).
30. Yang, W., Wang, L. & Zhou, J. Masked face recognition with residual attention and synthetic data augmentation, Proc. IEEE FG, pp. 1–8, (2021).
31. Yao, H., Liu, J. & Ma, X. High-resolution flood mapping using adaptive cropping and deep learning. *Remote Sens.* **12** (14), 2285 (2020).
32. Pham, H. & Ghimire, D. Urban water body segmentation using spatial cropping and convolutional neural networks. *J. Hydrol.* **597**, 126312 (2021).

Figures

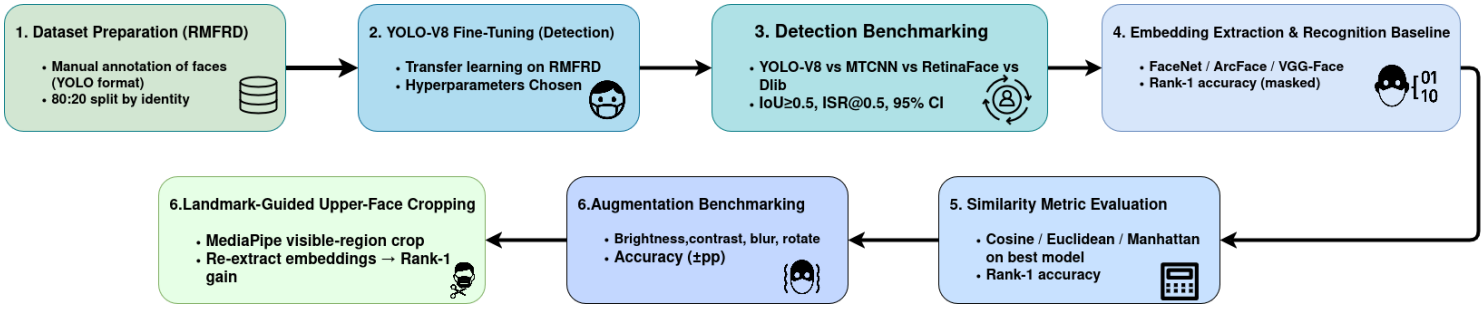


Figure 1

Methodology pipeline for end-to-end masked face recognition



Figure 2

Sample Images from the RMFRD Dataset



```
[-0.16597123  0.11949642  0.10122138 -0.12254387 -0.11838274 -0.02759848
  0.01384261 -0.07288168  0.06527785 -0.13963827  0.17559825 -0.08519795
 -0.25020906 -0.0102928  -0.01079375  0.21948609 -0.21553843 -0.19072869
 -0.13214052 -0.10856642  0.01477439  0.07908501 -0.05431245  0.11583221
 -0.1715932  -0.23868942 -0.10927209 -0.01475592  0.01510147 -0.08837818
 -0.00252565  0.03029939 -0.17696191  0.00717495  0.04937807  0.12862331
 -0.01091681 -0.09640069  0.1934087   0.01389313 -0.23863317  0.01766105
  0.15954995  0.26412117  0.21739453 -0.00336392 -0.03702538 -0.05539587
  0.10163558 -0.28075284  0.05129996  0.20316276  0.02803321  0.16820049
 -0.02042791 -0.12237336 -0.00515839  0.22946236 -0.21929839  0.03012473
  0.10334508 -0.19385433  0.010464   -0.1046315   0.159848    0.14616895
 -0.19181257 -0.22495562  0.19514428 -0.23470432 -0.11755507  0.07461189
 -0.0839375  -0.09592333 -0.35081255  0.03915626  0.32692659  0.20959754
 -0.10889891 -0.01162777 -0.09144997 -0.02747694  0.00839157  0.15324235
 -0.02081387 -0.0425766  -0.06370837  0.0559925   0.19876704  0.04837478
  0.00898859  0.25267416  0.04914283  0.0009048   0.00353496  0.14037418
 -0.15166469 -0.07702    -0.10442869  0.03603458 -0.10455432 -0.05006842
  0.01399595  0.18746354 -0.10759938  0.24044091 -0.03882578 -0.01190661
 -0.12068785  0.03638472 -0.02094681 -0.02857181  0.10984001 -0.22919109
  0.14735889  0.20415246 -0.00772358  0.09734803  0.10852756  0.12523061
 -0.03349054 -0.08732443 -0.17586364 -0.09488015  0.04729999 -0.0654182
  0.10575557  0.07663471]
```

Figure 3

Image Converted to 128-Dimensional Embeddings By FaceNet

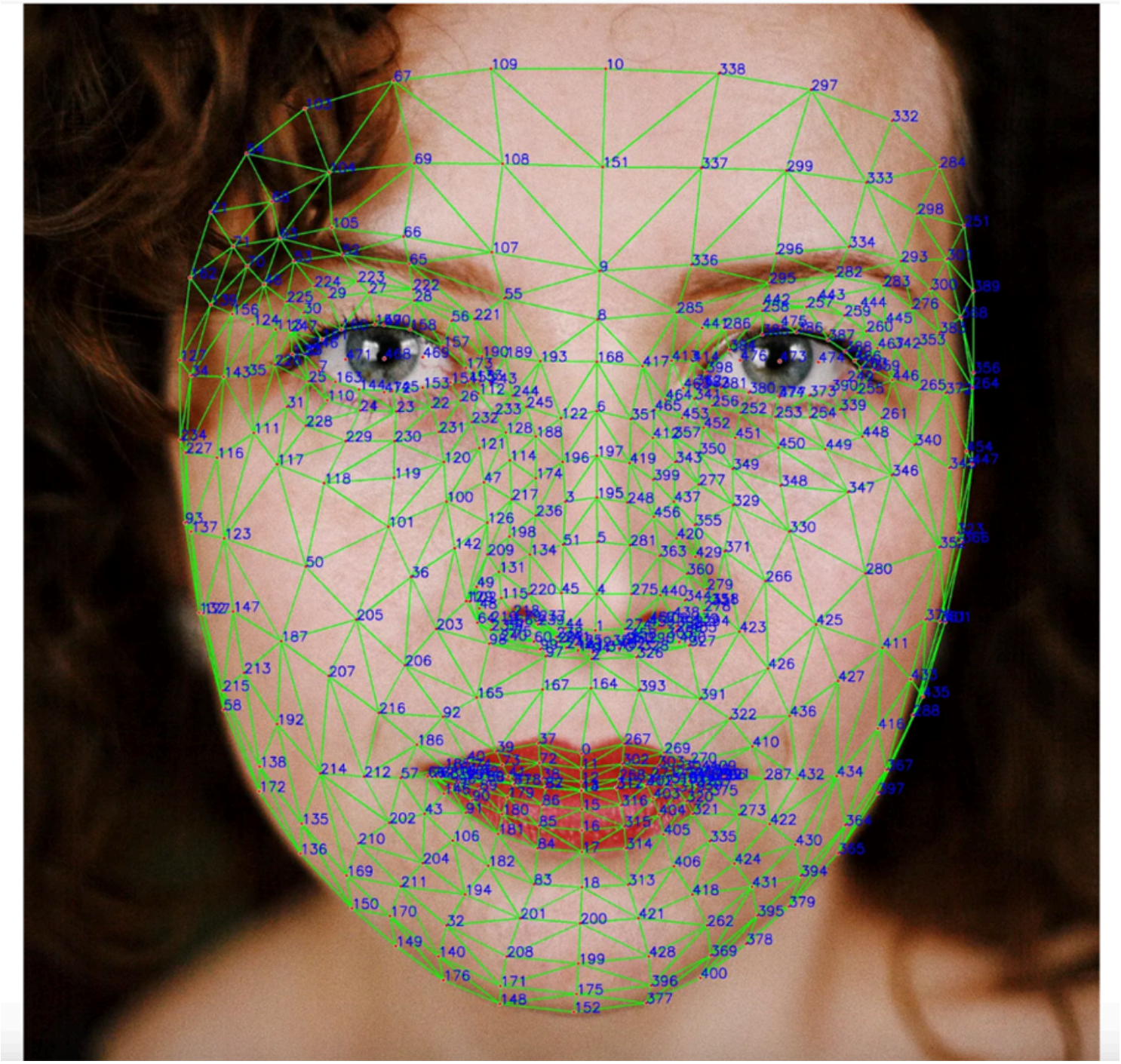


Figure 4

Indexed Facial Landmark Map Generated by MediaPipe

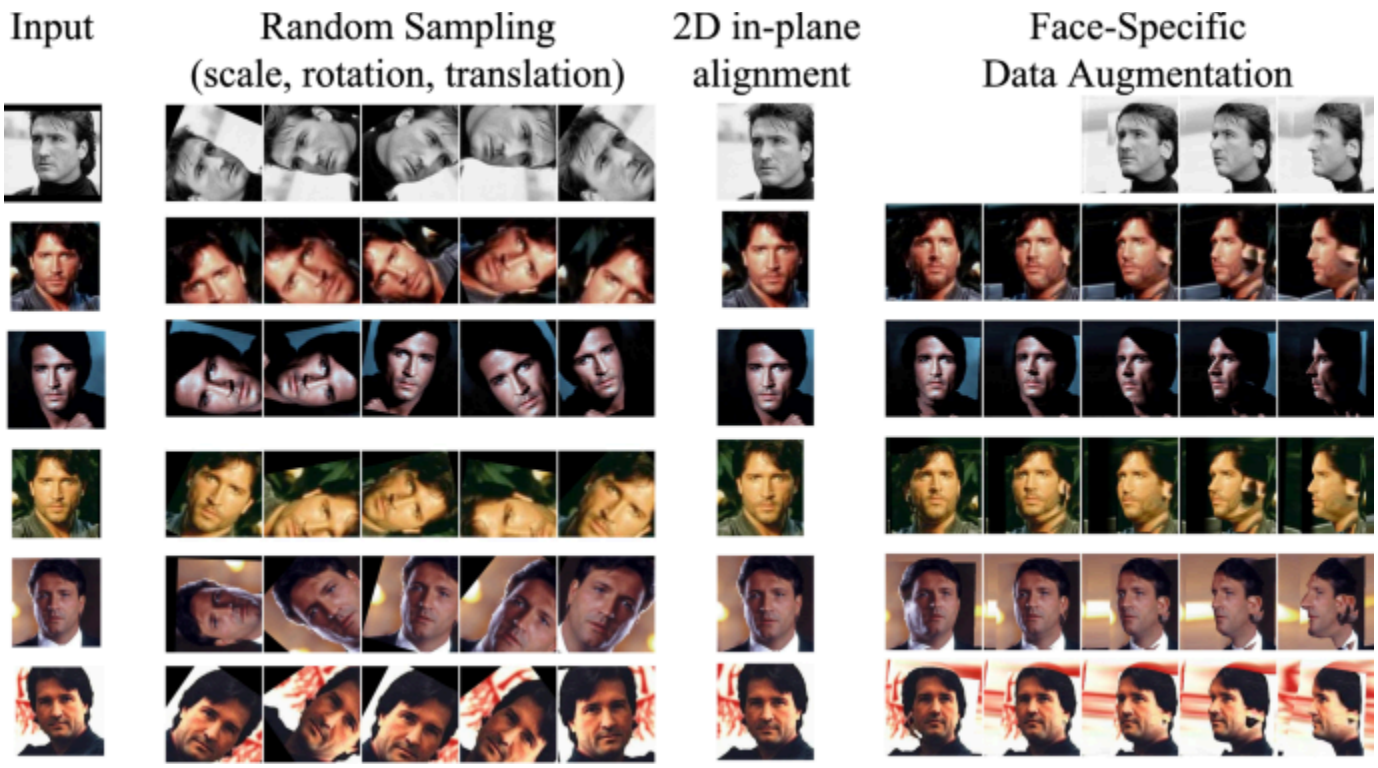


Figure 5

Examples Of Augmented Images