paravision

Face Recognition with Masks

Challenges and Considerations
NIST International Face Performance Conference (IFPC) 2020

Bhargav Avasarala, Director of Machine Learning

October 29, 2020

Masked Face Recognition is Challenging

- Critical features of the face may be completely occluded
- Additional variations in pose, lighting, accessories & garments, etc compound errors dramatically
- Increased prevalence of masks makes this a problem worth addressing

Overview

Problem

1. Effect of Masks on (Paravision) FR

Proposed Solutions

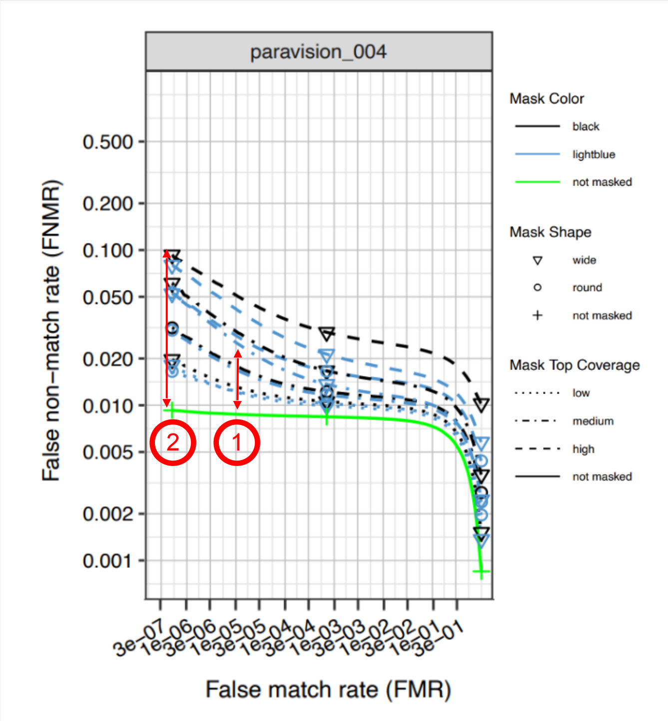
1. Mitigating Errors with Face Quality Filtering
2. FR Model Improvements

Effect of Masks on FR

NIST FRVT - Specification & Results

Reference

Probe



1 FRVT leaderboard measured at FMR=1e-5, lightblue, wide, medium

2 In the worst case, FNMR can increase up to nearly 10% for masked probes

Test details

- 1:1 Face verification
- References - **Application images**
  - Good conformance with ISO / IEC 19794-5 Full Frontal specs
- Probes - **Webcam images**
  - Yaw / Pitch / Roll variations
  - Perspective distortion, under-exposure possible
  - Poor conformance with ISO / IEC 19794-5 Full Frontal specs
  - Variety of synthetic masks (size, coverage, color)

[*Source*] **Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face recognition accuracy with masks using pre-COVID-19 algorithms**: https://pages.nist.gov/frvt/reports/facemask/frvt_facemask_report.pdf

Internal Dataset Collection

Test details

- 1:1 Face Verification
- **849** IDs (*483* M / *366* F)
- **50K** genuine pairs, **21.4M** imposter pairs, matched by gender, race
- Reference face image is not masked
- Probe face image can be either
  - Non-masked
  - Masked, Real
  - Masked, Synthetic (multi-colored)
- Sources of variation (by design)
  - Yaw/Pitch/Roll
  - Prescription glasses, sunglasses

Out of scope *(ongoing)*

- 1:N metrics
- Mask-to-mask pairs
- Race, age effects
- FMR shift & Differential Performance

Genuine pairs | **Imposter pairs**

Non-masked

Masked, Real

Masked, Synthetic

Internal Dataset Collection

Reference Images

Details

- 1 image per subject
- Acquired at eye-level, full frontal, neutral expression, no harsh lighting / blur / etc
- All subjects collected in South Asia
- Variety of age groups
- Women wear head coverings (scarves, shawls)

Probe Videos / Images

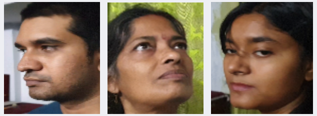
Details

- 1 masked and 1 non-masked video per subject, same-day
- Heavy variations in yaw/pitch/roll
- Mask types - medical, cloth (multi-colored, textured)
- Sunglasses, eyeglasses

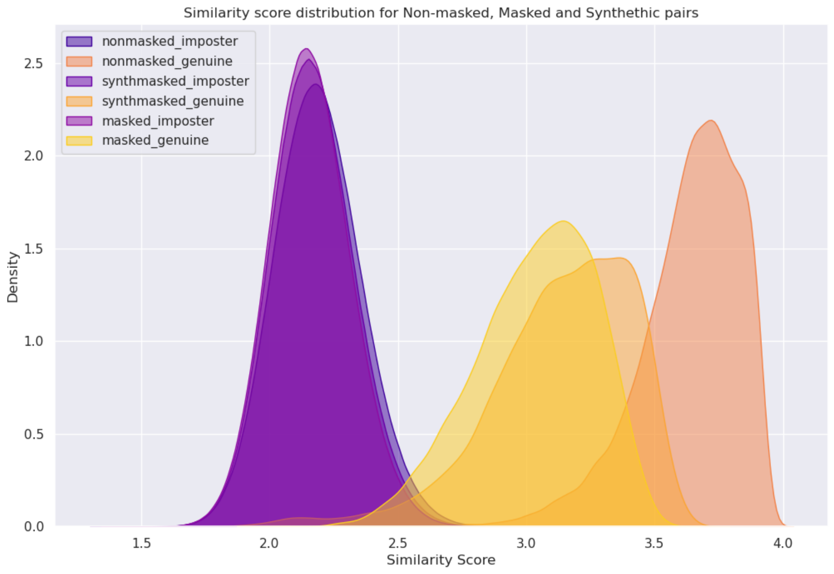
Non-mask

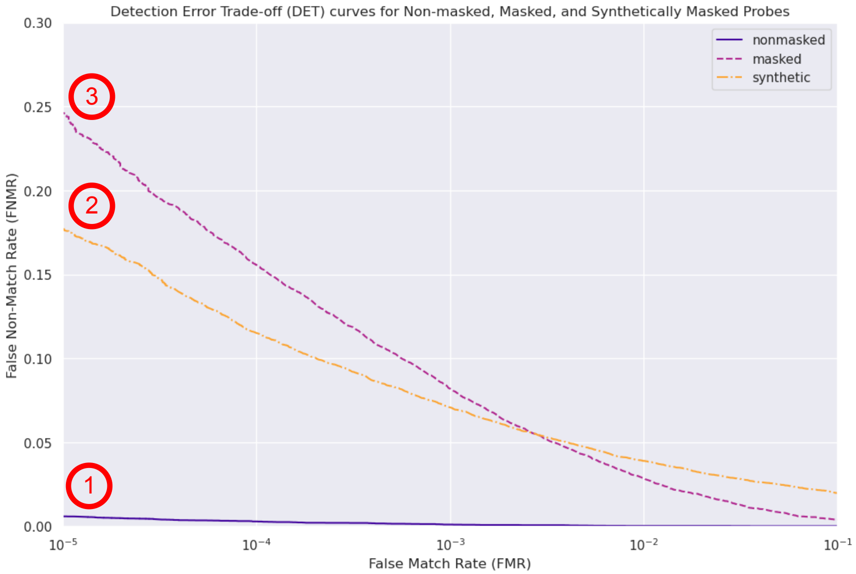
Mask

Cleaning, removal of near-duplicates, outliers (e.g. false positive detections, extremely poor quality images, etc)



Results





For mask pairs:

- Larger intra-class variance in genuine pairs
- Lower mean similarity score in genuine pairs
- Slightly lower mean similarity score in imposter pairs

Non-masked FNMR@FMR=1e-5	**Synthetic Masks** FNMR@FMR=1e-5	**Real Masks** FNMR@FMR=1e-5
0.0059[1]	0.1766[2]	0.2459[3]

Mask Failure Cases

False matches | False non-matches

Real Masks

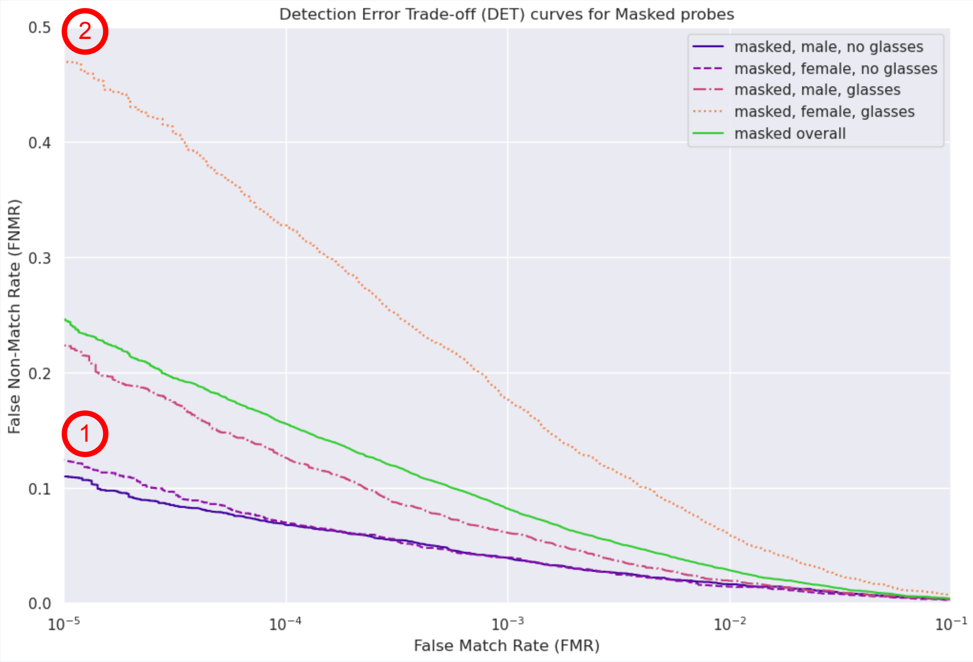
Synthetic Masks

Random sampling of probe failure cases at the **FMR=1e-5** level

- **False matches**: Largely frontal faces, mainly women wearing headwear.
- **False non-matches:** Large mask coverage, closed eyes, presence of glasses, off-frontal

Note: Alignment failures can tamper with synthetic mask application

Masks and Other Sources of Variation



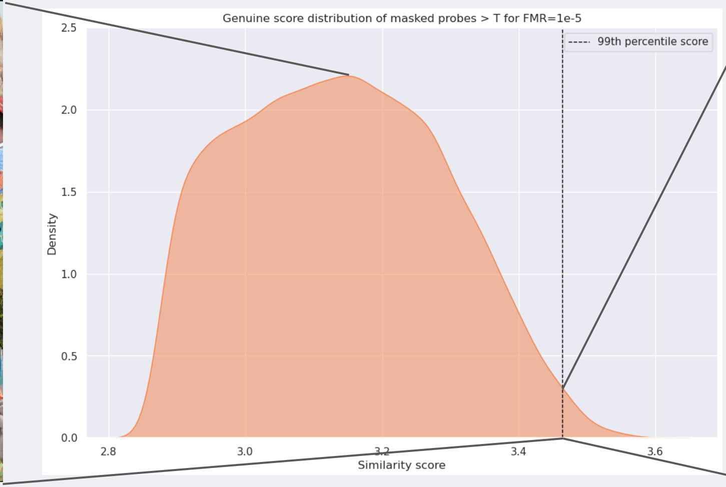
DET curves for *real* masked probes only, conditioned on gender and presence of glasses

Observations:

1. Without glasses, FNMR for both male and female subsets are reduced
2. Much larger gap in FNMR for women compared to men when conditioned on glasses

➔ Presence of head covering likely the cause

True Matches for Masked Probes



Random sampling of true matches below the 99th percentile

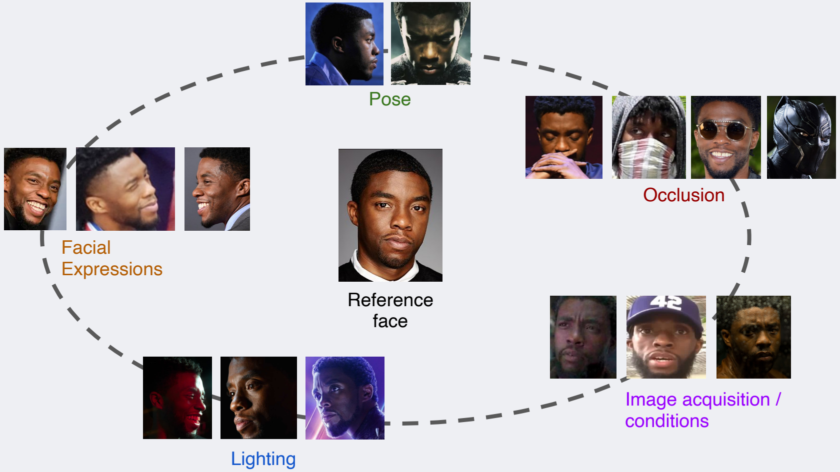
Random sampling of true matches in the 99th percentile in similarity score.

- Frontal faces with visible eyes yield high similarity scores
- ...but some faces with undesirable characteristics can still match
- Difficult to explain why a face matches when it does

So… what do we do about this?

Mitigating Errors with Face Quality Filtering

Face Quality



- Every face can be characterized by its **quality**, or its likelihood to match with its reference face
- Masked faces need not be treated any differently under this definition

*[*Source*] **Ongoing Face Recognition Vendor Test (FRVT) Part 5: Face Image Quality Assessment**
https://pages.nist.gov/frvt/reports/quality/frvt_quality_report.pdf

Face Quality

Low quality
(Q < 0.3)

Medium quality
(0.3 < Q < 0.6)

High quality
(Q > 0.6)

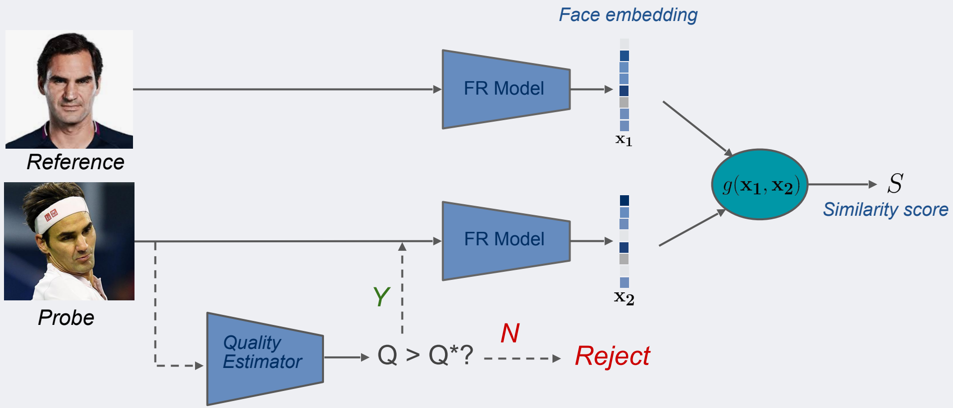
Masked probes

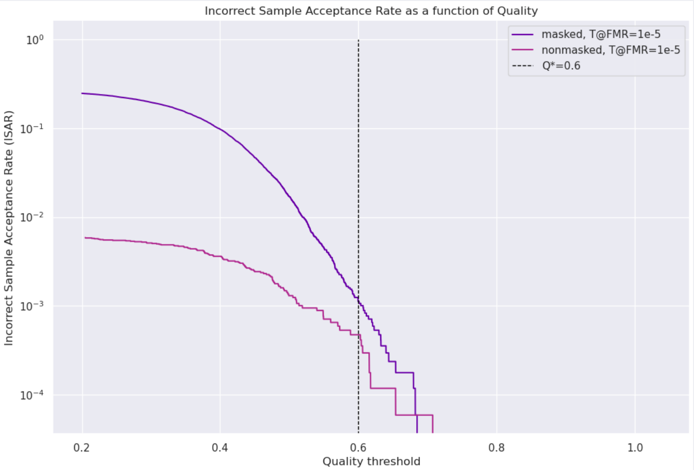
Non-masked probes

- Characteristics of faces in each quality bucket are qualitatively different for masked vs. non-masked faces

Mitigating Errors with Face Quality

Filtering by quality isolates images which are most likely to match, for both masked and non-masked probes



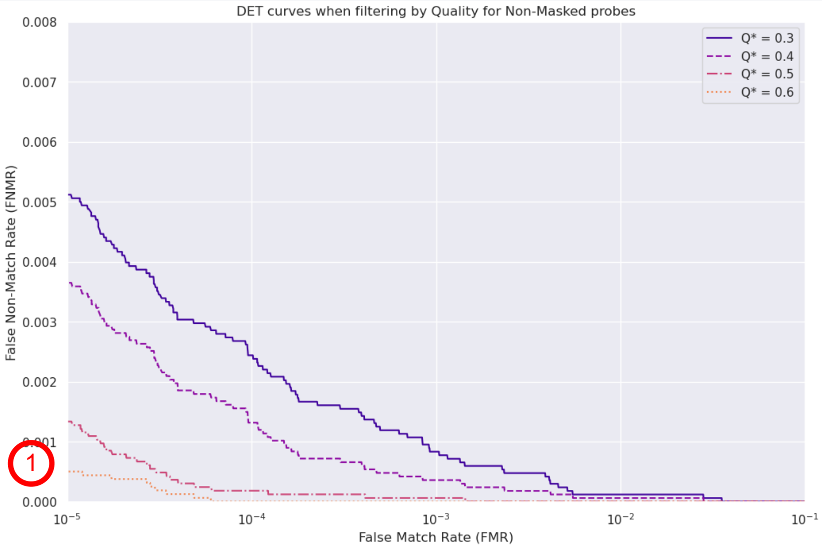


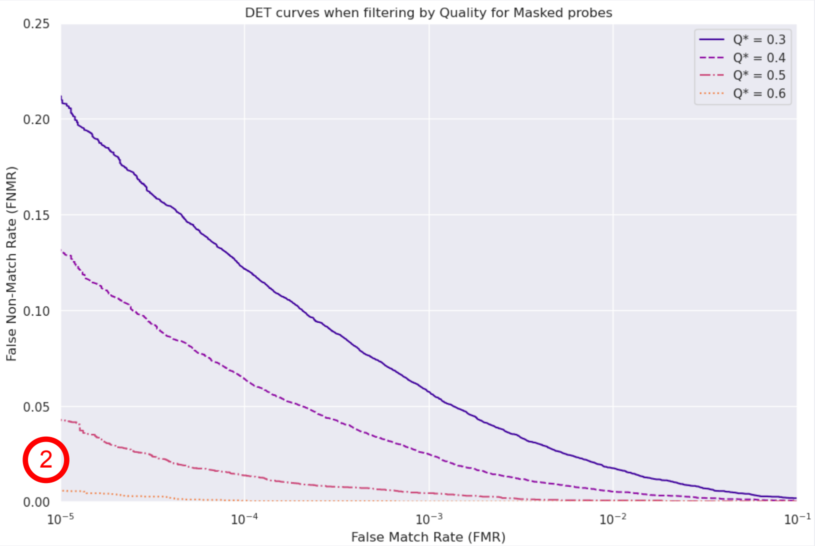
Q*=0.6 yields ISAR ≤ 1e-3 for both masked and non-masked probes

***Incorrect Sample Acceptance**: Quality of probe is greater than threshold (accepted) but FR algorithm fails to match with reference

*[*Source*] **Ongoing Face Recognition Vendor Test (FRVT) Part 5: Face Image Quality Assessment** https://pages.nist.gov/frvt/reports/quality/frvt_quality_report.pdf

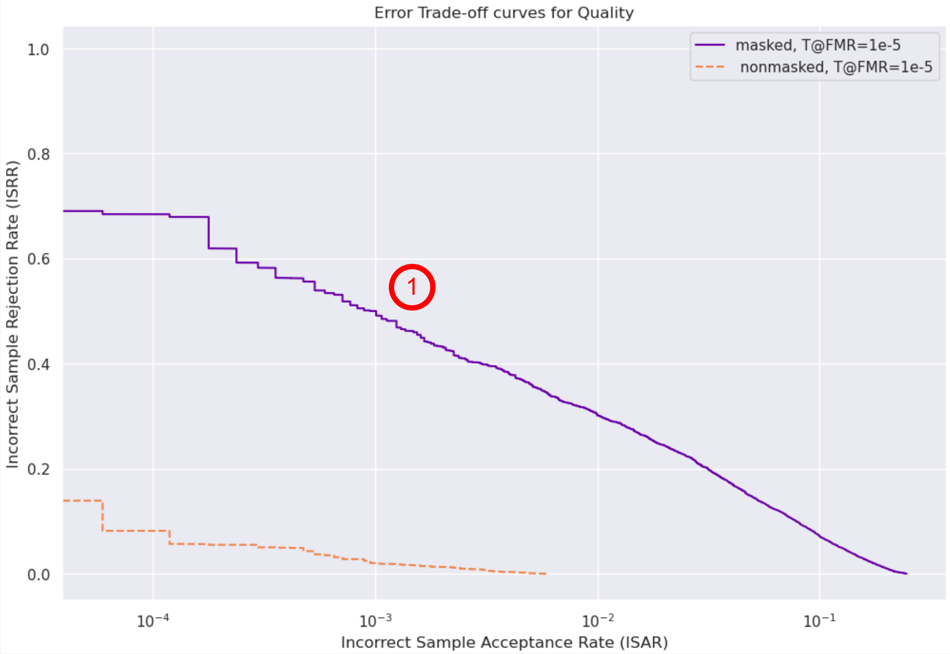
Mitigating Errors with Face Quality





Non-masked FNMR@FMR=1e-5, Q* = 0.6	**Real Masks** FNMR@FMR=1e-5, Q* = 0.6
0.05%[1]	0.56%[2]

Error Tradeoff Curves for Quality



Error-tradeoff of ISAR and ISRR at a similarity score threshold that achieves FMR=1e-5, separated by Masks and Non-masked probes

1 Needs improvement. Roughly half of the faces that would have matched are rejected at ISAR=1e-3

Non-masked ISRR@ISAR=1e-3	**Real Masks** ISRR@ISAR=1e-3
0.020	0.493

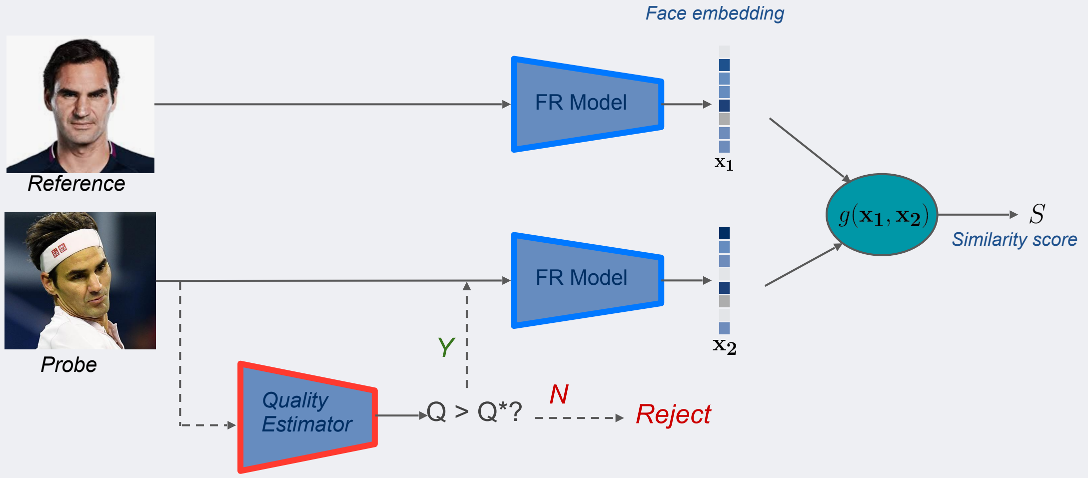
***Incorrect Sample Acceptance**: Quality of probe is greater than threshold (accepted) but FR algorithm fails to match with reference
***Incorrect Sample Rejection**: Quality of probe is less than threshold (rejected) but FR algorithm would have matched with reference

*[*Source*] **Ongoing Face Recognition Vendor Test (FRVT) Part 5: Face Image Quality Assessment** https://pages.nist.gov/frvt/reports/quality/frvt_quality_report.pdf

FR Model Improvements

Areas of Focus

Improving FR embedding robustness in the presence of masks is the primary focus



Improving the quality estimator helps mitigate errors in FR, but does not fundamentally address the problem

Improving FR Robustness to Masks

Continuing to make face embeddings more robust *generally* helps with accuracy on masks by extension

Training with multi-colored synthetic masks can help accuracy on *real* masks without significantly impacting non-mask accuracy **(still actively being investigated)**

	Latest collection		Existing non-masked internal sets	
Model	Non-masked FNMR@FMR=1e-5	Real Masks FNMR@FMR=1e-5	Internal set - Standard FNMR@FMR=1e-7	Internal Set - Hard FNMR@FMR=1e-6
Gen. 3 (paravision_004)	0.0059	0.2459	0.0027	0.0861
Gen. 4	**0.0039**	0.1948	**0.0018**	**0.0532**
Gen. 4 + synth mask data augmentation	0.0042	**0.1739**	0.0023	0.0577

Bold values indicate the highest accuracy per column.

Summary

1. Frontal faces work best for masked FR

   ...but additional occlusion rapidly increases FNMR.

2. Reasoning about faces, masked or otherwise, in terms of quality is a useful abstraction.

3. Filtering by face quality helps mitigate errors in FR

   ...but needs greater specificity for masks.

4. Synthetic masks as data augmentation helps improve FR accuracy on real masks

   ...but comes with some tradeoffs...
   …and still has a long way to go.

Thank you!

Bhargav R. Avasarala
Director of ML
Paravision
bhargav@paravision.ai

Acknowledgements

The wonderful team at Paravision

Dr. Tong (Martin) Man
Senior ML Engineer

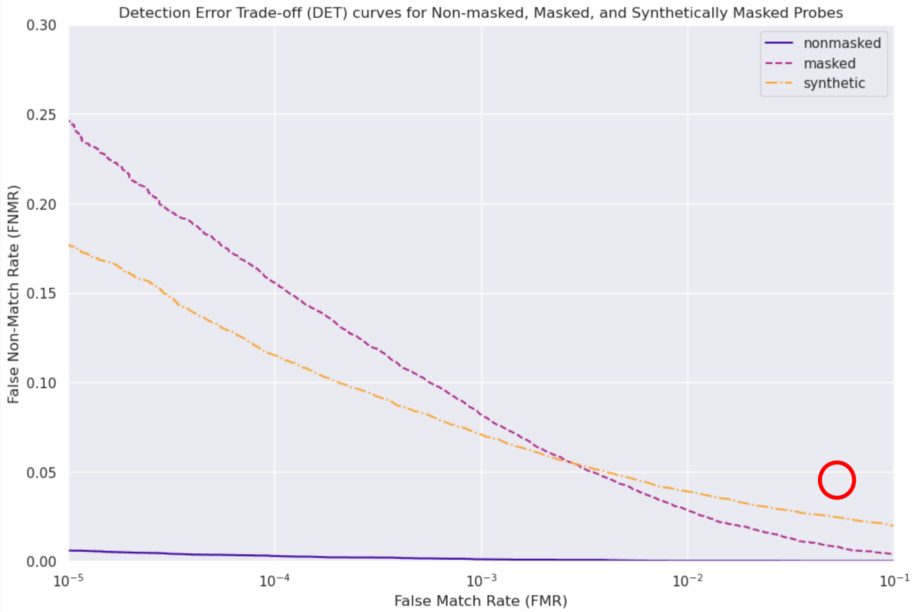
Kavya Padmesh
Senior ML Engineer

Dr. Maryam Arani
Senior ML Engineer

Charles E. Rice
CTO

Supplemental Material

Synthetic mask FNMR issues



- Synthetic mask application failures constitute the bulk of noise that raises the FNMR floor at high FMR values

Operational thresholds and FMR Shift

We observe differential performance between masks and non-masked probes, e.g. using an operational threshold based on non-masked faces results in an FMR shift for masked faces.

Filtering by quality does not remove this effect