NIST IR 8485 DRAFT SUPPLEMENT

Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment

Specific Image Defect Detection

Joyce Yang
Patrick Grother
Mei Ngan
Kayee Hanaoka
Austin Hom

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8485

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NIST IR 8485 DRAFT SUPPLEMENT

Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment

Specific Image Defect Detection

Joyce Yang
Patrick Grother
Mei Ngan
Kayee Hanaoka
Austin Hom
Image Group
Information Access Division
Information Technology Laboratory
This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8485

August 2025

U.S. Department of Commerce
Howard W. Lutnick, Secretary

National Institute of Standards and Technology
Craig S. Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

Disclaimer

Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Institutional Review Board

The National Institute of Standards and Technology's Research Protections Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2023-09-05

How to cite this NIST Technical Series Publication:
Joyce Yang, Patrick Grother, Mei Ngan, Kayee Hanaoka, Austin Hom (2025) Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment. (National Institute of Standards and Technology, Gaithersburg, MD), NIST IR 8485 DRAFT SUPPLEMENT. https://doi.org/10.6028/NIST.IR.8485

Contact Information
frvt@nist.gov

August 2025

Abstract

This report summarizes the results of the FATE Quality Vector assessment track, which tests face image quality algorithms' ability to detect specific defects such as non-frontal pose and background non-uniformity in the context of facial images. All algorithms submitted have some success at measuring various quality-related parameters.

Keywords

Face; defect; detection; image; quality; quality component; quality measure; specific.

EXECUTIVE SUMMARY

This report summarizes results from the Face Analysis Technology Evaluation (FATE) Quality Specific Image Defect Detection (SIDD) activity. All algorithms submitted have some success at measuring various quality-related parameters. The measures that were implemented most frequently include total faces present, pitch, yaw, roll, eyes open, and inter-eye distance. We will continue to add, replace, and extend test cases and test datasets to identify core capability. As this report is scrutinized by developers and end-users, comments and participation from other developers are welcomed.

RELEASE NOTES

2025-08-07: The FATE Quality SIDD track is closed until September 9th. Developers should wait four months from their last (successful) submission to submit a new submission.

- ▷ We have added results for daon_000, innovatrics_000, and pixelall_002.

2025-06-17: The FATE Quality SIDD track remains open. Developers should wait four months from their last (successful) submission to submit a new submission.

- ▷ We have added a set for the measure Distance from Eyes to Edges, containing images with varying pose and background uniformity in which one eye is very close to the edge of the image, or partially out of the frame.

2025-05-22: The FATE Quality SIDD track remains open. Developers should wait four months from their last (successful) submission to submit a new submission.

- ▷ We have added results for mobbl_003 and veridium_003.
- ▷ We have added a Resolution set of images with blur caused by air turbulence. This is an example of physical blur from a natural cause, rather than synthetic blur. See section 3.14.

2025-04-15:

- ▷ We have added additional plots showing the dependence of the Resolution quality measure on the width of Gaussian blur normalized by inter-eye distance.

2025-03-14:

- ▷ We have added results for pixelall_001.

2025-01-21:

- ▷ We have added results for igd_005, veridium_002, mobbl_002, roc_008, and kasikornlabs_000.
- ▷ We have grouped demographic violin plots and CDF plots by measure (EyesOpen2, MouthOpen2, Overexposure, Underexposure, Resolution, UnifiedQualityScore).

2024-12-09:

- ▷ We have added results for viante_001.

2024-11-15:

- ▷ In Section 4, we have separated demographic violin plots by sex as well as region-of-birth and introduced cumulative distribution plots to highlight demographic differentials. In addition, we have added demographic analysis for Unified Quality Score.

- ▷ We removed the entry-type set for demographic differentials, so that demographic differentials are evaluated on only application-type images.

2024-10-04:

- ▷ We have added results for papil11_000.

2024-09-06: The FATE Quality SIDD track remains open. Developers should wait four months from their last (successful) submission to submit a new submission.

- ▷ We have added results for ediffiqal_000 and igd_004.
- ▷ We have added summary tables for Eyes Open 2 and Mouth Open 2.

2024-08-08:

- ▷ We have added results for mobbl_001 and veridium_001.

2024-07-03:

- ▷ We have added results for neurotechnology_005, qazsmartvisionai_000, and idemia_003.
- ▷ We have added Section 4 to address two topics: thresholds on quality measures, and demographic variability in quality measures. The two topics are related because the use of a threshold could lead to inequitable sample rejection outcomes across demographic groups. The new section includes violin plots by subject region of birth for five quality measures: Underexposure, Overexposure, Resolution, Eyes Open 2, and Mouth Open 2.

2024-05-13: The FATE Quality SIDD track remains open. Developers should wait four months from their last (successful) submission to submit a new submission.

- ▷ We have added results for cu-face_001.

2024-04-26: The FATE Quality SIDD track remains open. Developers should wait four months from their last (successful) submission to submit a new submission.

- ▷ We have added results for two new submissions: frpkauai_001 and secunet_005.

2024-04-16: The FATE Quality SIDD track remains open. Developers should wait four months from their last (successful) submission to submit a new submission.

- ▷ We have added results for two new submissions: igd_003 and roc_007.
- ▷ We have reduced the range of the y-axis in the Pitch Set 1, Pitch Set 2, and Roll plots so that the violins are more clearly readable.

2024-04-05: Starting today, we are implementing a four-month waiting period; developers should wait four months from their last (successful) submission to submit a new submission.

2024-03-29: The FATE Quality SIDD track remains open. Developers should wait until one week following the release of the results of their last submission to submit a new submission.

- ▷ Due to a systematic offset in the determination of ground truth for Pitch Set 2, we have increased ground truth pitch values by 10 degrees; that is, each violin's x-coordinate has increased by 10 in Figure 9. Table 5 has been updated accordingly. The Frankfurt Horizon may have been used to determine pitch in the initial collection, causing the offset.

2024-03-08: The FATE Quality SIDD track remains open. Developers should wait until one week following the release of the results of their last submission to submit a new submission.

- ▷ We have added results for one new submission: mobbl_000.

2024-02-14: The FATE Quality SIDD track remains open. Developers should wait until one week following the release of the results of their last submission to submit a new submission.

- ▷ We have added results for two new submissions: roc_006 and viante_000.

2024-02-02: The FATE Quality SIDD track remains open. Developers should wait until one week following the release of the results of their last submission to submit a new submission.

- ▷ We have added results for one submission: igd_002.
- ▷ We have changed the EyeGlassesPresent violin plot to a confusion matrix to show detection and error rates based on a threshold.

2023-12-29: The FATE Quality SIDD track remains open. Developers should wait until one week following the release of the results of their last submission to submit a new submission.

- ▷ We have added results for two new submissions: datech_000 and vsoft_002.

2023-12-15: The FATE Quality SIDD track remains open. Developers should wait until one week following the release of a report to submit a new submission.

- ▷ We have added results for two new submissions: seamfix_002 and neurotechnology_004.
- ▷ The Unified Quality Score panel for digidata_001 was previously calculated based on an FNMR average from only two recognition algorithms; it has been updated to be based on an FNMR average from 15 recognition algorithms, consistent with the other panels.

- ▷ We have added a set for Inter-Eye Distance with non-zero yaw: Inter-Eye Distance Set 3.

2023-12-01: The FATE Quality SIDD track remains open. From this point forward, developers should wait until one week following the release of a report to submit a new submission.

- ▷ We have added results for two new submissions: secunet_003 and secunet_004.
- ▷ There are two images for Eyes Open for which we have corrected the ground truth values.
- ▷ We have added three measures: Eyes Open2, MouthOpen 2, and FaceOcclusion 2. As described in our API Concept Document, these modify the Eyes Open, Mouth Open, and Face Occlusion measures and are aligned with ISO/IEC:29794-5.

  The original measures, Eyes Open, Mouth Open, and Face Occlusion should no longer be implemented, and will not be evaluated if they are implemented from this point forward.

2023-11-06: The FATE Quality SIDD track remains open.

- ▷ We have added results for three new submissions: Dermalog, IDEMIA, and SeamFix.
- ▷ For quality measures for which we report error with respect to ground truth, we are now reporting mean absolute error instead of median absolute error. The mean absolute error includes a penalty for failures to detect a face. The penalty is described in the caption of the plots when such failures occur.
- ▷ We have added two sets for manually estimated pitch and yaw: Pitch Set 3 and Yaw Set 3.

2023-10-13: The FATE Quality SIDD track remains open.

- ▷ We have added results for one new submission: Fraunhofer IGD.
- ▷ Yaw Set 2 previously contained three copies of each image at different sizes. We have fixed this issue, so the largest size of each image is now present, and others have been removed.
- ▷ The matrix showing performance for Total Faces Present has been updated so that when the estimated face count is not given, we report an estimated count of zero faces.
- ▷ We have updated the Face Occlusion plot with the count of instances when the algorithm did not return an estimate.

2023-09-19: The FATE Quality SIDD track remains open.

▷ This document is the first release of the Quality SIDD report. It contains results for seven submissions from five participants: Digidata, FRP, Secunet, Neurotechnology, and Rank One.

The procedure and format of submissions to the evaluation can be found in the API document [PDF].

Table of Contents

List of Tables

List of Figures

Acknowledgments

The authors would like to thank the U.S. Department of Homeland Security Office of Biometric Identity Management (DHS OBIM) and the U.S. Department of Homeland Security's Science and Technology Directorate (S&T) for their collaboration and contributions to this activity. The authors are also grateful to staff in the NIST Biometrics Research Laboratory for infrastructure supporting rapid evaluation of algorithms.

Other Relevant Reports

Results from the Face Recognition Technology Evaluation (FRTE) and Face Analysis Technology Evaluation (FATE) activities appear in the series of NIST Interagency Reports tabulated below. From 1999 to May 2024, FRTE and FATE were collectively known as FRVT.

DATE	PROG.	NISTIR	TITLE
2014-03-20	FATE	7995	Performance of Automated Age Estimation Algorithms
2015-04-20	FATE	8052	Performance of Automated Gender Classification Algorithms
2014-05-21	FRTE	8009	Performance of Face Identification Algorithms
2017-03-07	FRTE	8173	FIVE - Face In Video Evaluation: Face Recognition of Non-Cooperative Subjects
2017-11-23	FRTE	8197	FRPC - The 2017 IARPA Face Recognition Prize Challenge
2020-01-03	FRTE	Draft	Part 1: Verification
2019-09-11	FRTE	8271	Part 2: Identification
2019-12-11	FRTE	8280	Part 3: Demographic Effects
2020-03-04	FATE	8292	Part 4: MORPH - Performance of Automated Face Morph Detection
2020-03-06	FATE	Draft	Part 5: Face Image Quality Assessment
2020-07-24	FRTE	8311	Part 6A: Face Recognition Accuracy with Face Masks using Pre-COVID-19 Algorithms
2022-01-20	FRTE	8331	Part 6B: Face Recognition Accuracy with Face Masks using Post-COVID-19 Algorithms
2022-07-13	FRTE	8381	Part 7: Identification for Paperless Travel and Immigration
2022-09-30	FRTE	8429	Part 8: Summarizing Demographic Differentials
2022-09-30	FRTE	8439	Part 9A: Face Recognition Verification Accuracy on Distinguishing Twins
2023-09-20	FATE	8491	Part 10: Performance of Passive Software-based Presentation Attack Detection (PAD) Algorithms
2023-09-20	FATE	8485	Part 11: Face Image Quality Vector Assessment: Specific Image Defect Detection
2024-05-29	FATE	8525	Part 12: Face Analysis Technology Evaluation: Age Estimation and Verification

Details appear on pages linked from https://www.nist.gov/programs-projects/face-projects.

1. Introduction

Consider the procedure of taking a passport photo: whether for renewal or obtaining a visa, there are formal standards that the capture subject and photographer must follow in order to take an acceptable photo. These standards vary in the required photo size, but commonly require a frontal viewpoint, open eyes, a neutral expression, a uniform background, and other criteria to be fulfilled; for a detailed discussion of these criteria, see Annex D1 of ISO/IEC 39794-5:2019 . These standards and the practices needed to conform to them are intended to support highly accurate face recognition by ensuring that the captured photo can serve as a high quality reference photo in a machine readable travel document (e.g. passport) or in a reference database (e.g. the IDENT system in the US, or the EU-VIS BMS system in Europe).

This Face Analysis Technology Evaluation (FATE) track, Specific Image Defect Detection (SIDD), is being conducted to support quality assessment in general, and to support assessment of quality component algorithms that implement the quality checks of ISO/IEC 29794-5:2024. That standard enumerates checks on face photos that derive from ISO/IEC 19794-5:2011, which established photographic and subject appearance requirements for enrollment images in the European Entry-Exit-System (according to EU-EES implementing decision 2019/329), and ISO/IEC 39794-5:2019, which refined and extended photograph specification and will be used for e-Passports from 2030 onwards.

The existing FATE Quality Summarization Track is an ongoing track that examines the relation between quality score and false non-match rates in order to gauge how well a quality component algorithm can predict false negative errors. However, it does not differentiate between different factors (i.e. quality components) that affect quality. In the SIDD track, we delve deeper into a nuanced discussion of quality measures.

The procedure and format of submissions to our evaluation are described in the Quality SIDD Assessment API document.

2. Test Sets

2.1. Development

The Quality SIDD assessment proceeds by passing photographs to algorithms using a NIST-defined C++ API. The test sets consist of images sequestered at NIST, i.e. developers do not have access to the images. NIST has curated sets specifically to evaluate the performance of algorithms measuring the quantities given in Table 3. For example, there are various sets of frontal and non-frontal images to evaluate pose estimation accuracy; the images in these sets have known pitch and yaw angles.

We formulate ground truth for test sets using three approaches:

- Camera placement: At the time of capture, a camera is placed at a specific angle to the subject.
- Manual labeling: We determine ground truth by human inspection, by measuring the desired quantity using software such as GIMP.
- Synthetic degradation: We generate images with different degrees of a defect (blur, overexposure and underexposure) by applying varying amounts of a defect to a natural image.

This section contains results from all algorithms submitted from the inception of the Quality SIDD evaluation in July 2022.

2.2. Limitations

For several measures, such as inter-eye distance and mouth aperture, we use ground truth that is determined by human inspection. This style of testing, in which ground truth is a continuous variable with some measurement error, means that the software can never be perfect. This is in contrast to a recognition test, for example, where labels are discrete and, ideally, error-free.

2.3. Test Set Sizes

Table 1 lists the number of images in each test set for the quality measures that were implemented so far.

Table 1. Set sizes. This table presents the quality measures in the SIDD track and the number of images in each test set.

Dataset	Number of Images
TotalFacesPresent	92
SubjectPosePitch-1	6291
SubjectPosePitch-2	7145
SubjectPosePitch-3	288
SubjectPoseYaw-1	6267
SubjectPoseYaw-2	11338
SubjectPoseYaw-3	288
SubjectPoseRoll	12000
EyesOpen	107
EyesOpen2	107
InterEyeDistance-1	40
InterEyeDistance-2	39
InterEyeDistance-3	1800
MouthOpen	145
MouthOpen2	145
BackgroundUniformity	229
Resolution-1	8000
Resolution-2	1772
Underexposure	250
Overexposure	250
EyeGlassesPresent	279
SunGlassesPresent	40
CompressionArtifacts	500
FaceOcclusion	30
FaceOcclusion2	30
MotionBlur	6000
PixelsFromEyeToLeftEdge-1	40
PixelsFromEyeToRightEdge-1	40
PixelsFromEyesToTop-1	40
PixelsFromEyesToBottom-1	40
PixelsFromEyeToLeftEdge-2	36
PixelsFromEyeToRightEdge-2	36
PixelsFromEyesToTop-2	36
PixelsFromEyesToBottom-2	36

3. Algorithms and Results

3.1. Algorithms

Table 2 lists the participants who submitted algorithms to the Quality SIDD Assessment.

Table 2. Quality SIDD Assessment Participants

Participant Name	Short Name	Sequence Number	Submission Date
Digidata	digidata	001	2022.09.29
FRP LLC	frpkauai	000	2022.10.28
Secunet Security Networks AG (part of OFIQ)	secunet	001	2023.02.16
Secunet Security Networks AG (part of OFIQ)	secunet	002	2023.04.21
Neurotechnology	neurotechnology	002	2023.07.10
Rank One Computing	rankone	005	2023.07.14
Neurotechnology	neurotechnology	003	2023.08.10
Fraunhofer IGD	igd	001	2023.09.18
Idemia	idemia	002	2023.10.16
Dermalog	dermalog	002	2023.10.19
Seamfix Limited	seamfix	001	2023.10.23
Secunet Security Networks AG (part of OFIQ)	secunet	004	2023.11.17
Secunet Security Networks AG (part of OFIQ)	secunet	003	2023.11.24
Seamfix Limited	seamfix	002	2023.11.24
Neurotechnology	neurotechnology	004	2023.12.06
Dactionable Technologies	datech	000	2023.12.08
Vsoft	vsoft	002	2023.12.11
Fraunhofer IGD	igd	002	2024.01.22
ROC	roc	006	2024.02.05
Viante.AI	viante	000	2024.02.06
Mobbeel Solutions	mobbl	000	2024.02.22
Fraunhofer IGD	igd	003	2024.03.28
ROC	roc	007	2024.04.05
FRP LLC	frpkauai	001	2024.04.15
Secunet Security Networks AG (part of OFIQ)	secunet	005	2024.04.24
Cu-Face	cu-face	001	2024.04.30
Neurotechnology	neurotechnology	005	2024.05.30
QazSmartVision.AI	qazsmartvisionai	000	2024.06.04
Idemia	idemia	003	2024.06.14
Ediffiqal	ediffiqal	000	2024.07.22
Mobbeel Solutions	mobbl	001	2024.07.23

Table 2. Quality SIDD Assessment Participants *(continued)*

Participant Name	Short Name	Sequence Number	Submission Date
Veridium	veridium	001	2024.07.30
Fraunhofer IGD	igd	004	2024.08.14
PAPIL11 S.R.O.	papil11	000	2024.09.24
Viante.AI	viante	001	2024.12.02
Fraunhofer IGD	igd	005	2024.12.16
Veridium	veridium	002	2024.12.19
Mobbeel Solutions	mobbl	002	2024.12.20
ROC	roc	008	2024.12.23
Kasikorn Labs	kasikornlabs	000	2024.12.23
Guangzhou Pixel Solutions	pixelall	001	2025.03.04
Mobbeel Solutions	mobbl	003	2025.05.08
Veridium	veridium	003	2025.05.13
Innovatrics	innovatrics	000	2025.07.03
Guangzhou Pixel Solutions	pixelall	002	2025.07.10
Daon	daon	000	2025.08.05

3.2. Quality Measures Supported

Table 3 lists the quality measures defined in our API and the algorithms that implement them. The first row indicates whether the quality measure must be checked for an image to be used as a reference image in a machine-readable travel document (MRTD) such as a passport, which is Use Case 1 (UC1) as listed in ISO/IEC 29794-5:2024.

Table 3. Quality Measures Supported. This table presents the participating algorithm name and which SIDD quality measures were implemented. A 'Y' (for 'Yes') indicates that the quality measure is implemented; a blank space indicates that it is not implemented. The first row indicates which quality measures are required to be checked for use as a reference photo in a machine-readable travel document (MRTD) according to ISO/IEC 29794-5:2024.

Algorithm	TotalFacesPresent	SubjectPosePitch	SubjectPoseYaw	SubjectPoseRoll	EyesOpen	EyesOpen2	InterEyeDistance	MouthOpen	MouthOpen2	BackgroundUniformity	Resolution	Underexposure	Overexposure	PixelsFromEyeToLeftEdge	PixelsFromEyeToRightEdge	PixelsFromEyesToTop	PixelsFromEyesToBottom	EyeGlassesPresent	SunGlassesPresent	CompressionArtifacts	FaceOcclusion	FaceOcclusion2	MotionBlur	UnifiedQualityScore
Required for MRTD	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y						Y	Y
digidata-001	Y	Y	Y	Y	Y		Y	Y			Y	Y	Y											
frpkauai-000	Y	Y	Y	Y	Y		Y			Y	Y													
secunet-001	Y	Y	Y	Y	Y		Y	Y		Y		Y	Y											
secunet-002	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y											Y
neurotechnology-002	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y
rankone-005	Y	Y	Y	Y	Y		Y	Y			Y			Y	Y	Y	Y	Y	Y	Y	Y		Y	Y
neurotechnology-003	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y
igd-001	Y	Y	Y	Y			Y				Y	Y	Y	Y	Y	Y	Y							
idemia-002	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y
dermalog-002	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y					Y
seamfix-001	Y	Y	Y	Y				Y		Y	Y	Y	Y					Y	Y					
secunet-004	Y	Y	Y	Y			Y			Y	Y									Y		Y		Y
secunet-003	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y	Y
seamfix-002	Y	Y	Y	Y						Y	Y	Y	Y					Y	Y				Y	Y

Table 3. Quality Measures Supported. This table presents the participating algorithm name and which SIDD quality measures were implemented. A 'Y' (for 'Yes') indicates that the quality measure is implemented; a blank space indicates that it is not implemented. The first row indicates which quality measures are required to be checked for use as a reference photo in a machine-readable travel document (MRTD) according to ISO/IEC 29794-5:2024. *(continued)*

Algorithm	TotalFacesPresent	SubjectPosePitch	SubjectPoseYaw	SubjectPoseRoll	EyesOpen	EyesOpen2	InterEyeDistance	MouthOpen	MouthOpen2	BackgroundUniformity	Resolution	Underexposure	Overexposure	PixelsFromEyeToLeftEdge	PixelsFromEyeToRightEdge	PixelsFromEyesToTop	PixelsFromEyesToBottom	EyeGlassesPresent	SunGlassesPresent	CompressionArtifacts	FaceOcclusion	FaceOcclusion2	MotionBlur	UnifiedQualityScore
neurotechnology-004	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
datech-000	Y																							Y
vsoft-002																								Y
igd-002	Y	Y	Y	Y		Y	Y		Y		Y	Y	Y	Y	Y	Y	Y							
roc-006	Y	Y	Y	Y		Y	Y				Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
viante-000	Y	Y	Y	Y		Y	Y		Y		Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y
mobbl-000	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
igd-003	Y	Y	Y	Y		Y	Y		Y		Y	Y	Y	Y	Y	Y	Y		Y			Y		
roc-007	Y	Y	Y	Y		Y	Y				Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
frpkauai-001	Y	Y	Y	Y		Y	Y			Y	Y			Y	Y	Y	Y							Y
secunet-005												Y	Y							Y				
cu-face-001	Y					Y			Y					Y	Y	Y	Y	Y	Y					
neurotechnology-005	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
qazsmartvisionai-000	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
idemia-003	Y	Y	Y	Y			Y				Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y

Table 3. Quality Measures Supported. This table presents the participating algorithm name and which SIDD quality measures were implemented. A 'Y' (for 'Yes') indicates that the quality measure is implemented; a blank space indicates that it is not implemented. The first row indicates which quality measures are required to be checked for use as a reference photo in a machine-readable travel document (MRTD) according to ISO/IEC 29794-5:2024. *(continued)*

Algorithm	TotalFacesPresent	SubjectPosePitch	SubjectPoseYaw	SubjectPoseRoll	EyesOpen	EyesOpen2	InterEyeDistance	MouthOpen	MouthOpen2	BackgroundUniformity	Resolution	Underexposure	Overexposure	PixelsFromEyeToLeftEdge	PixelsFromEyeToRightEdge	PixelsFromEyesToTop	PixelsFromEyesToBottom	EyeGlassesPresent	SunGlassesPresent	CompressionArtifacts	FaceOcclusion	FaceOcclusion2	MotionBlur	UnifiedQualityScore
ediffiqal-000																								Y
mobbl-001	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
veridium-001	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y
igd-004	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y		Y			Y		
papil11-000	Y	Y	Y	Y			Y							Y	Y	Y	Y							Y
viante-001	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y
igd-005		Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y		Y			Y		Y
veridium-002	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
mobbl-002	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
roc-008	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
kasikornlabs-000	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
pixelall-001	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
mobbl-003	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
veridium-003	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	
innovatrics-000	Y	Y	Y	Y			Y							Y	Y	Y	Y							Y

Table 3. Quality Measures Supported. This table presents the participating algorithm name and which SIDD quality measures were implemented. A 'Y' (for 'Yes') indicates that the quality measure is implemented; a blank space indicates that it is not implemented. The first row indicates which quality measures are required to be checked for use as a reference photo in a machine-readable travel document (MRTD) according to ISO/IEC 29794-5:2024. *(continued)*

Algorithm	TotalFacesPresent	SubjectPosePitch	SubjectPoseYaw	SubjectPoseRoll	EyesOpen	EyesOpen2	InterEyeDistance	MouthOpen	MouthOpen2	BackgroundUniformity	Resolution	Underexposure	Overexposure	PixelsFromEyeToLeftEdge	PixelsFromEyeToRightEdge	PixelsFromEyesToTop	PixelsFromEyesToBottom	EyeGlassesPresent	SunGlassesPresent	CompressionArtifacts	FaceOcclusion	FaceOcclusion2	MotionBlur	UnifiedQualityScore
pixelall-002	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y
daon-000	Y						Y							Y	Y	Y	Y							Y

3.3. Timing

The duration of execution of quality algorithm (QA) software is important in those applications where fast quantification is needed to support usability by providing usable feedback to a capture subject. It may be important also, for example, in running QA software over large legacy collections. This section gives duration of the various implementations running on a common hardware platform.

Figure 1 shows the timing performance for the participants who submitted algorithms to the Quality SIDD Assessment.

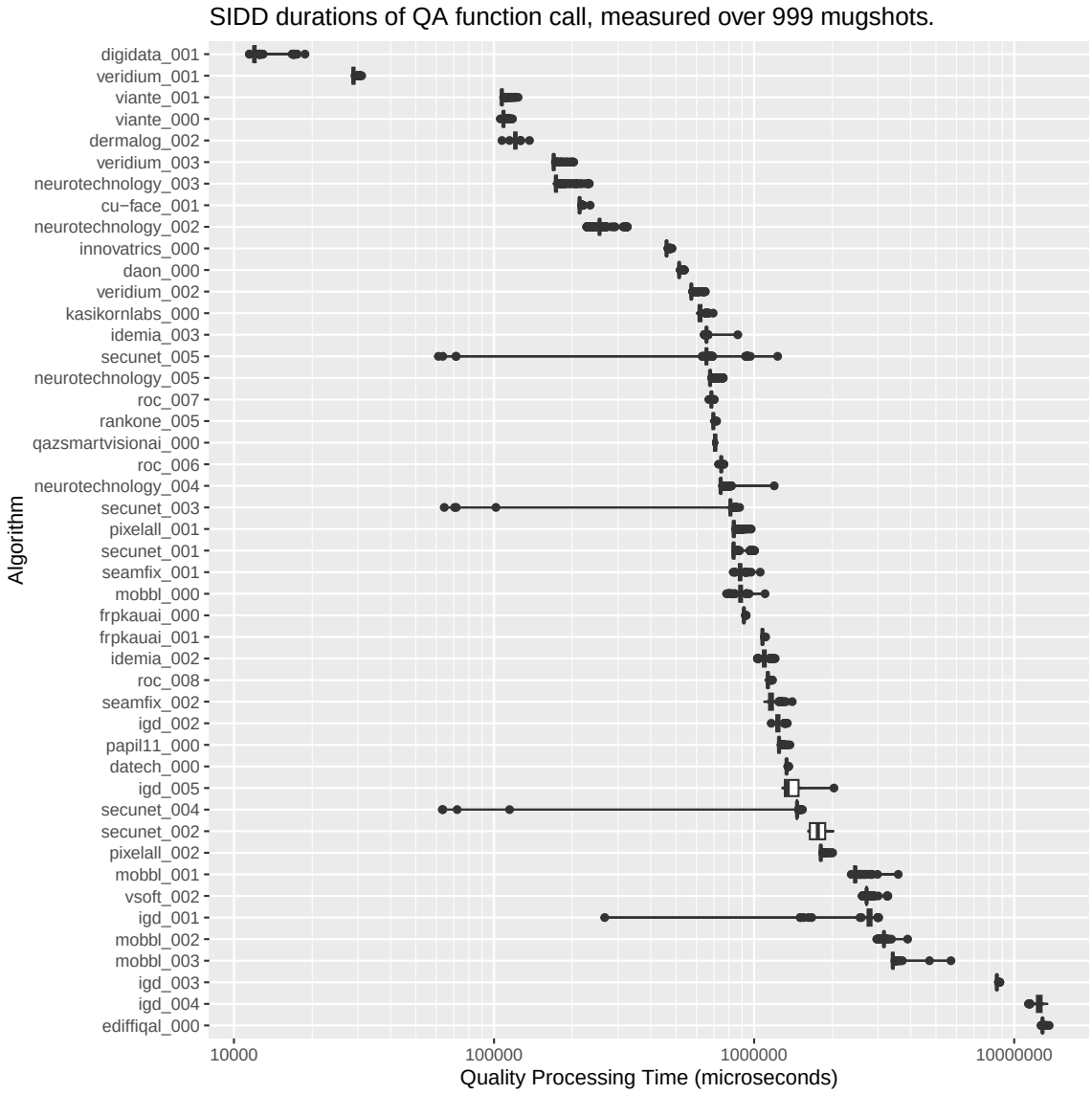


Fig. 1. Distribution of time required for the quality algorithm (QA) function call, measured over 999 mugshots. The implementations vary, in part, because they are computing different quality components – see Table 3. Durations are measured on a fixed Intel Xeon Gold 6140 CPU running at 2.30 GHz. Durations are measured by wrapping the function call in a high resolution timer.

3.4. Total Faces Present

3.4.1. Images Used

The images in the Total Faces Present dataset are captured in a border-crossing setting with a variety of poses and some background non-uniformity. The input images generally have one primary face that is larger than the others. We count faces manually, where a face is counted if its inter-eye distance is estimated to be larger than 0.02 times the width of the image.

3.4.2. Results for Total Faces Present

Figure 2 summarizes the performance of all algorithms that implemented the Total Faces Present measure. Note that there are more missed detections (below the diagonal) than false detections (above the diagonal). Missed detection rate is the number of missed faces divided by the total number of faces; false detection rate is the average number of wrong detections per image. For both false detection rate and missed detection rate, lower values are better.

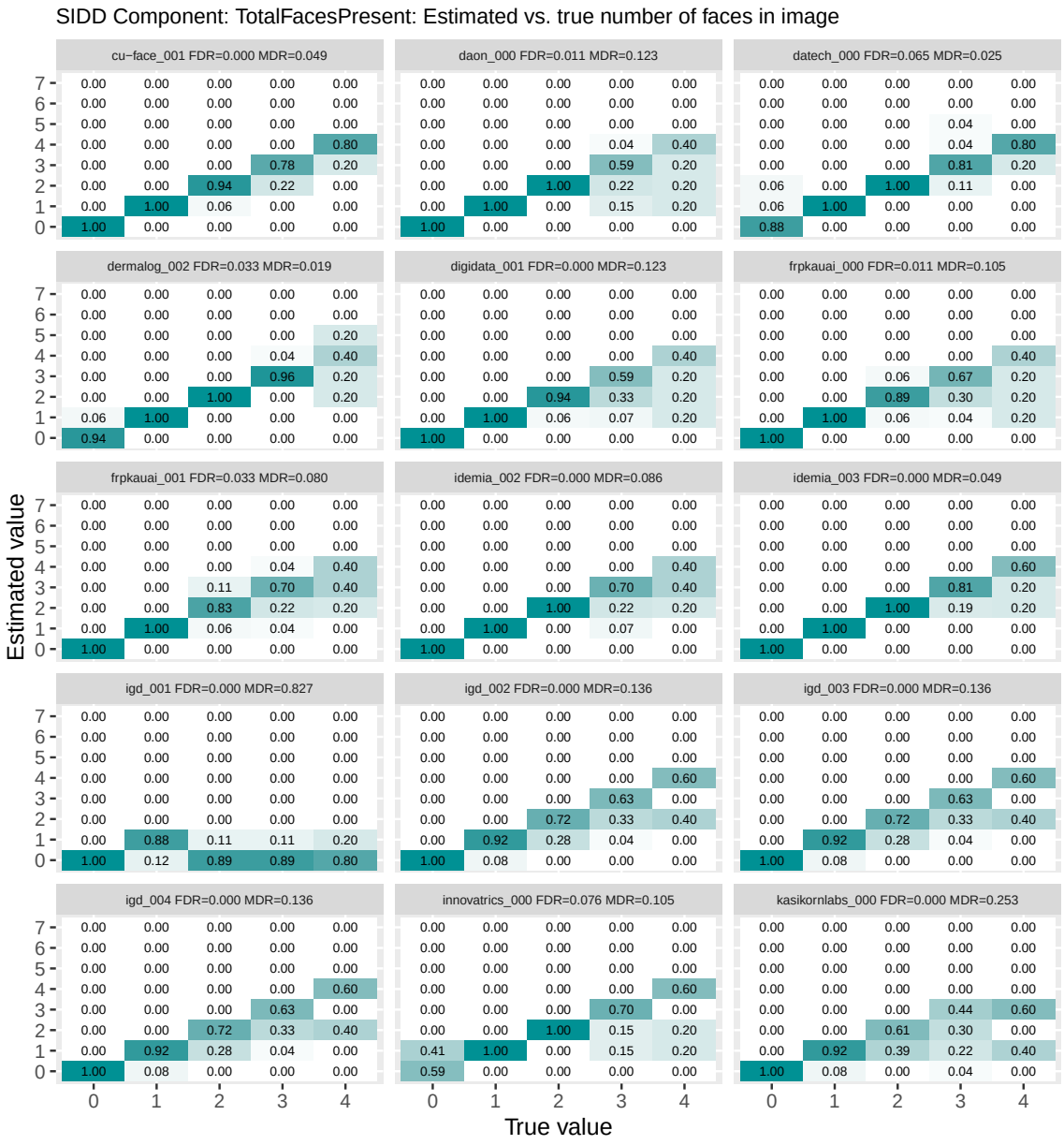


Fig. 2. Matrix of estimated vs. known number of faces, with darker shading indicating larger values. Perfect performance corresponds to zero on off-diagonal entries and 1 on each of the diagonal entries.

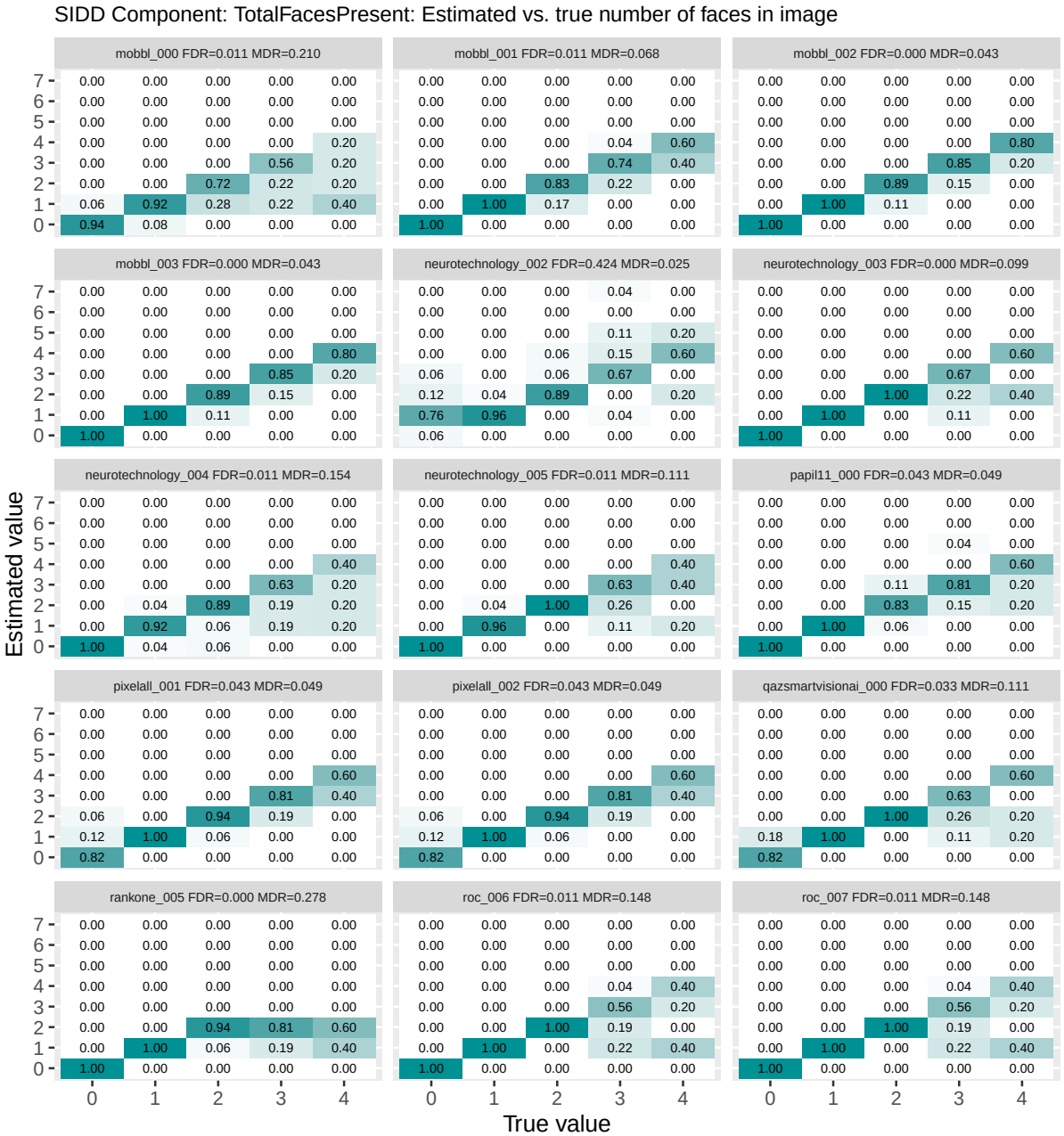


Fig. 3. Matrix of estimated vs. known number of faces, with darker shading indicating larger values. Perfect performance corresponds to zero on off-diagonal entries and 1 on each of the diagonal entries.

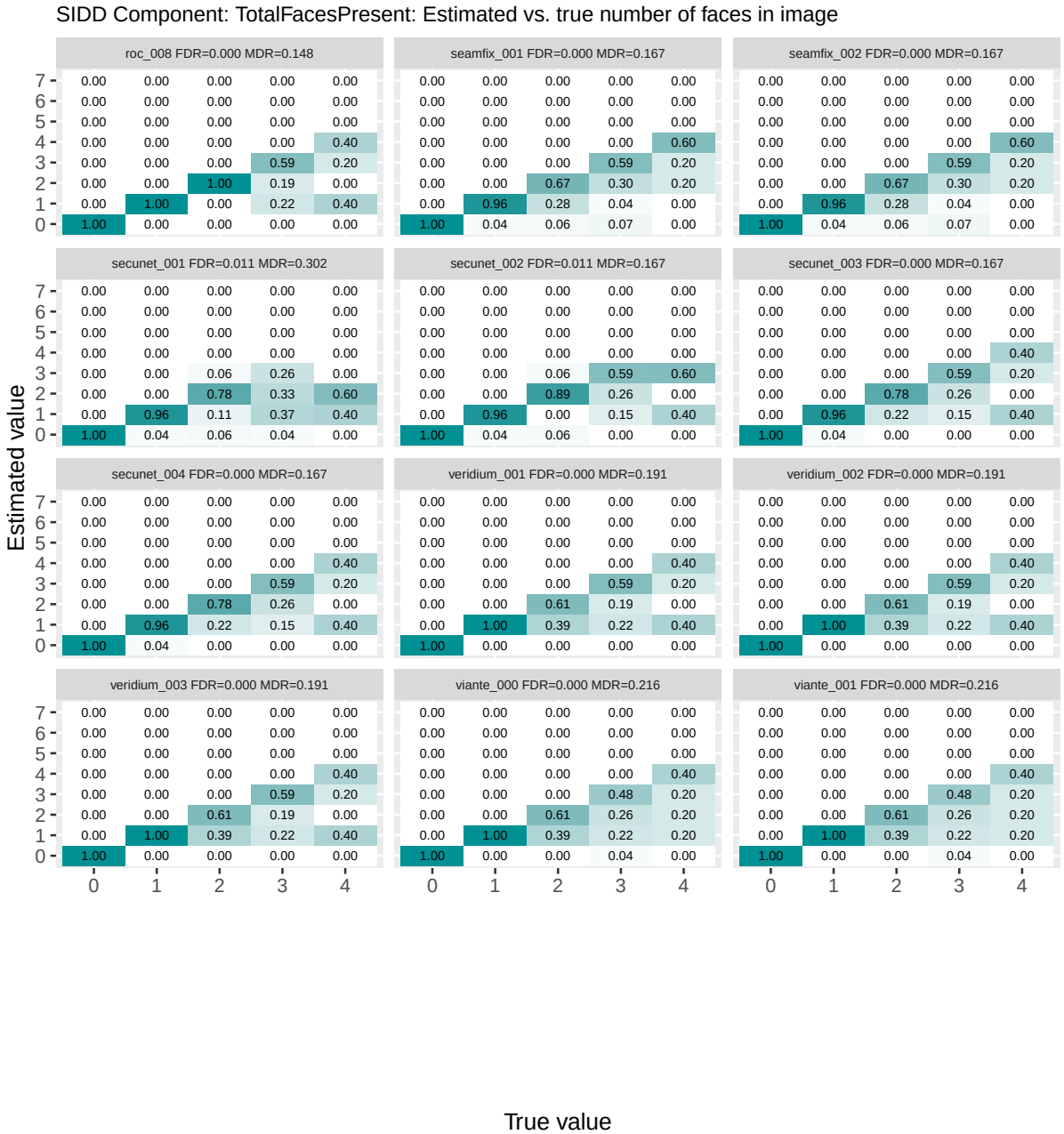


Fig. 4. Matrix of estimated vs. known number of faces, with darker shading indicating larger values. Perfect performance corresponds to zero on off-diagonal entries and 1 on each of the diagonal entries.

3.5. Yaw Angle

3.5.1. Images Used

The images for the Yaw quality measure are from three sets of sequestered photos. The images in these sets have a well-illuminated setting with a uniform background.

1. For the first set, at the time of capture, a camera is placed to the right or left of the subject at varying angles, with the subject remaining frontal and stationary. Yaw is recorded at the time of collection.
2. For the second set, at the time of capture, the subject turns the head to look at a target to the left or right. Yaw is recorded at the time of collection.
3. For the third set, a camera is placed to the right or left of the subject at varying angles; the subject remains stationary and frontal. Ground truth yaw values are determined by rotating a computer-generated model of a head to match the subject's pose, and taking the corresponding yaw reading.

Camera placement to the subject's right corresponds to yaw being positive. Camera placement to the subject's left corresponds to yaw being negative. This sign convention is consistent with the ISO/IEC 39794-5:2019 standard.

3.5.2. Results for Yaw Angle

Table 4, Figure 5, Figure 6, and Figure 7 summarize algorithm performance. For images for which the algorithm does not detect a face, the error is 45 degrees. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 4. SIDD PoseYaw Mean Absolute Error.

Algorithm	Yaw Set 1	Yaw Set 2	Yaw Set 3
roc_008	5.0	9.6	5.1
viante_001	6.3	11.3	4.5
viante_000	6.3	11.3	4.5
roc_006	6.7	10.8	5.5
roc_007	6.7	10.8	5.5
papil11_000	7.0	11.8	3.8
frpkauai_001	7.3	11.8	3.7
veridium_003	7.4	12.4	4.4
frpkauai_000	7.6	11.8	3.7
veridium_001	7.6	12.5	4.4
veridium_002	7.6	12.5	4.4
dermalog_002	7.8	13.2	5.8
idemia_003	8.7	12.5	2.9
idemia_002	8.8	12.5	2.9
qazsmartvisionai_000	8.9	13.2	3.2
neurotechnology_003	9.5	12.6	3.6
neurotechnology_002	9.6	12.4	3.6
mobbl_002	10.1	11.8	4.4
mobbl_003	10.1	11.8	4.4
digidata_001	11.3	20.3	5.1
secunet_004	11.4	11.3	4.0
secunet_002	11.6	14.6	4.2
rankone_005	11.8	12.1	5.3
kasikornlabs_000	13.1	16.1	3.5
secunet_003	13.8	12.4	3.6
neurotechnology_005	18.0	12.9	3.0
mobbl_000	18.2	12.3	4.2
neurotechnology_004	19.6	14.1	3.2
innovatrics_000	21.0	19.6	2.4
seamfix_001	23.4	43.7	13.0
seamfix_002	23.4	43.7	13.0
igd_001	26.8	27.0	9.7
igd_004	28.6	22.5	3.6
igd_005	30.7	22.8	3.4
secunet_001	41.2	14.7	4.2
pixelall_001	42.5	37.6	5.8

Table 4. SIDD PoseYaw Mean Absolute Error. *(continued)*

Algorithm	Yaw Set 1	Yaw Set 2	Yaw Set 3
pixelall_002	42.5	37.6	5.8
igd_003	47.8	37.1	12.3
igd_002	54.7	39.6	16.8
mobbl_001	69.6	11.4	8.8

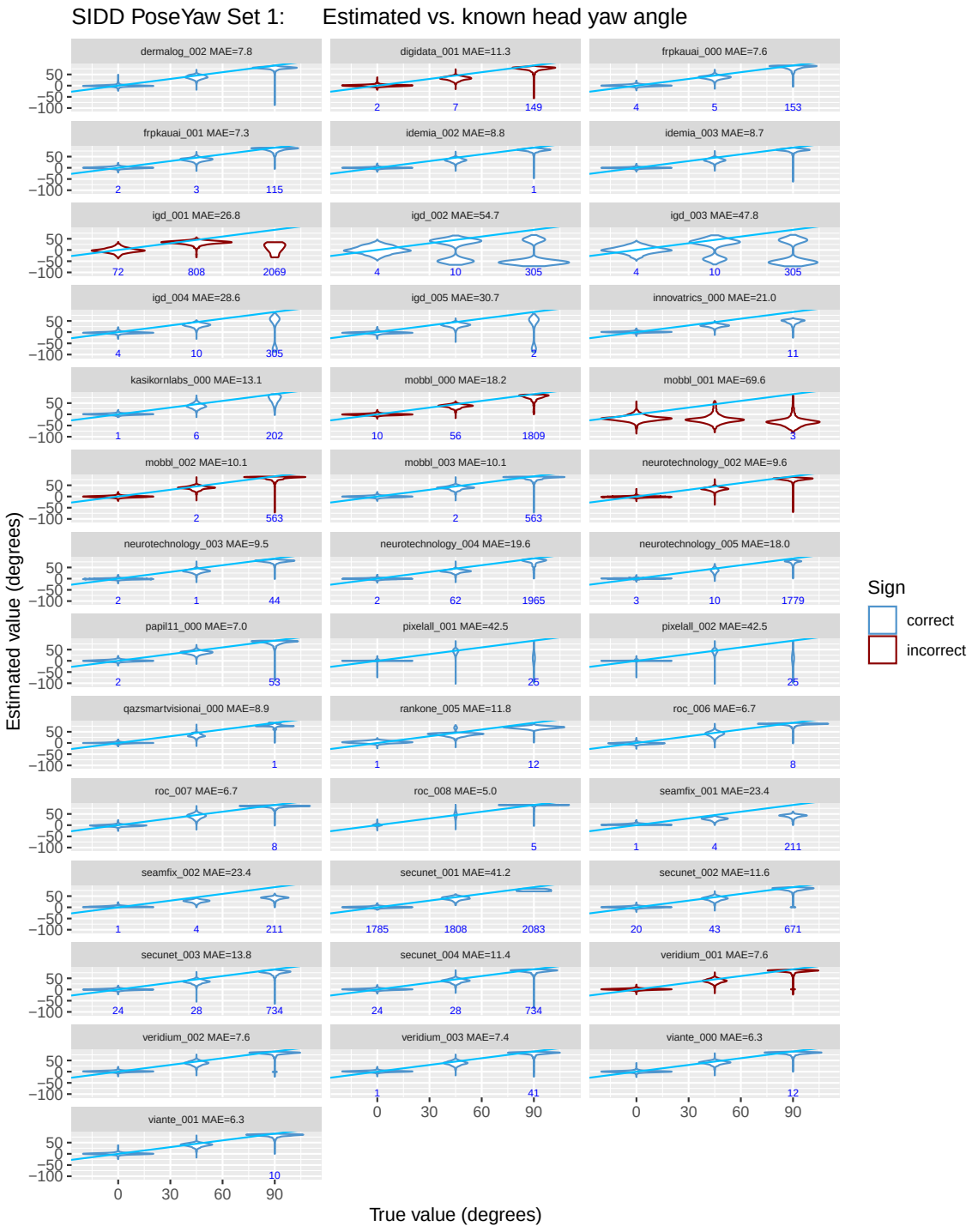


Fig. 5. Estimated vs. known values of yaw angle. For Yaw Set 1, ground truth yaw values are determined by the placement of the camera at the time of capture; the subject remains stationary and frontal. The blue line ($y = x$) represents perfect performance. The plot shows violins at true yaw values with the tails extending to the minimum and maximum estimated values. The small blue numbers represent the count of images for which the software did not detect a face; for such images, the error is set to 45 degrees. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

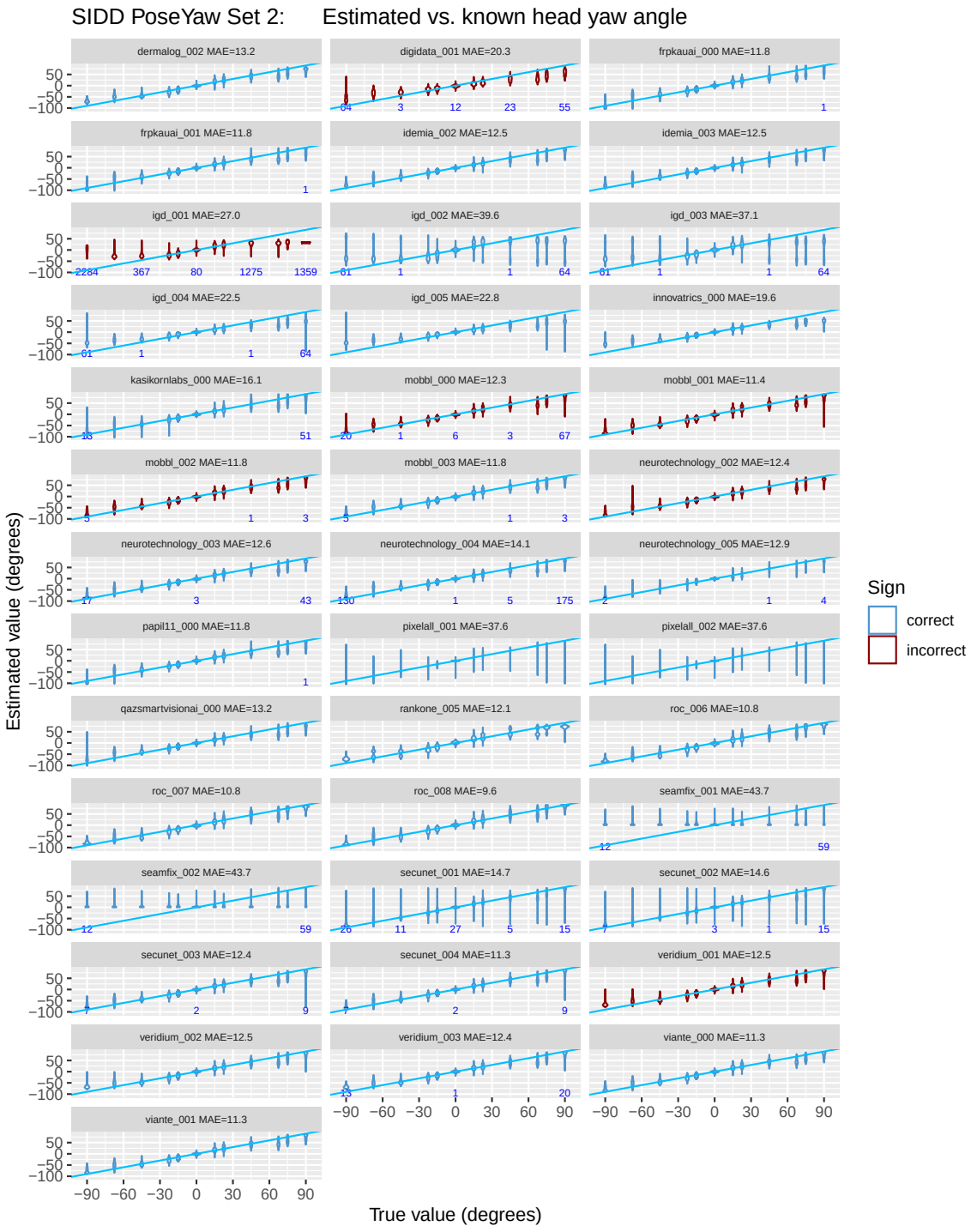


Fig. 6. Estimated vs. known values of yaw angle. For Yaw Set 2, the subject turns the head to look at targets placed to the right or left. The blue line ($y = x$) represents perfect performance. The plot shows violins at true yaw values with the tails extending to the minimum and maximum estimated values. The small blue numbers represent the count of images for which the software did not detect a face; for such images, the error is set to 45 degrees. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

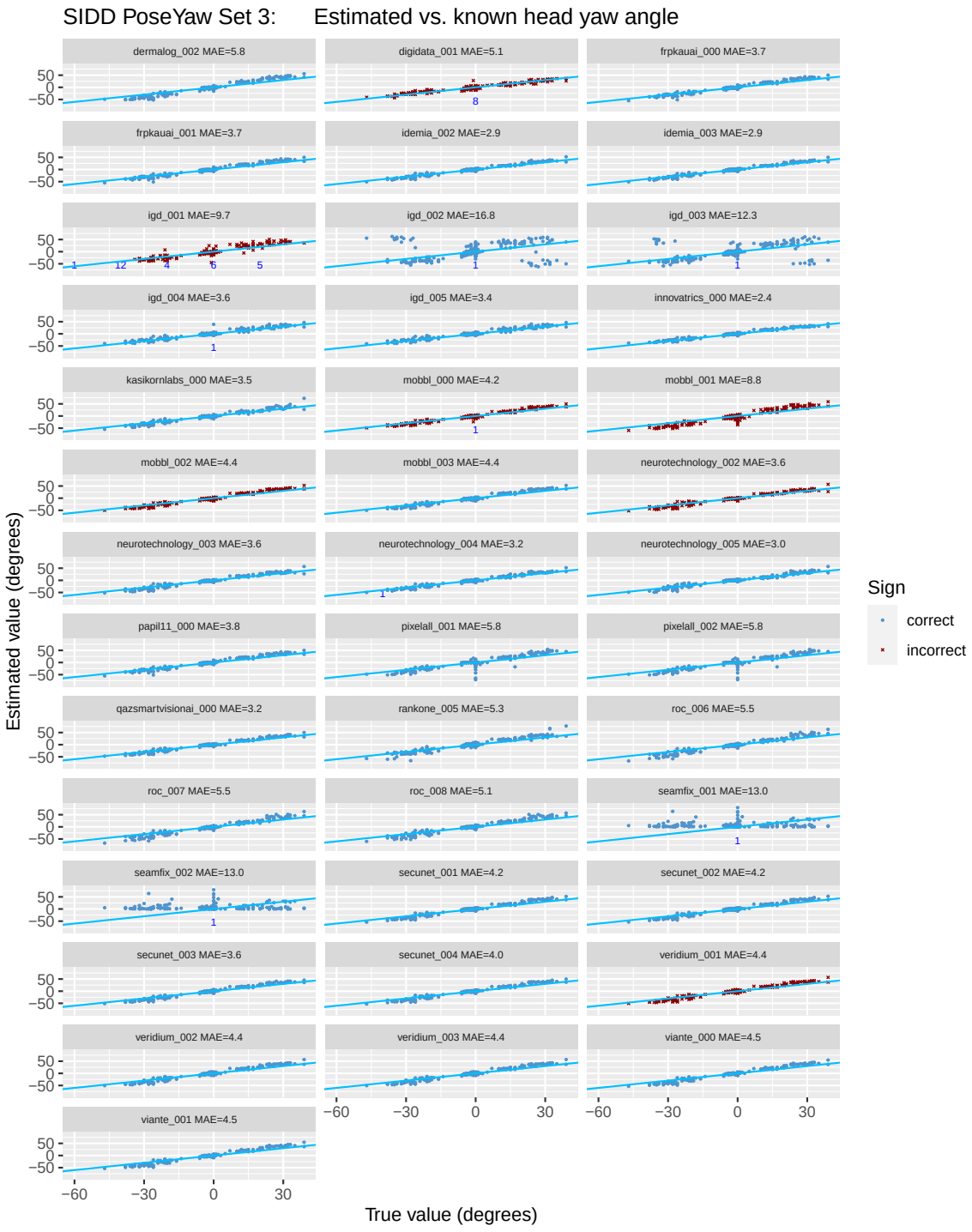


Fig. 7. Estimated vs. known values of yaw angle. For Yaw Set 3, a camera is placed to the right or left of the subject at varying angles; the subject remains stationary and frontal. Ground truth yaw values are determined by rotating a computer-generated model of a head to match the subject's pose, and taking the corresponding yaw reading. The blue line ($y = x$) represents perfect performance. The small blue numbers represent the count of images for which the software did not detect a face; for such images, the error is set to 45 degrees. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

3.6. Pitch Angle

3.6.1. Images Used

The images for the Pitch quality measure are from three sets of sequestered photos.

1. In Set 1, the subject generally has a neutral position and is standing against a mostly uniform background, with some shadows behind the subject. At the time of capture, a camera is placed at varying heights; the subject is asked to be frontal. Pitch is recorded at the time of collection.
2. In Set 2, the subject is seated and is against a uniform background. At the time of capture, a camera is placed at varying heights; the subject is asked to be frontal. Pitch is recorded at the time of collection.
3. In Set 3, the subject raises or lowers the head to look at targets at varying heights. Ground truth pitch values are determined by rotating a computer-generated model of a head to match the subject's pose, and taking the corresponding pitch reading.

Camera placement above the subject, with the top of the head being more exposed, corresponds to pitch being positive, and placement below the subject, with the chin being more exposed, corresponds to pitch being negative. This sign convention is consistent with the ISO/IEC 39794-5:2019 standard.

3.6.2. Results for Pitch Angle

Table 5, Figure 8, Figure 9, and Figure 10 summarize the performance of the algorithms in our evaluation when estimating pitch angle. For images for which the algorithm does not detect a face, the error is 30 degrees. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Note that the definition of zero-pitch is not as well-defined as zero-roll and zero-yaw.

Table 5. SIDD PosePitch Mean Absolute Error.

Algorithm	Pitch Set 1	Pitch Set 2	Pitch Set 3
veridium_001	6.1	13.6	5.6
veridium_002	6.1	13.6	5.6
veridium_003	6.1	13.7	5.6
qazsmartvisionai_000	6.5	7.9	4.8
secunet_004	6.6	14.5	5.3
secunet_002	6.6	13.0	4.9
neurotechnology_004	6.7	5.8	5.1
mobbl_002	6.8	8.2	5.1
mobbl_003	6.8	8.2	5.1
viante_001	7.0	7.5	4.7
viante_000	7.0	7.5	4.7
neurotechnology_003	7.1	5.2	4.8
dermalog_002	7.3	7.0	5.1
mobbl_000	7.4	8.6	5.4
neurotechnology_002	7.5	5.2	5.0
secunet_003	7.9	10.6	4.8
digidata_001	8.0	7.4	5.6
seamfix_001	8.7	8.6	14.2
seamfix_002	8.7	8.6	14.2
kasikornlabs_000	9.1	8.1	5.1
papil11_000	9.1	7.8	4.4
roc_008	9.7	8.5	8.0
idemia_002	10.2	9.7	4.8
idemia_003	10.4	9.7	4.8
roc_006	10.6	13.1	9.5
roc_007	10.6	13.1	9.5
innovatrics_000	10.6	5.8	4.0
neurotechnology_005	10.6	7.7	5.7
igd_004	11.2	8.2	9.0
igd_005	11.7	7.8	8.6
frpkauai_001	12.6	8.8	5.5
frpkauai_000	12.7	8.8	5.5
igd_001	13.2	20.0	10.2
rankone_005	14.4	11.9	12.2
igd_002	15.0	12.1	11.4
igd_003	15.4	13.8	8.0

Table 5. SIDD PosePitch Mean Absolute Error. *(continued)*

Algorithm	Pitch Set 1	Pitch Set 2	Pitch Set 3
mobbl_001	16.6	9.4	6.3
pixelall_001	25.3	46.1	39.8
pixelall_002	27.0	46.3	50.2
secunet_001	28.5	13.1	4.9

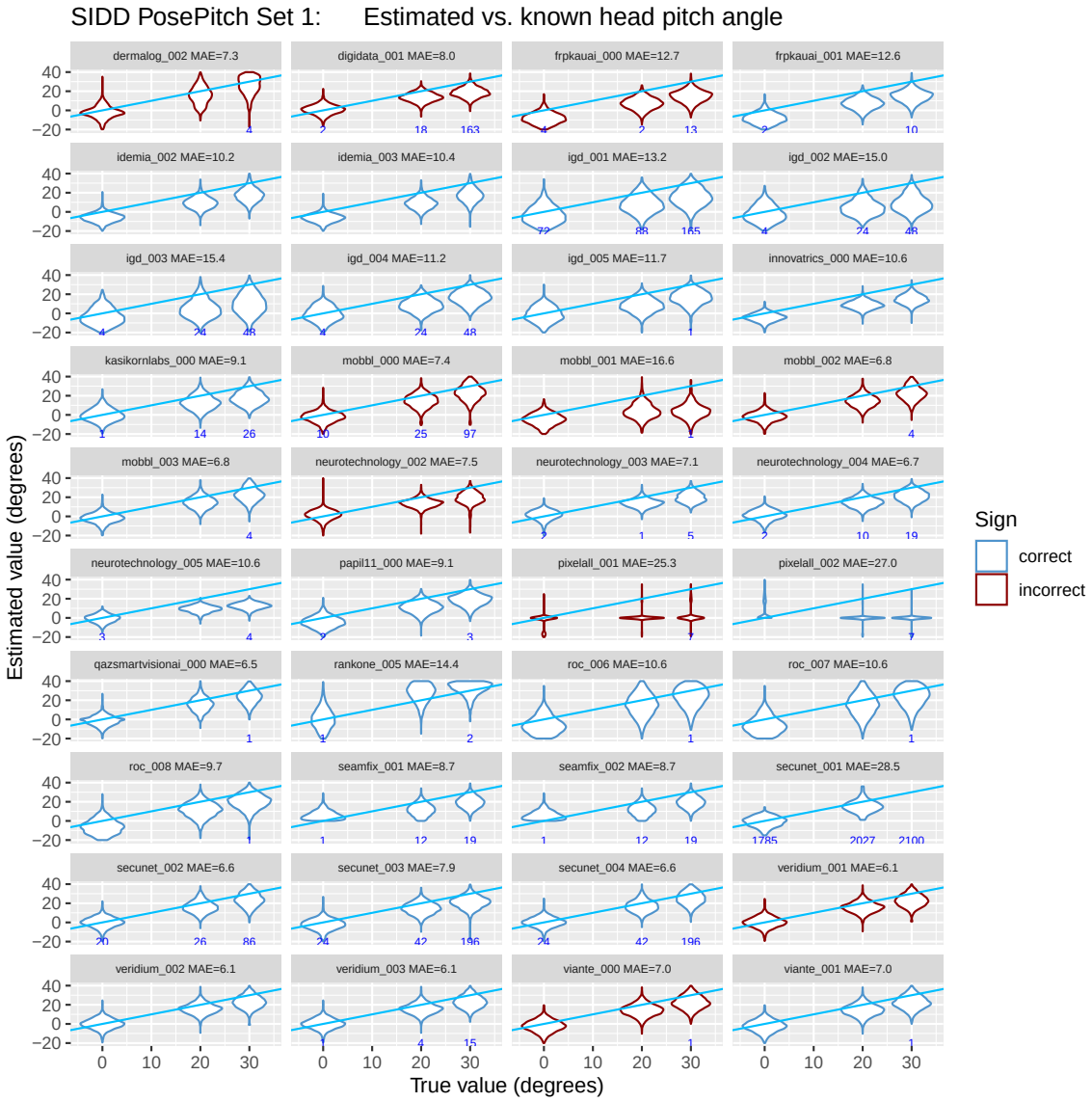


Fig. 8. Estimated vs. known values of pitch angle. For Pitch Set 1, ground truth pitch values are determined by the placement of the camera at the time of capture; the subject remains stationary and frontal. The blue line ($y = x$) represents perfect performance. The small blue numbers represent the count of images for which the software did not detect a face; for such images, the error is set to 30 degrees. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

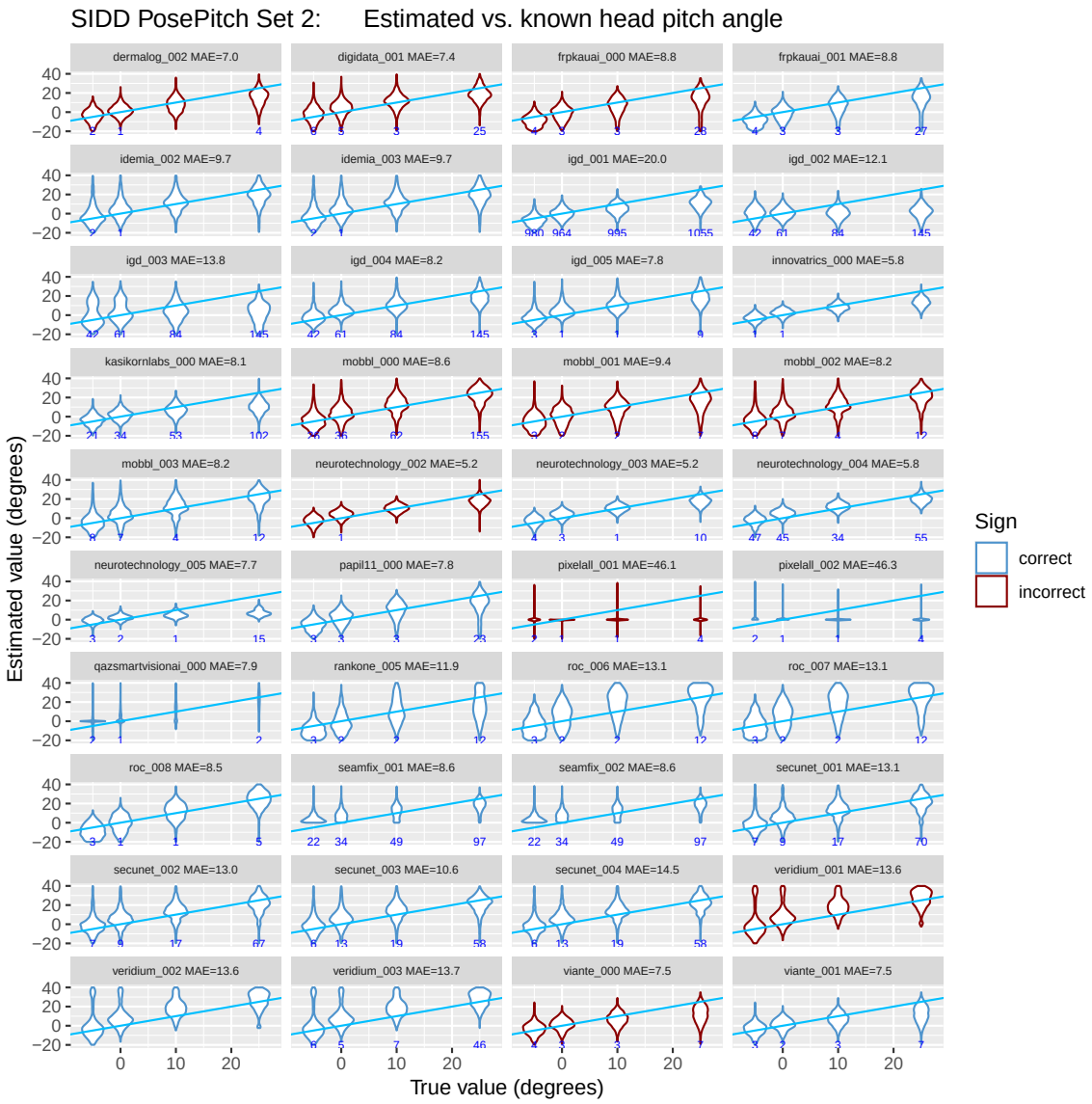


Fig. 9. Estimated vs. known values of pitch angle. For Pitch Set 2, ground truth pitch values are determined by the placement of the camera at the time of capture; the subject remains stationary and frontal. The blue line ($y = x$) represents perfect performance. The plot shows violins at true pitch values with tails extending to the minimum and maximum estimated values. The small blue numbers represent the count of images for which the software did not detect a face; for such images, the error is set to 30 degrees. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

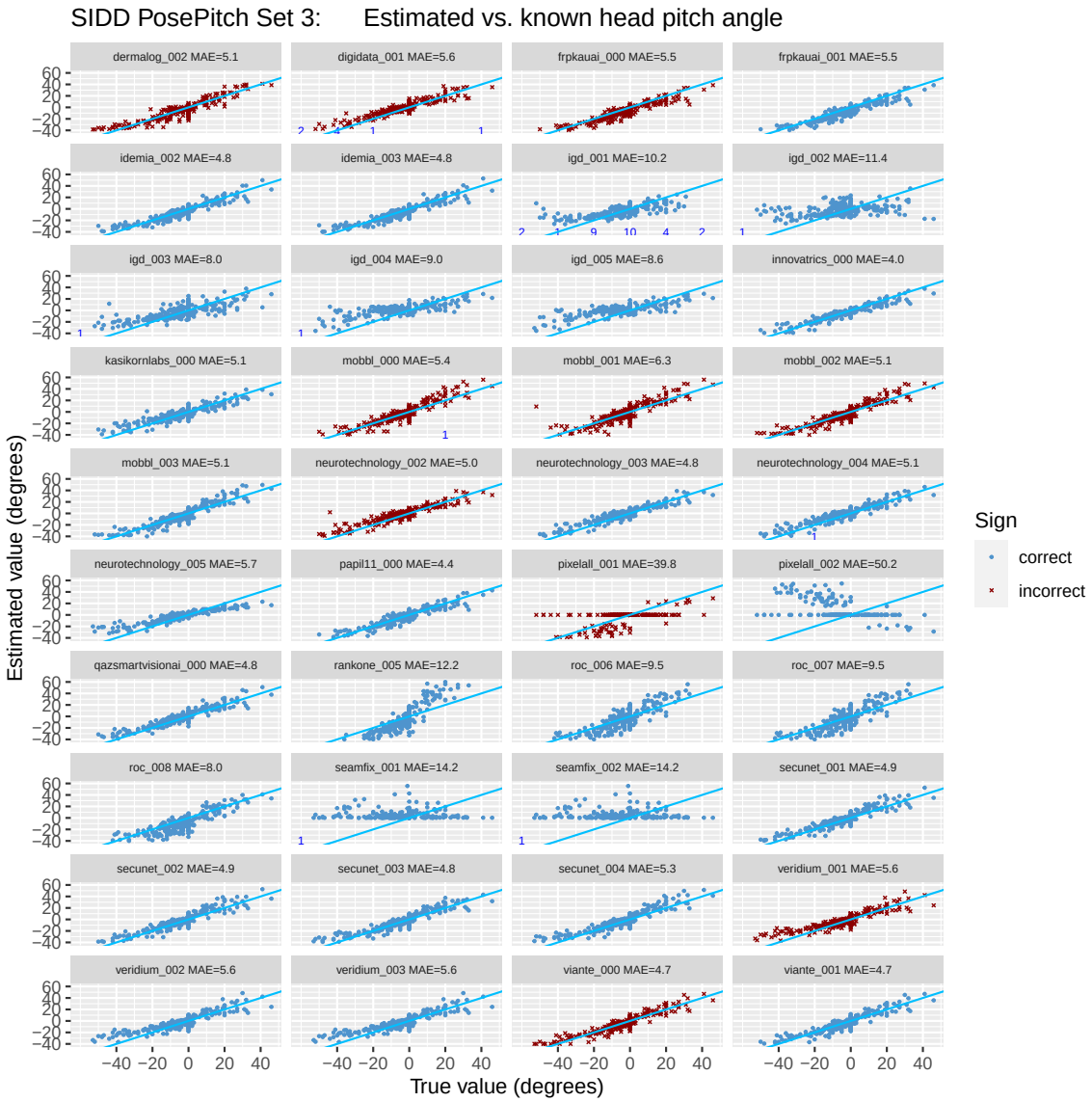


Fig. 10. Estimated vs. known values of pitch angle. For Pitch Set 3, the subject raises or lowers the head to look at targets at varying heights. Ground truth pitch values are determined by rotating a computer-generated model of a head to match the subject's pose, and taking the corresponding pitch reading. The blue line ($y = x$) represents perfect performance. The plot shows violins at true pitch values with the tails extending to the minimum and maximum estimated values. The small blue numbers represent the count of images for which the software did not detect a face; for such images, the error is set to 30 degrees. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate the scores in the next submission.

3.7. Roll Angle

3.7.1. Images Used

The images in the Roll dataset are mugshots that are rotated by a roll angle ranging from −30 to 30 degrees. In particular, we do not include images with a roll angle of 90 degrees. Rotation towards the subject's right shoulder corresponds to a positive roll angle. Rotation towards the subject's left shoulder corresponds to a negative roll angle. This sign convention is consistent with the ISO/IEC 39794-5:2019 standard.

3.7.2. Results for Roll Angle

Table 6 and Figure 11 summarize the performance of the algorithms in our evaluation when estimating roll angle. For images for which the algorithm did not detect a face, the error is 45 degrees. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 6. SIDD PoseRoll Mean Absolute Error

Algorithm	MAE (in degrees)
neurotechnology_004	0.9
innovatrics_000	1.1
igd_005	1.2
igd_004	1.2
idemia_002	1.2
neurotechnology_005	1.3
idemia_003	1.3
viante_001	1.3
viante_000	1.3
secunet_003	1.3
roc_006	1.4
roc_007	1.4
qazsmartvisionai_000	1.4
secunet_004	1.4
frpkauai_000	1.5
frpkauai_001	1.5
rankone_005	1.5
papil11_000	1.5
roc_008	1.5
veridium_001	1.5
veridium_002	1.5

Table 6. SIDD PoseRoll Mean Absolute Error *(continued)*

Algorithm	MAE (in degrees)
veridium_003	1.5
dermalog_002	1.7
pixelall_001	1.7
mobbl_002	1.7
mobbl_003	1.7
neurotechnology_003	1.8
secunet_002	1.8
neurotechnology_002	1.8
secunet_001	1.8
mobbl_000	2.1
digidata_001	2.2
igd_002	2.6
igd_003	3.9
mobbl_001	4.2
kasikornlabs_000	5.2
seamfix_001	8.2
seamfix_002	8.2
igd_001	9.9

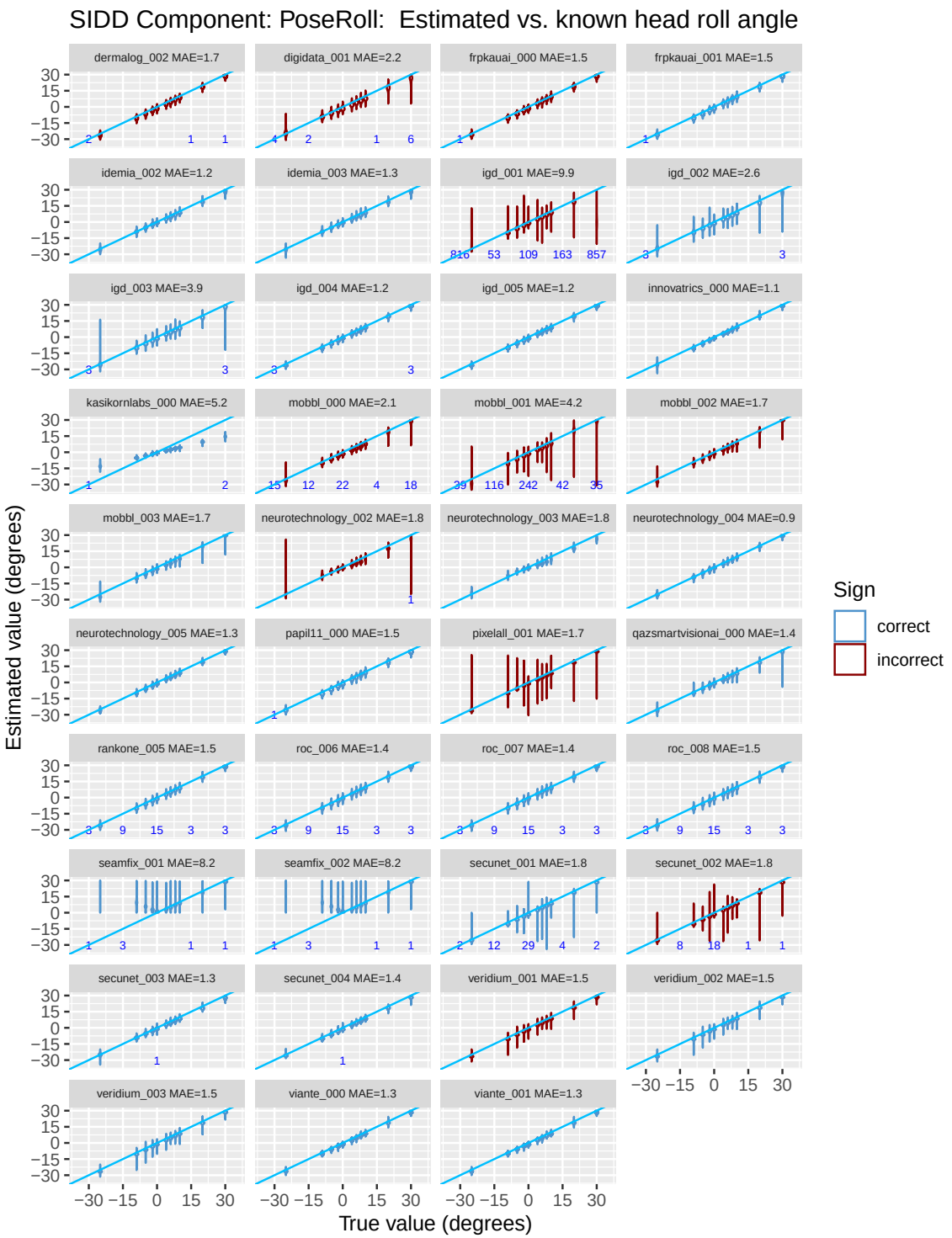


Fig. 11. Estimated vs. known values of roll angle. Ground truth roll values were determined by rotating mugshot images by a known angle. The blue line ($y = x$) represents perfect performance. The plot shows violins at true roll values with the tails extending to the minimum and maximum estimated values. The small blue numbers represent the count of images for which the software did not detect a face; for such images, the error is set to 45 degrees. The dark red color-coding indicates that the developer uses the opposite sign convention, and should negate their scores in the next submission.

3.8. Eyes Open

3.8.1. Images Used

The images for the Eyes Open test are mugshot images. We calculate the EyesOpen measure by comparing the left and right maximum apertures of the eyes as shown in Fig. 12, taking the minimum of the two values, and dividing the result by the inter-eye distance. This procedure correctly assigns a ground truth value of zero for eyes that are closed.

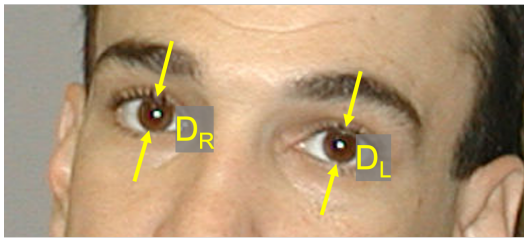


Fig. 12. The EyesOpen measure is computed by comparing the left and right maximum apertures of the eyes, taking the minimum of the two values, and dividing the result by inter-eye distance. Image from NIST Special Database 32, MEDS.

3.8.2. Results for Eyes Open

Table 7 and Figure 13 summarize algorithm performance. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 7. SIDD EyesOpen Mean Absolute Error.

Algorithm	MAE (dimensionless)
secunet_003	0.02
idemia_002	0.03
neurotechnology_002	0.03
neurotechnology_003	0.03
secunet_002	0.03
secunet_001	0.03
digidata_001	0.03
rankone_005	0.15
dermalog_002	0.21
frpkauai_000	0.74

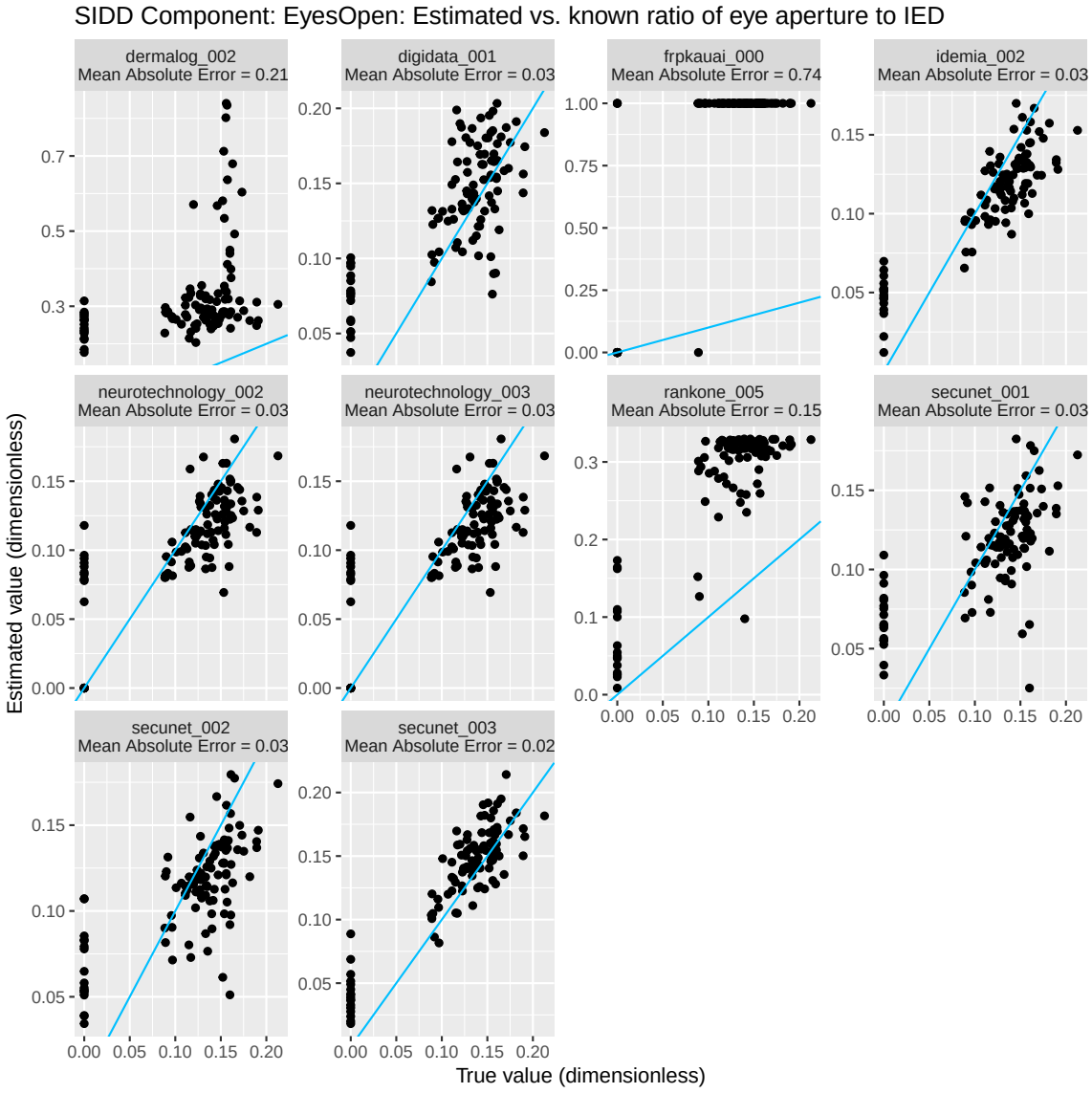


Fig. 13. Estimated vs. known values of the ratio of eye aperture to inter-eye distance. Ground truth preparation is discussed in Section 3.8.1. The blue line ($y = x$) represents perfect performance. The vertical line of dots at true value zero corresponds to closed eyes. Note that the plots have different y-axis ranges.

3.9. Eyes Open 2

3.9.1. Images Used

The images for the Eyes Open 2 measure are mugshot images. We calculate the EyesOpen2 measure by comparing the left and right maximum apertures of the eyes as shown in Fig. 14, taking the minimum of the two values, and dividing the result by the distance from the chin to the midpoint of the eye-centers. This procedure assigns a ground truth value of zero for eyes that are closed and is consistent with the ISO/IEC 29794-5 standard.

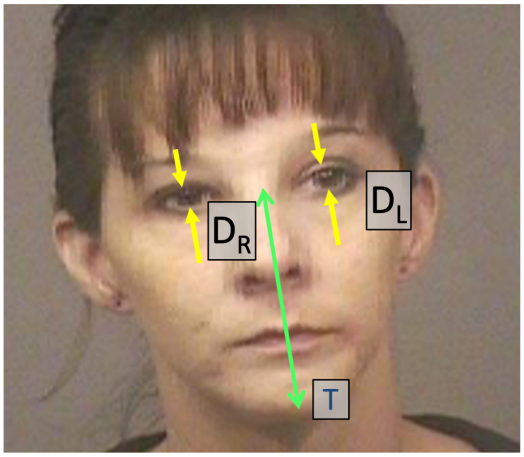


Fig. 14. The EyesOpen2 measure is computed by comparing the left and right maximum apertures of the eyes, taking the minimum of the two values, and dividing the result by the T-metric, the distance from the chin to the midpoint of the eye-centers. The eye apertures and the T-metric are computed perpendicular to the distance between the eyes. Image from NIST Special Database 32, MEDS.

3.9.2. Results for Eyes Open 2

Table 8 and Figure 15 summarizes algorithm performance. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 8. SIDD EyesOpen2 Mean Absolute Error.

Algorithm	MAE (dimensionless)
viante_001	0.01
viante_000	0.01
secunet_003	0.01
igd_004	0.01
igd_005	0.01
neurotechnology_005	0.01
cu-face_001	0.01

Table 8. SIDD EyesOpen2 Mean Absolute Error. *(continued)*

Algorithm	MAE (dimensionless)
mobbl_001	0.01
veridium_003	0.02
frpkauai_001	0.02
igd_003	0.02
neurotechnology_004	0.02
mobbl_002	0.02
igd_002	0.02
mobbl_003	0.02
kasikornlabs_000	0.02
veridium_001	0.03
veridium_002	0.03
qazsmartvisionai_000	0.03
pixelall_001	0.04
pixelall_002	0.04
roc_008	0.06
roc_006	0.06
roc_007	0.06
mobbl_000	0.84

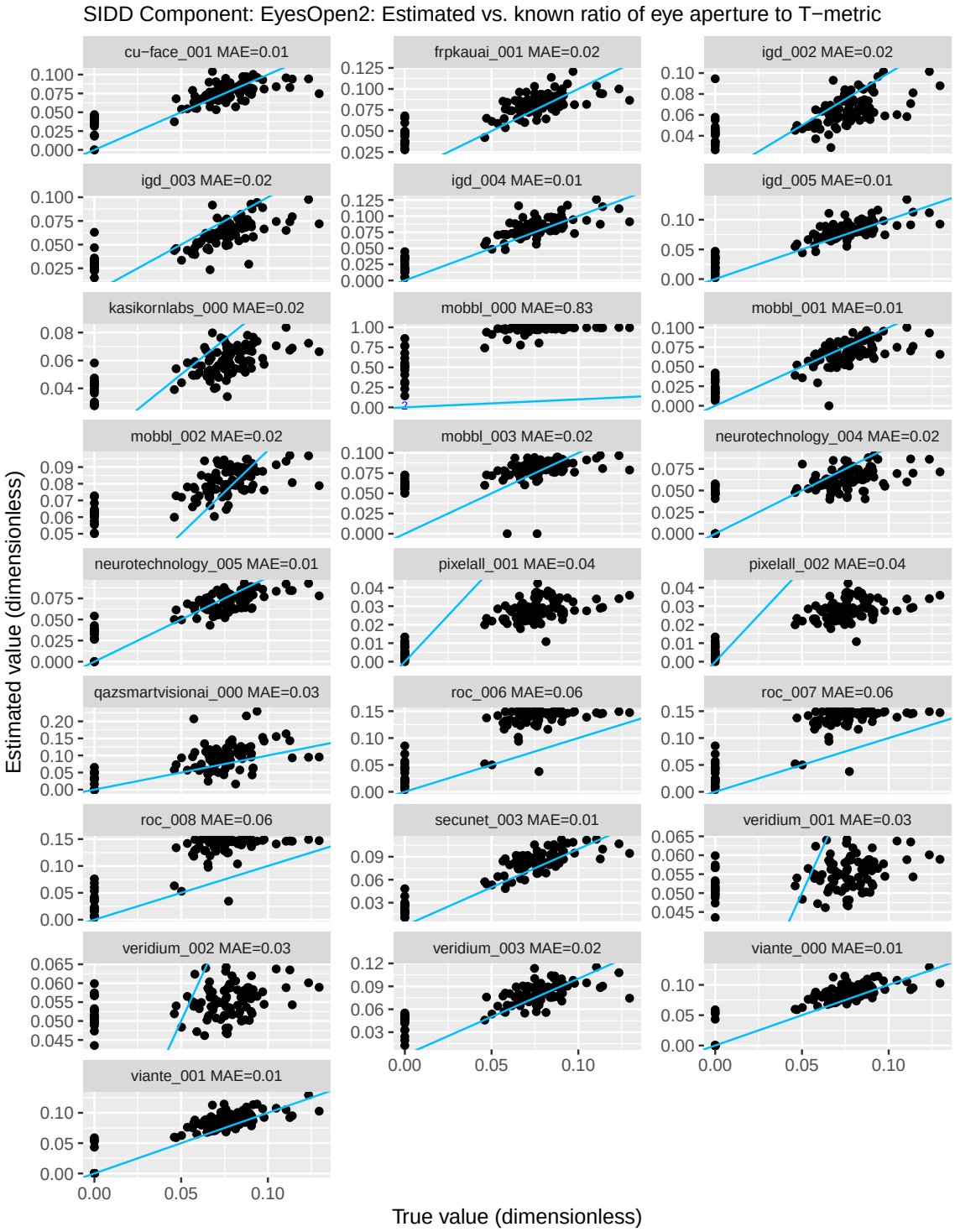


Fig. 15. Estimated vs. known values of the ratio of eye aperture to the distance T between the chin and midpoint of the eye-centers. Ground truth preparation is discussed in Section 3.9.1. The blue line ($y = x$) represents perfect performance. The vertical line of dots at true value zero corresponds to closed eyes. The small blue numbers along the x-axis represent instances when the algorithm did not detect a face; for such images, the error was set to 0.1.

3.10. Inter-Eye Distance

3.10.1. Images Used

The images for the Inter-Eye Distance test are from three sets: two sets of frontal images, and one set with varying yaw angles.

1. In the first set, image sizes range from 310 to 1000 pixels in width, and 240 to 1330 pixels in height. Subjects are generally frontal.
2. In the second set, image sizes range from 720 to 5200 pixels in width, and 1080 to 3500 pixels in height. Subjects are generally frontal.
3. In the third set, there are mated sets of images in which the same person rotates to take on different yaw angles while maintaining the same distance from the camera. The images in this set are grayscale.

In order to determine ground truth for the first and second set, we manually find the eye-centers by determining the two points where eyelids meet for each eye and averaging the two points. The distance between the two eye-centers is used as the ground truth inter-eye distance, as shown in Fig. 16. Note that subjects may have properties that make detection of eyelid corners challenging. Examples include drooping eyelids, makeup on the eyes, and long eyelashes. As a result, there may be times when ground truth is imperfect.

Fig. 16. The 2-dimensional inter-eye distance is calculated by averaging the canthi for each eye and taking the distance of the two resulting points. Image from NIST Special Database 32, MEDS.

For the third set, the ground truth for yaw is determined at the time of capture. This set includes images with yaw ranging from -60 to 60 degrees. The reported inter-eye distance should be the same with yaw and without yaw; one possibility for determining the implied 3-dimensional inter-eye distance is dividing the 2-dimensional inter-eye distance by the cosine of yaw. This is shown in Fig. 17.

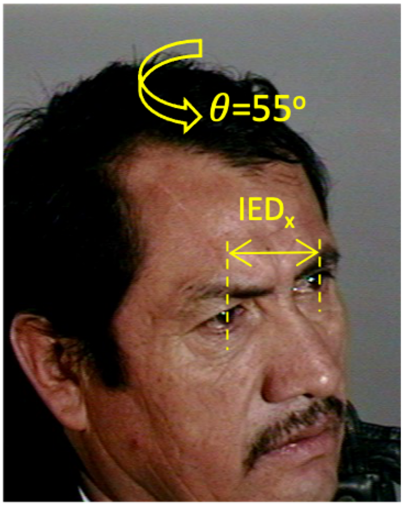


Fig. 17. The 3-dimensional inter-eye distance may be determined by calculating the 2-dimensional inter-eye distance and dividing by the cosine of yaw; $IED_{3D} = IED_x / cos(\theta_{Yaw})$. Image from NIST Special Database 32, MEDS.

3.10.2. Results for Inter-Eye Distance

Table 9, Figures 18-21, and Figure 22 summarize algorithm performance. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 9. SIDD InterEyeDistance Mean Absolute Error

Algorithm	IED Set 1	IED Set 2
viante_001	2.8	13.4
viante_000	2.8	13.6
secunet_004	3.0	9.8
mobbl_002	3.0	13.2
mobbl_003	3.0	13.2
secunet_003	3.0	10.0
igd_005	3.0	16.7
neurotechnology_005	3.1	13.2
kasikornlabs_000	3.2	10.3
igd_004	3.2	17.6
qazsmartvisionai_000	3.3	11.0
rankone_005	3.4	13.4
dermalog_002	3.6	9.7
veridium_001	3.6	14.4
veridium_002	3.6	14.4
veridium_003	3.6	14.4
roc_006	3.6	17.5
roc_007	3.6	17.5
pixelall_001	3.7	17.3
pixelall_002	3.7	17.3
daon_000	3.8	18.0
secunet_001	3.9	25.0
mobbl_000	3.9	9.5
idemia_003	3.9	8.0
igd_001	3.9	31.1
igd_003	4.0	16.1
secunet_002	4.1	23.4
igd_002	4.3	15.4
roc_008	4.5	20.8
mobbl_001	4.6	33.4
innovatrics_000	5.0	19.8
frpkauai_000	5.0	20.8
frpkauai_001	5.1	20.9
papil11_000	5.3	21.8
idemia_002	5.7	21.3
neurotechnology_004	6.1	19.1

Table 9. SIDD InterEyeDistance Mean Absolute Error *(continued)*

Algorithm	IED Set 1	IED Set 2
neurotechnology_002	7.2	12.9
neurotechnology_003	7.2	12.9
digidata_001	10.8	28.6

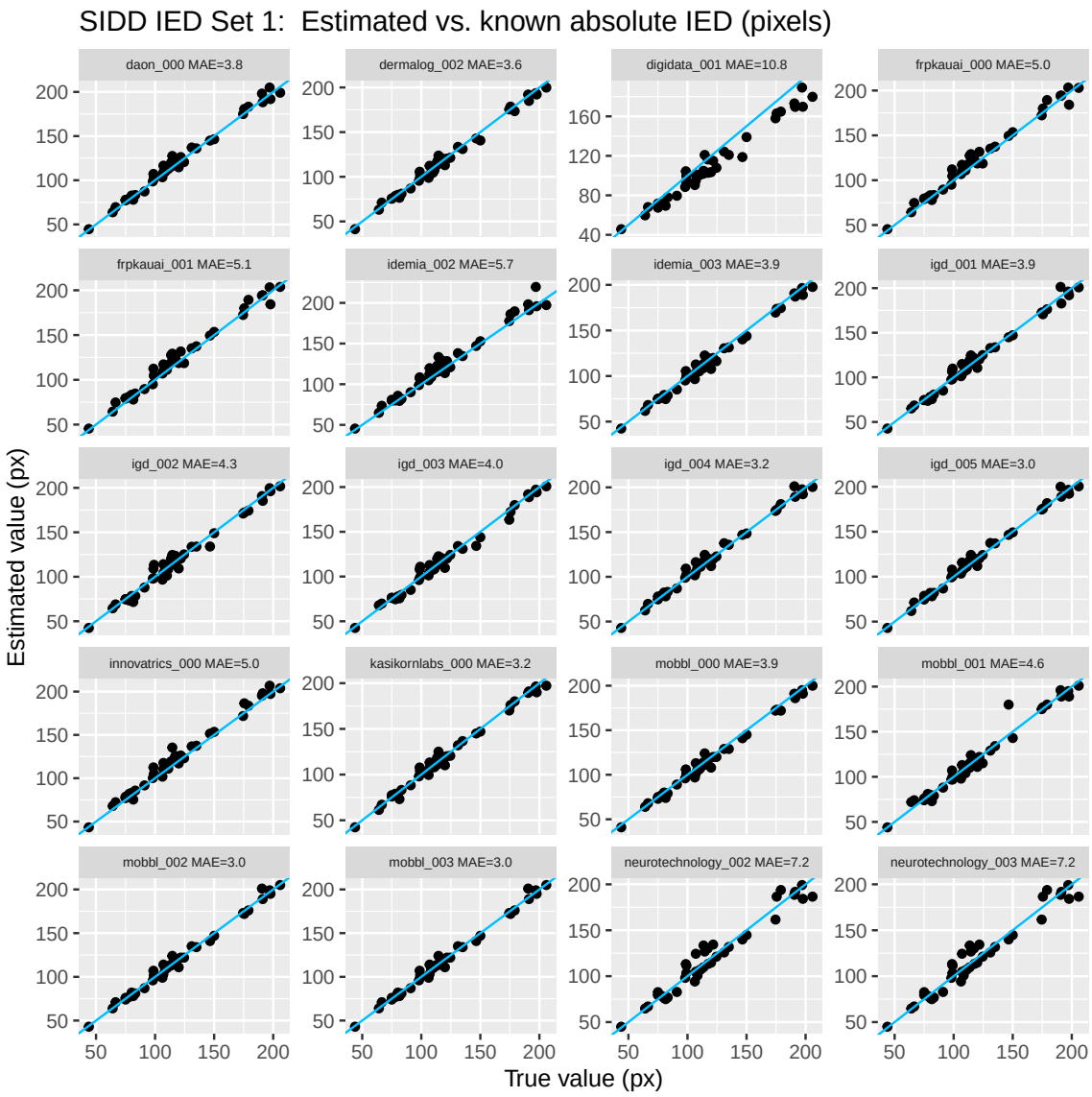


Fig. 18. Estimated vs. known values of inter-eye distance for Set 1.

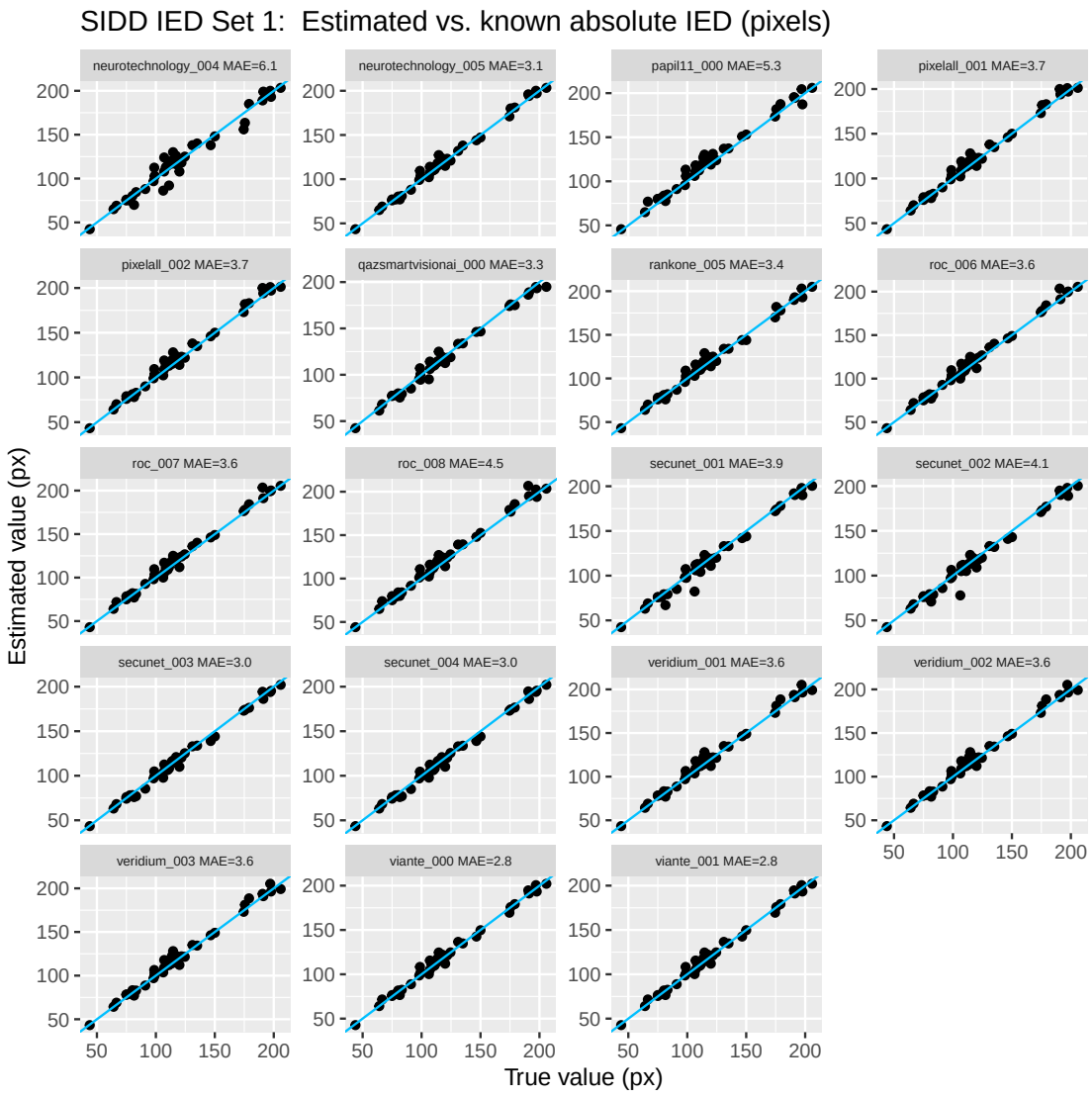


Fig. 19. Estimated vs. known values of inter-eye distance for Set 1.

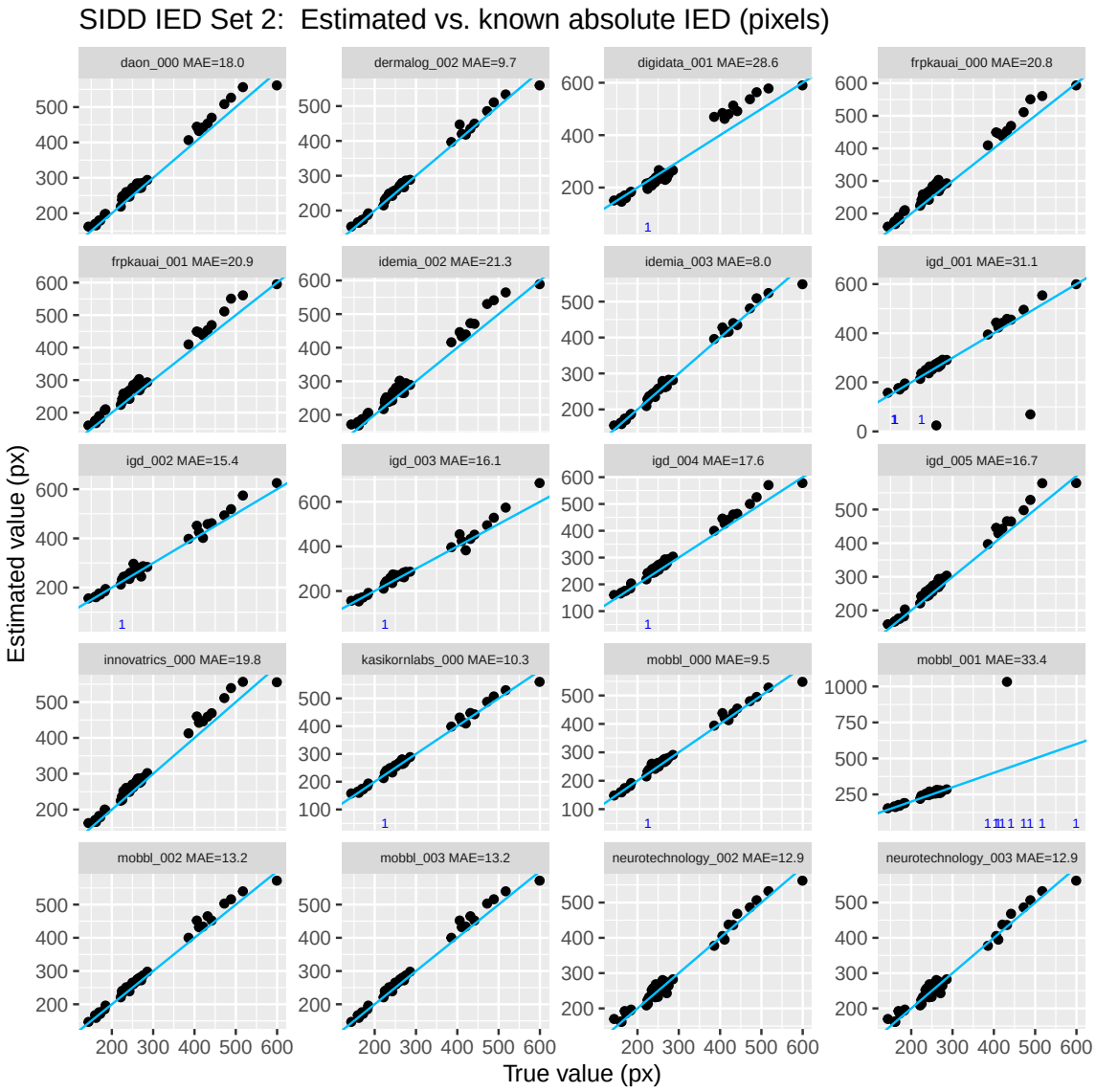


Fig. 20. Estimated vs. known values of inter-eye distance for Set 2.

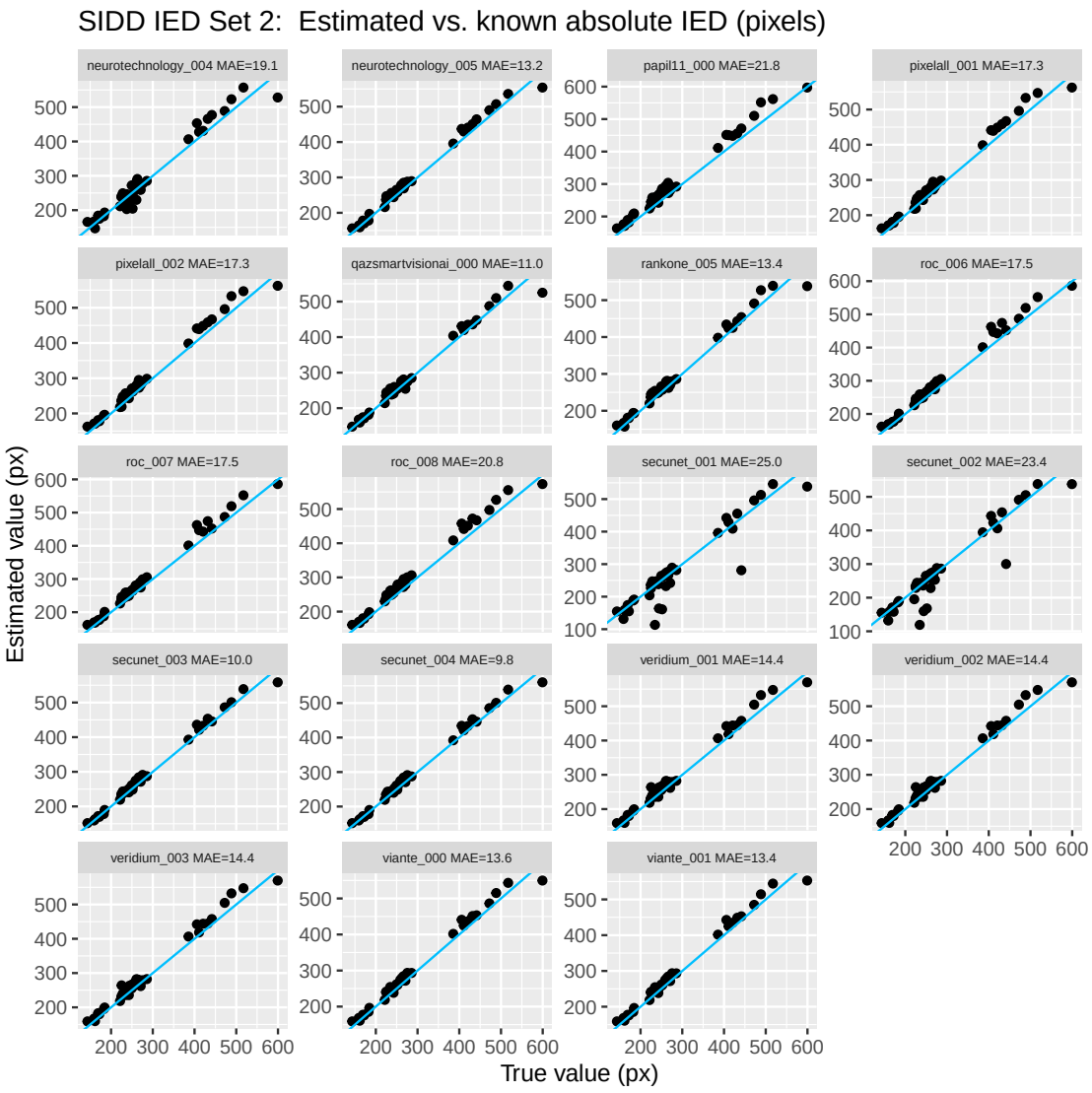


Fig. 21. Estimated vs. known values of inter-eye distance for Set 2.

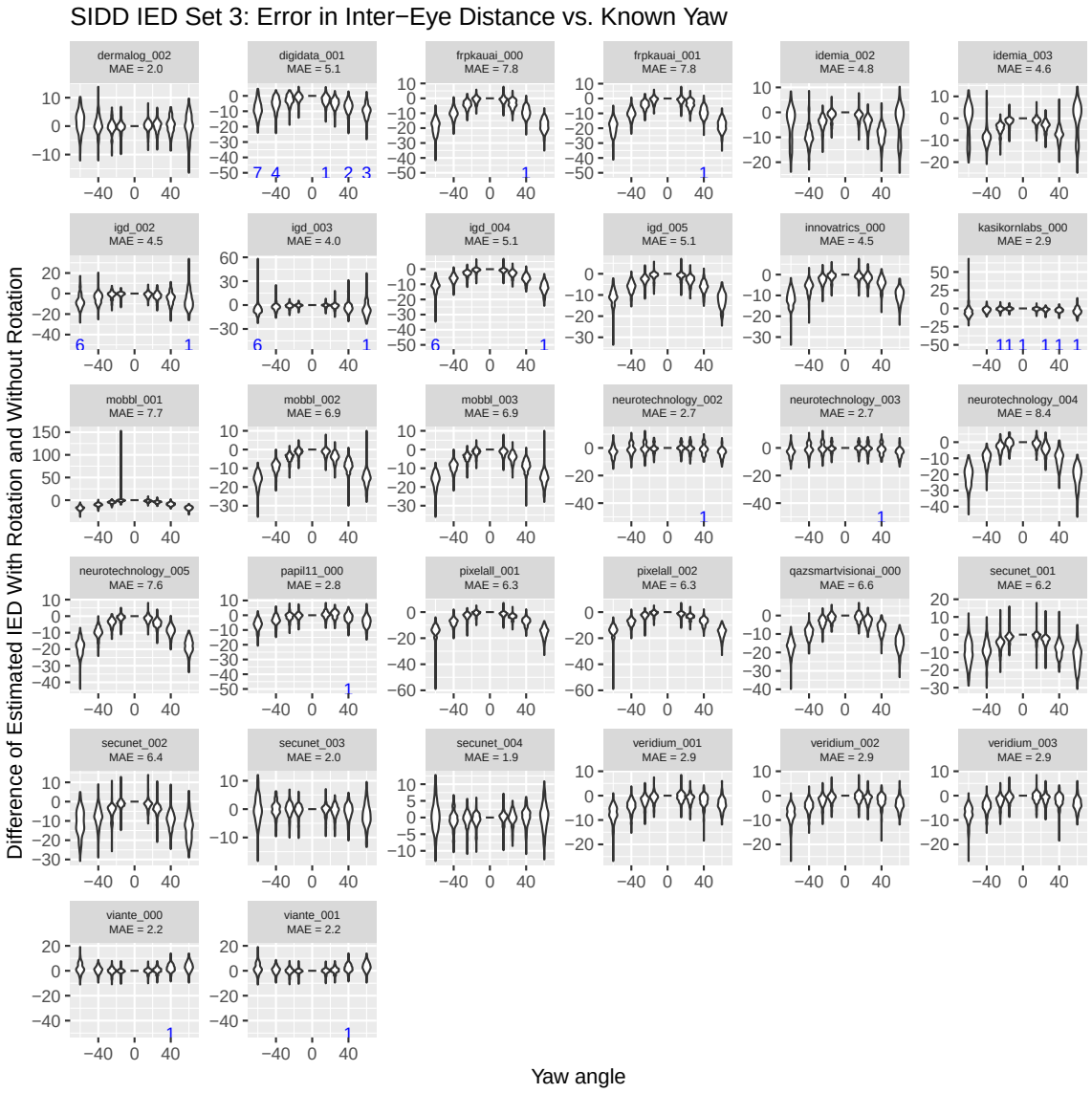


Fig. 22. Difference in reported inter-eye distance (IED) with yaw and without yaw. Perfect performance corresponds to violins centered at $y = 0$ with height 0 across all yaw values. The small blue numbers along the x-axis represent instances when the algorithm did not detect a face; for such images, the error was set to 20. The Mean Absolute Error (MAE), where error is computed as the difference in IED with yaw and IED without yaw for a given capture session, is shown for each algorithm. Lower MAE is better.

3.11. Mouth Open

3.11.1. Images Used

We use mugshot images for the Mouth Open 2 measure. The maximum distance from the bottom of the upper lip to the top of the lower lip is measured, then divided by the inter-eye distance to determine ground truth, as shown in Figure 23. This procedure assigns a ground truth value of zero for mouths that are closed.

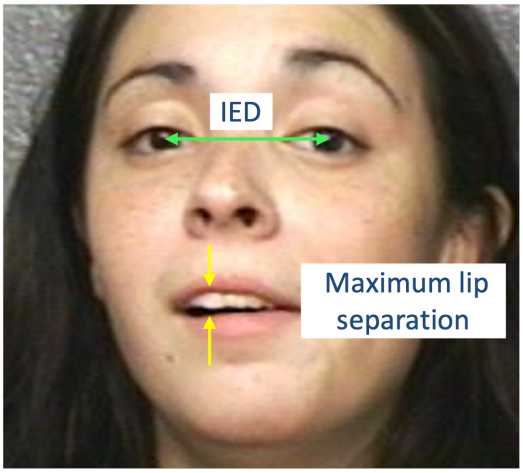


Fig. 23. The MouthOpen measure is computed by taking the maximum distance from the bottom of the upper lip to the top of the lower lip, and dividing the result by inter-eye distance. Image from NIST Special Database 32, MEDS.

3.11.2. Results for Mouth Open

Table 10 and Figure 24 summarize algorithm performance. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 10. SIDD MouthOpen Mean Absolute Error.

Algorithm	MAE (dimensionless)
neurotechnology_002	0.01
neurotechnology_003	0.01
idemia_002	0.02
secunet_003	0.02
rankone_005	0.03
seamfix_001	0.04
secunet_002	0.04
secunet_001	0.04
digidata_001	0.05
dermalog_002	0.30

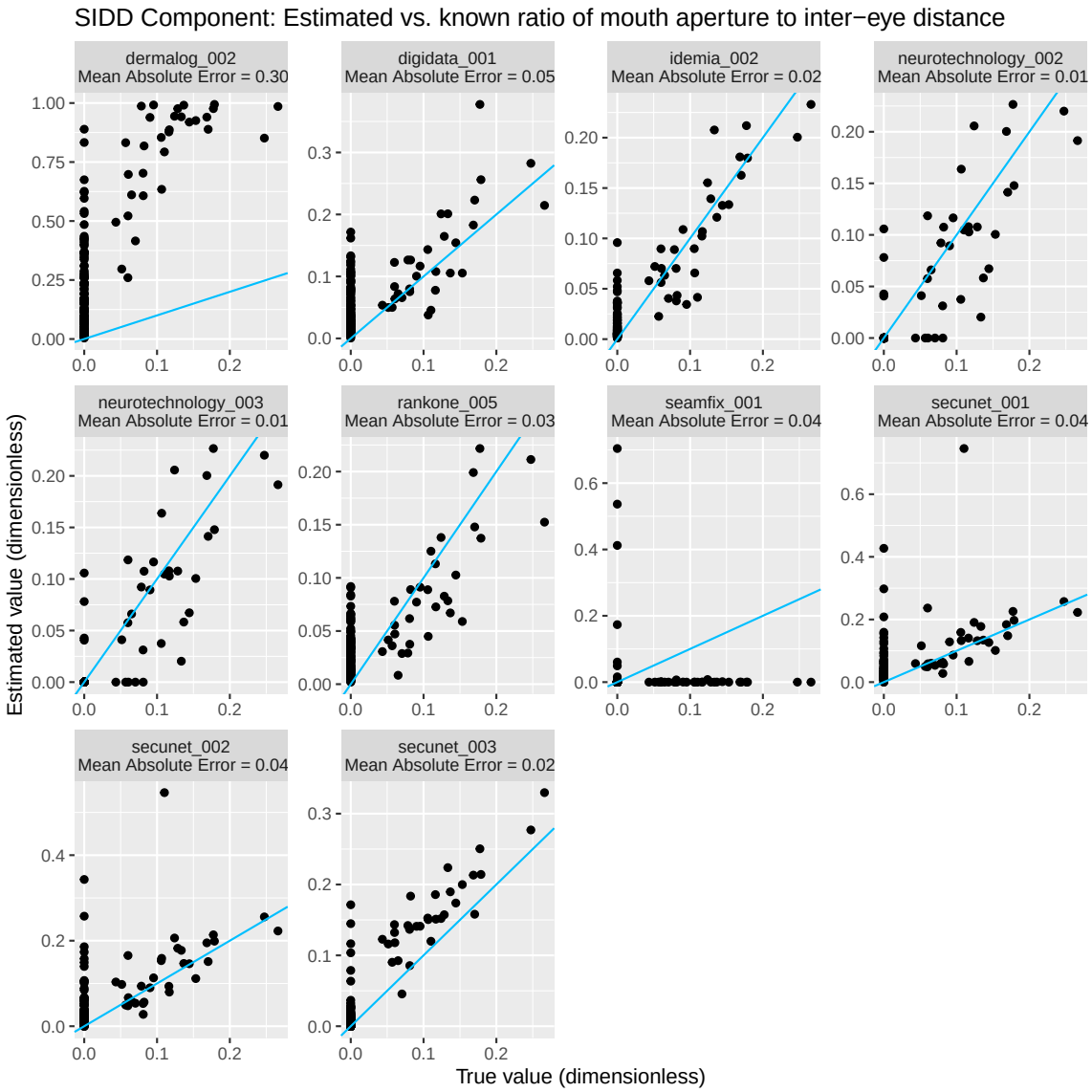


Fig. 24. Estimated vs. known values of the ratio of mouth aperture to inter-eye distance. Ground truth preparation is discussed in Section 3.11.1. The blue line ($y = x$) represents perfect performance. The vertical line of dots at true value zero corresponds to closed mouths.

3.12. Mouth Open 2

3.12.1. Images Used

We use mugshot images for the Mouth Open measure. The maximum distance from the bottom of the upper lip to the top of the lower lip is measured, then divided by the T-metric, the distance from the chin to the midpoint of the eye-centers, as shown in Figure 25. This procedure assigns a ground truth value of zero for mouths that are closed and is consistent with the ISO/IEC 29794-5 standard.

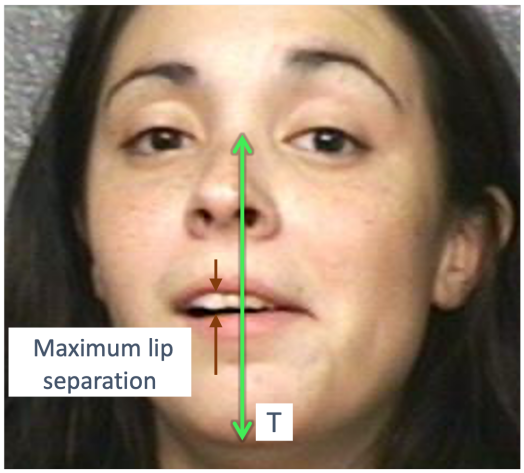


Fig. 25. The MouthOpen2 measure is computed by taking the maximum distance from the bottom of the upper lip to the top of the lower lip, and dividing the result by the T-metric, the distance from the chin to the midpoint of the eye-centers. Image from NIST Special Database 32, MEDS.

3.12.2. Results for Mouth Open 2

Table 11 and Figure 26 summarize algorithm performance. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

Table 11. SIDD MouthOpen2 Mean Absolute Error.

Algorithm	MAE (dimensionless)
neurotechnology_005	0.00
neurotechnology_004	0.01
mobbl_002	0.01
mobbl_003	0.01
igd_003	0.01
secunet_003	0.01
pixelall_001	0.01
pixelall_002	0.01
kasikornlabs_000	0.01
igd_004	0.01
cu-face_001	0.01
igd_005	0.01
viante_000	0.01
viante_001	0.01
roc_008	0.02
mobbl_001	0.02
qazsmartvisionai_000	0.02
igd_002	0.02
veridium_003	0.03
veridium_001	0.03
veridium_002	0.03
mobbl_000	0.08

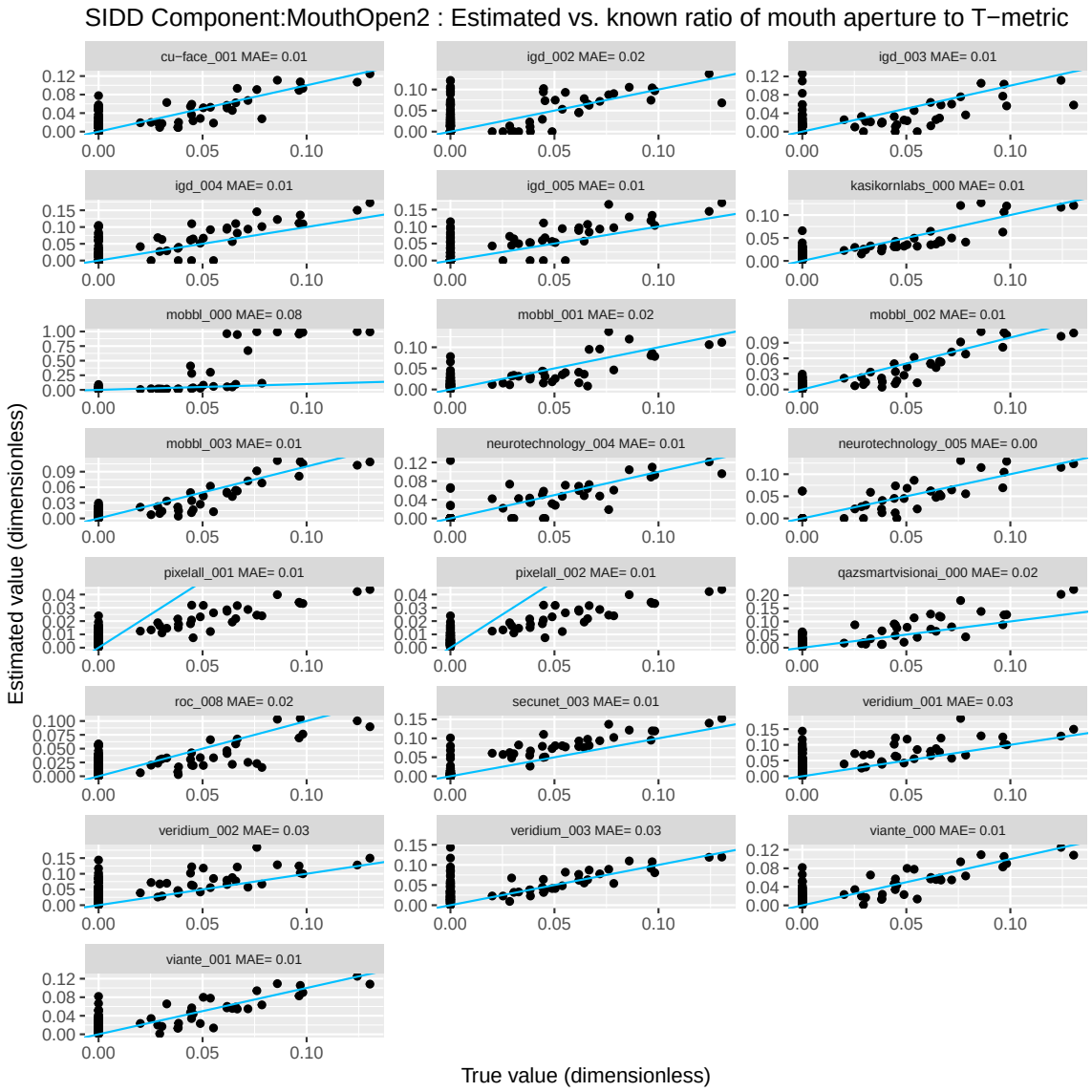


Fig. 26. Estimated vs. known values of the ratio of mouth aperture to T-metric. Ground truth preparation is discussed in Section 3.12.1. The blue line ($y = x$) represents perfect performance. The vertical line of dots at true value zero corresponds to closed mouths.

3.13. Background Uniformity

3.13.1. Images Used

We use mugshot images for the Background Uniformity measure. We categorize images into three categories: Uniform, Attempt at Uniform, and Cluttered.

Uniform images have a plain background, with no brick or shadows behind the subject. Images in the Attempt at Uniform category might have a background with concrete or brick texture. Alternatively, they may have shadows behind the subject, but no other significant non-uniformity. We categorize all other images as Cluttered. The images in the Cluttered category include backgrounds containing furniture, walls with writing behind the subject, and significant variation in background color. Examples are in Table 12.

Table 12. Images in order of increasing background uniformity; the first and third images are used with the permission of the subject. The second image is from NIST Special Database 32, MEDS

Category	Cluttered	Attempt at Uniformity	Uniform
Example			

3.13.2. Results for Background Uniformity

Figure 27 summarizes algorithm performance for background uniformity.

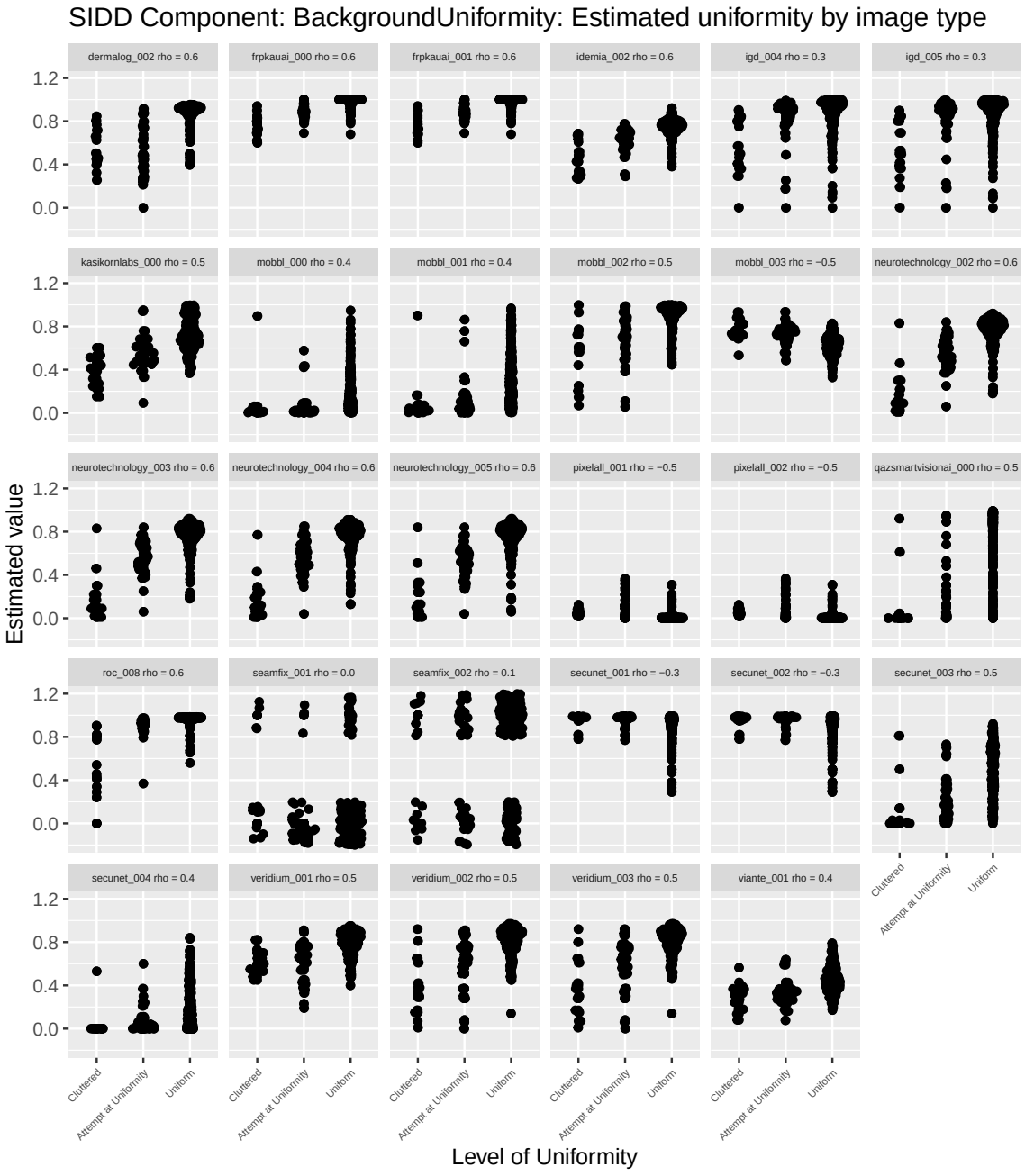


Fig. 27. Estimated degree of background uniformity by category (Cluttered, Attempt at Uniformity, Uniform). Perfect performance corresponds to clusters that shift upward as uniformity increases, and a rank correlation value (rho) of 1.

3.14. Resolution

3.14.1. Images Used

For Resolution, we use two sets of images.

1. Synthetic Blur: For the first resolution set, images are produced by blurring mugshots ranging in size from 128 to 3456 pixels in width and 120 to 2719 pixels in height. We use the *convert* command from the ImageMagick package with the argument *gaussian-blur*, as illustrated in Table 13. This command convolves each pixel in the input image with a Gaussian kernel. Higher values of the σ parameter, the standard deviation of the Gaussian, correspond to lower resolution. The inter-eye distance (IED) in the images ranges from approximately 15 to 600 pixels. We use eight values of σ, ranging from 0 to 7. The highest value of sigma, 7, corresponds to not being able to discern the canthi accurately, but still being able to detect that the eyes are open. Note that the resolution perceived by a reader of this report depends on the handling of the image by LaTeX, the device used to display the image, and other optical factors.

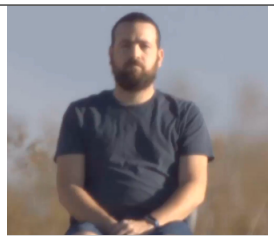
   To ensure that blur is quantified relative to the size of the face, we introduce the relative blur B, defined as σ divided by IED of the unblurred image. The first image in table 13 is unblurred and has IED 188 (before re-sizing) and $B = 0$. The subsequent images have $B = 0.011$ and $B = 0.027$.

2. Natural blur: For the second resolution set, the images are from outdoor video frames with non-uniform lighting. The blur in the image is from turbulence from air movement due to thermal heating of the ground. The turbulence is quantified by the refractive index structure parameter, C_n^2. The images are illustrated in Table 14. Higher values of turbulence correspond to lower resolution.

Table 13. Resolution Set 1 Illustration. Images are used with the permission of the subject.

Standard deviation σ	0	2	5
Result of convert -gaussian-blur 0xσ			
$B = \sigma/IED$	0	0.011	0.027

Table 14. Resolution Set 2 Illustration. The images have two levels of turbulence (low and high). Here, turbulence refers to the distortions in an image caused by movement of air due to thermal heating of the ground.

Turbulence C_n^2 (refractive index structure parameter)	5.2×10^{-16}	8.7×10^{-13}
Example Image		
Amount of blur (log of turbulence)	-15.28	-12.06

3.14.2. Effect of Gaussian Blur on IED Error

Figure 28 shows a violin plot of the IED error at 8 different values of σ, ranging from 0 to 7. Error is measured with respect to developer-reported IED, so that the error at σ=0 corresponds to the developer measurement of IED for an original, unblurred image.

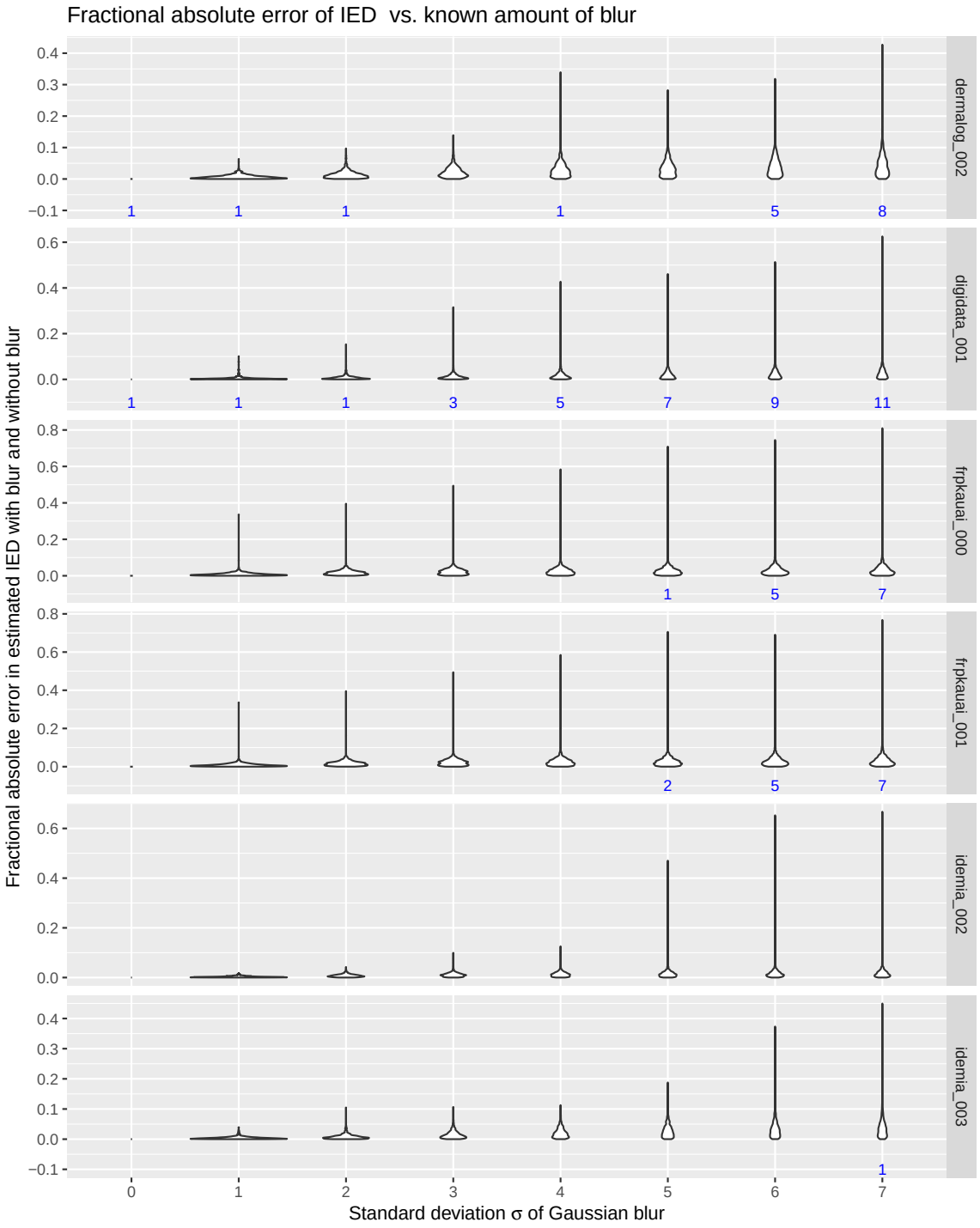


Fig. 28. Change in IED with Gaussian Blur

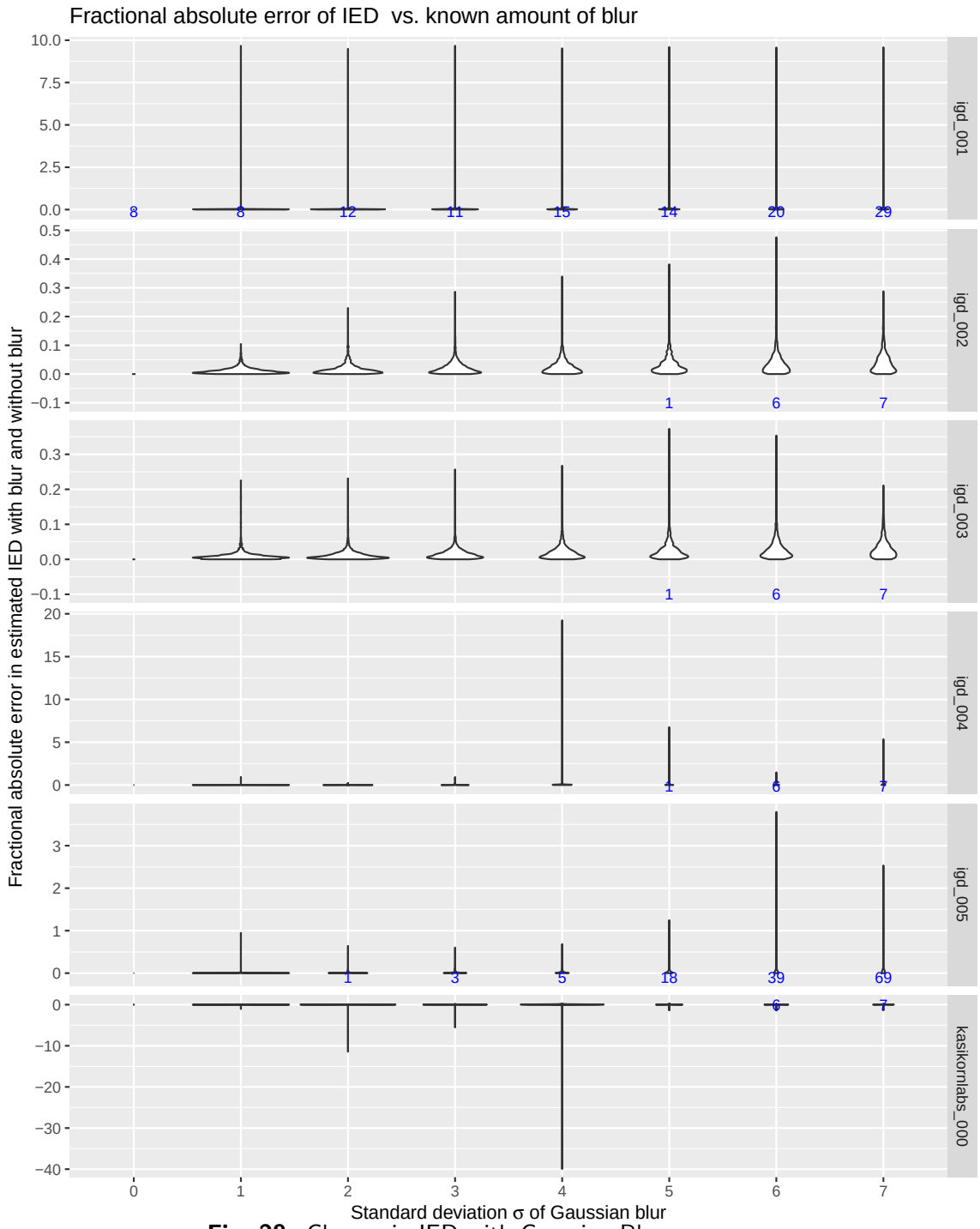


Fig. 29. Change in IED with Gaussian Blur

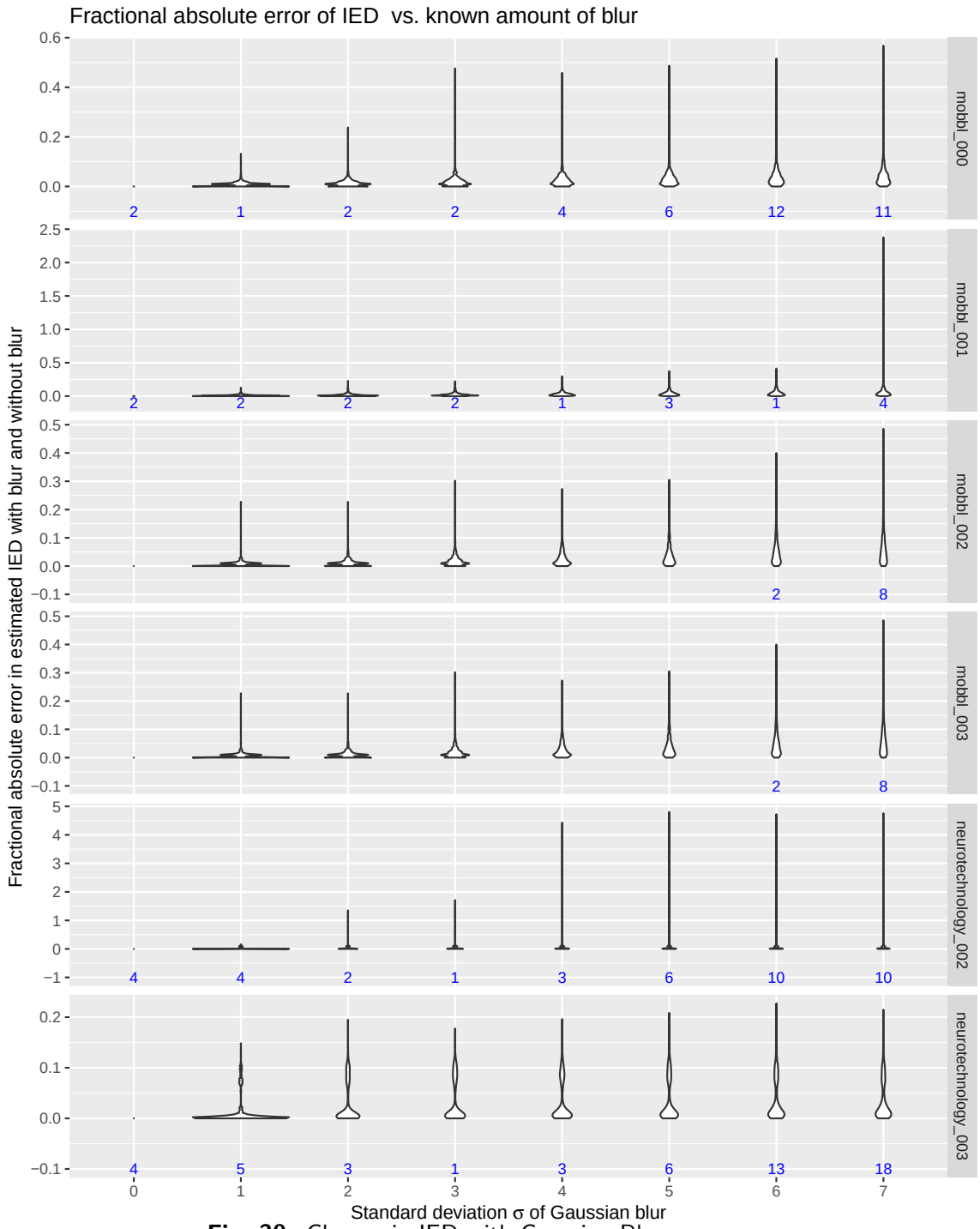


Fig. 30. Change in IED with Gaussian Blur

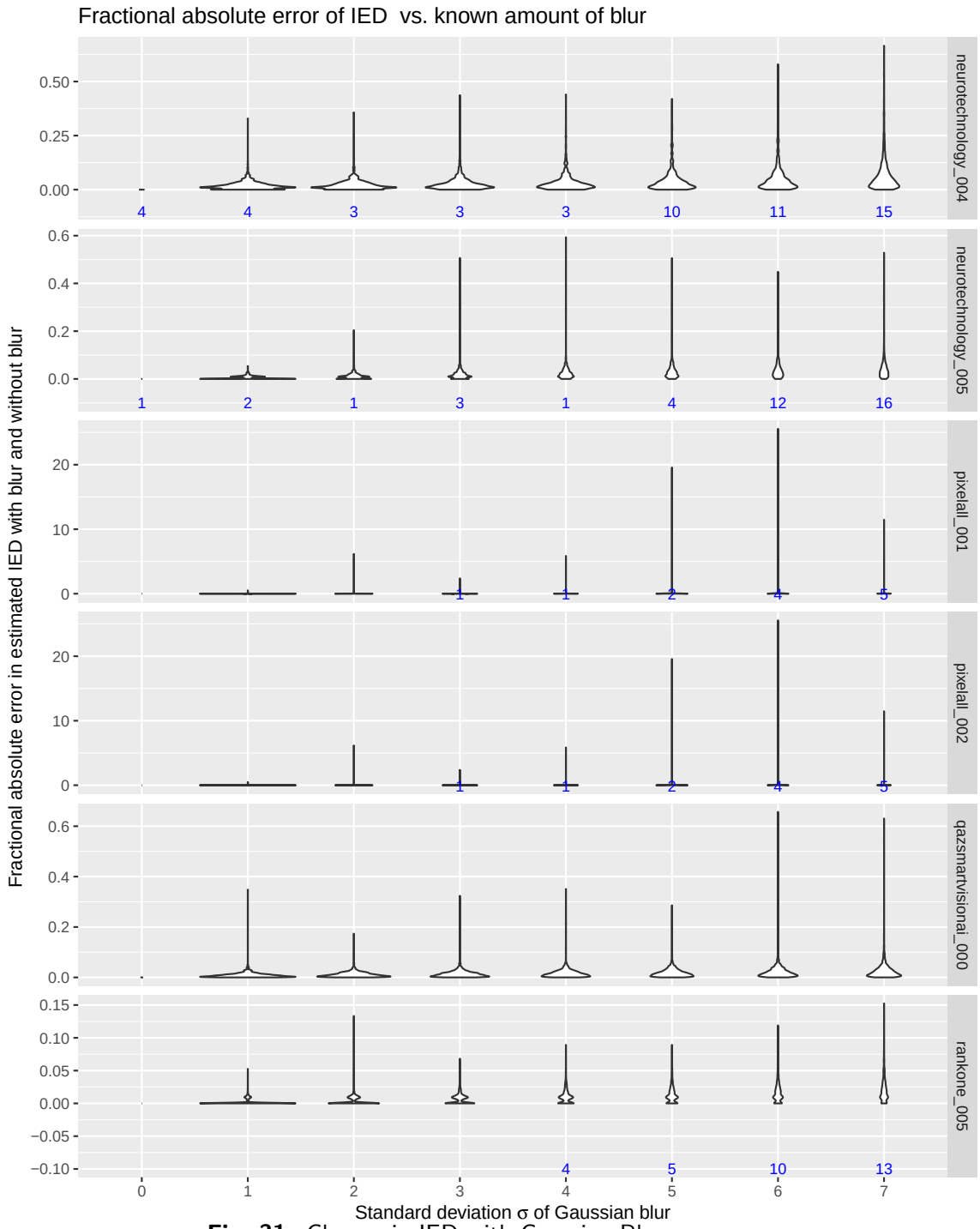


Fig. 31. Change in IED with Gaussian Blur

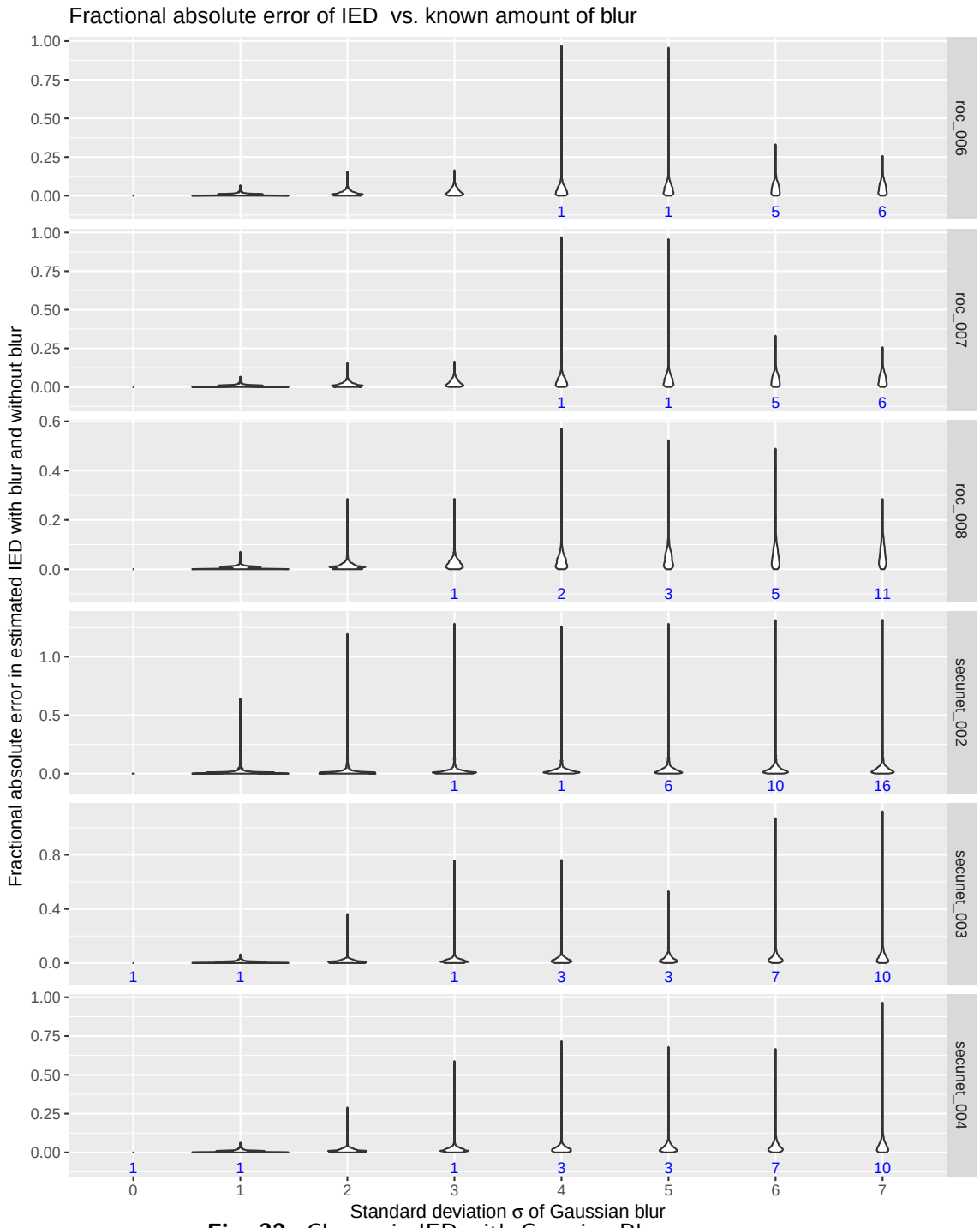


Fig. 32. Change in IED with Gaussian Blur

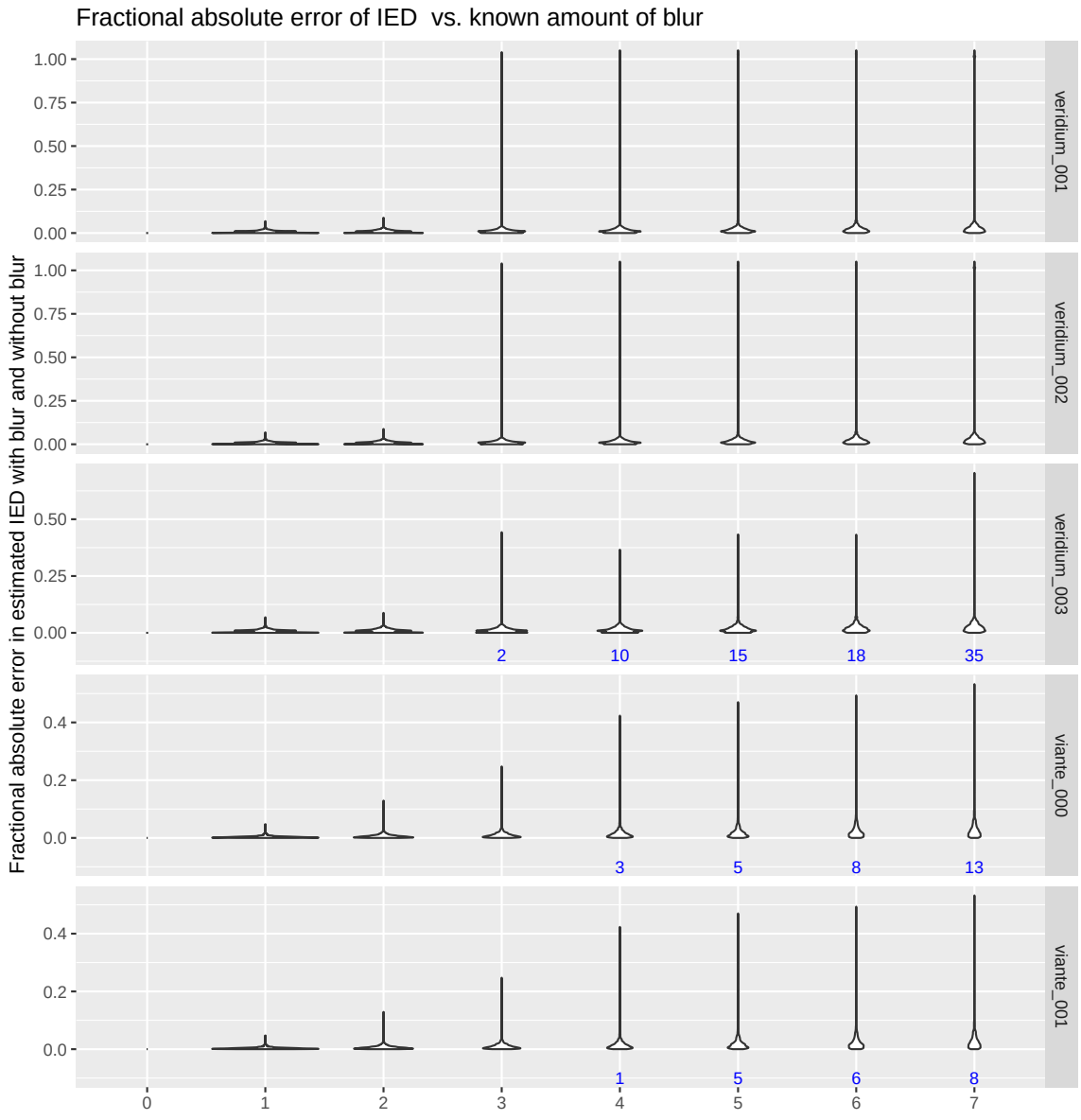


Fig. 33. Change in IED with Gaussian Blur

3.14.3. Results for Resolution

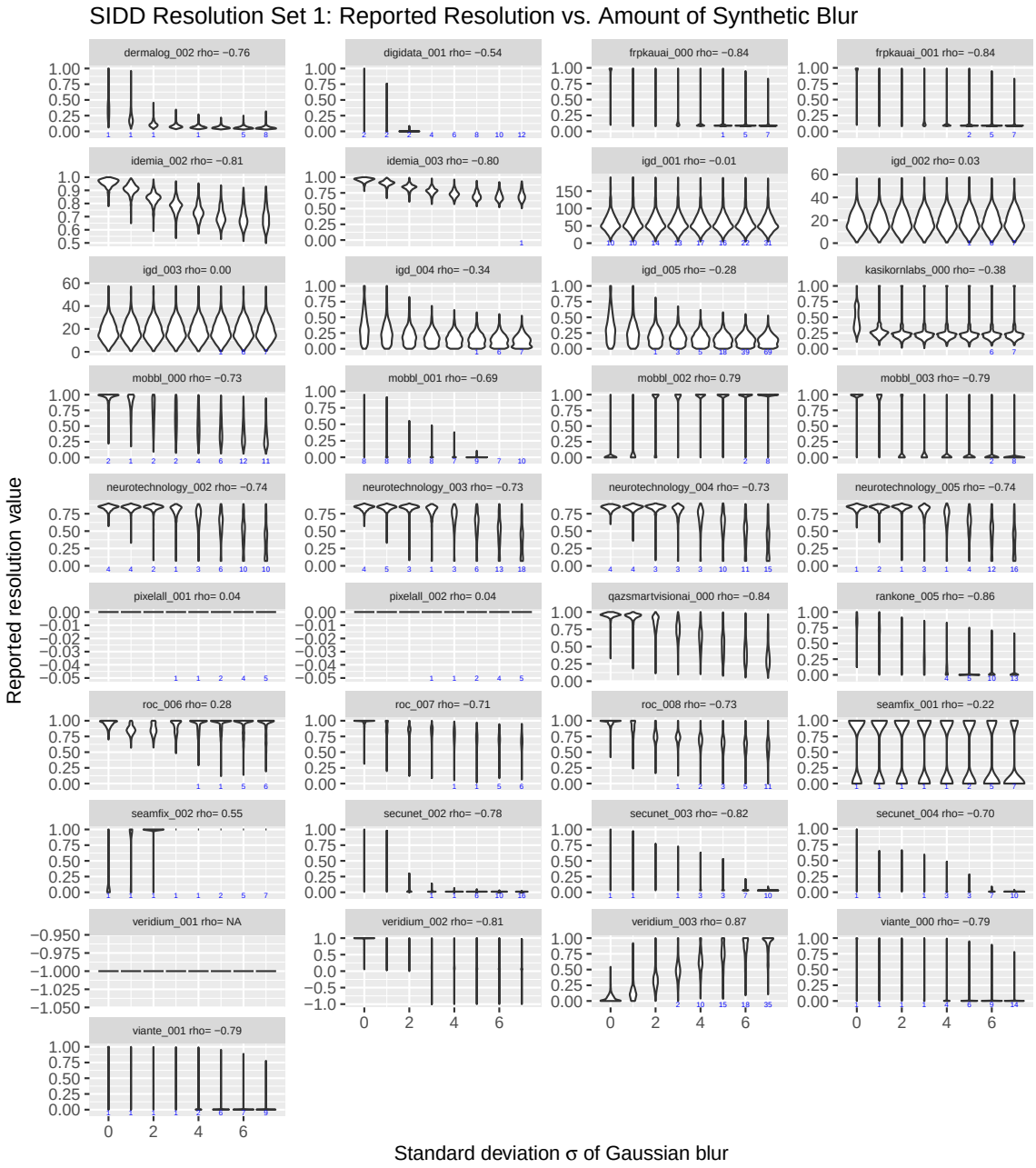


Fig. 34. Distribution of estimated resolution vs. σ parameter of Gaussian blur applied to a set of mugshot images. The higher the σ, the more extreme the blur. The small blue numbers at $y = 0$ represent the count of images for which the software did not return an estimate, for instance, when it did not detect a face. Perfect performance would correspond to monotonically decreasing resolution estimates as σ increases, and a rank correlation value (rho) of -1. Note that the plots have different y-axis ranges.

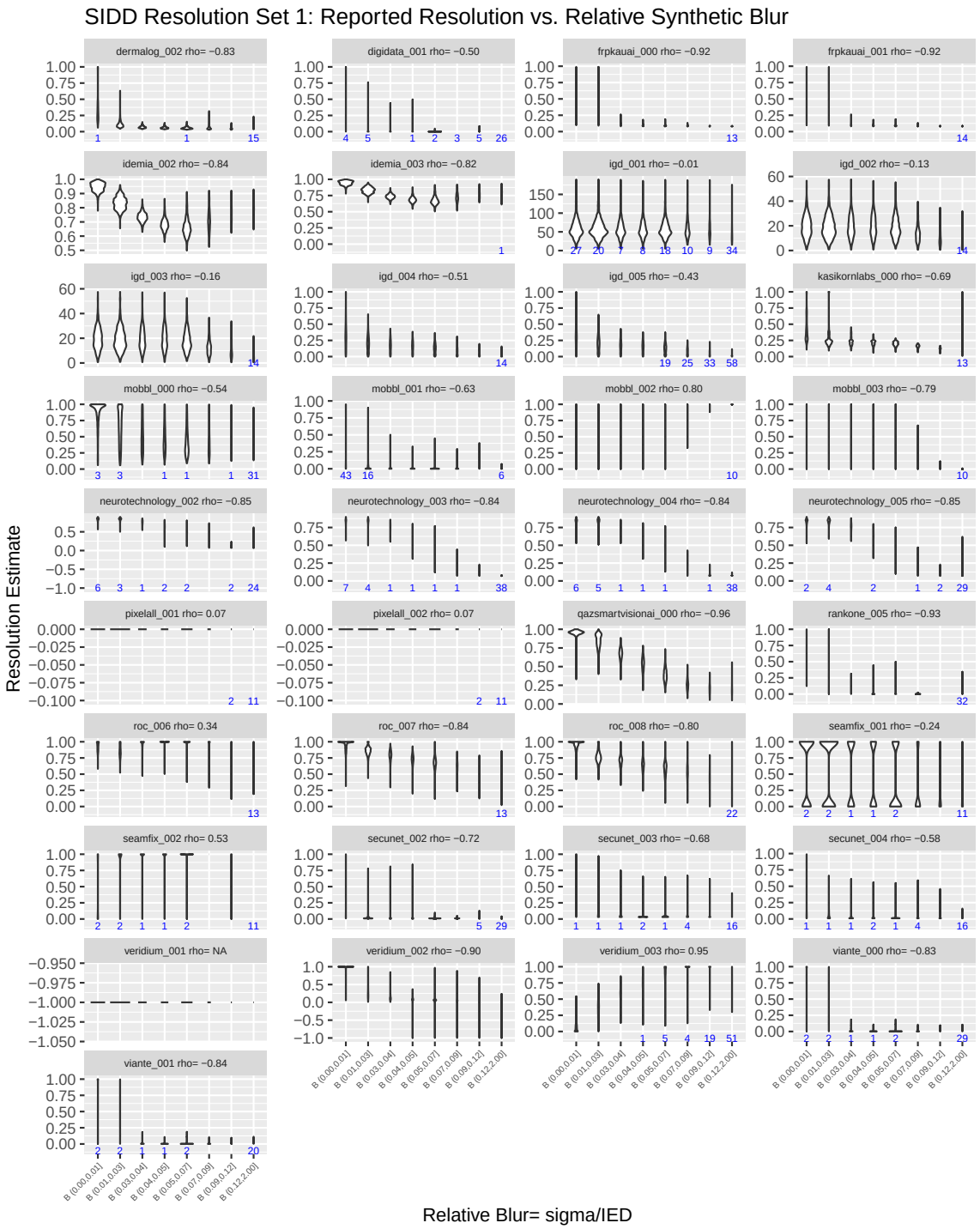


Fig. 35. Distribution of estimated resolution vs. bin values of relative blur $B = \sigma/IED$ applied to a set of mugshot images. The higher the value of B, the more extreme the blur. The small blue numbers below the x-axis represent the count of images for which the software did not return an estimate, for instance, when it did not detect a face. Perfect performance would correspond to monotonically decreasing resolution estimates as B increases, and a rank correlation value (rho) of -1. Note that the plots have different y-axis ranges.

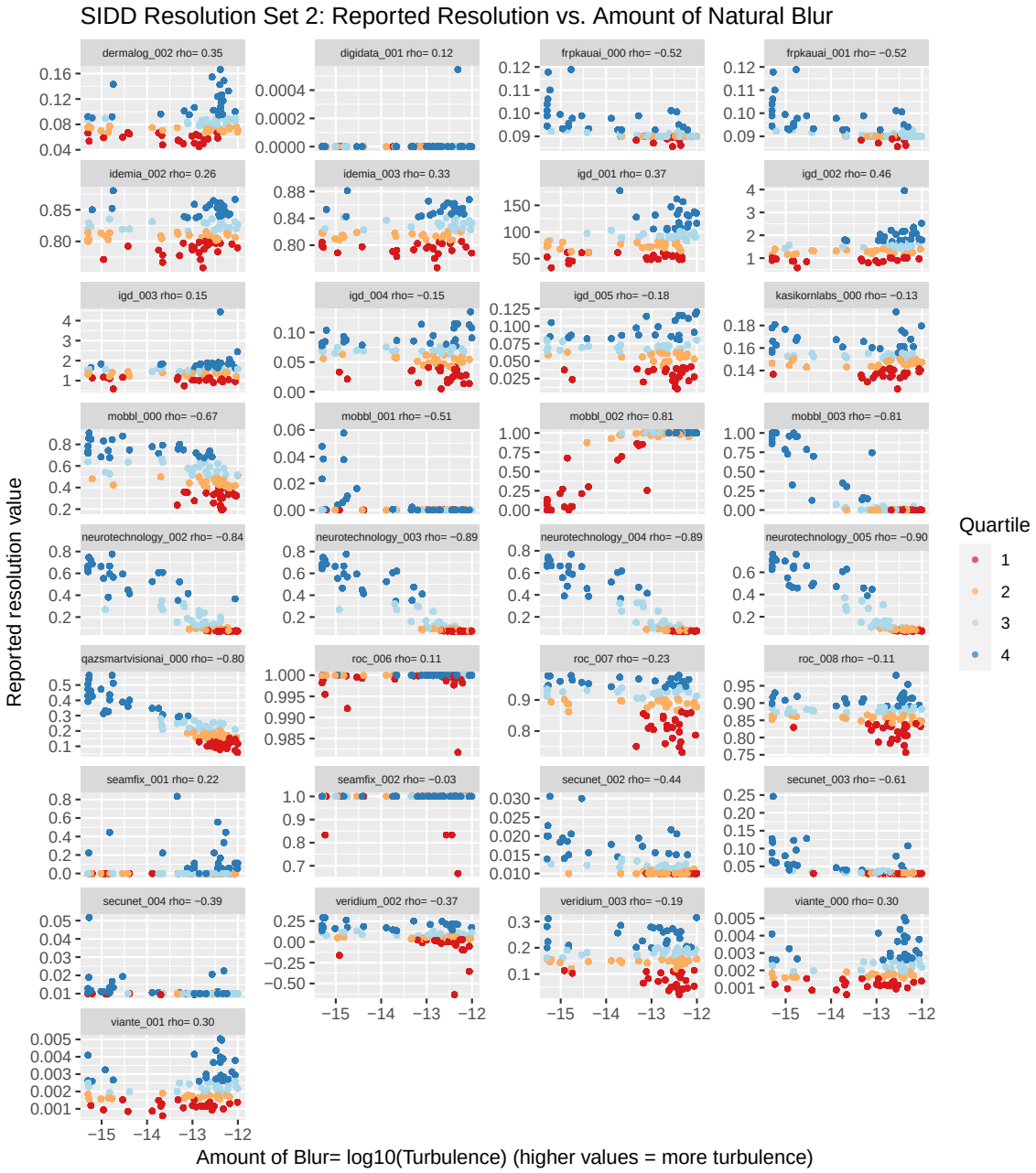


Fig. 36. Distribution of estimated resolution vs. the log of turbulence. The higher the log of turbulence, the higher blur in the image. Perfect performance would correspond to monotonically decreasing resolution estimates as blur increases and a rank correlation value (rho) of -1. The resolution quartiles are color-coded, with blue representing the highest quartile and red representing the lowest. The highest values (blue) of estimated resolution would ideally lie to the left of the lower values (red). Note that the plots have different y-axis ranges.

3.15. Underexposure

3.15.1. Images Used

To generate ground truth for the underexposure measure, we start by using mugshot images. We use the *convert* command from the ImageMagick package with argument *brightness-contrast*, as illustrated in Table 15. We use five values of d ranging from 0 to 32. For this measure, more negative values correspond to more underexposure.

Note that the two parameters for brightness and contrast d_1 and d_2 are both inputs, separated by the symbol x. We use $d_1 = -d_2$ to ensure that the two values are inversely proportional and have equal ranges of values.

Table 15. Underexposure Illustration. Images are used with the permission of the subject.

Brightness and contrast (d_1, d_2)	(0,0)	(-16,16)	(-32,32)
Result of convert -brightness-contrast d_1xd_2			

3.15.2. Results for Underexposure

Fig. 37. Distribution of estimated vs. true underexposure. The x-values represent the contrast and magnitude of decreased brightness of an image. The higher the x value, the more extreme the underexposure. Perfect performance corresponds to clusters shifting upward as x increases, and a rank correlation (rho) value of 1. The small numbers under the x-axis represent failures to detect a face.

3.16. Overexposure

3.16.1. Images Used

We start with images from mugshot sets for the overexposure measure. We then use the *convert* command from the ImageMagick package with the argument *brightness-contrast*, as illustrated in Table 16. We use five values of d ranging from 0 to 40. For this measure, higher brightness corresponds to more overexposure.

Note that the two parameters for brightness and contrast d_1 and d_2 are both inputs, separated by the symbol x. We use $d_1 = d_2$ to ensure that the two values increase linearly with each other and lie on the same range.

Table 16. Overexposure Illustration. Images are used with the permission of the subject.

Brightness and contrast (d_1, d_2)	(0,0)	(20,20)	(40,40)
Result of convert -brightness-contrast d_1xd_2			

3.16.2. Results for Overexposure

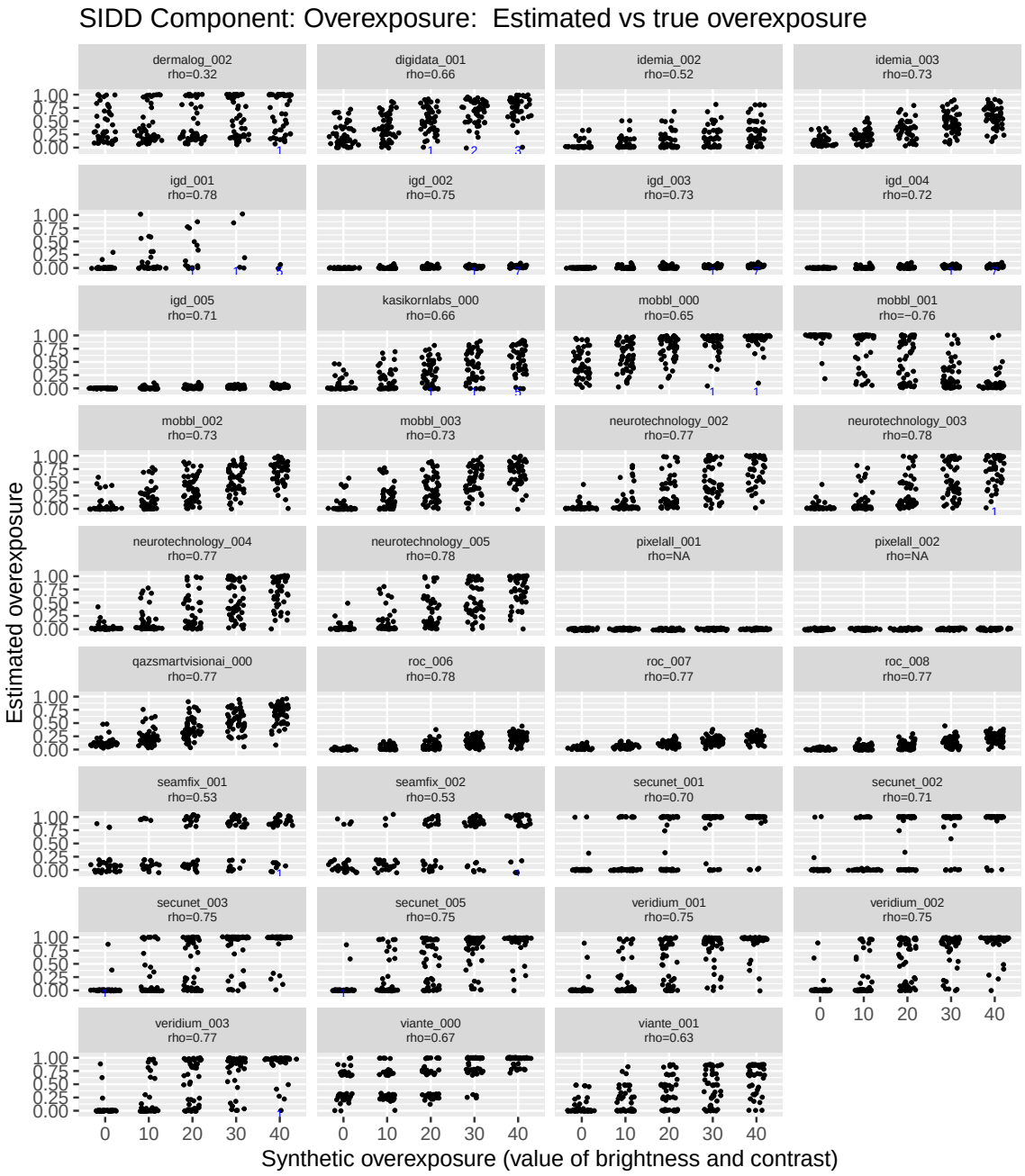


Fig. 38. Distribution of estimated vs. true overexposure. The x-values represent the contrast and magnitude of brightness applied to an image. The higher the x value, the more extreme the overexposure. Perfect performance corresponds to clusters shifting upward as x increases, and a rank correlation (rho) value of 1. The small numbers under the x-axis represent failures to detect a face.

3.17. Eyeglasses Present

3.17.1. Images Used

The images in the Eyeglasses set are mugshot images, in which pose is generally frontal and background is generally uniform. We assign ground truth value of Eyeglasses for images in which the subject is wearing eyeglasses (transparent or sunglasses), and No Eyeglasses otherwise. For our analysis, developers' estimated category is Eyeglasses if the estimated probability of eyeglasses is greater than or equal to a threshold of 0.5, and No Eyeglasses otherwise.

3.17.2. Results for Eyeglasses Present

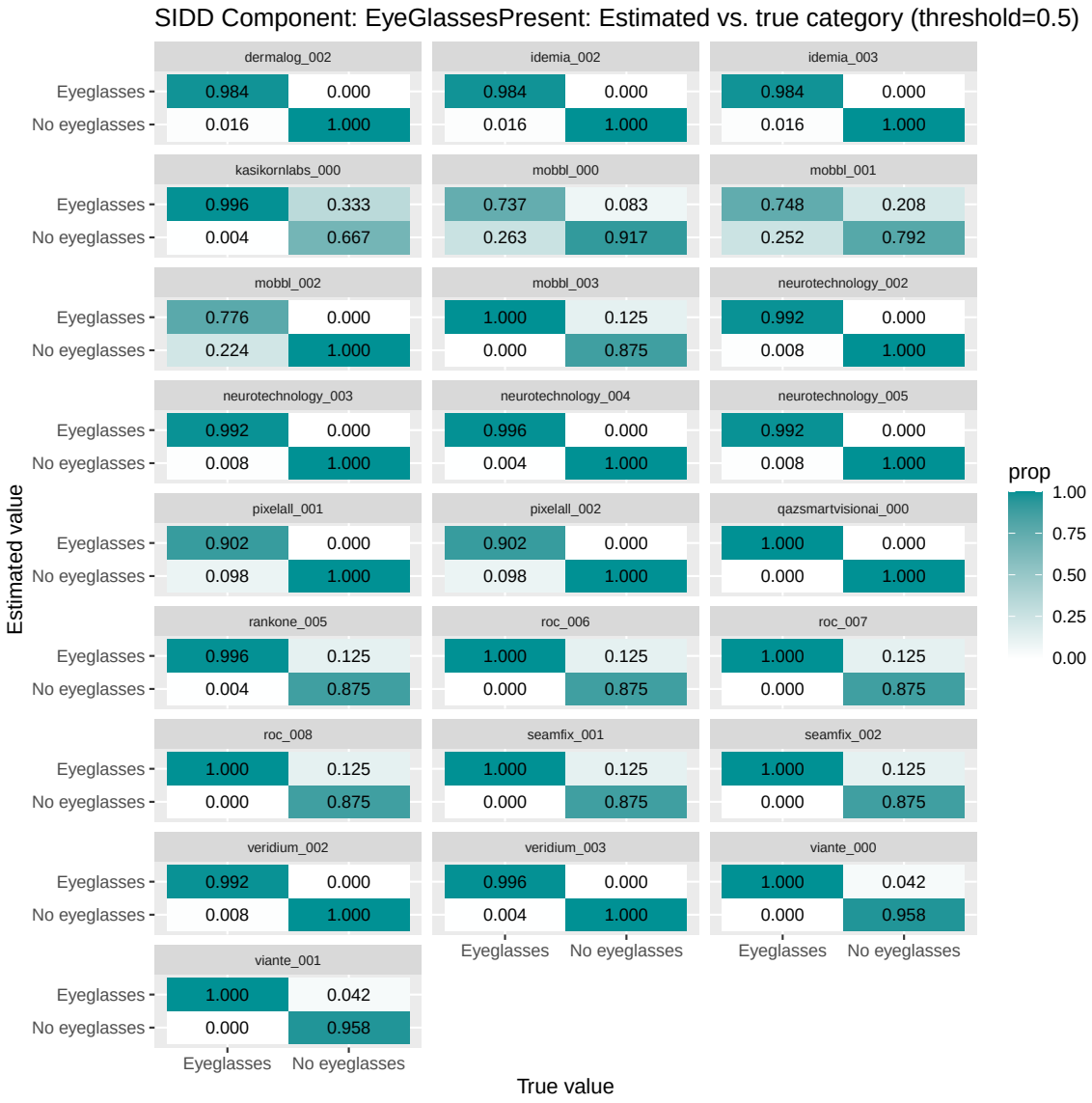


Fig. 39. Estimated vs. known presence of eyeglasses. Perfect performance corresponds to values of 1 at the upper left (for eyeglasses) and the lower right (no eyeglasses), and 0 elsewhere. The values are normalized by the number of images in each category so that each column sums to 1.

3.18. Sunglasses Present

3.18.1. Images Used

The images in the Sunglasses set are images in a natural setting, including non-frontal poses and non-uniform background. We evaluate submissions on images from three categories: opaque, semi-opaque, and transparent.

3.18.2. Results for Sunglasses Present

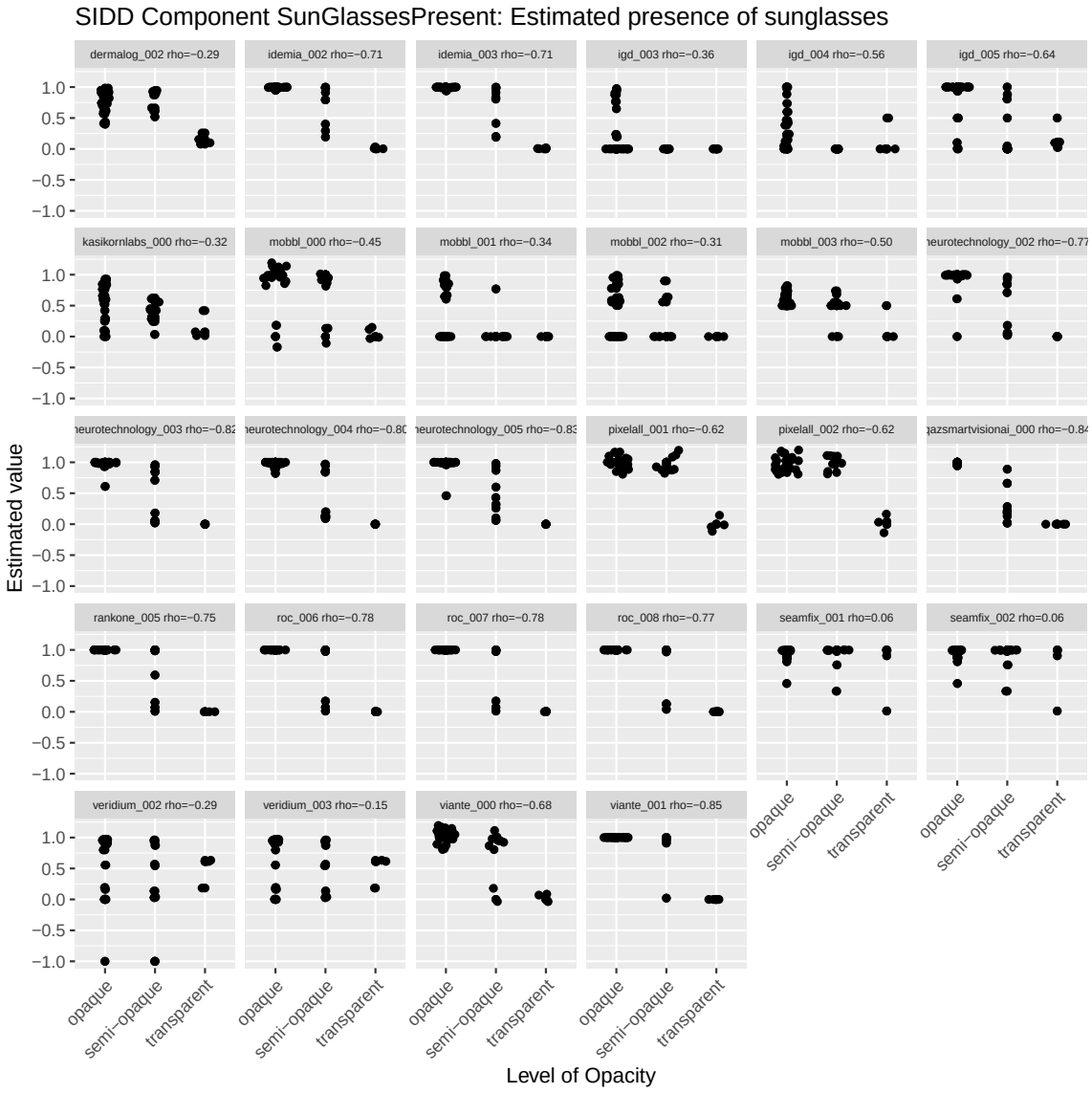


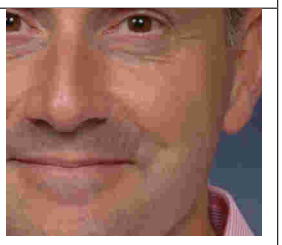
Fig. 40. Estimated vs. known presence of sunglasses. Perfect performance would correspond to monotonically decreasing clusters across the three categories, and a rank correlation (rho) value of -1.

3.19. Compression Artifacts

3.19.1. Images Used

We start by using mugshots for the Compression Artifacts set. We then use the imageMagick *convert* function with the argument *-quality* to apply JPEG compression to the original images. Table 17 shows the effect of blur at three values of compression d.

Table 17. Compression Artifacts Illustration. Images are used with the permission of the subject.

Compression parameter d	90	40	10
Result of convert -quality d			

3.19.2. Results for Compression Artifacts

Figure 41 summarizes the performance of the algorithms who have implemented detection of compression artifacts.

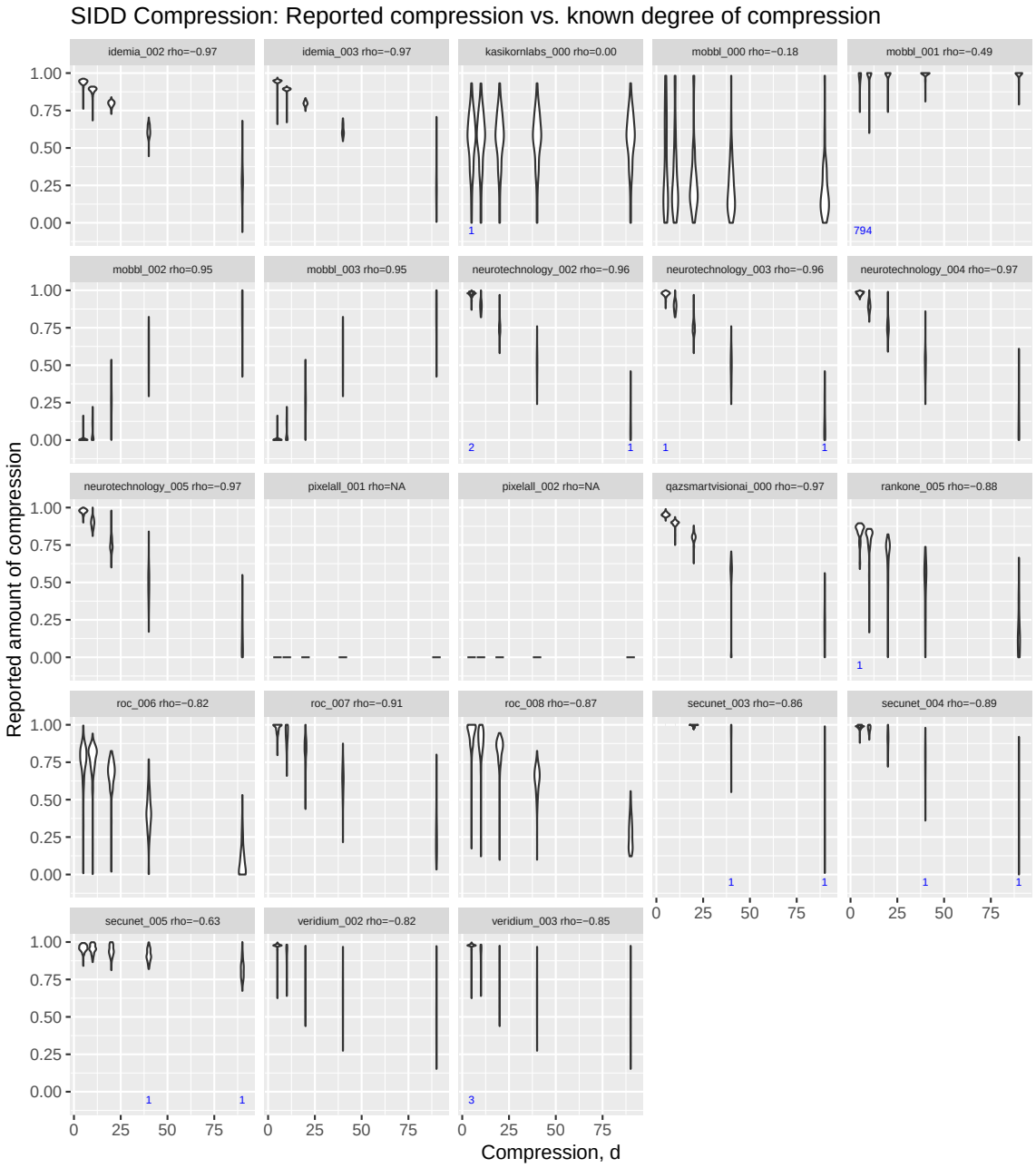


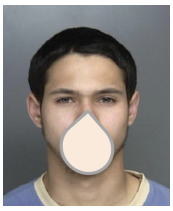
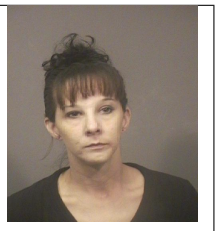
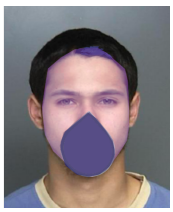
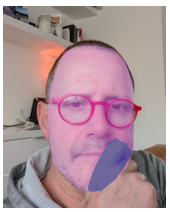
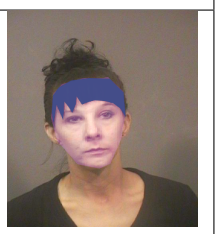
Fig. 41. Distribution of estimated amount of compression vs. d parameter of compression applied to a set of mugshot images. The small blue numbers along the x-axis represent the count of images for which the software did not return an estimate, for instance, when it did not detect a face. The higher the value of d, the lower the compression value. Perfect performance would correspond to monotonically decreasing estimates as d increases, and a rank correlation value (rho) of -1.

3.20. Face Occlusion

3.20.1. Images Used

For the Face Occlusion set, we use images that are generally frontal and well-illuminated. We then compute the occluded area and take the ratio of the occluded area to the total area of the facial region, as described in our API document. Table 18 illustrates values for three example images.

Table 18. Face Occlusion Illustration. The first and third images are from NIST Special Database 32, MEDS; the second image is used with permission of the subject.

Original image			
Image with occluded area shown in blue			
Ratio of occluded area to total area	0.27	0.11	0.36

3.20.2. Results for Face Occlusion

Figure 42 summarizes the performance of the algorithms who have implemented face occlusion.

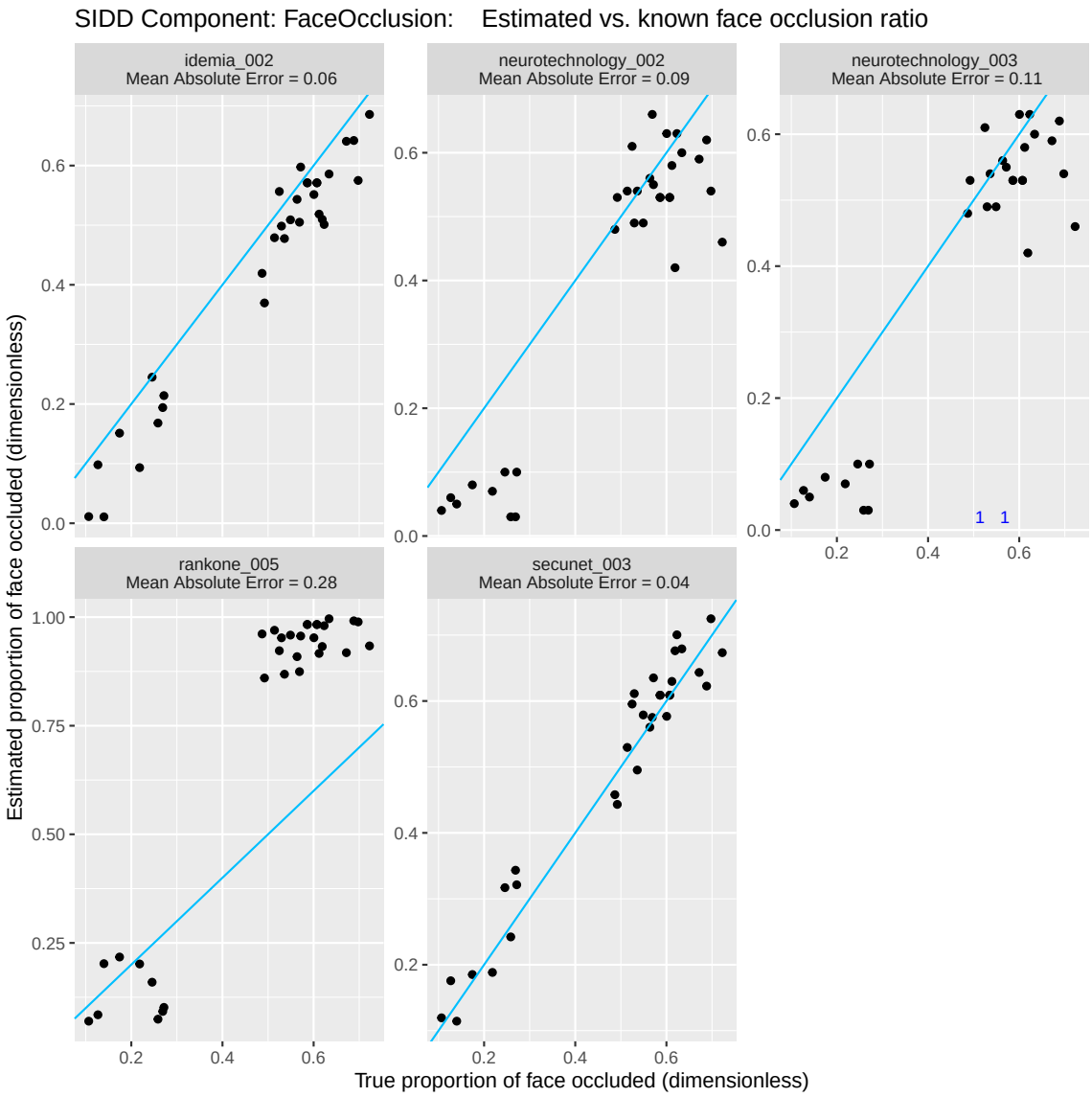


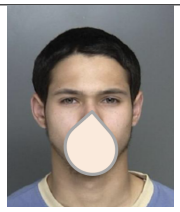
Fig. 42. Estimated vs. known ratio of occluded area to total area of the face. The blue line represents perfect performance. The small blue numbers above the x-axis represent the count of images for which the software did not detect a face; for such images, the error is set to 0.5. Note that the plots have different y-axis ranges. When points fall significantly above or below the blue line, the developer is likely implementing a different definition of the occluded area; for example, including beards or frames of eyeglasses when they should not be considered occlusion.

3.21. Face Occlusion 2

3.21.1. Images Used

For the Face Occlusion 2 set, we use images that are generally frontal and well-illuminated. We then compute the occluded area and take the ratio of the occluded area to the area of the facial region, where the facial region is a convex shape bounded by the eyebrows, chin, and sides of the face, as described in our API document. Note that the facial region does not include the forehead. This definition is consistent with the ISO/IEC 29794-5 standard. Table 19 illustrates values for two example images.

Table 19. Face Occlusion 2 Illustration. Images are from NIST Special Database 32, MEDS.

Original image		
Image with occluded area shown in gray		
Ratio of occluded area (gray) to total area (pink+gray)	0.41	0.22

3.21.2. Results for Face Occlusion 2

Figure 43 summarizes the performance of the algorithms who have implemented Face Occlusion 2.

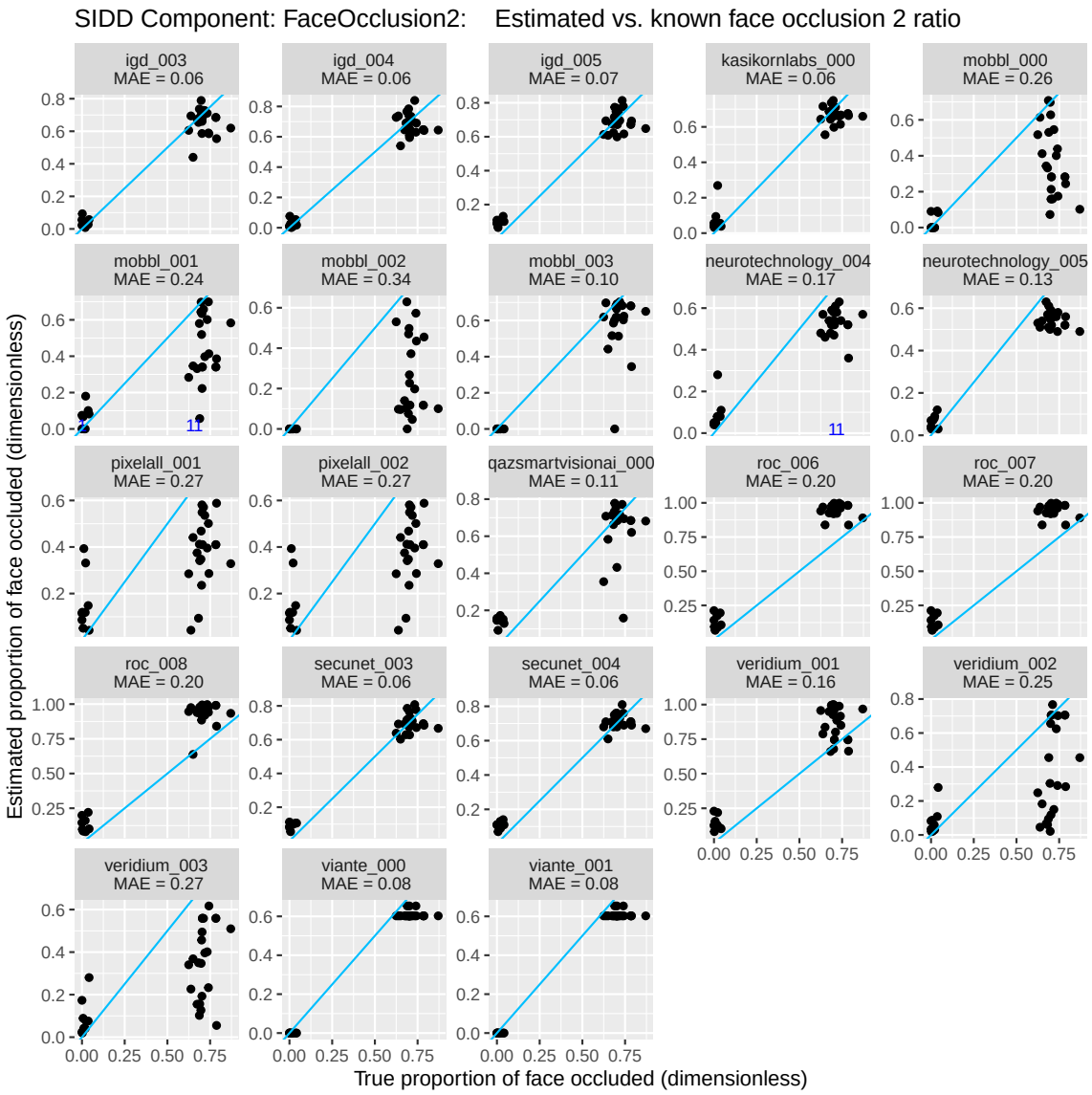


Fig. 43. Estimated vs. known ratio of occluded area to total area of the face. The blue line represents perfect performance.

3.22. Motion Blur

3.22.1. Images Used

We start by using mugshots for the Motion Blur set. We then use the imageMagick *convert* function with the argument *-motion-blur* to apply motion blur to the original images, which are selected to have no visible blur, motion blur, or compression artifacts to begin with. Table 20 shows the effect of blur at three values of displacement d. For our test we use six values of displacement ranging from 0 to 20.

Table 20. Motion Blur Illustration. Images are used with the permission of the subject.

Displacement d	0	8	16
Result of convert -motion-blur 0xd			

3.22.2. Results for Motion Blur

Figure 44 summarizes the performance of the algorithms who have implemented motion blur.

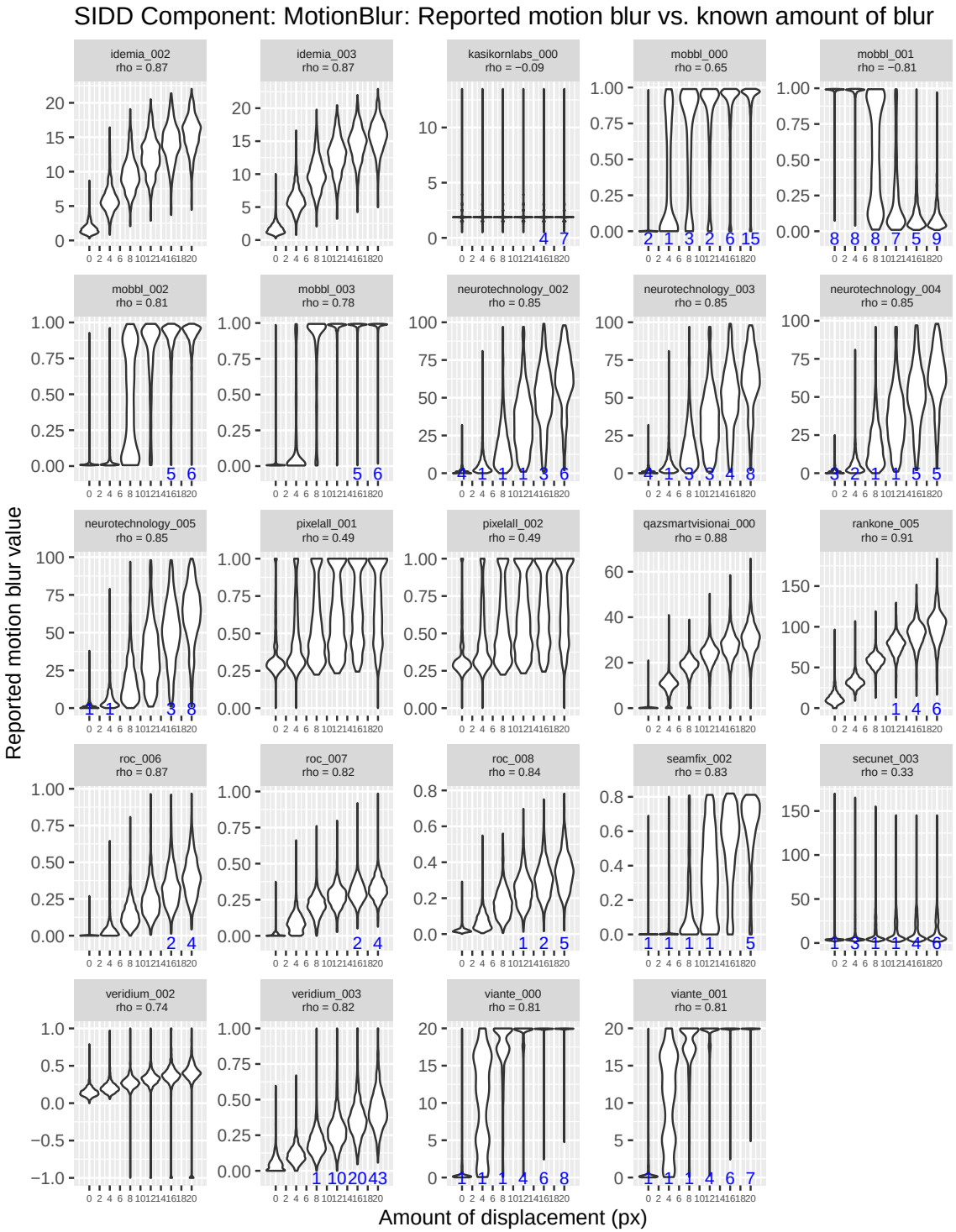


Fig. 44. Distribution of estimated motion blur vs. d parameter of motion blur applied to a set of mugshot images. The higher the value of d, the more extreme the blur.The small blue numbers along the x-axis represent the count of images for which the software did not return an estimate, for instance, when it did not detect a face. Perfect performance would correspond to monotonically increasing estimates as d increases, and a rank correlation value (rho) of 1.

3.23. Distance from Eyes to Edges

3.23.1. Images Used

We use two set of images for the four distance-from-eye-to-edge quality measures.

1. Mugshot: Pose is generally frontal and backgrounds are generally uniform. Images with a face approximately centered in the frame are cropped to varying degrees. The images range from 311 to 1000 pixels in width and from 240 to 1330 pixels in height. Figure 45 shows an example.
2. Kiosk: There are a variety of pose angles and lighting conditions. In many of the images, the pitch is nonzero and the face is off-center. For these images, one eye is either very close to the edge of the image, or partially out of the frame. The images are generally 320 pixels in width and 240 pixels in height. Figure 46 shows an example (in which the subject's left eye is close to the right edge of the image).

In order to determine ground truth, we manually find the eye-centers by determining the two points where eyelids meet for each eye and averaging the two points. For each image,we calculate the following:

1. The distance from the left edge to the closest eye-center
2. The distance from the right edge to the closest eye-center
3. The distance from the top edge to the average of the eye-centers
4. The distance from the bottom edge to the average of the eye-centers

These quantities are shown in figure 45.

Fig. 45. Image from NIST Special Database 32, MEDS.

This procedure is consistent with that described in the ISO/IEC 29794-5:2024 standard.

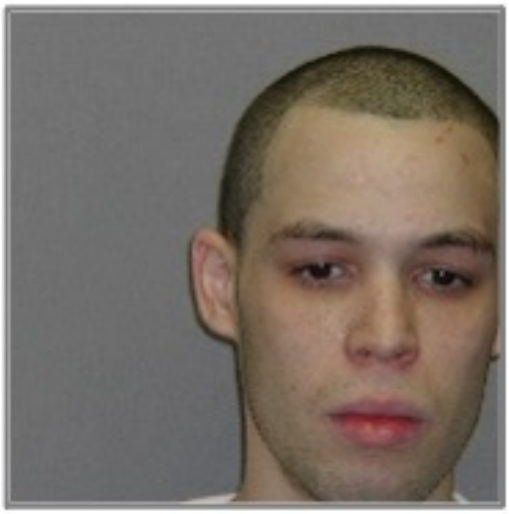


Fig. 46. Image from NIST Special Database 32, MEDS.

3.23.2. Results for Distance from Eyes to Edges

Figure 51, 52, 53, and 54 summarize algorithm performance. The Mean Absolute Error (MAE), where error is computed as the difference between ground truth and reported value, is shown for each algorithm. Lower MAE is better.

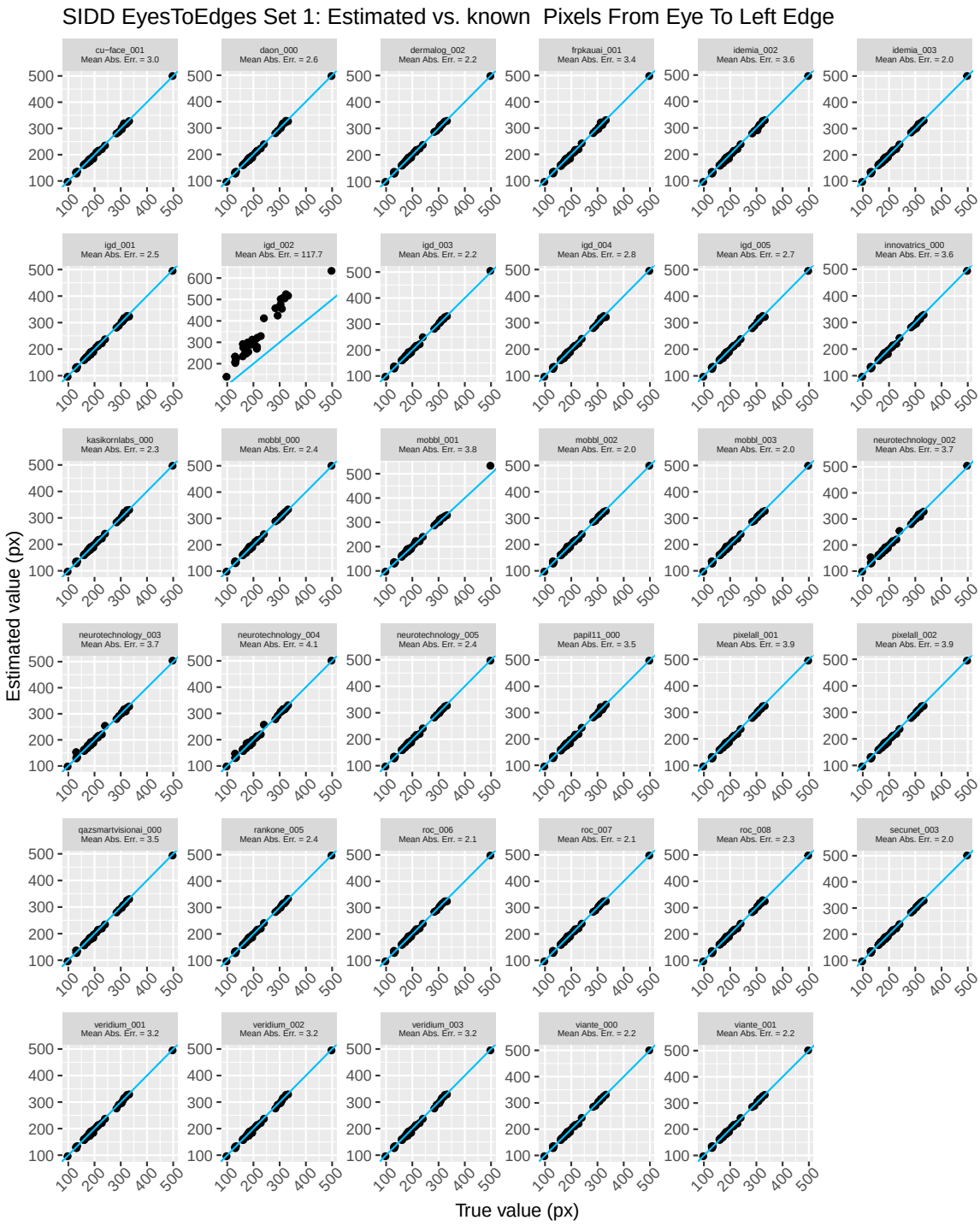


Fig. 47. Estimated vs. known pixels from left edge to the closest eye center. The blue line represents perfect performance.

Fig. 48. Estimated vs. known pixels from right edge to the closest eye center. The blue line represents perfect performance.

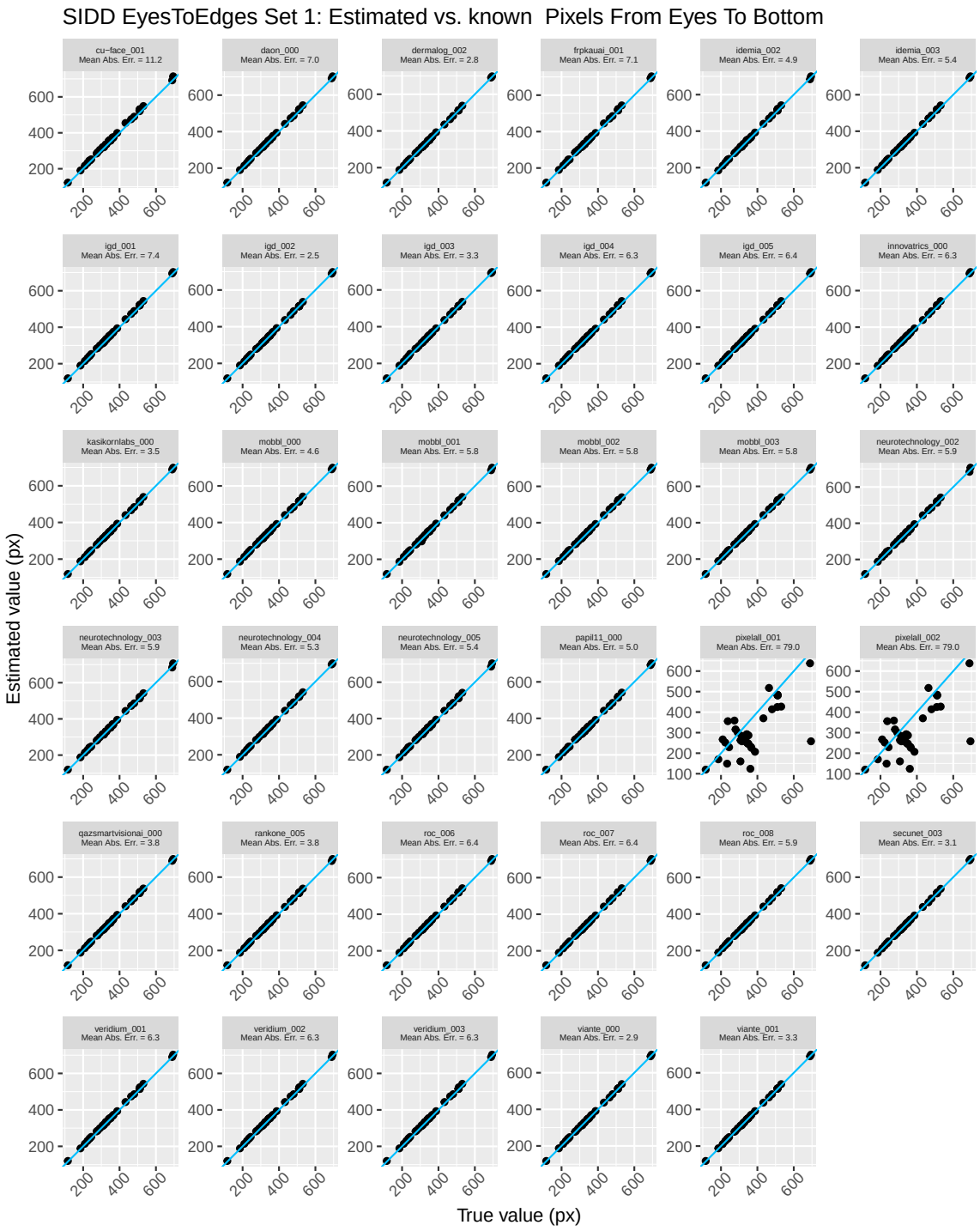


Fig. 49. Estimated vs. known pixels from center of eyes to the bottom of the image. The blue line represents perfect performance.

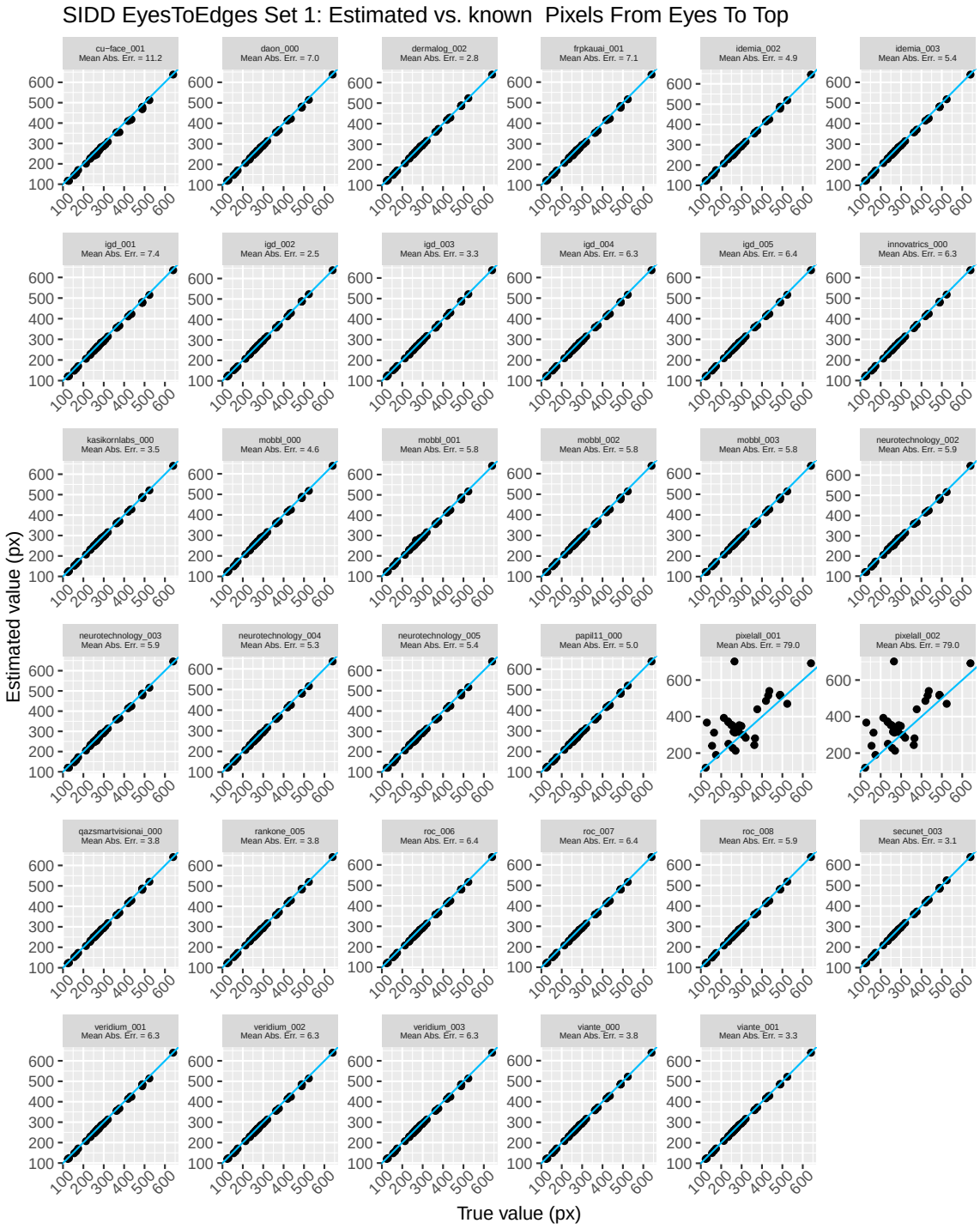


Fig. 50. Estimated vs. known pixels from center of eyes to the top of the image. The blue line represents perfect performance.

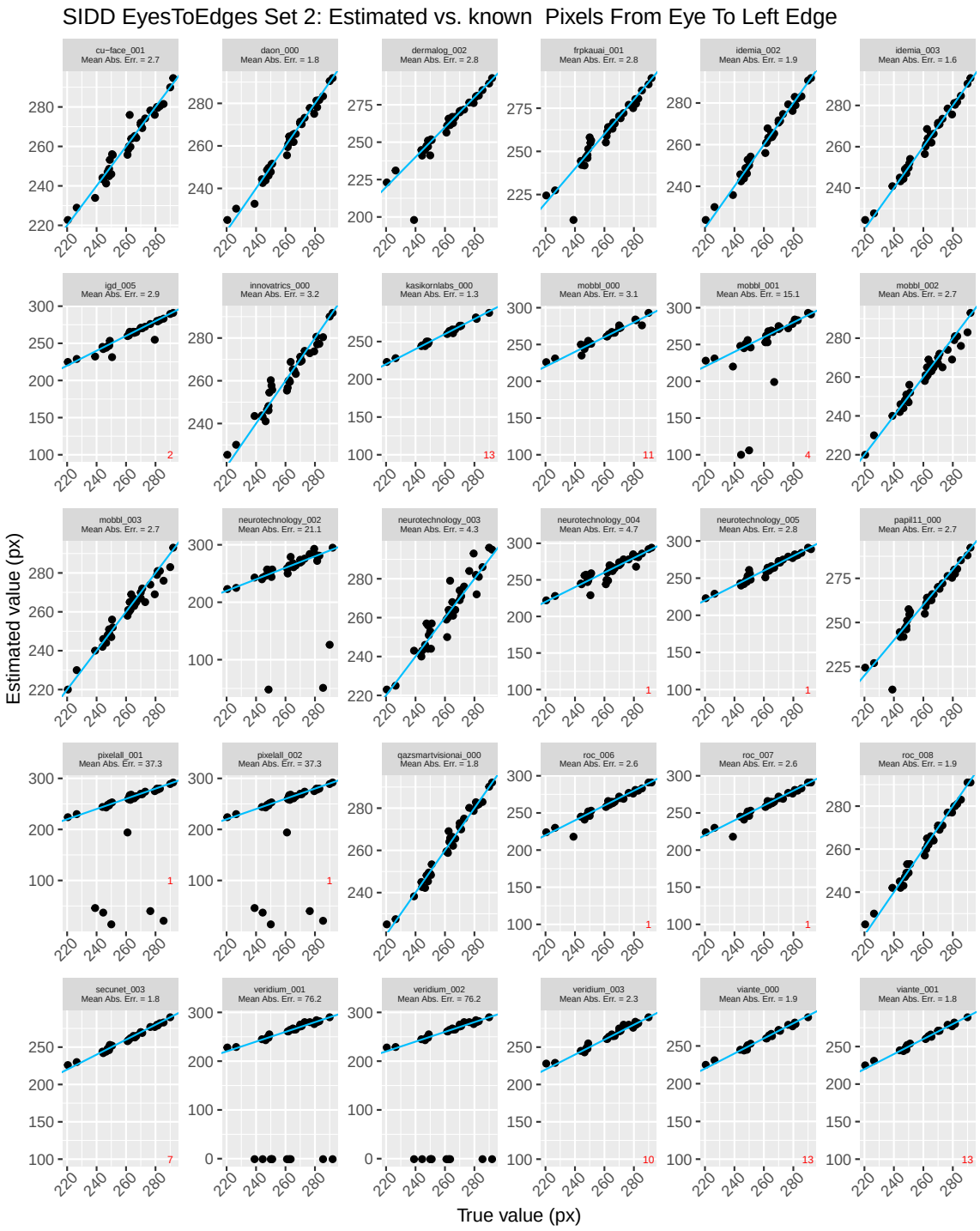


Fig. 51. Estimated vs. known pixels from left edge to the closest eye center. The blue line represents perfect performance. The small red numbers at the bottom of the plot represent failures to detect a face.

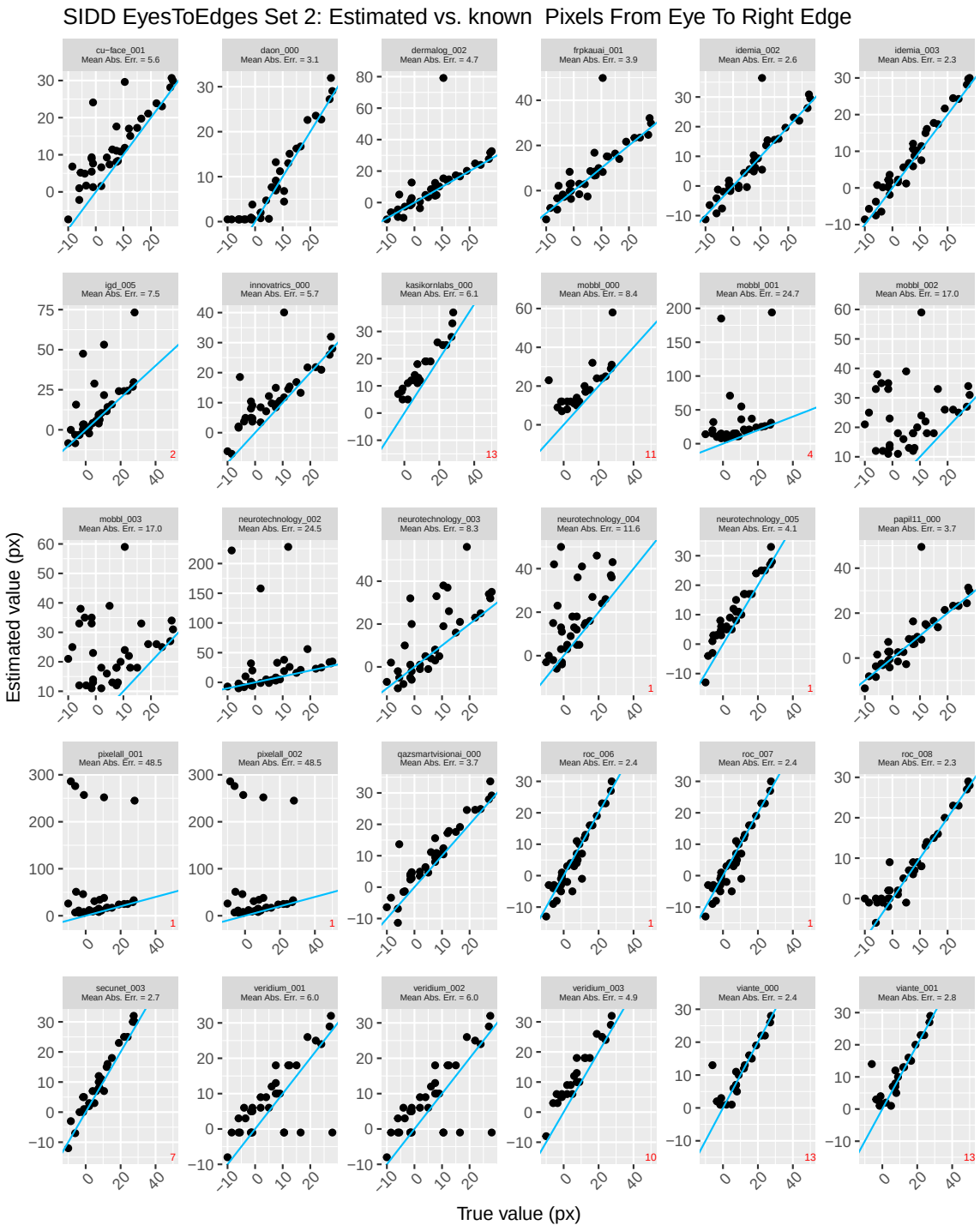


Fig. 52. Estimated vs. known pixels from right edge to the closest eye center. The blue line represents perfect performance. The small red numbers at the bottom of the plot represent failures to detect a face.

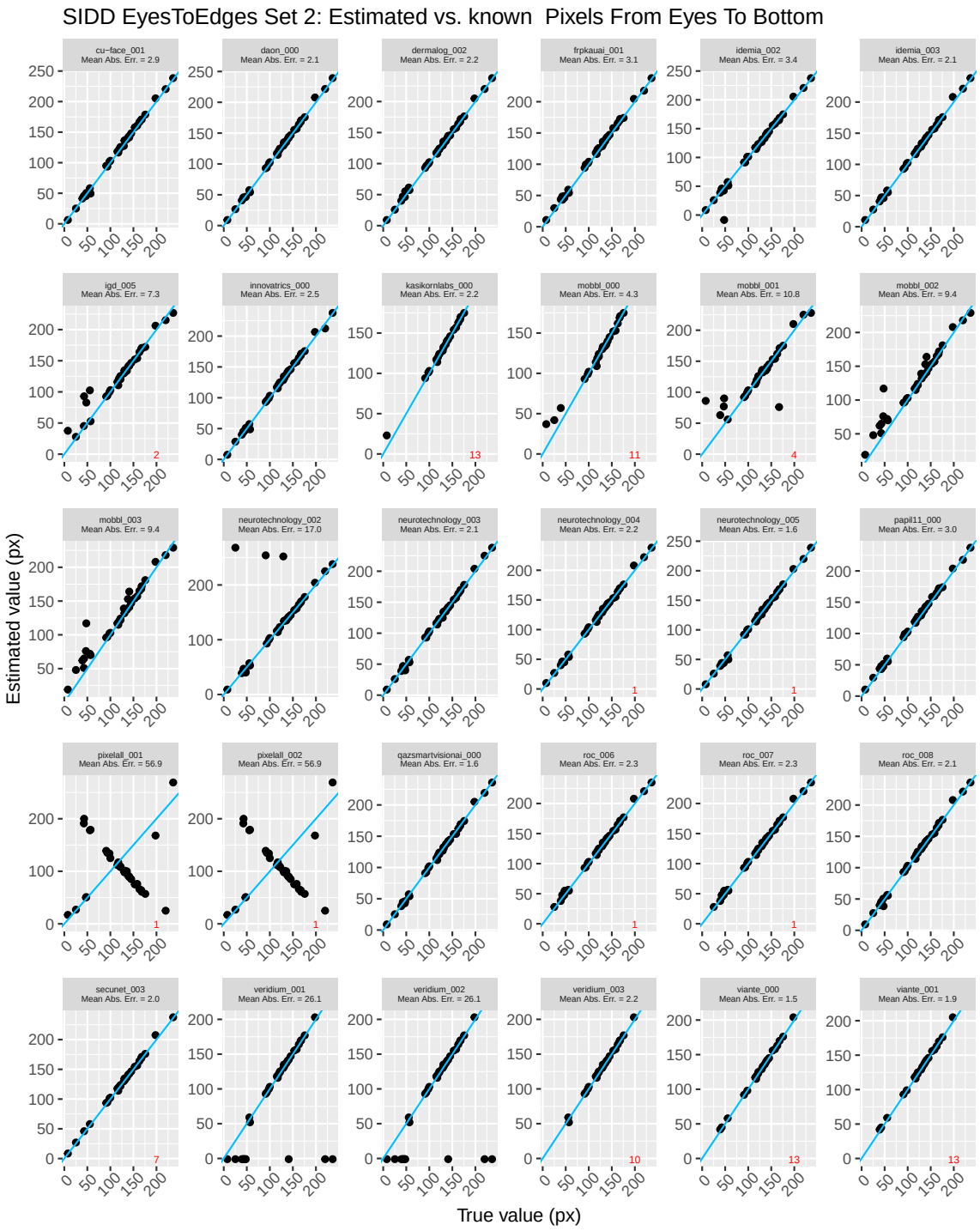


Fig. 53. Estimated vs. known pixels from center of eyes to the bottom of the image. The blue line represents perfect performance.The small red numbers at the bottom of the plot represent failures to detect a face.

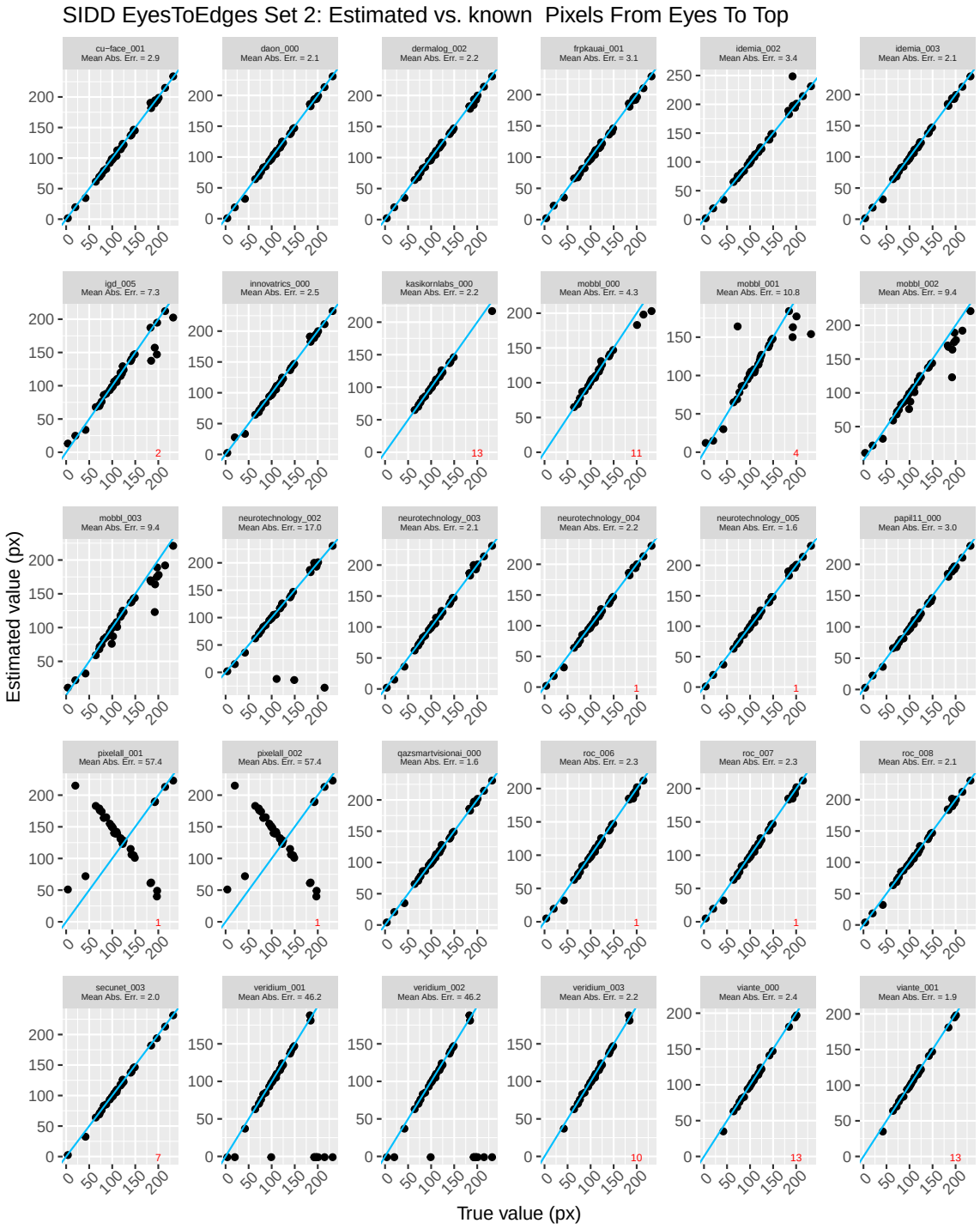


Fig. 54. Estimated vs. known pixels from center of eyes to the top of the image. The blue line represents perfect performance.The small red numbers at the bottom of the plot represent failures to detect a face.

3.24. Unified Quality Score

For analysis of Unified Quality Score, we examine the reduction in FNMR as Unified Quality Score is used to discard low quality images. The FNMR values are computed by taking the mean across 15 high-performing recognition algorithms from 15 unique developers.

3.24.1. Results for Unified Quality Score

Figure 55 shows false non-match rate (FNMR) gains as a function of the fraction of lowest Unified Quality Score images discarded, for four initial FNMR values: 0.5%, 1%, 2% and 5%.

Similarity scores are generated from mated comparison of high quality visa-like application photos with medium quality airport arrival webcam photos. Quality is computed only on the webcam photos. For images for which the algorithm did not detect a face, the quality is set to the minimum quality assigned by the algorithm.

The horizontal dashed line gives the initial FNMR value; the curved dashed line represents perfect performance.

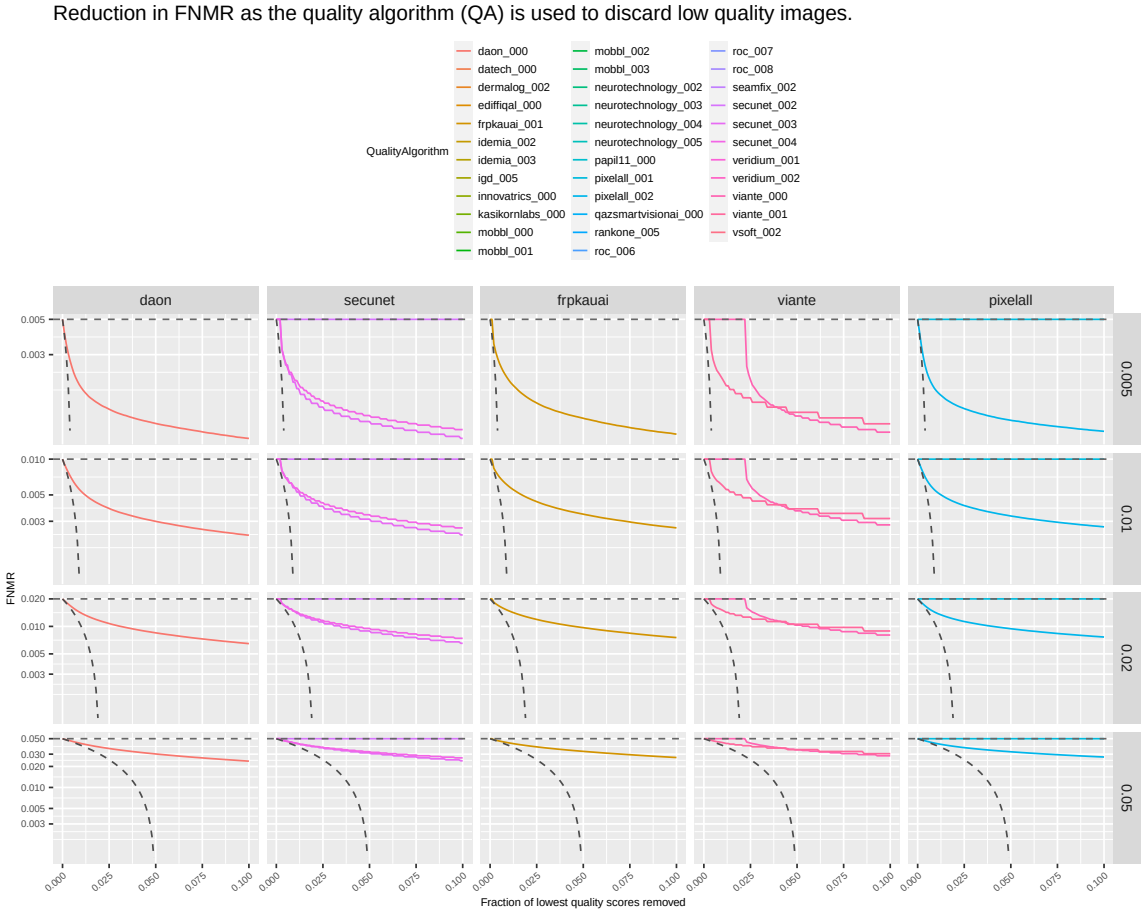


Fig. 55. Reduction in FNMR as the quality algorithm (QA) is used to discard low quality images. The curved dashed line represents perfect performance.

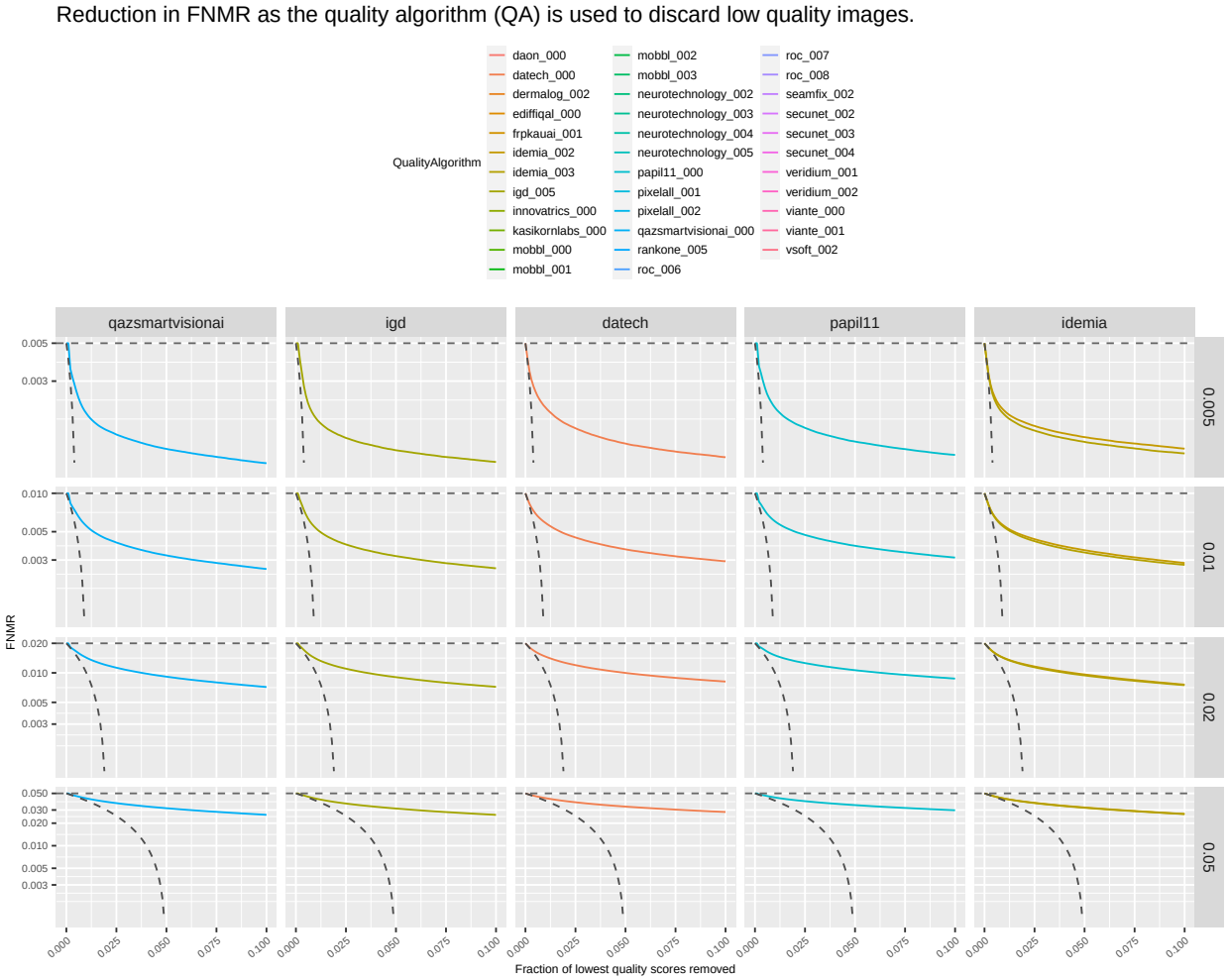


Fig. 56. Reduction in FNMR as the quality algorithm (QA) is used to discard low quality images. The curved dashed line represents perfect performance.

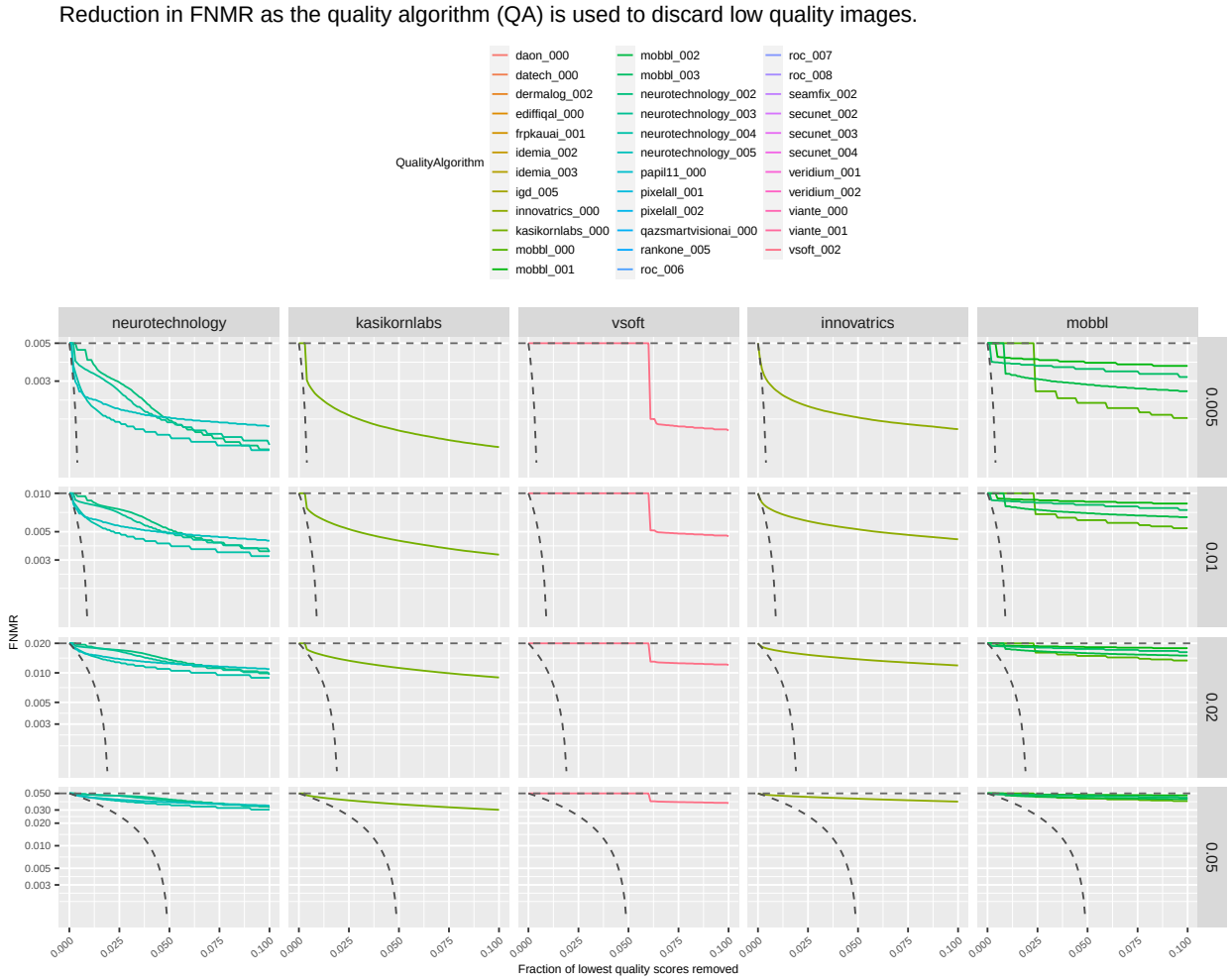


Fig. 57. Reduction in FNMR as the quality algorithm (QA) is used to discard low quality images. The curved dashed line represents perfect performance.

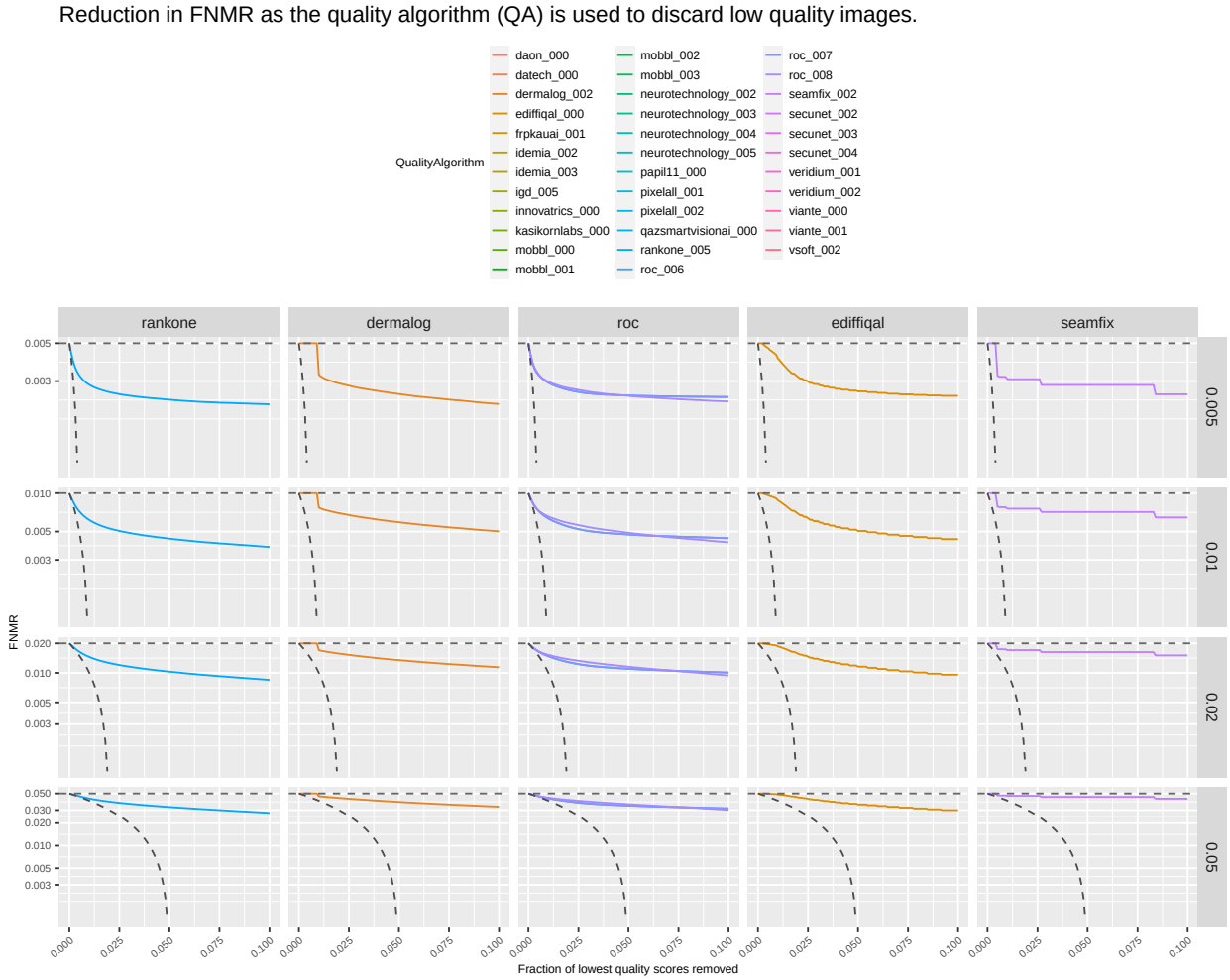


Fig. 58. Reduction in FNMR as the quality algorithm (QA) is used to discard low quality images. The curved dashed line represents perfect performance.

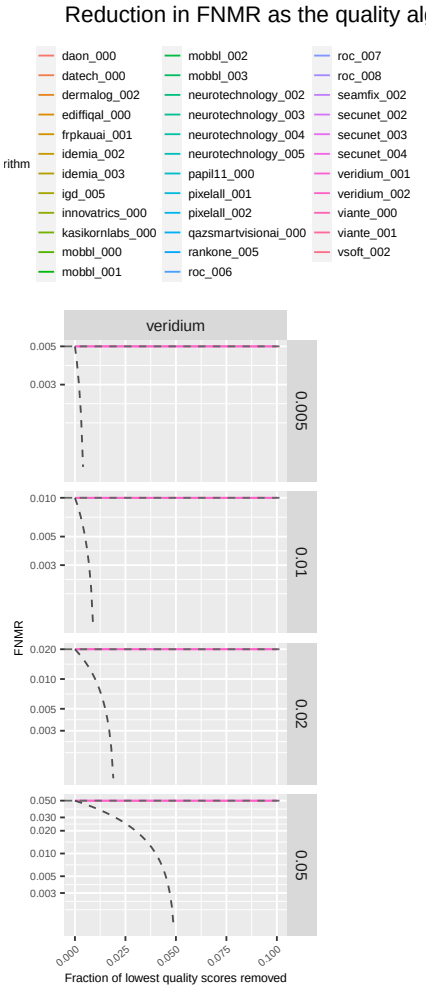


Fig. 59. Reduction in FNMR as the quality algorithm (QA) is used to discard low quality images. The curved dashed line represents perfect performance.

4. Quality Measures By Demographic Group

A key question when any AI or ML-based algorithm is applied to human data is whether it gives equitable outputs across demographic groups. In cooperative biometric applications, such as face enrollment, an image quality assessment algorithm should not yield values that would cause photo rejection rates to be higher in one group than another.

Operationally quality algorithms would be configured with a fixed threshold for each quality component.

Depending on local policy, any photo assessed to have a quality component worse than threshold would, in principle, be sent for human review or even rejected outright. A quality algorithm should be insensitive to demographic membership and sensitive only to imaging conditions and, for things like head pose, the behavior of the individual.

This section employs a large high-quality dataset to quantify variations in quality outputs across demographics and to inform what the threshold values could be.

4.1. Datasets

We use high-quality *immigration application* photos for this study.

The *application* images are collected in an attended setting where a subject is imaged with a dedicated photo capture device after having been asked to remove glasses, maintain a neutral expression, and face the camera. As such the images have high quality and consistency with few, visually small, quality variations. This is reflected in the distribution of values for the quality values we examined. These images are suitable for informing thresholds because most of the images are of excellent quality. If a quality assessment algorithm produces outputs that vary across demographic groups, this will not be due to imaging system or environment, or subject behavior.

We define demographic groups in terms of region of birth. The region of birth for each subject is determined according to the country of birth, as follows:

- ▷ East Africa: Ethiopia, Kenya, Somalia, Sudan, Uganda, Tanzania
- ▷ East Asia: South Korea, China, Japan, Taiwan
- ▷ East Europe: Poland, Ukraine, Russia, Hungary, Romania, Czechia
- ▷ South Asia: India, Pakistan, Bangladesh, Afghanistan, Myanmar, Nepal
- ▷ South East Asia: Philippines, Vietnam, Cambodia, Indonesia, Malaysia, Laos, Thailand
- ▷ West Africa: Nigeria, Liberia, Sierra Leone, Ghana, Benin, Mali, Senegal, Togo

These countries and regions are selected because we consider being born in this country will serve as reasonable proxy for ethnicity because these countries have seen low levels of transcontinental migration. Countries such as France, United Kingdom and the United States are not included because they do not fall into this category.

4.2. Quality Measures by Region of Birth

We consider algorithm outputs across the demographic groups. Ideal performance would correspond to identically distributed quality values for each region of birth. This would be apparent in distributional plots such as violin- and box-plots if they appear similar in height and shape. For cumulative distribution plots, ideal performance corresponds to plots with no separation between the different-colored curves.

Demographic results are reported for each measure that is implementd by a given algorithm.

Below is a summary of the distribution of Eyes Open 2, Mouth Open 2, Underexposure, Overexposure, Resolution, and Unified Quality Score for six different regions of birth (East Africa, East Asia, East Europe, South Asia, South East Asia, and West Africa), and two sexes (Male and Female).

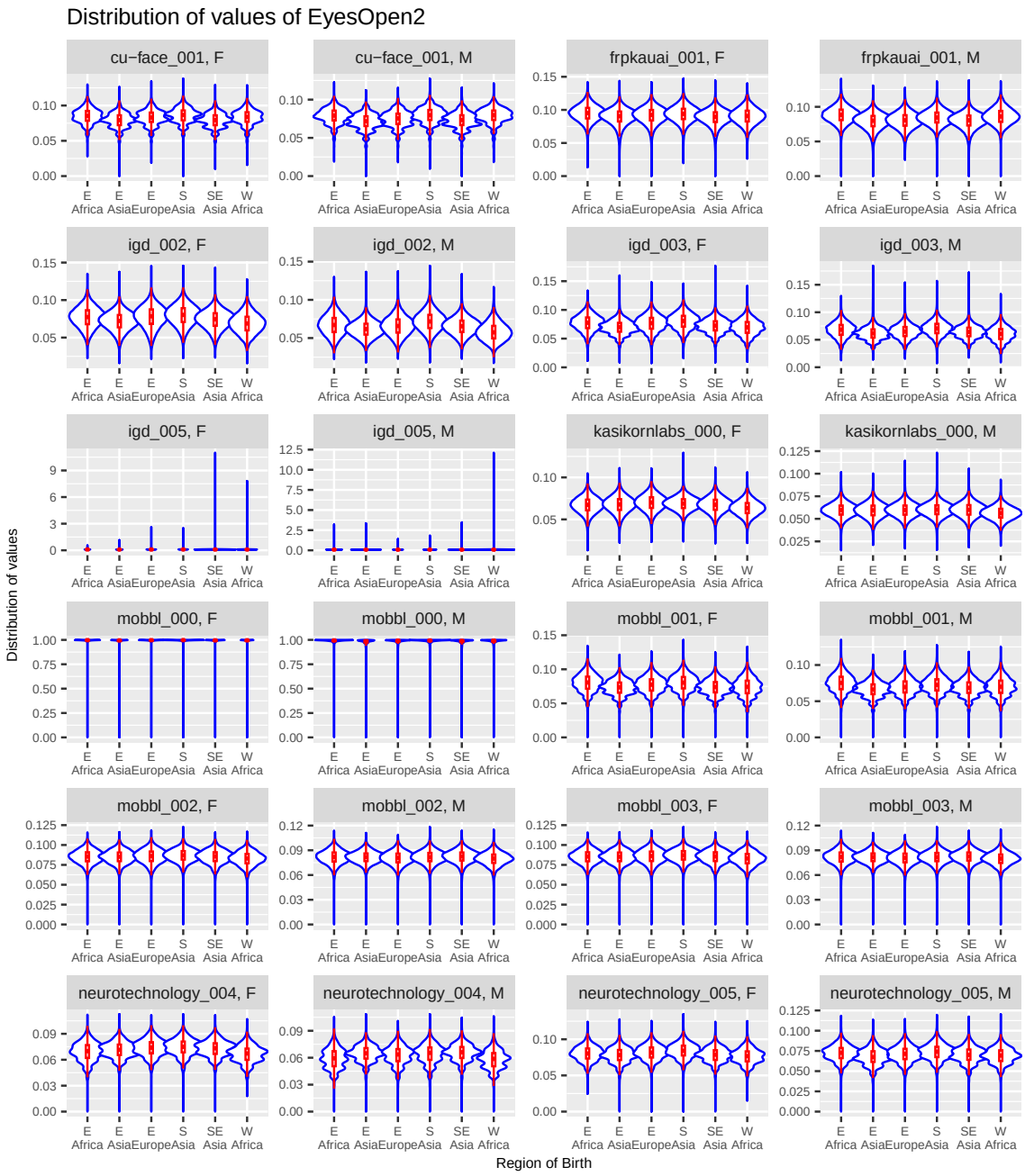


Fig. 60. Distribution of EyesOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

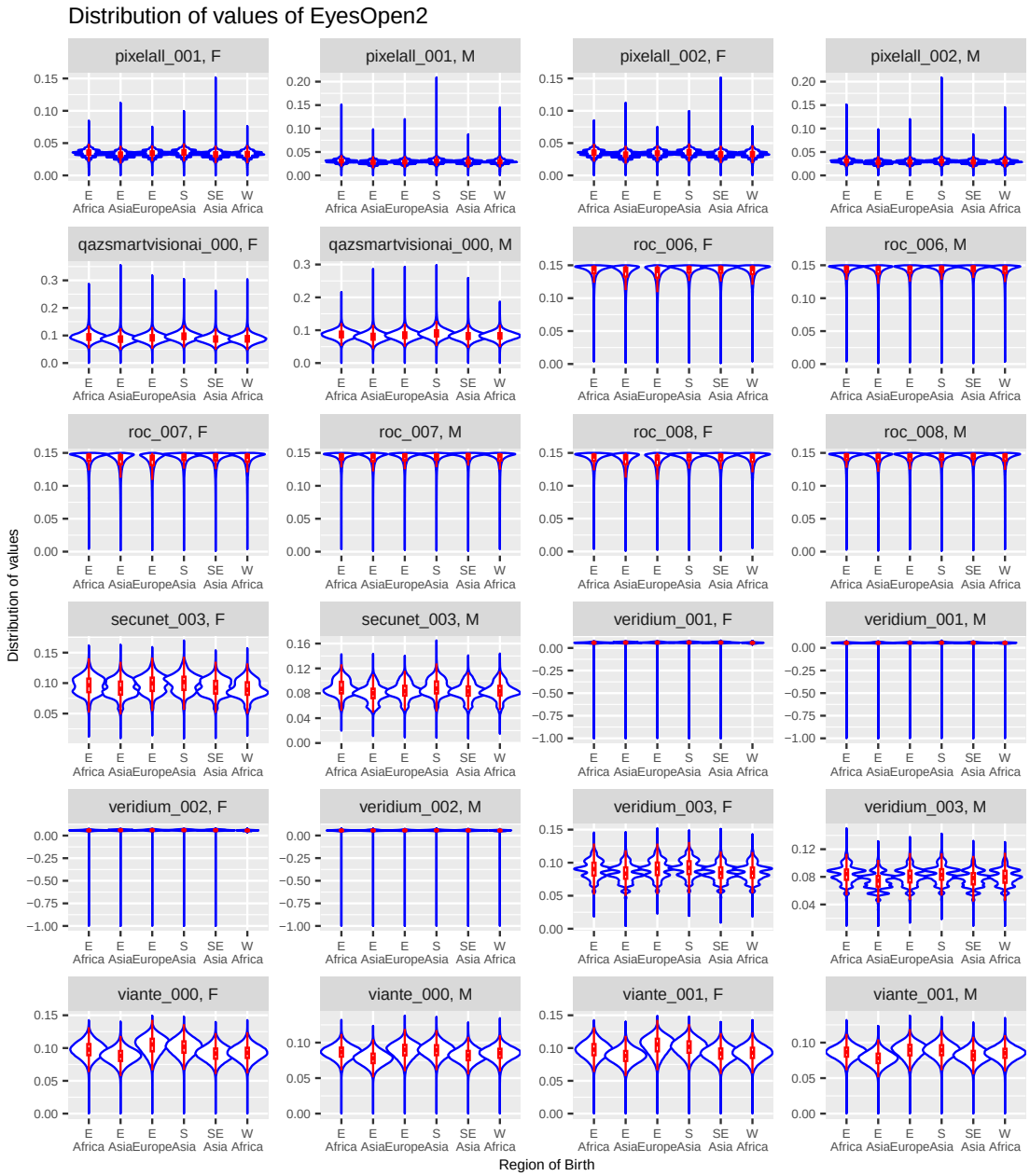


Fig. 61. Distribution of EyesOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 62. Distribution of MouthOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 63. Distribution of MouthOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 64. Distribution of Underexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 65. Distribution of Underexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 66. Distribution of Underexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

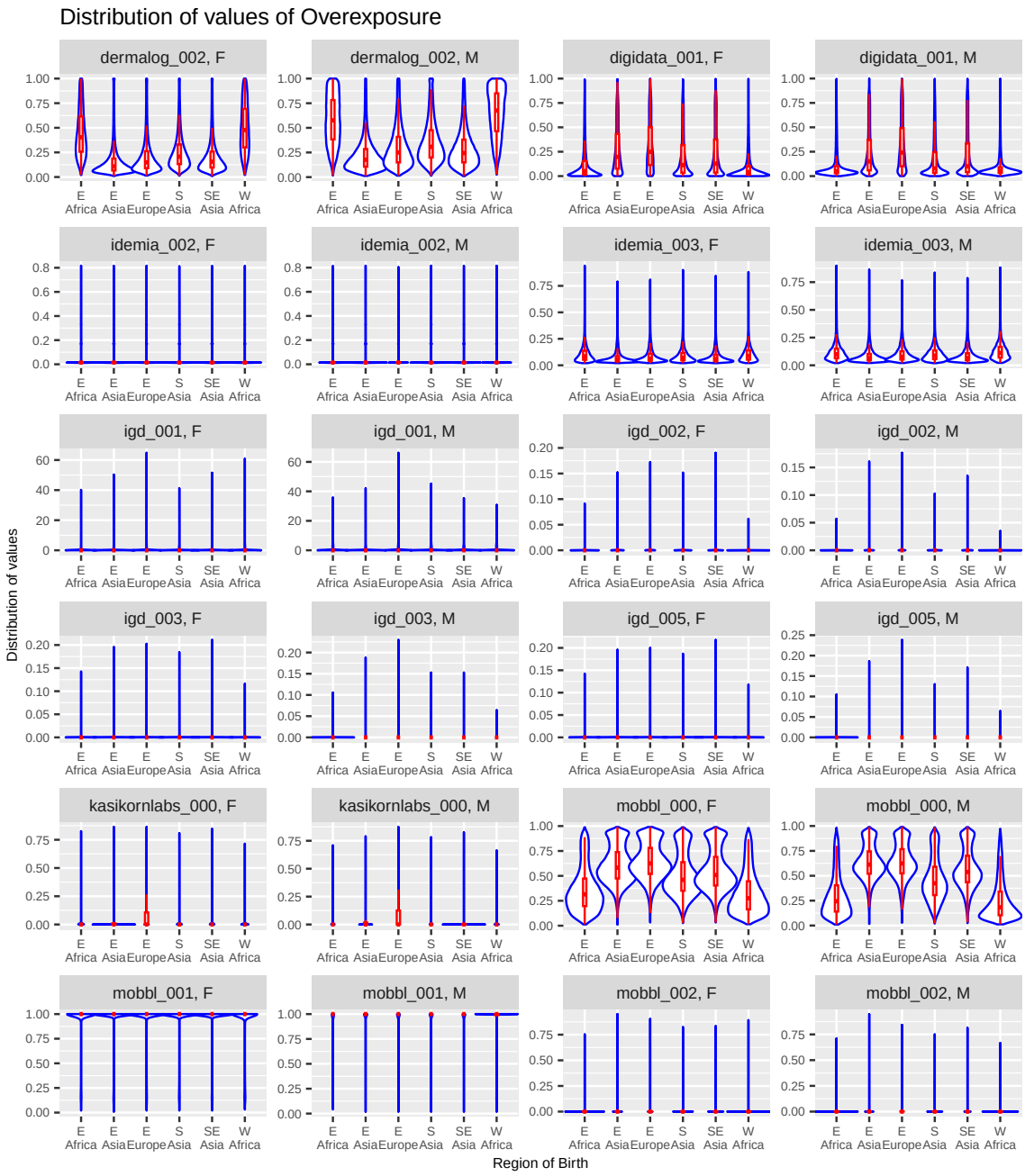


Fig. 67. Distribution of Overexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 68. Distribution of Overexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

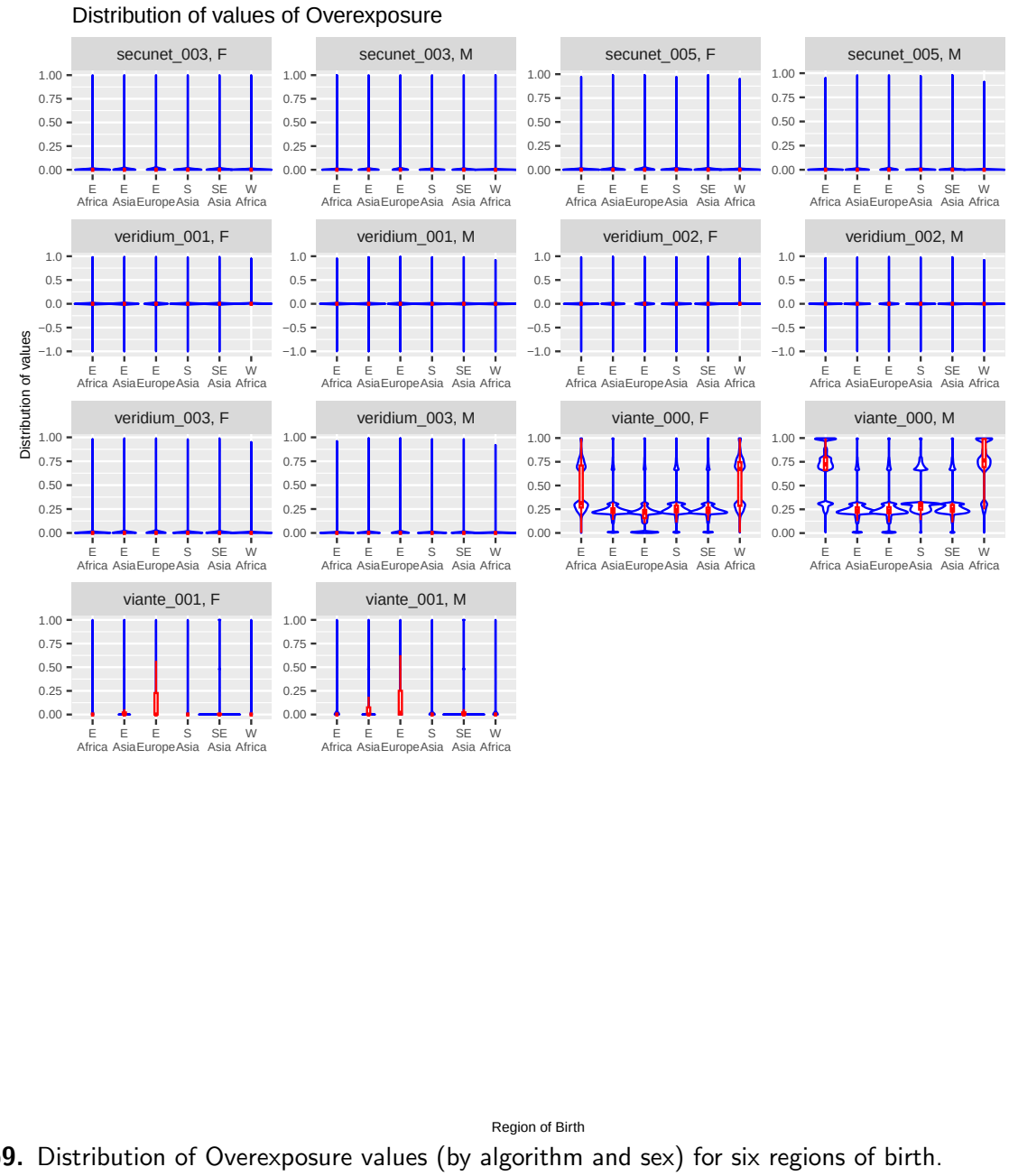


Fig. 69. Distribution of Overexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

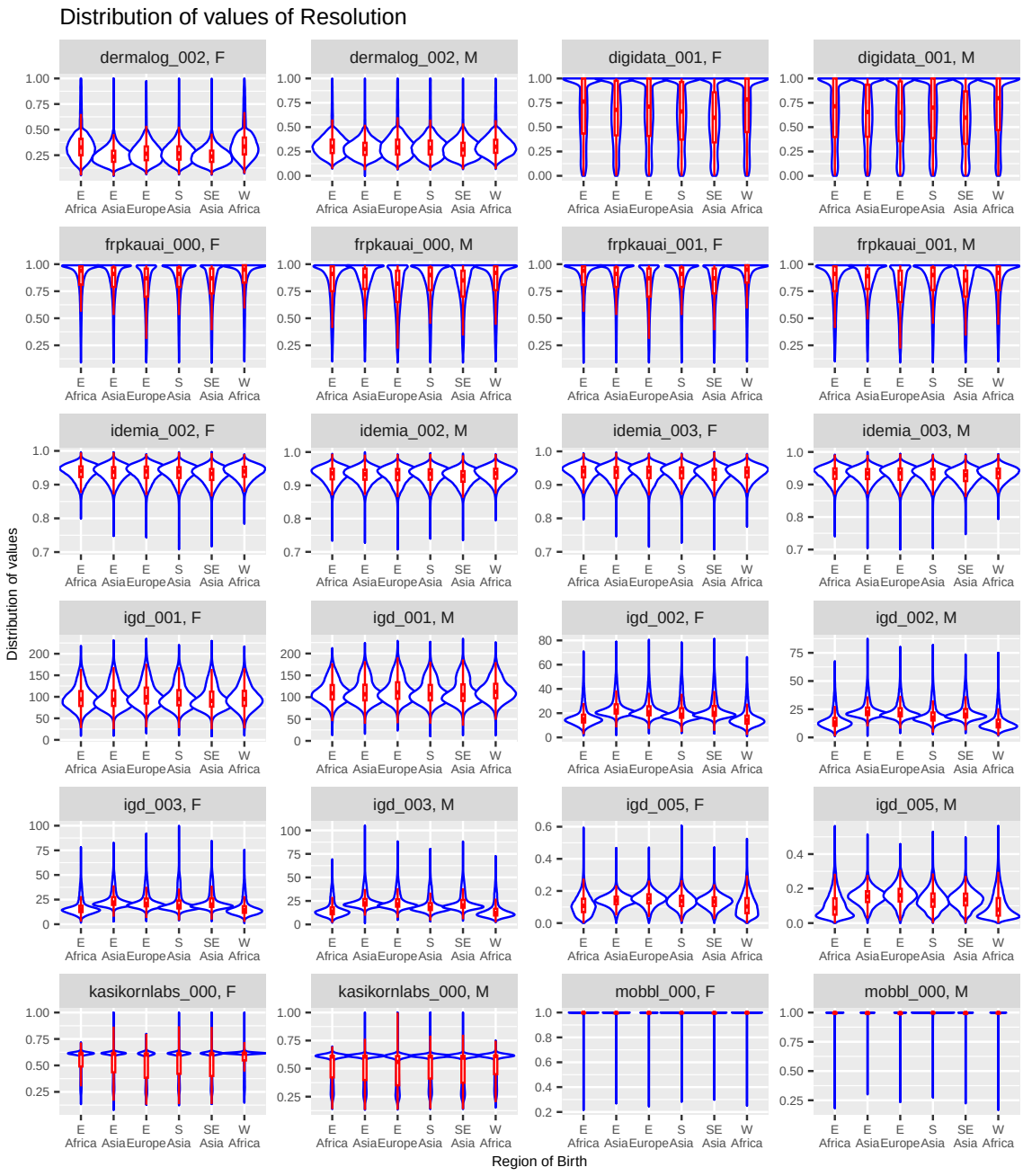


Fig. 70. Distribution of Resolution values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 71. Distribution of Resolution values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

Fig. 72. Distribution of Resolution values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

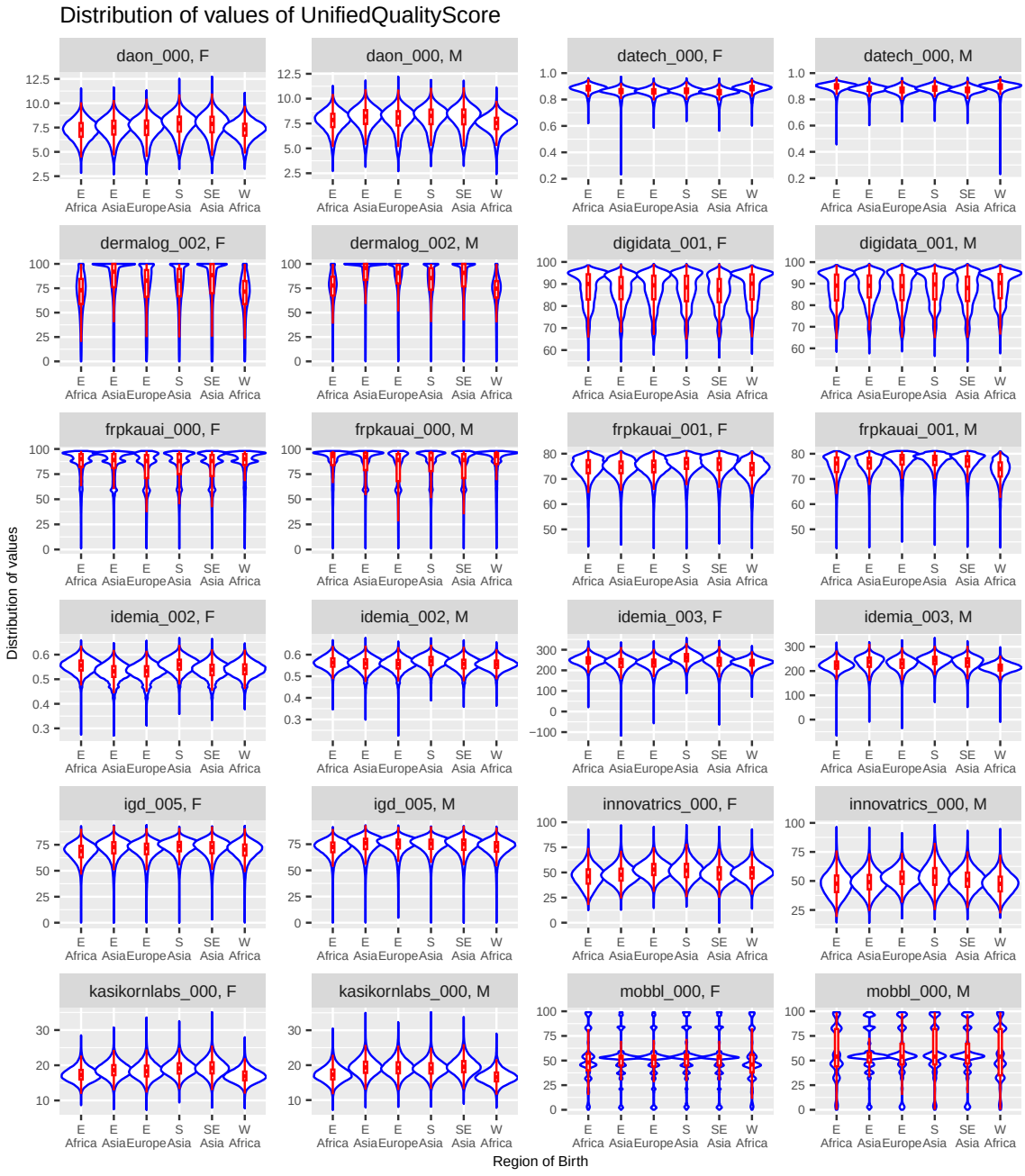


Fig. 73. Distribution of Unified Quality Score values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

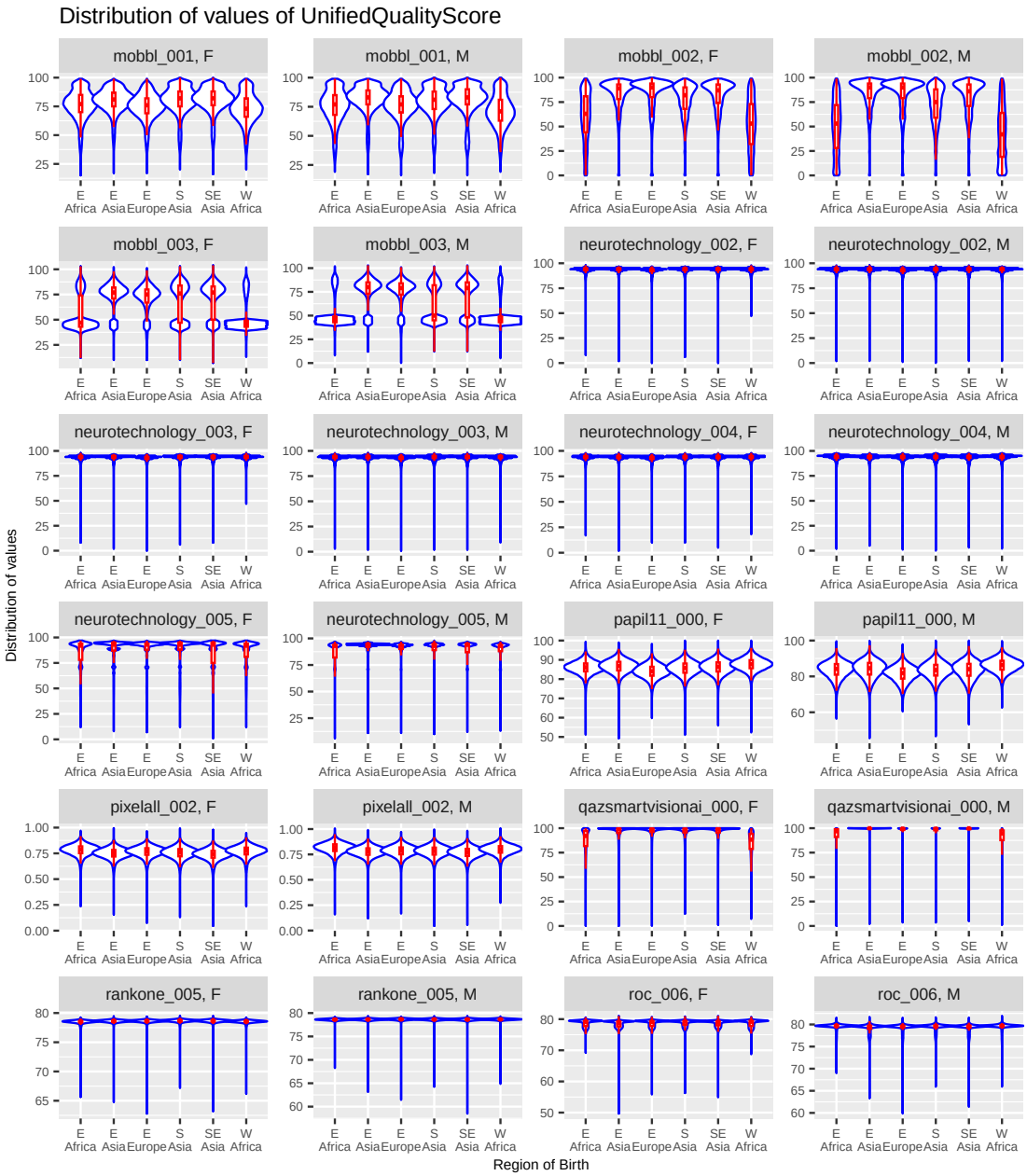


Fig. 74. Distribution of Unified Quality Score values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

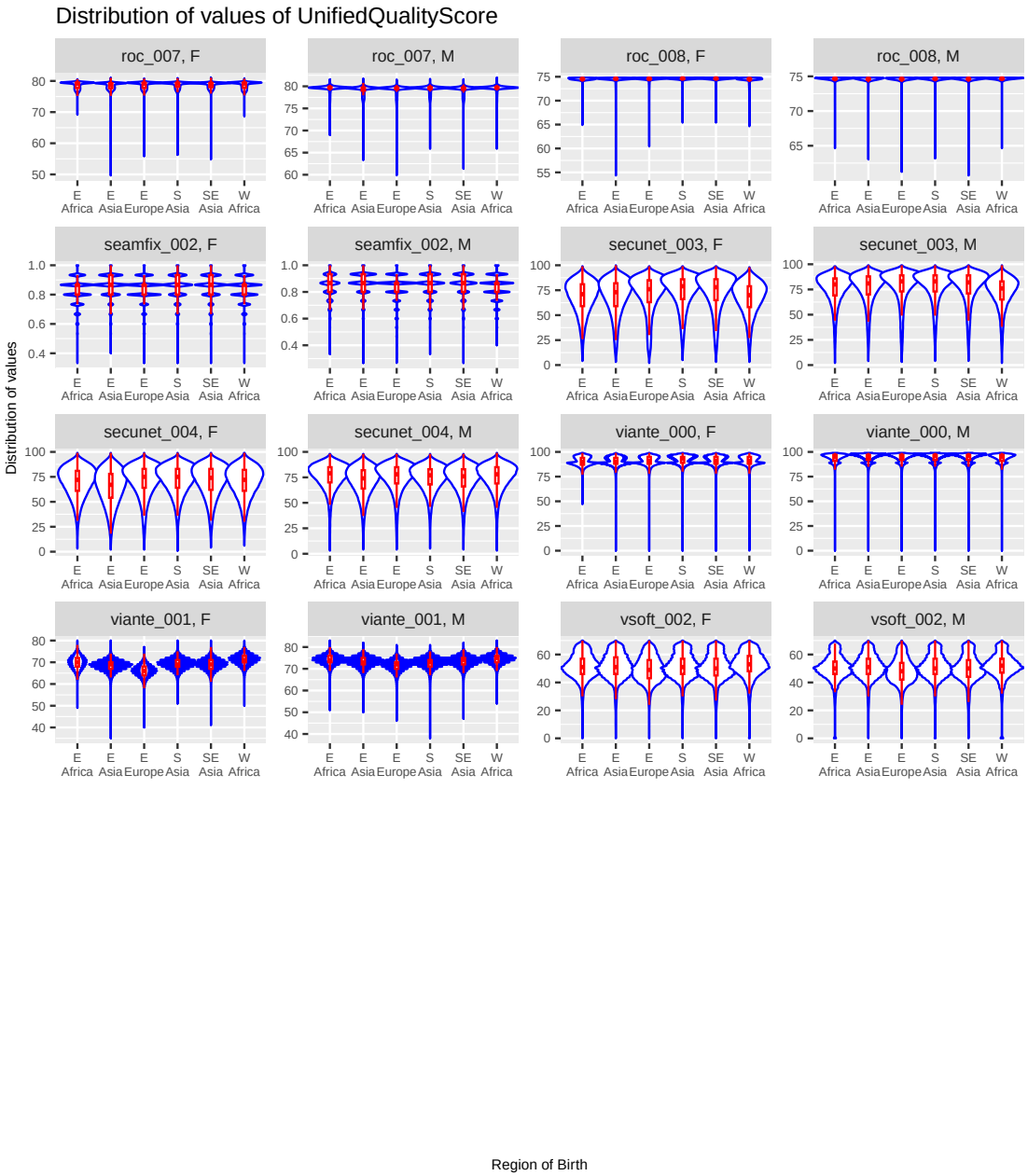


Fig. 75. Distribution of Unified Quality Score values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to violins that are all at the same vertical position and all have the same distribution.

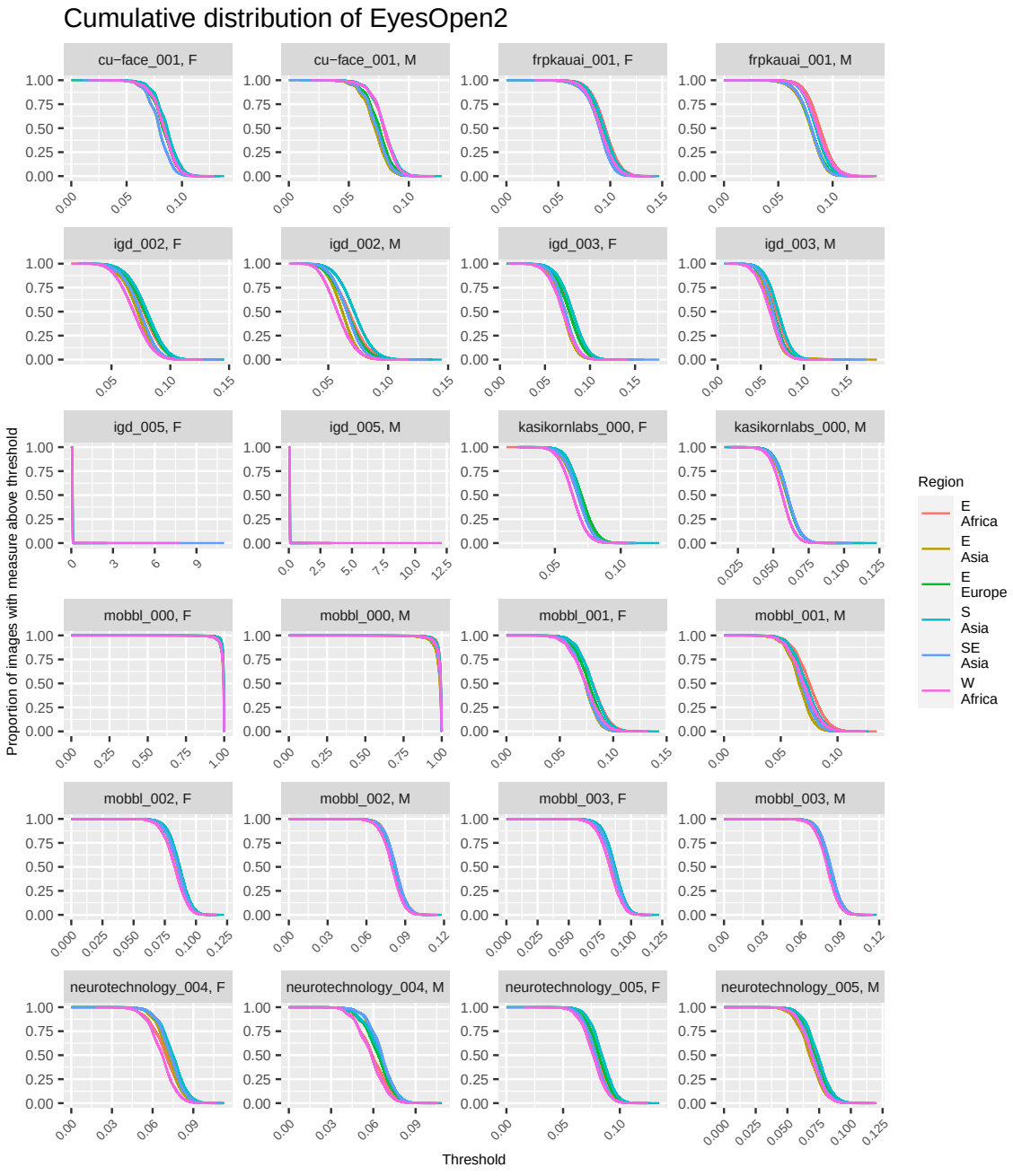


Fig. 76. Cumulative distribution of EyesOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

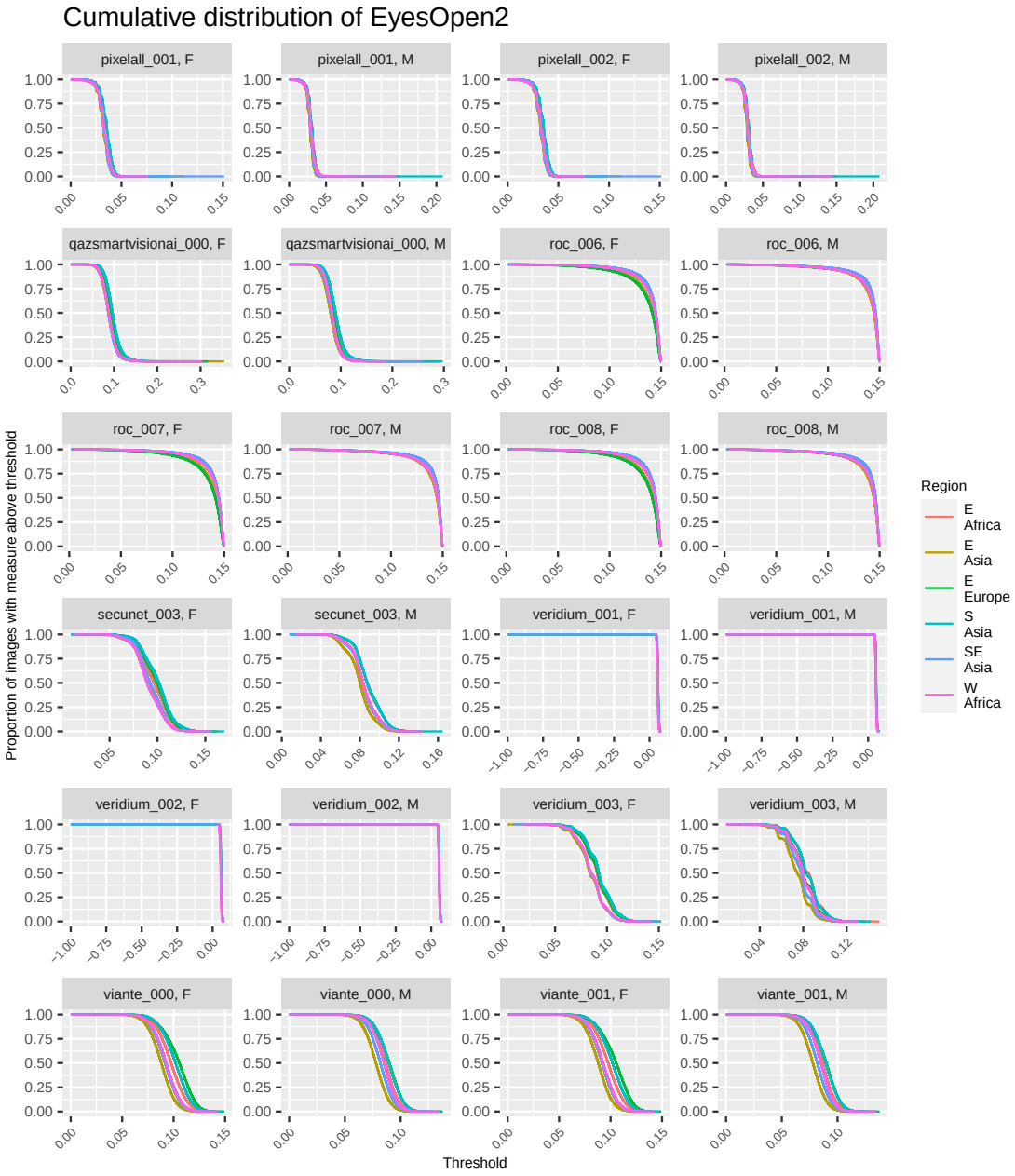


Fig. 77. Cumulative distribution of EyesOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

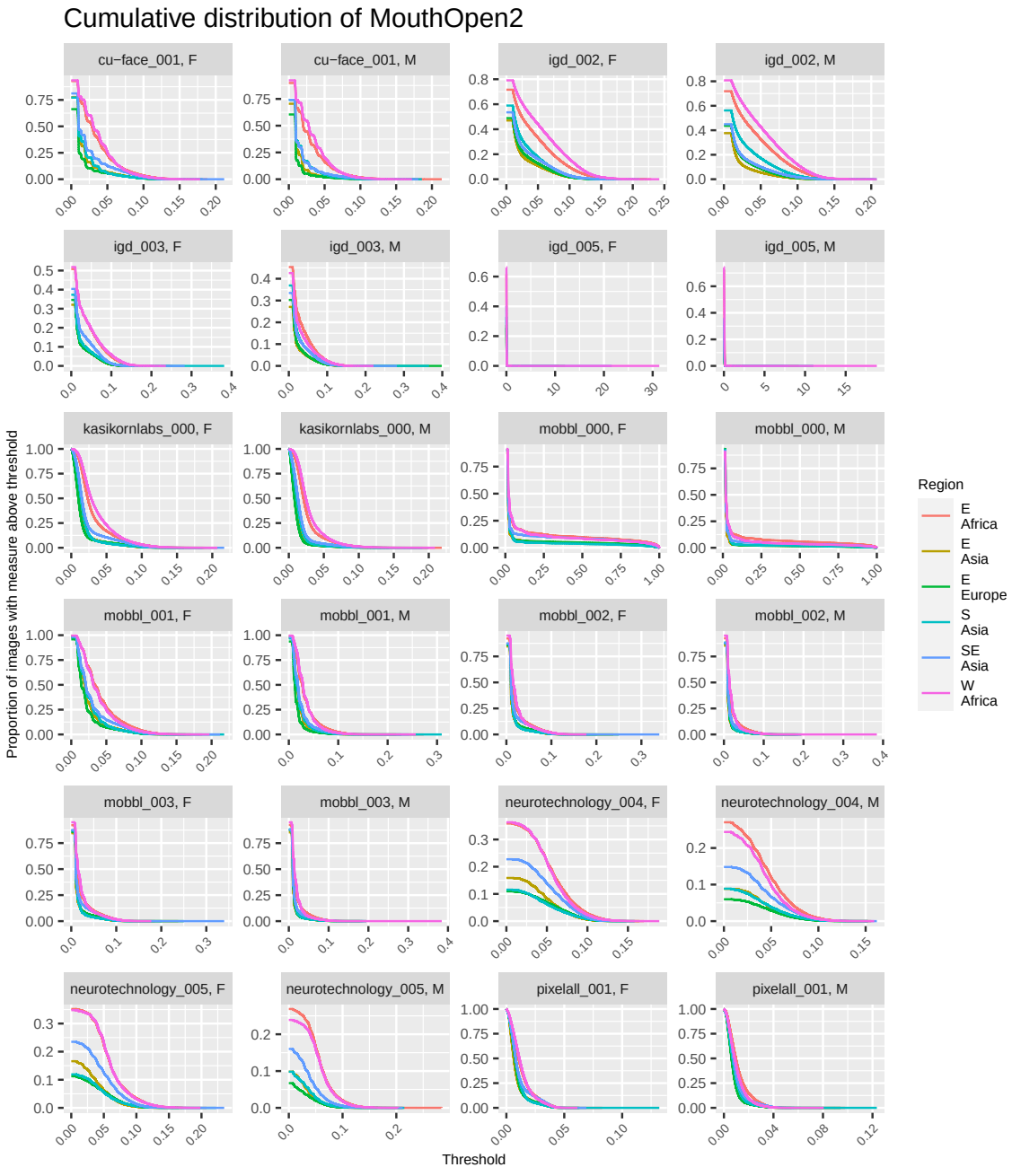


Fig. 78. Cumulative distribution of MouthOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

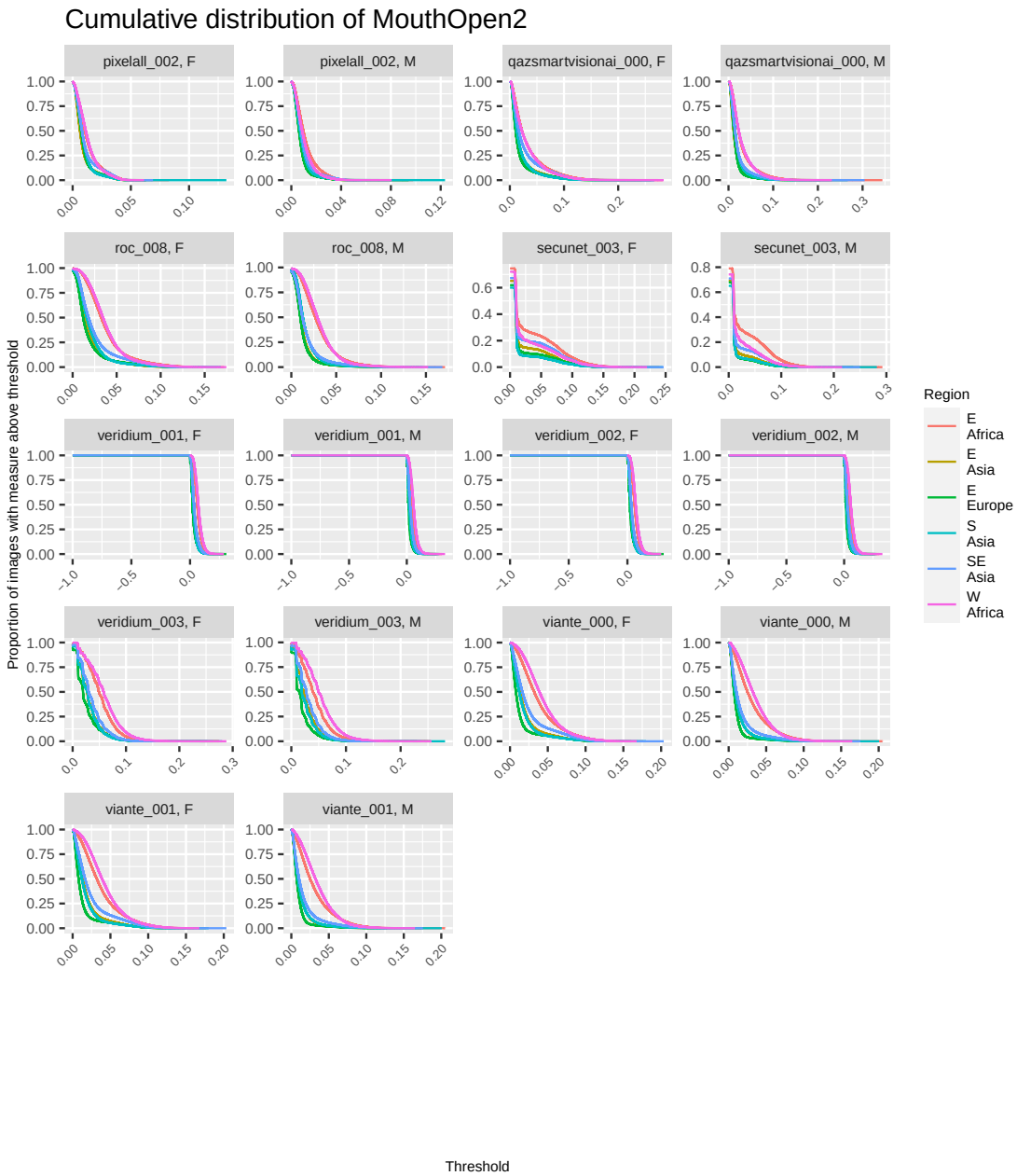


Fig. 79. Cumulative distribution of MouthOpen2 values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

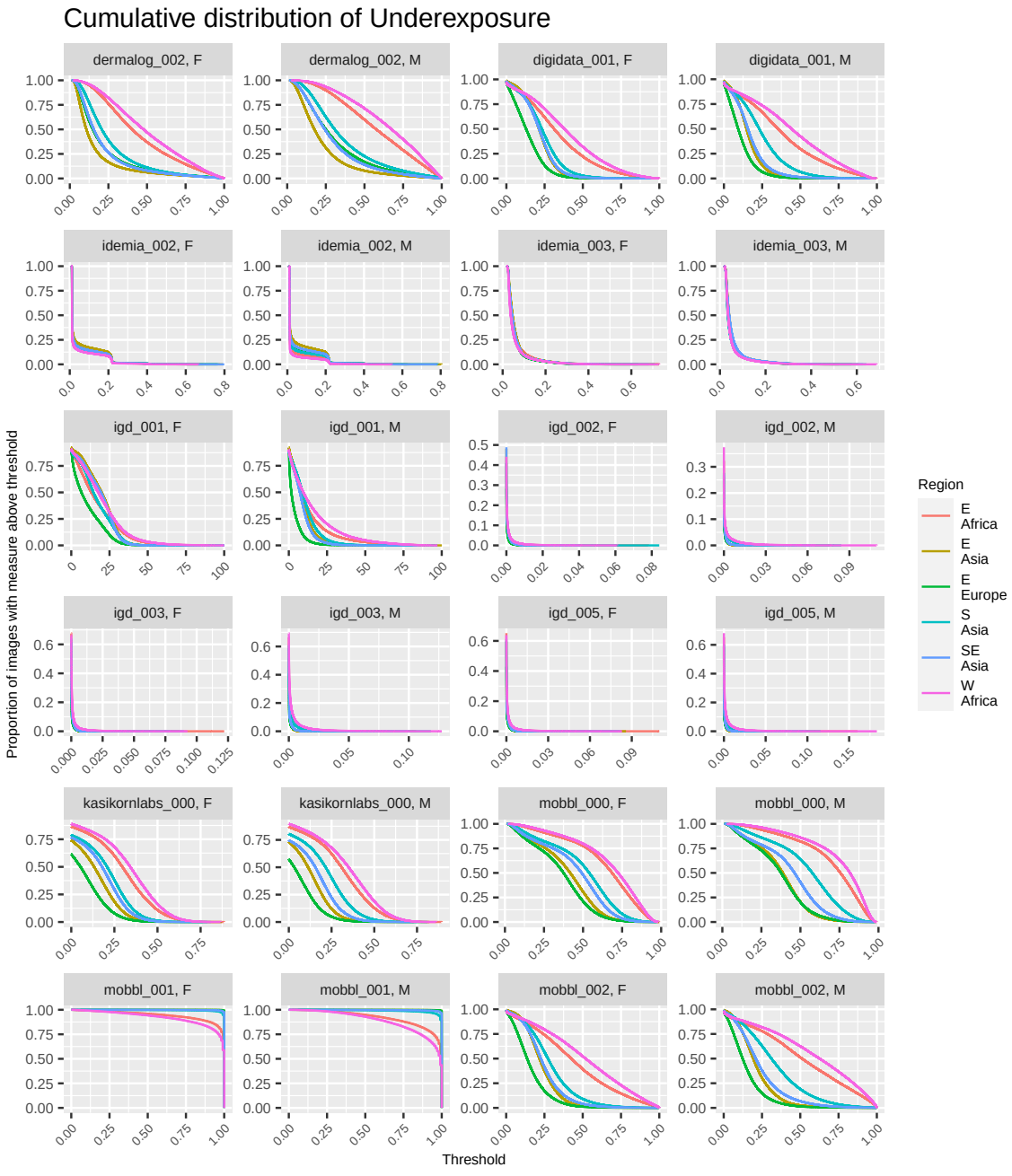


Fig. 80. Cumulative distribution of Underexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

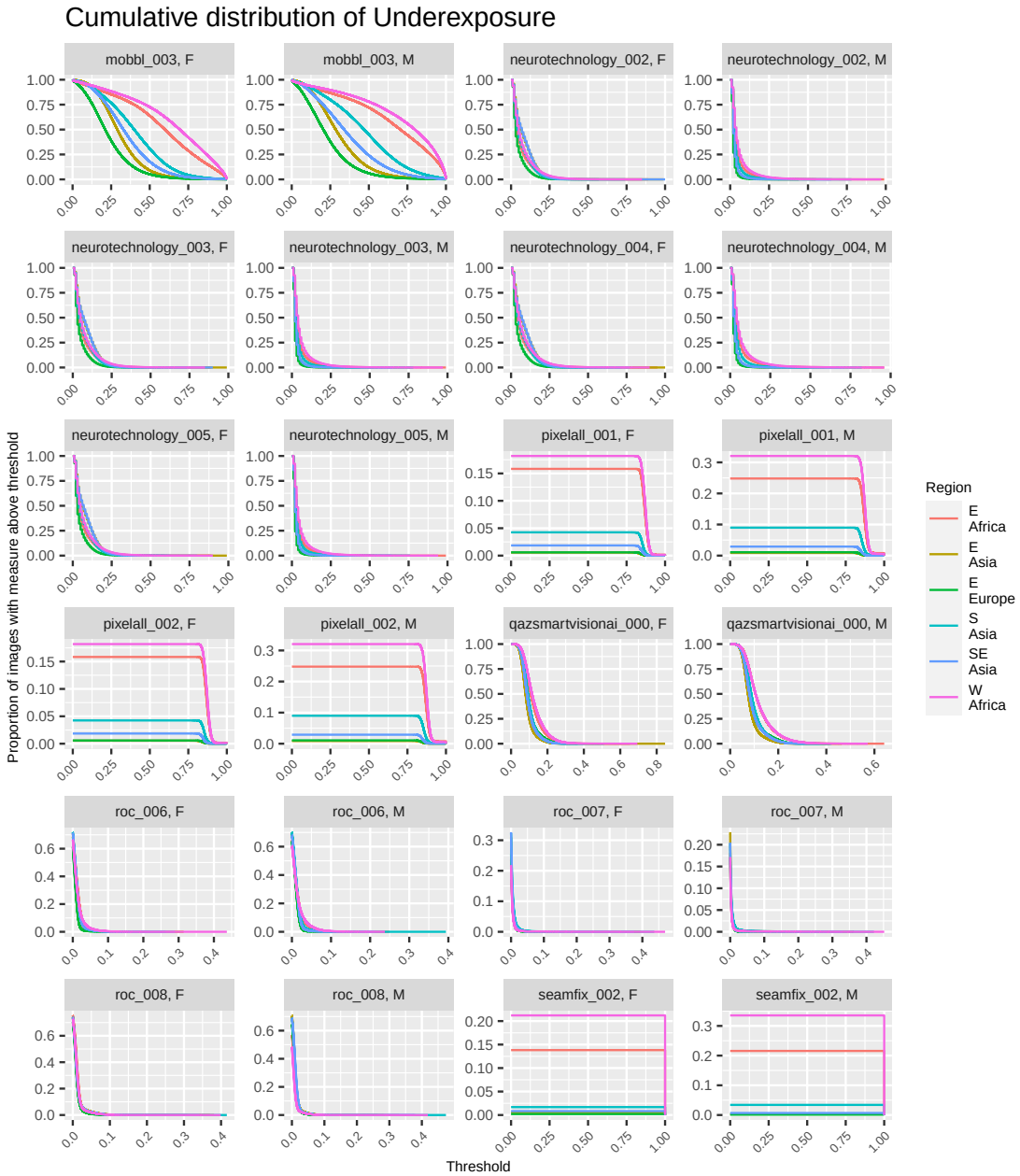


Fig. 81. Cumulative distribution of Underexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

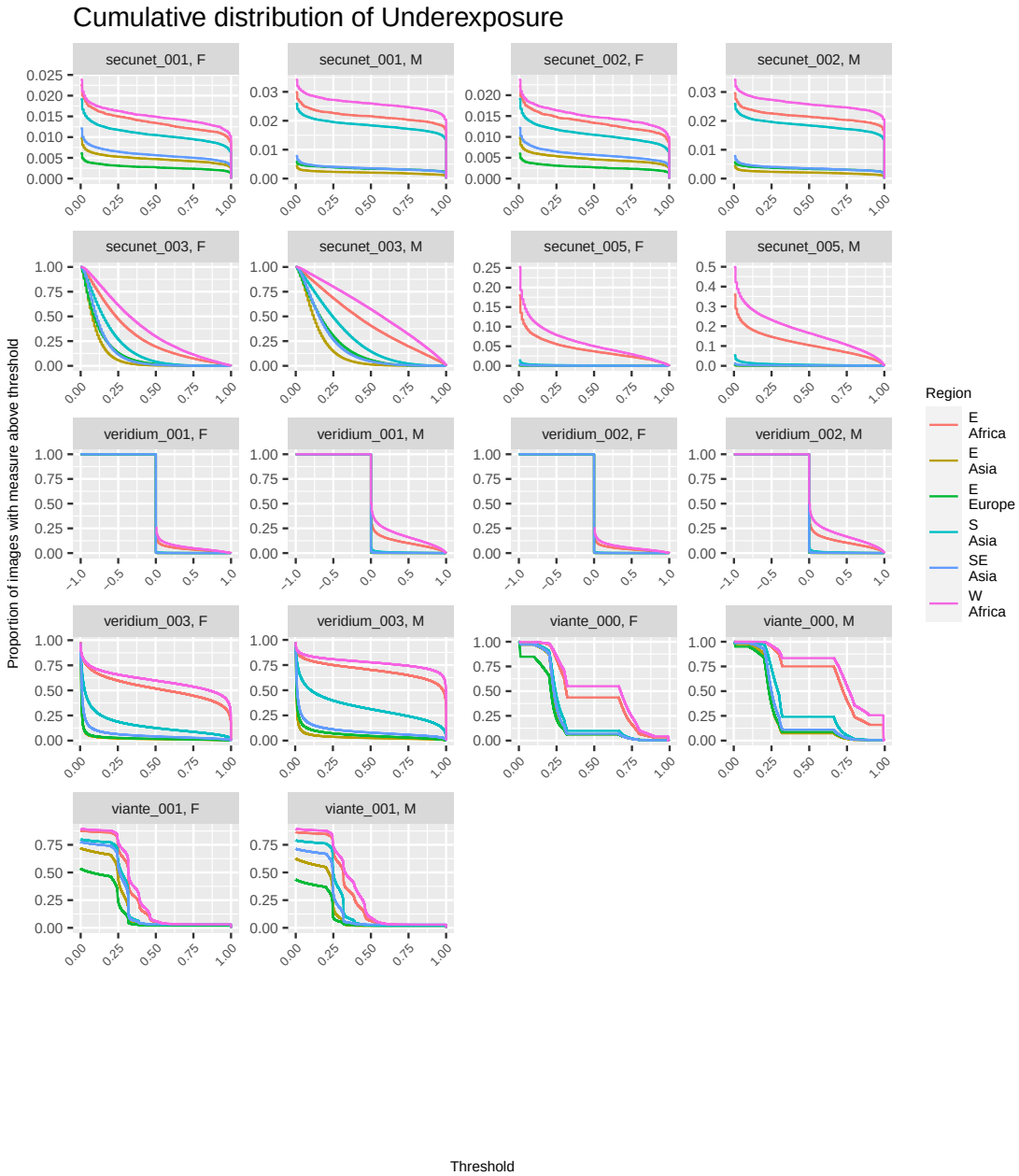


Fig. 82. Cumulative distribution of Underexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

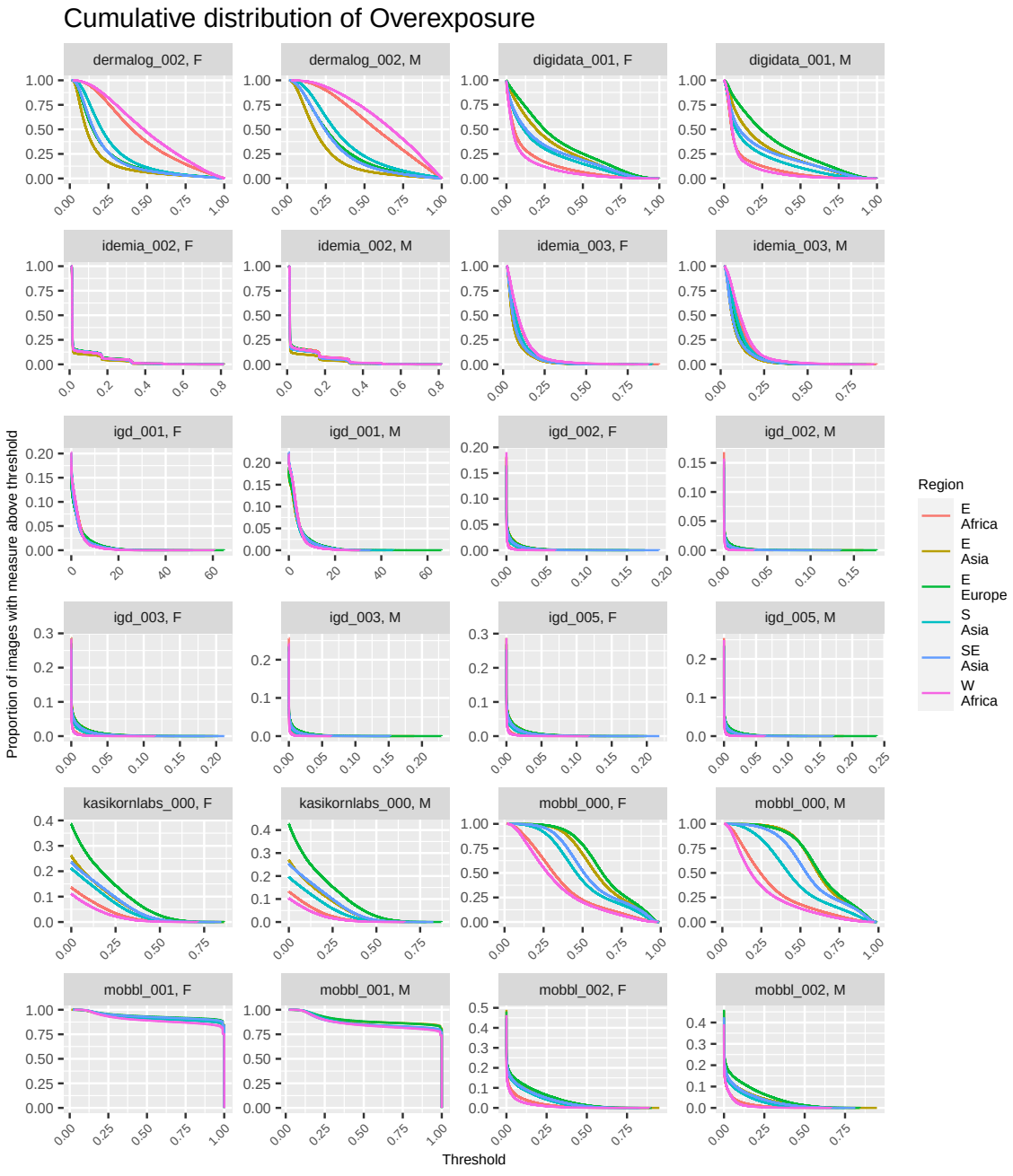


Fig. 83. Cumulative distribution of Overexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

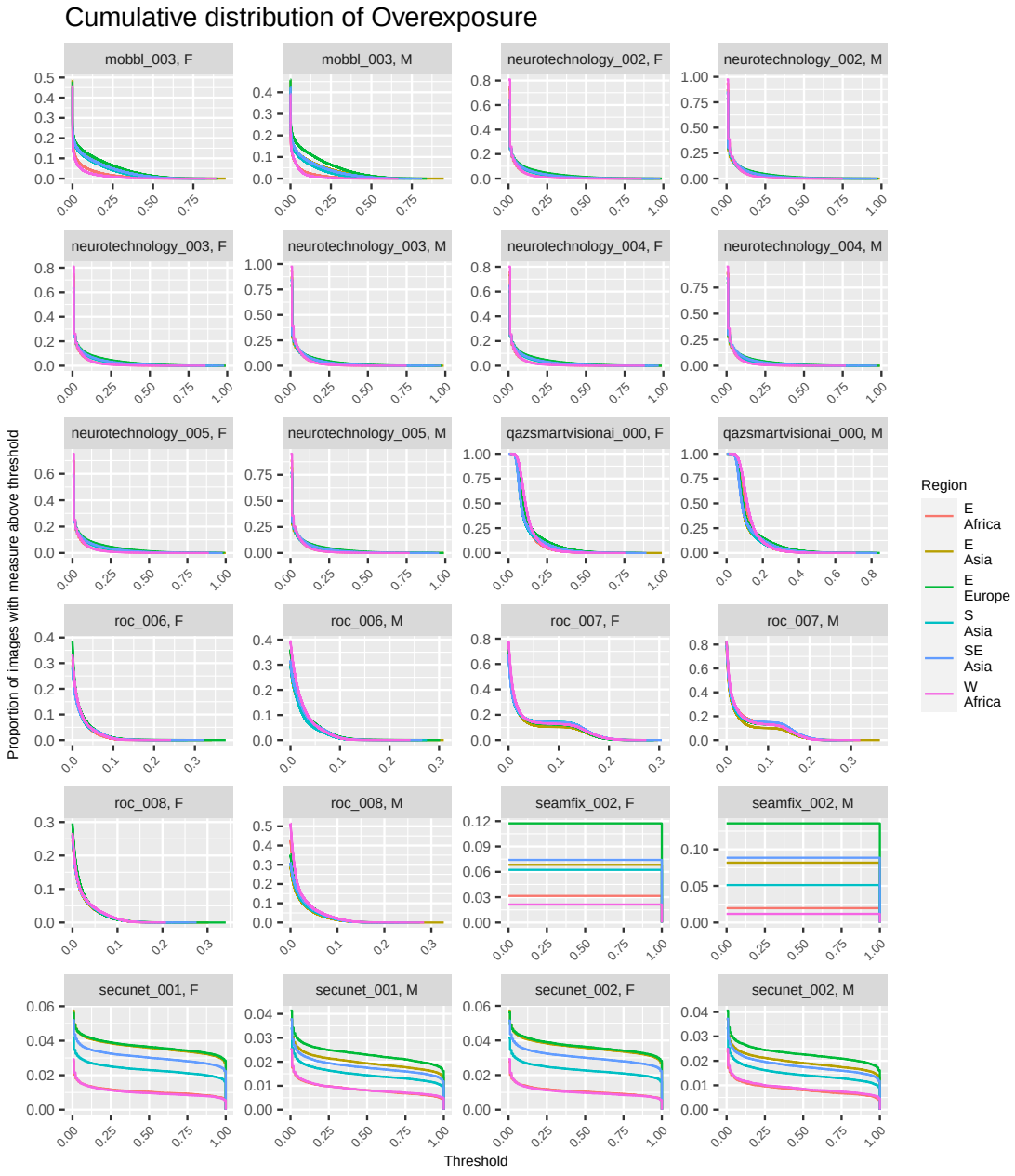


Fig. 84. Cumulative distribution of Overexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

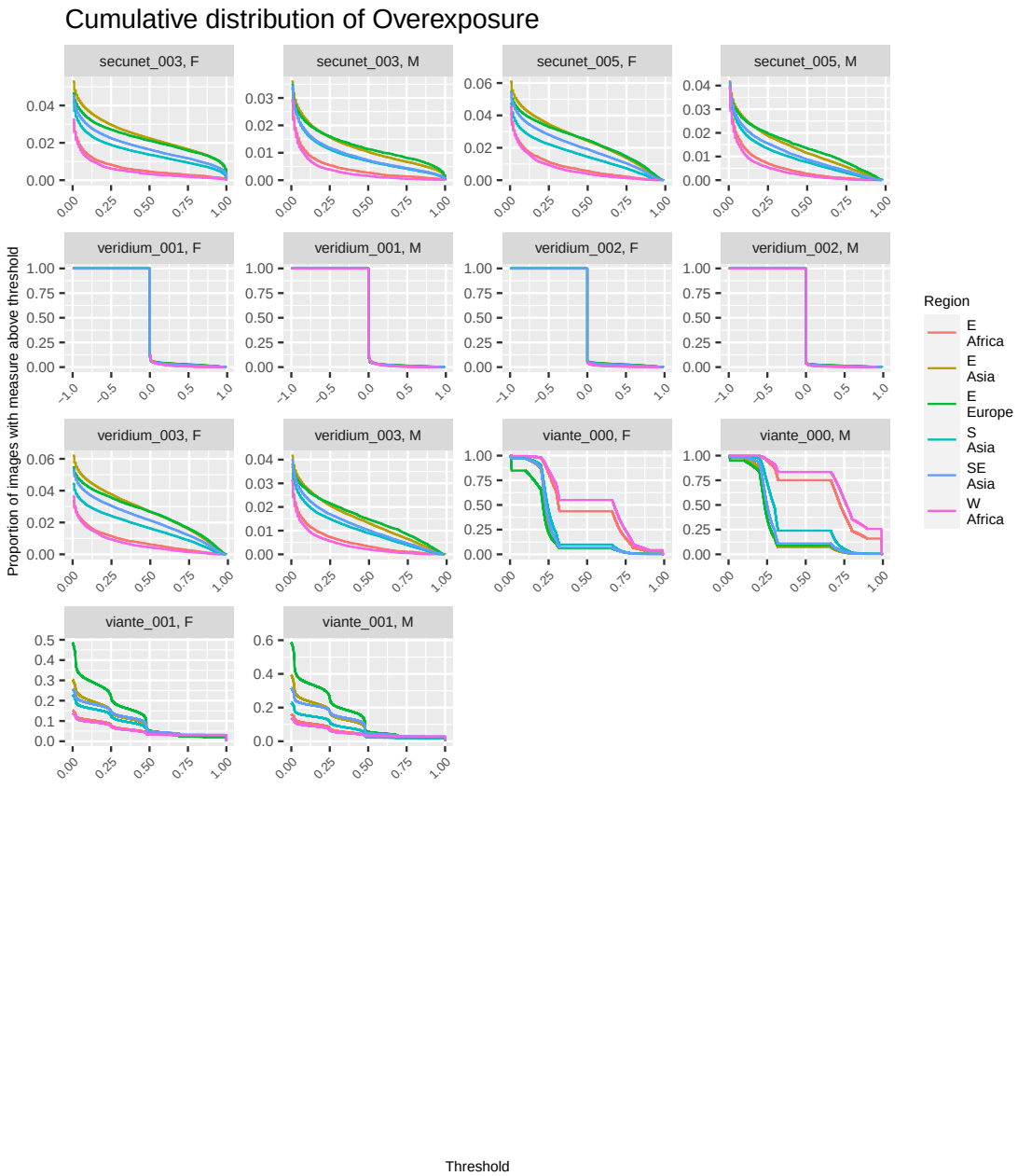


Fig. 85. Cumulative distribution of Overexposure values (by algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

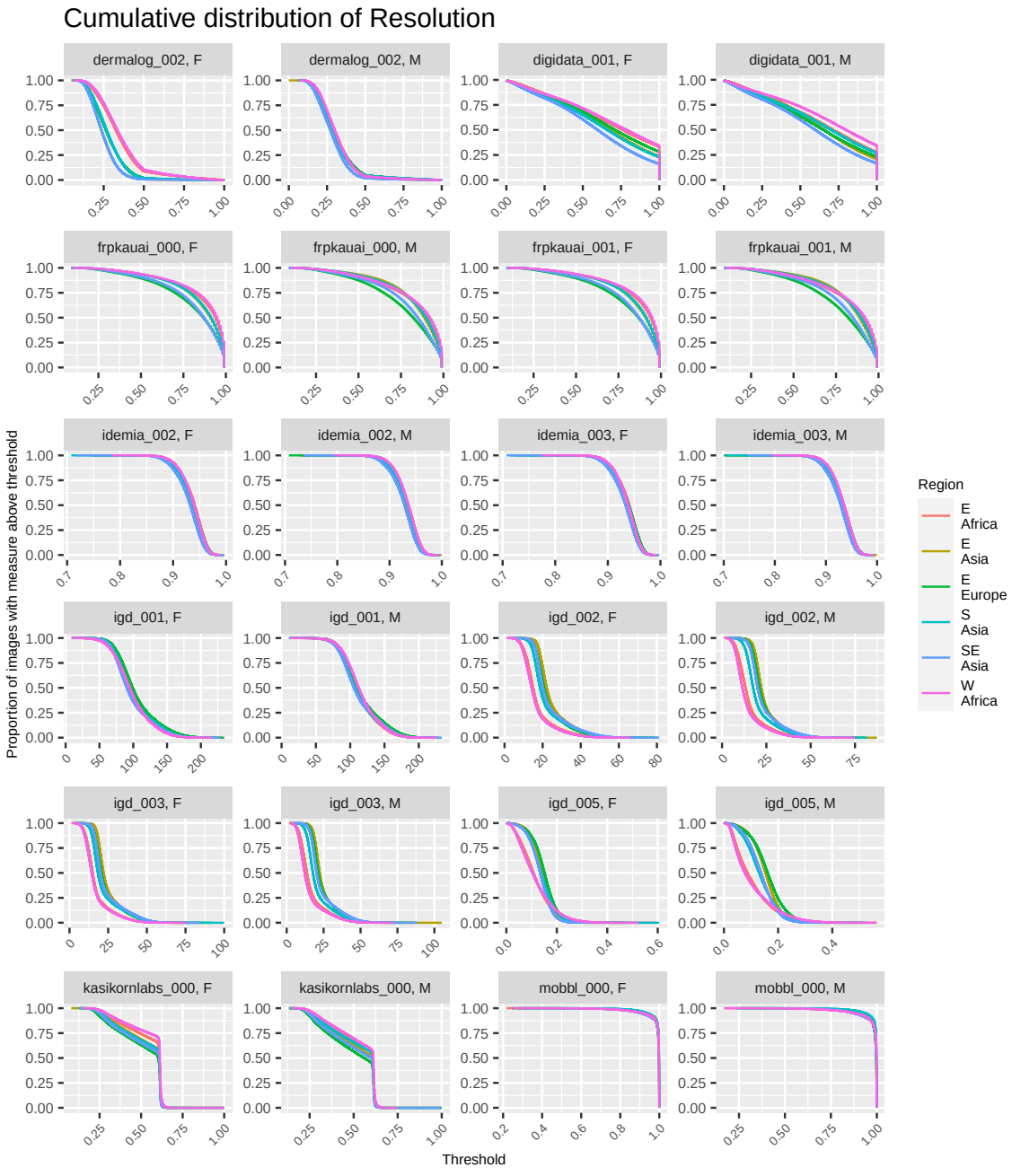


Fig. 86. Cumulative distribution of Resolution values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

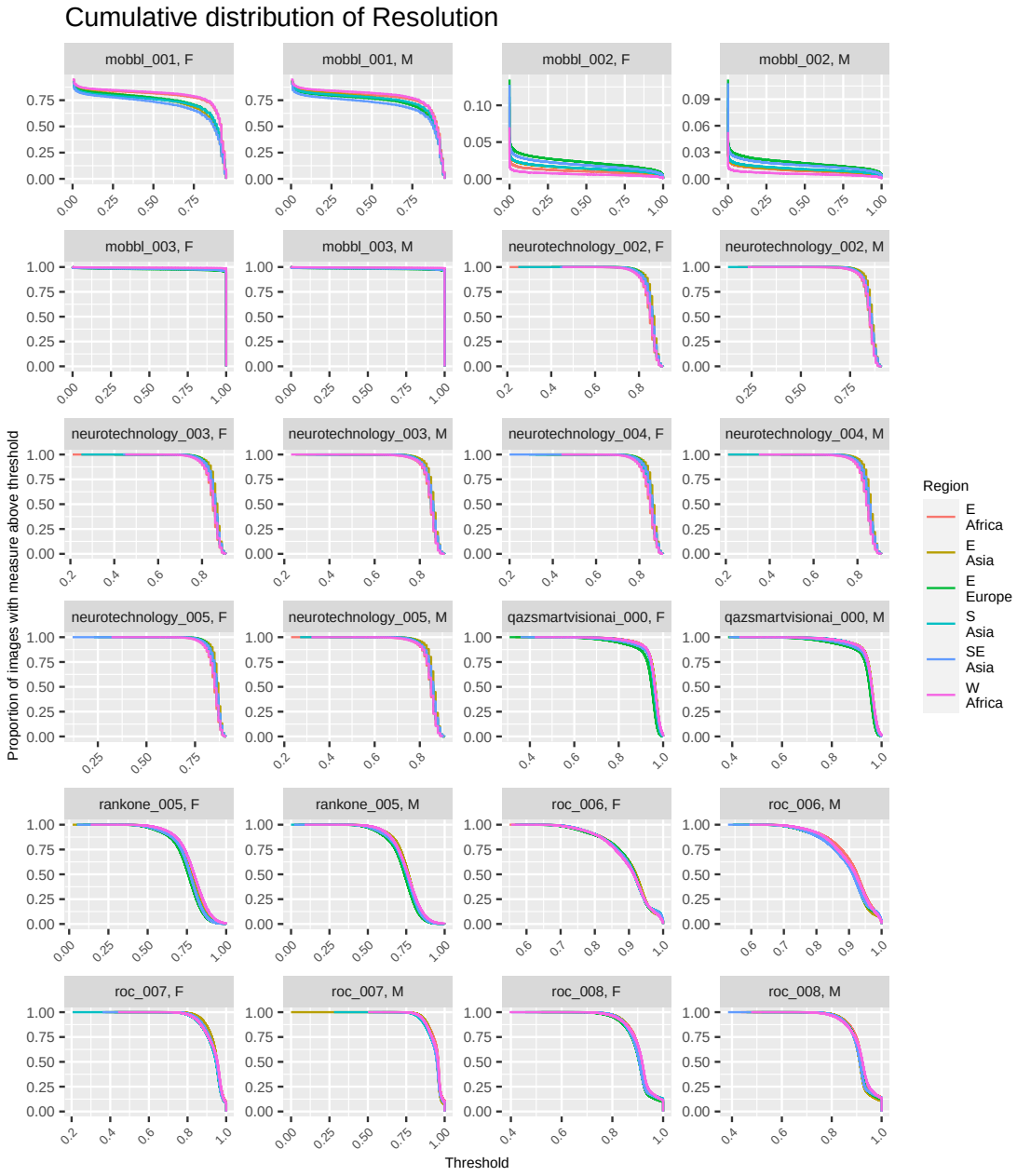


Fig. 87. Cumulative distribution of Resolution values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

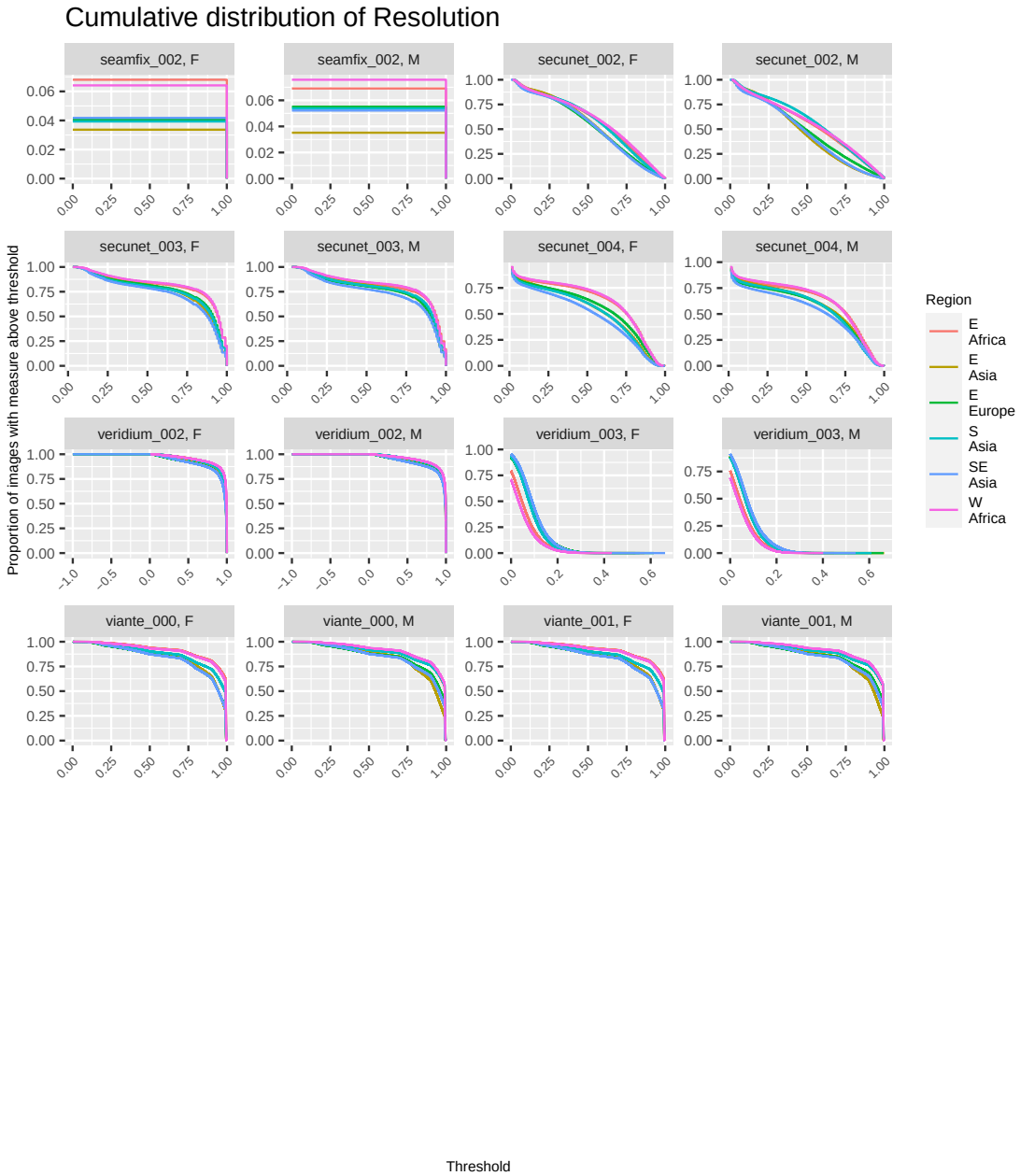


Fig. 88. Cumulative distribution of Resolution values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

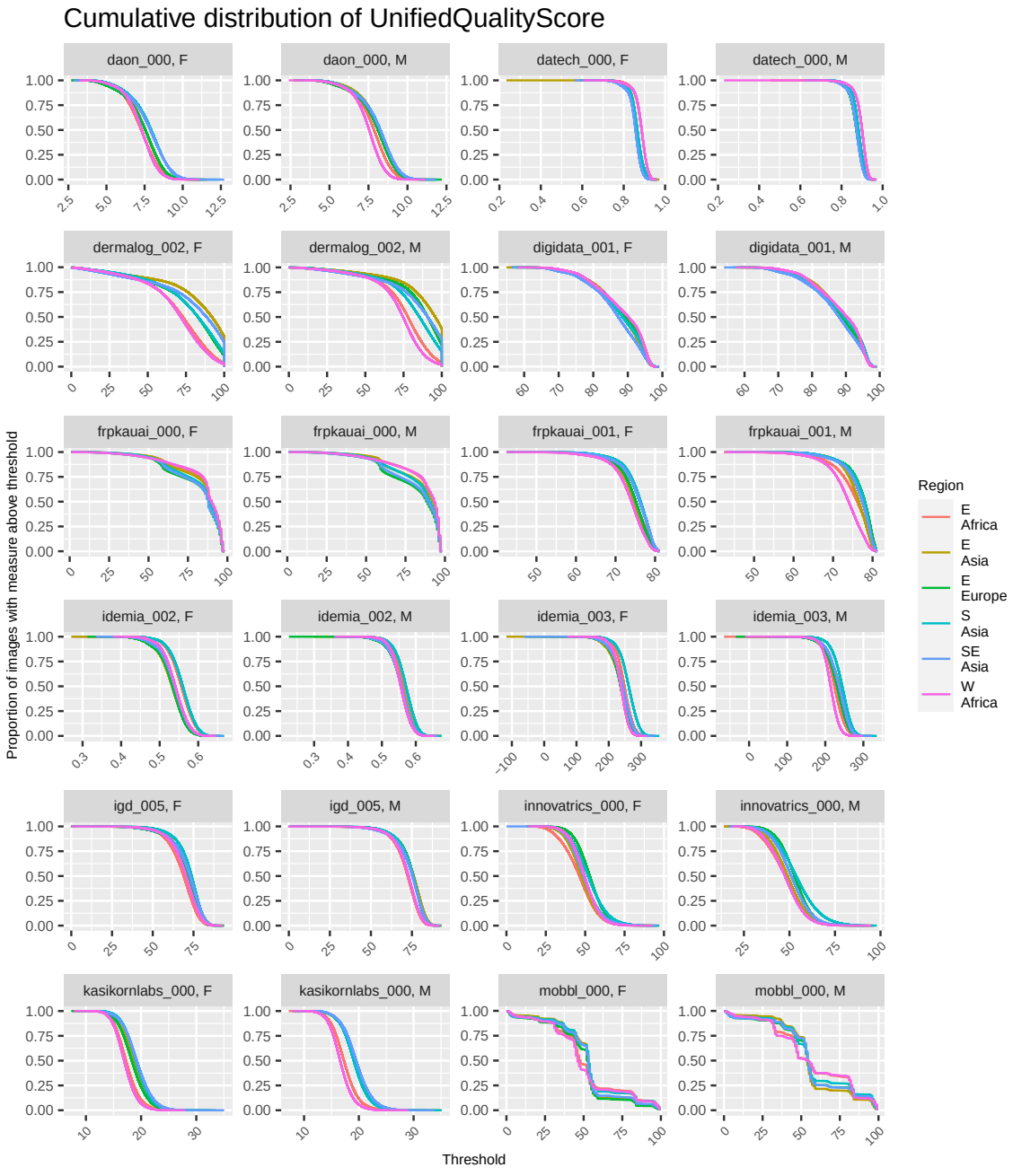


Fig. 89. Cumulative distribution of Unified Quality Score values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

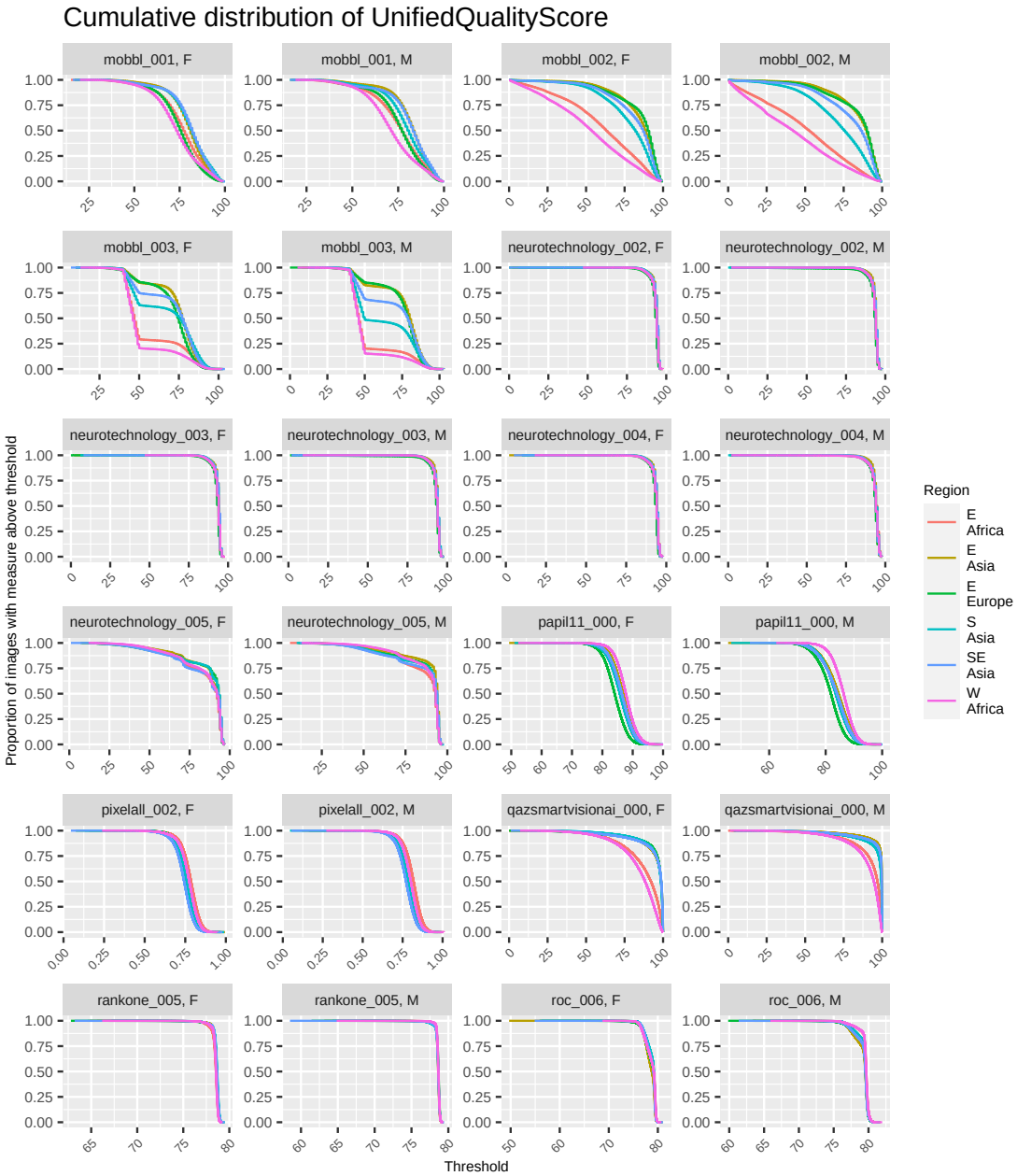


Fig. 90. Cumulative distribution of Unified Quality Score values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.

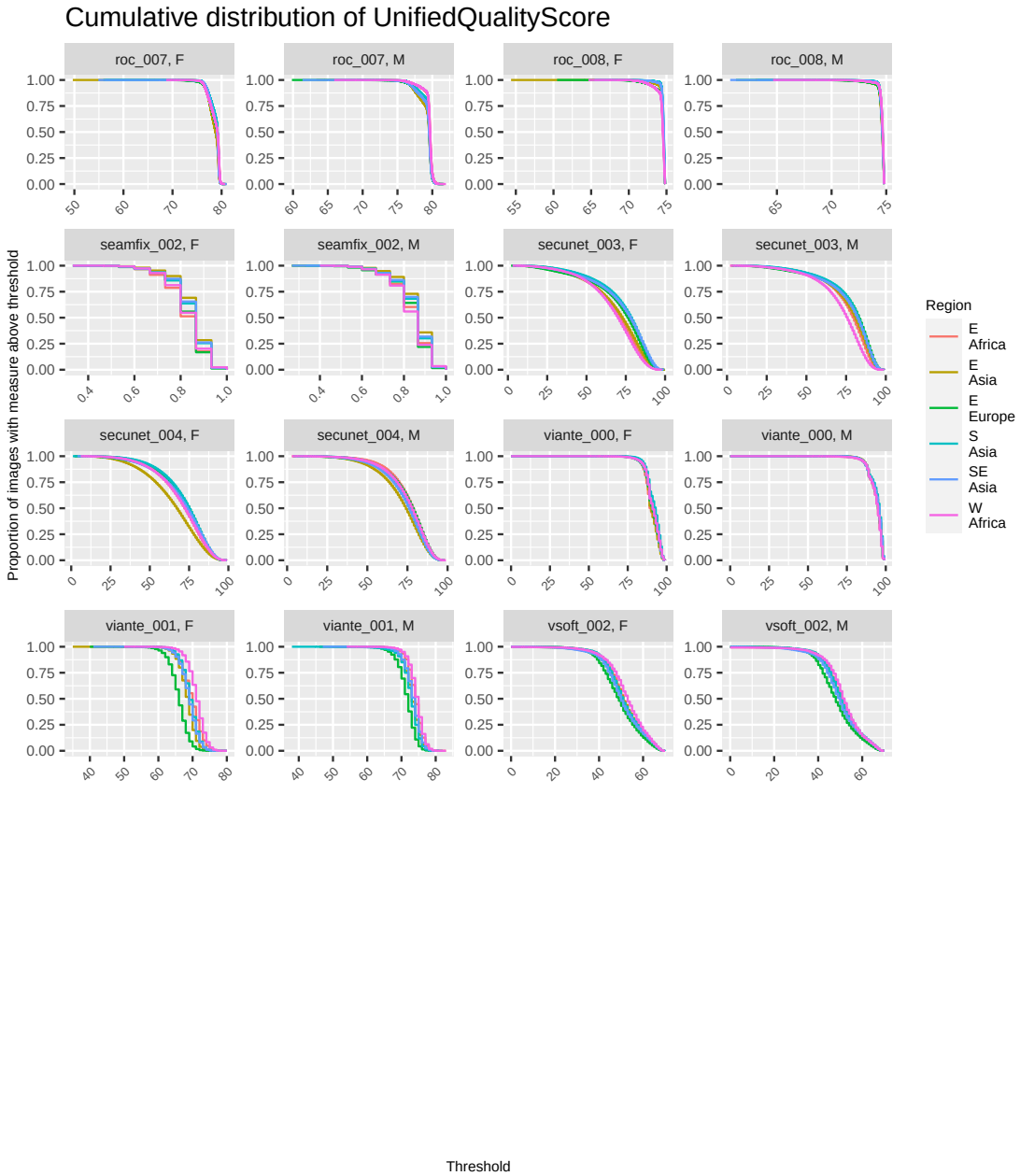


Fig. 91. Cumulative distribution of Unified Quality Score values (over algorithm and sex) for six regions of birth. Ideal performance corresponds to no separation between the curves in a single facet.