FHIBE Bias Evaluation Report

Task: Keypoint Estimation

Model: hrnet_2025June30

Dataset version: DEC24

Reported generated on June 30, 2025

Task Description

The keypoint estimation task evaluates how well a model can predict the location of major body parts in an image of a person's entire body, including their face. We adopt the 17 keypoints from the COCO dataset, and each image is only evaluated on the subset of visible keypoints. This task is evaluated on the full body FHIBE dataset. Each image contains one or two people, with one ground truth set of keypoints per person.

Evaluation Dataset

The model was evaluated using the downsampled FHIBE dataset. This dataset consists of 10319 images comprising one or two people in a background setting. Each image is resized such that the larger dimension has 2048 pixels, while maintaining the aspect ratio of the original, higher resolution image.

Accompanying the dataset is a rich set of metadata. The metadata include self-reported demographic information, such as age, ancestry, and gender pronouns of image subjects, as well as annotator-provided metadata for more objective physical attributes such as pose and subject actions. The metadata also characterize the scene via lighting and weather descriptions. Drawn annotations are also provided for person bounding boxes, face bounding boxes, full body keypoints, person segmentation masks, and face segmentation masks.

The metadata specific to this task include the full body keypoint annotations.

The following attributes were used to aggregate results in this report. For descriptions and definitions of these attributes, see the supplement.

Attributes: pronoun, age, ancestry, apparent_skin_color

Evaluation Results

Here we report the results for each metric that was evaluated.

Metric 1. AR_OKS: Average recall over object keypoint similarity.

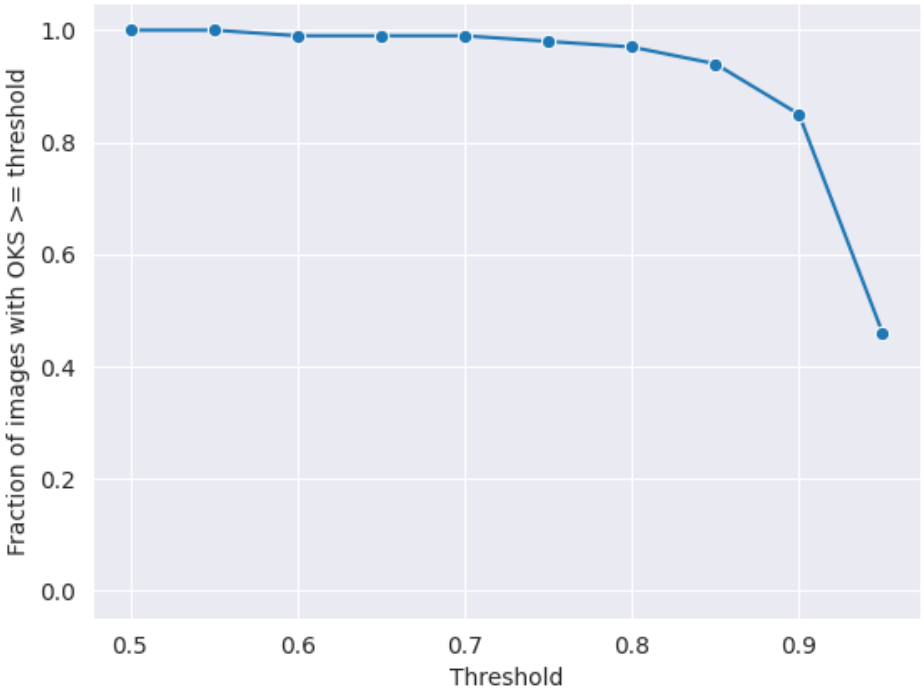
Description:

The object keypoint similarity (OKS) is calculated as in the COCO evaluation dataset: https://cocodataset.org/#keypoints-eval. It has a minimum value (worst) of 0 and a maximum value (best) of 1. At each value in a list of OKS thresholds between 0 and 1, each image is given a value of 1 (correct) or 0 (incorrect) based on whether OKS ≥ threshold. Using these binary outcomes, the recall is calculated. The average recall over all thresholds is reported.

Thresholds: 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95

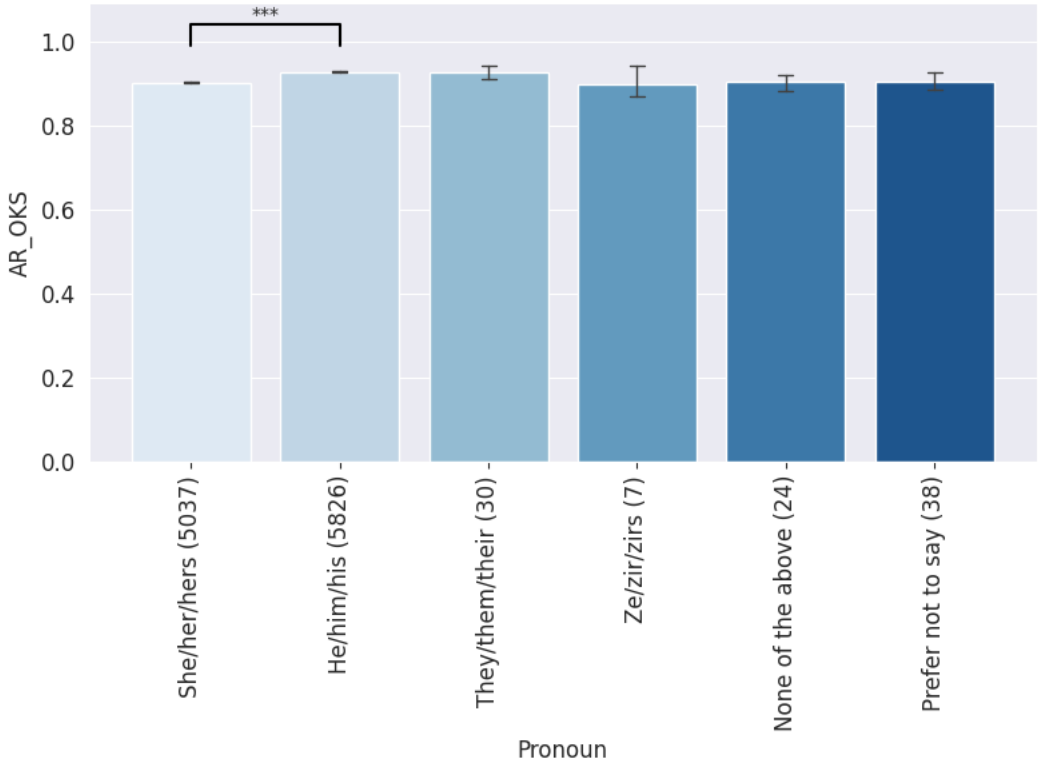
Results:

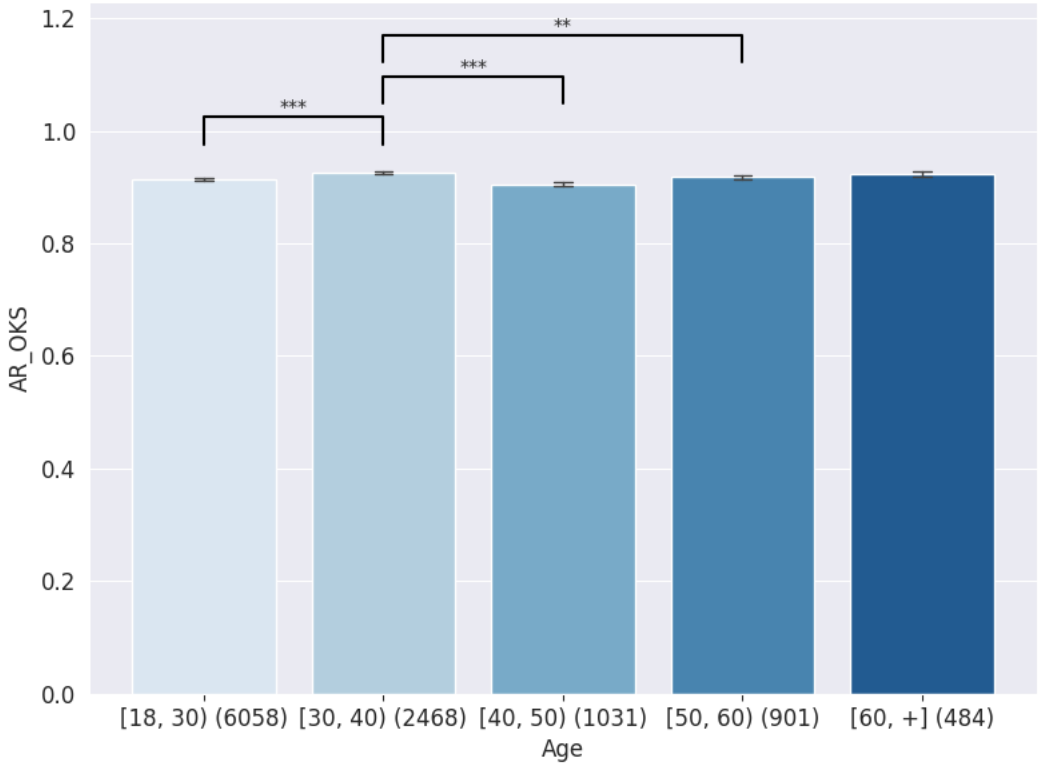
OKS vs. threshold curve calculated over all images

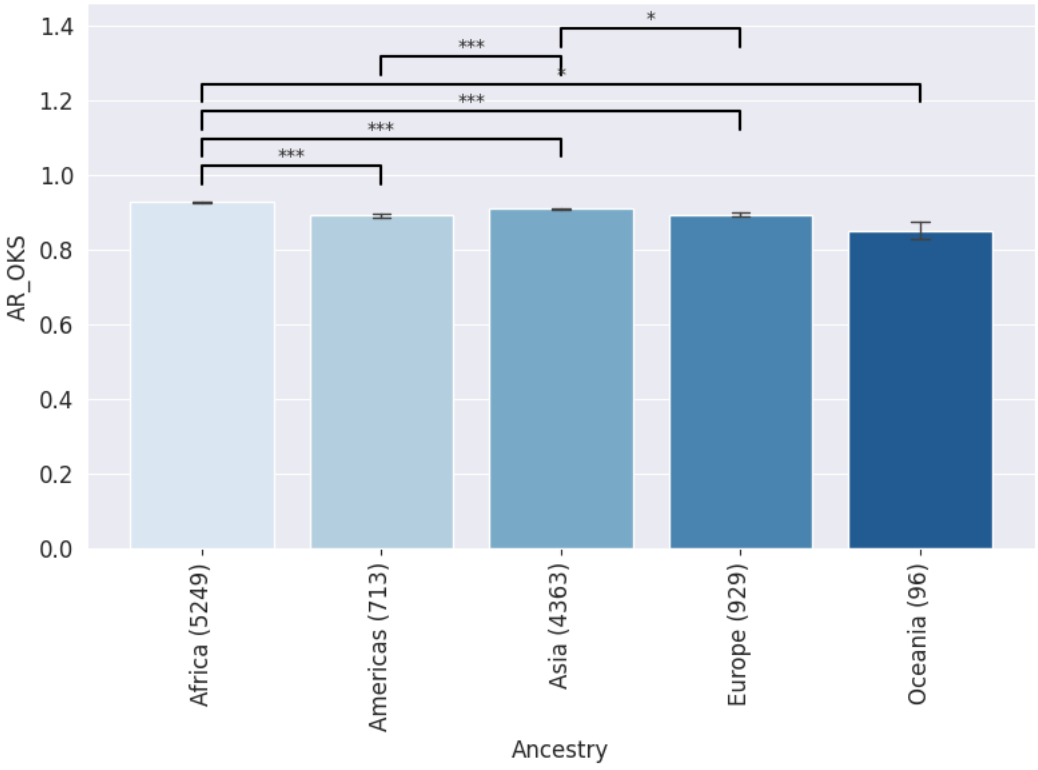


Metric performance aggregated in demographic groups

Bar plots show mean and 68% confidence intervals. Group sizes are included in parentheses in the x-axis labels. Significant differences are indicated with brackets and asterisks, where * indicates p<0.05/m, ** indicates p<0.01/m, and *** indicates p<0.001/m, where p is the p-value and and m is the number of attribute pairs tested for significance (Bonferroni correction). Groups with size < 20 are not included in significance calculations.


***
1.0
0.8
0.6
0.4
0.2
0.0
AR_OKS
She/her/hers (5037)
He/him/his (5826)
They/them/their (30)
Ze/zir/zirs (7)
None of the above (24)
Prefer not to say (38)
Pronoun


**
***
***
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AR_OKS
[18, 30) (6058)
[30, 40) (2468)
[40, 50) (1031)
[50, 60) (901)
[60, +] (484)
Age


AR_OKS
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
*
***
*
***
***
***
Africa (5249)
Americas (713)
Asia (4363)
Europe (929)
Oceania (96)
Ancestry

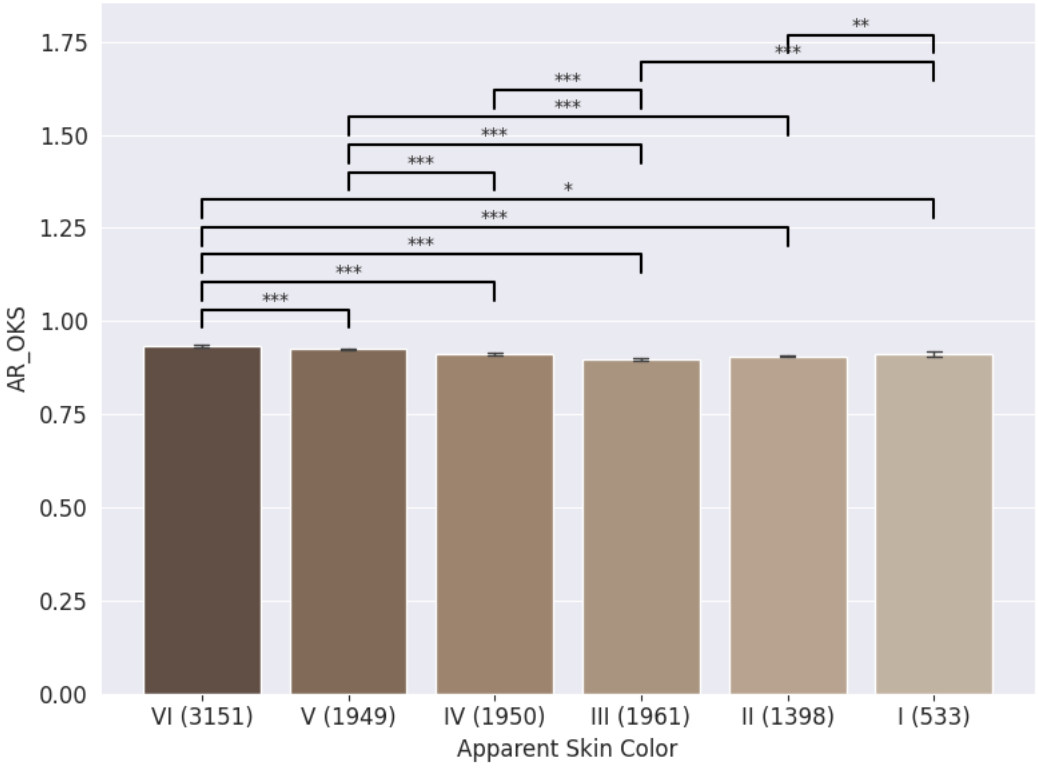

AR_OKS
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
**
***
***
***
***
***
*
***
***
***
VI (3151)
V (1949)
IV (1950)
III (1961)
II (1398)
I (533)
Apparent Skin Color

Table of highest disparity in metric performance across all attribute intersections

Disparity is defined as 1 - (AVG(worst group) / AVG(best group))

Only statistically significant results are shown. Significance is determined via p < alpha/m, where p is the p value, alpha=0.05, and m is the number of attribute pair for a single attribute (or attribute combination).

Attributes with highest disparities in AR_OKS performance							
Attribute(s)	**Worst group**	**Worst mean/median (class size)**	**Best group**	**Best mean/median (class size)**	**Mean/median disparity**	**P value**	**Effect size**
• Pronoun	• She/her/hers	0.90/0.90 (5037)	• He/him/his	0.93/1.00 (5826)	0.03 / 0.10	< 0.001	13.69
• Age	• [18, 30)	0.91/0.90 (6058)	• [30, 40)	0.93/1.00 (2468)	0.01 / 0.10	< 0.001	5.65
• Age	• [40, 50)	0.91/0.90 (1031)	• [30, 40)	0.93/1.00 (2468)	0.02 / 0.10	< 0.001	5.78
• Age	• [50, 60)	0.92/0.90 (901)	• [30, 40)	0.93/1.00 (2468)	0.01 / 0.10	< 0.001	3.20
• Ancestry	• Americas	0.89/0.90 (713)	• Africa	0.93/1.00 (5249)	0.04 / 0.10	< 0.001	10.51
• Ancestry	• Asia	0.91/0.90 (4363)	• Africa	0.93/1.00 (5249)	0.02 / 0.10	< 0.001	11.08
• Ancestry	• Europe	0.89/0.90 (929)	• Africa	0.93/1.00 (5249)	0.04 / 0.10	< 0.001	9.21
• Ancestry	• Oceania	0.85/0.90 (96)	• Africa	0.93/1.00 (5249)	0.08 / 0.10	0.003	2.70
• Apparent Skin Color	• V	0.92/0.90 (1949)	• VI	0.93/1.00 (3151)	0.01 / 0.10	< 0.001	4.98
• Apparent Skin Color	• IV	0.91/0.90 (1950)	• VI	0.93/1.00 (3151)	0.02 / 0.10	< 0.001	9.26
• Apparent Skin Color	• III	0.90/0.90 (1961)	• VI	0.93/1.00 (3151)	0.04 / 0.10	< 0.001	13.34
• Apparent Skin Color	• II	0.91/0.90 (1398)	• VI	0.93/1.00 (3151)	0.03 / 0.10	< 0.001	9.37
• Apparent Skin Color	• I	0.91/0.90 (533)	• VI	0.93/1.00 (3151)	0.02 / 0.10	0.003	2.68
• Ancestry	• Asia	0.91/0.90 (4363)	• Americas	0.89/0.90 (713)	-0.02 / 0.00	< 0.001	4.94
• Ancestry	• Europe	0.89/0.90 (929)	• Asia	0.91/0.90 (4363)	0.02 / 0.00	0.002	2.89
• Apparent Skin Color	• IV	0.91/0.90 (1950)	• V	0.92/0.90 (1949)	0.01 / 0.00	< 0.001	3.92
• Apparent Skin Color	• III	0.90/0.90 (1961)	• V	0.92/0.90 (1949)	0.03 / 0.00	< 0.001	7.77
• Apparent Skin Color	• II	0.91/0.90 (1398)	• V	0.92/0.90 (1949)	0.02 / 0.00	< 0.001	4.72
• Apparent Skin Color	• III	0.90/0.90 (1961)	• IV	0.91/0.90 (1950)	0.02 / 0.00	< 0.001	3.93
• Apparent Skin Color	• I	0.91/0.90 (533)	• III	0.90/0.90 (1961)	-0.02 / 0.00	< 0.001	5.06

• Apparent Skin Color	• I	0.91/0.90 (533)	• II	0.91/0.90 (1398)	-0.01 / 0.00	< 0.001	3.28

Metric 2. PCK: Percentage correct keypoints.

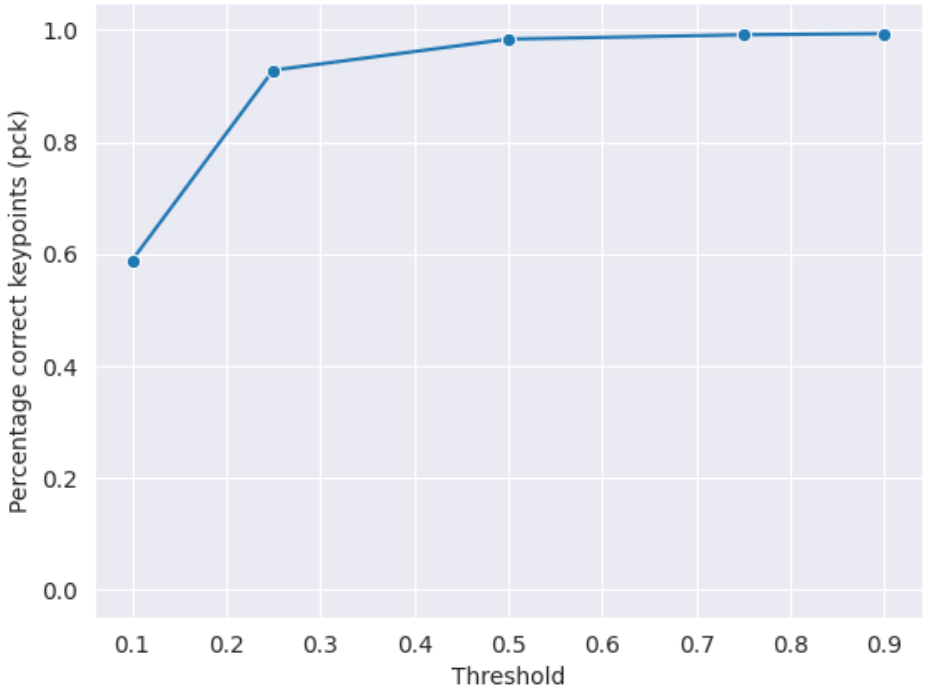
Description:

The distance between each ground truth keypoint and the closest predicted keypoint is compared to the product `thresh*face_bbox_diag`, where `thresh` is a threshold value and `face_bbox_diag` is the length of the diagonal of the ground truth face bounding box. If the distance is less than the product, a keypoint is considered correct. The fraction of correct keypoints in the set of ground truth keypoints in an image is the PCK at a single threshold for a single image. This is repeated for each threshold in a list of thresholds, and the mean over all thresholds is reported. PCK has a minimum value (worst) of 0 and maximum value (best) of 1.

Thresholds: 0.10, 0.25, 0.50, 0.75, 0.90

Results:

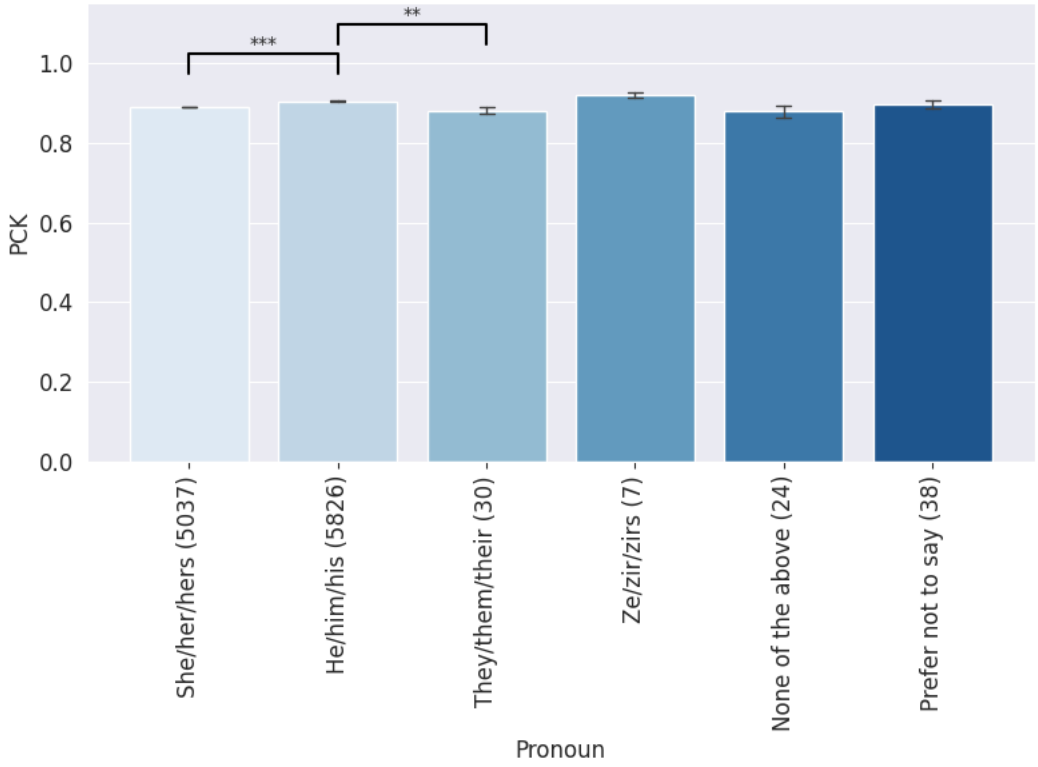
PCK vs. threshold curve calculated over all images

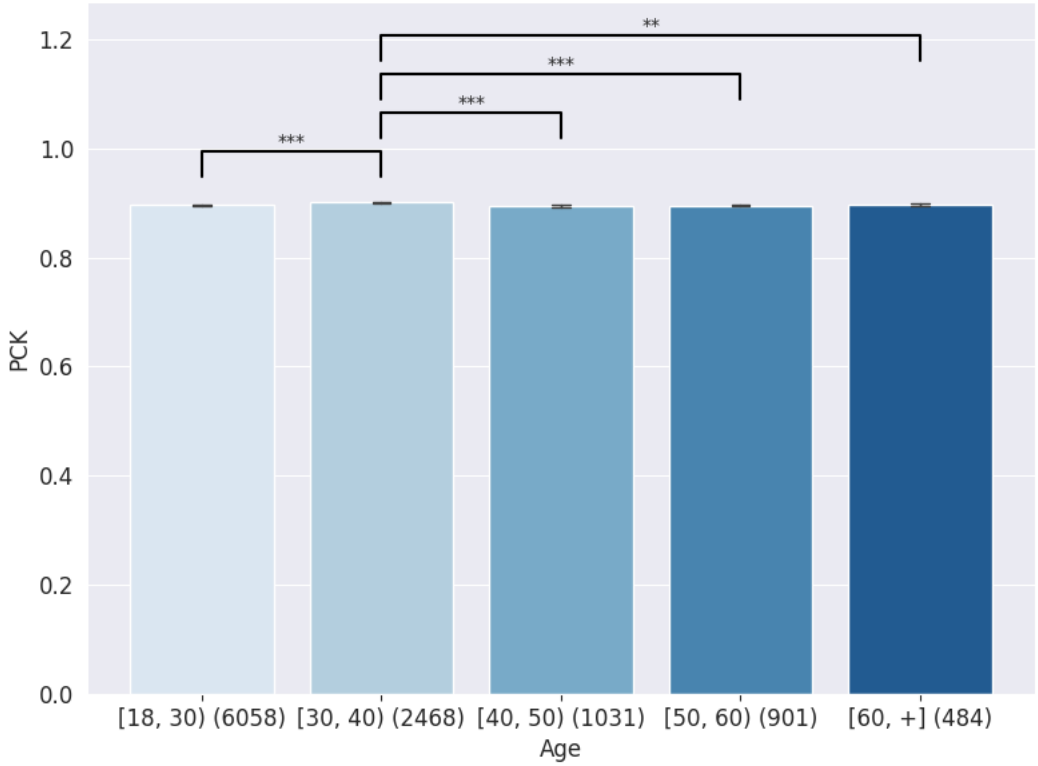


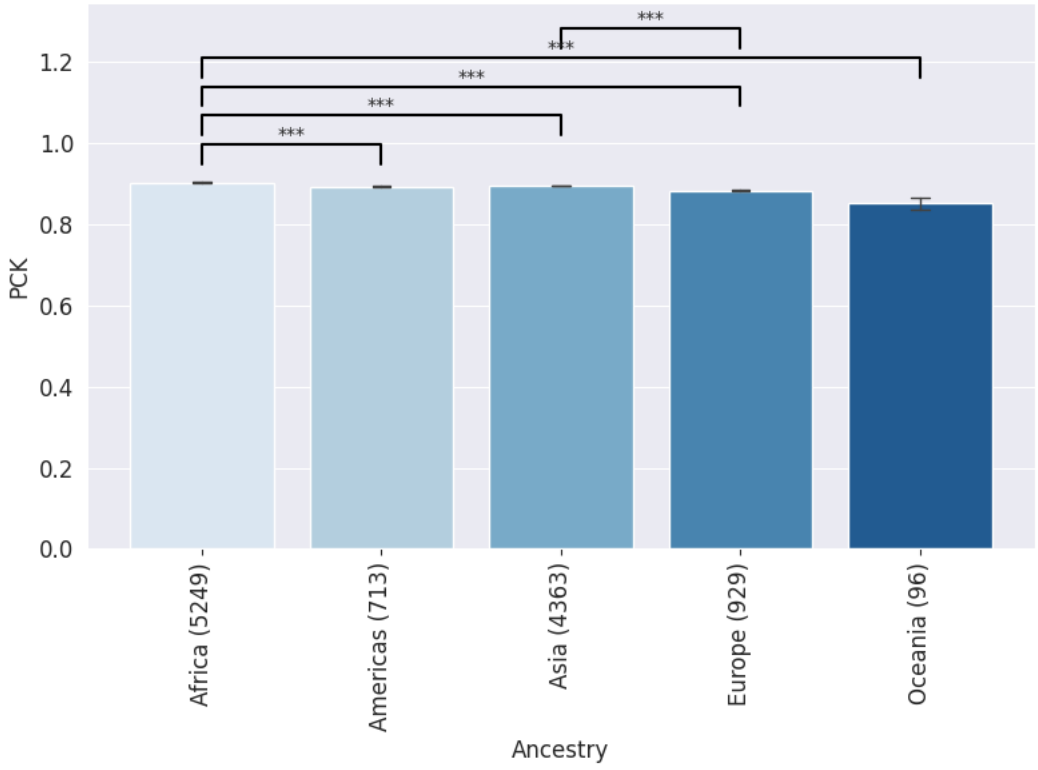
Metric performance aggregated in demographic groups

Bar plots show mean and 68% confidence intervals. Group sizes are included in parentheses in the x-axis labels. Significant differences are indicated with brackets and asterisks, where * indicates p<0.05/m, ** indicates p<0.01/m, and *** indicates p<0.001/m, where p is the p-value and and m is the number of attribute pairs tested for significance (Bonferroni correction). Groups with size < 20 are not

included in significance calculations.






***
***
***
***
***
1.2
1.0
0.8
0.6
0.4
0.2
0.0
PCK
Africa (5249)
Americas (713)
Asia (4363)
Europe (929)
Oceania (96)
Ancestry

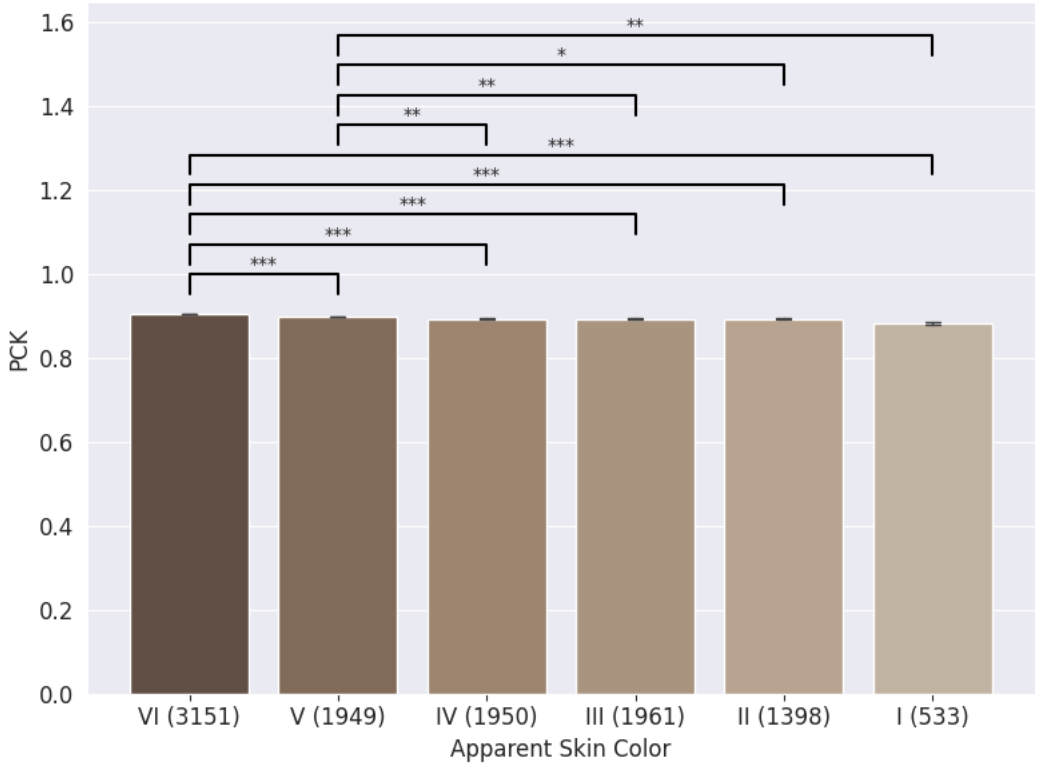

**
*
**
**
***
***
***
***
***
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
PCK
VI (3151)
V (1949)
IV (1950)
III (1961)
II (1398)
I (533)
Apparent Skin Color

Table of highest disparity in metric performance across all attribute intersections

Disparity is defined as 1 - (AVG(worst group) / AVG(best group))

Only statistically significant results are shown. Significance is determined via p < alpha/m, where p is the p value, alpha=0.05, and m is the number of attribute pair for a single attribute (or attribute combination).

Attributes with highest disparities in PCK performance							
Attribute(s)	**Worst group**	**Worst mean/median (class size)**	**Best group**	**Best mean/median (class size)**	**Mean/median disparity**	**P value**	**Effect size**
• Pronoun	• They/them/their	0.88/0.88 (30)	• He/him/his	0.90/0.91 (5826)	0.03 / 0.03	< 0.001	3.29
• Ancestry	• Oceania	0.85/0.89 (96)	• Africa	0.90/0.91 (5249)	0.06 / 0.02	< 0.001	4.50
• Ancestry	• Europe	0.88/0.89 (929)	• Africa	0.90/0.91 (5249)	0.02 / 0.02	< 0.001	11.41
• Apparent Skin Color	• IV	0.89/0.90 (1950)	• VI	0.91/0.92 (3151)	0.01 / 0.02	< 0.001	10.86
• Apparent Skin Color	• III	0.89/0.90 (1961)	• VI	0.91/0.92 (3151)	0.01 / 0.02	< 0.001	10.78
• Apparent Skin Color	• II	0.89/0.90 (1398)	• VI	0.91/0.92 (3151)	0.01 / 0.02	< 0.001	9.15
• Apparent Skin Color	• I	0.88/0.90 (533)	• VI	0.91/0.92 (3151)	0.02 / 0.02	< 0.001	7.98
• Pronoun	• She/her/hers	0.89/0.90 (5037)	• He/him/his	0.90/0.91 (5826)	0.02 / 0.02	< 0.001	17.71
• Age	• [40, 50)	0.89/0.90 (1031)	• [30, 40)	0.90/0.91 (2468)	0.01 / 0.01	< 0.001	5.10
• Ancestry	• Americas	0.89/0.90 (713)	• Africa	0.90/0.91 (5249)	0.01 / 0.01	< 0.001	7.16
• Ancestry	• Asia	0.89/0.90 (4363)	• Africa	0.90/0.91 (5249)	0.01 / 0.01	< 0.001	10.81
• Apparent Skin Color	• V	0.90/0.91 (1949)	• VI	0.91/0.92 (3151)	0.01 / 0.01	< 0.001	6.94
• Apparent Skin Color	• IV	0.89/0.90 (1950)	• V	0.90/0.91 (1949)	0.01 / 0.01	< 0.001	3.57
• Apparent Skin Color	• III	0.89/0.90 (1961)	• V	0.90/0.91 (1949)	0.01 / 0.01	< 0.001	3.77
• Apparent Skin Color	• II	0.89/0.90 (1398)	• V	0.90/0.91 (1949)	0.01 / 0.01	0.003	2.94
• Apparent Skin Color	• I	0.88/0.90 (533)	• V	0.90/0.91 (1949)	0.02 / 0.01	< 0.001	3.75
• Age	• [60, +]	0.90/0.91 (484)	• [30, 40)	0.90/0.91 (2468)	0.01 / 0.01	< 0.001	3.35
• Ancestry	• Europe	0.88/0.89 (929)	• Asia	0.89/0.90 (4363)	0.01 / 0.01	< 0.001	5.42
• Age	• [18, 30)	0.90/0.91 (6058)	• [30, 40)	0.90/0.91 (2468)	0.01 / 0.01	< 0.001	5.30
• Age	• [50, 60)	0.90/0.91 (901)	• [30, 40)	0.90/0.91 (2468)	0.01 / 0.01	< 0.001	4.23

Supplement

In this section, we describe the attributes that were used to aggregate results in this report. The following attributes were grouped at the continent level: ancestry.

- Gender pronouns: self-reported. Multiple selections are allowed, except when 'None of the above' or 'Prefer not to say' is selected. Subjects select from the following options:
  'She/her/hers', 'He/him/his', 'They/them/their', 'Ze/zir/zirs', 'None of the above', 'Prefer not to say'
- Age: self-reported age between 0 and 130. We aggregate results in five age bins:
  [18, 30), [30, 40), [40, 50), [50, 60), [60, +]
- Ancestry: self-reported. Subjects select from a list of 27 geographic regions. Multiple selections are allowed.
- Apparent skin color: self-reported at time of image capture. Subjects select a single value from the six-point Fitzpatrick skin type scale:
  I (Very light), II (Light), III (Intermediate), IV (Tan), V (Brown), VI (Dark)