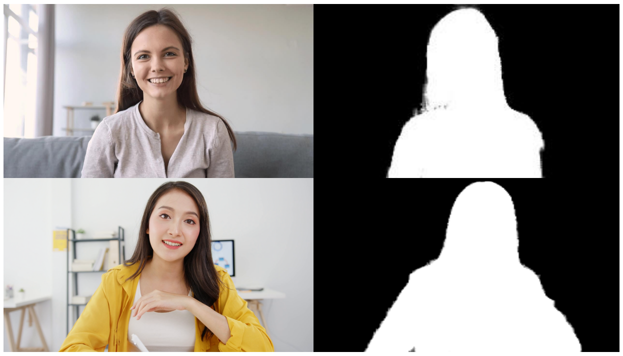
MODEL CARD

MediaPipe Selfie Segmentation

MODEL DETAILS

Two lightweight models (249KB size for general inputs, and 244KB size for landscape inputs) to segment the prominent humans[1] in the scene. Run in real-time via XNNPack TFLite backend on a laptop CPU or smartphone GPU.

Return a two class segmentation label (human or background) per pixel.



Left: Input frames. Right: Output person masks.

MODEL SPECIFICATIONS

Model Type
Convolutional Neural Network

Model Architecture
Convolutional Neural Network: MobileNetV3-like with customized decoder blocks for real-time performance.

Input(s)
General model: A frame of video or an image, represented as a 256 x 256 x 3 tensor.

Landscape model: A 144 x 256 x 3 tensor.

Channels order: RGB with values in [0.0, 1.0].

Output(s)
Generaal model: 256 x 256 x 1 tensor with a mask of person, where values are in range [0, 1.0].

Landscape model: 144 x 256 x 1 tensor with a mask of person, where values are in range [0, 1.0].

AUTHORS

Who created this model?

Tingbo Hou, Google
Siargey Pisarchyk, Google
Karthik Raveendran, Google

DATE
May 6, 2021

LICENSED UNDER

Apache License, Version 2.0

[1] If multiple people of similar scale are present, the model may include some/all of them in the person mask.

Intended Uses

APPLICATION

Human segmentation from videos in interactive applications.

DOMAIN AND USERS

- Augmented reality
- Video conferencing

OUT-OF-SCOPE APPLICATIONS

- Multiple people across different scales.
- People too far away from the camera (e.g. further than 14 feet / 4 meters).
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology.

Limitations

PRESENCE OF ATTRIBUTES

This model may segment multiple humans present in the scene particularly if they are of similar size. Some thin features of humans such as fingers might occasionally be missed in the mask.

TRADE-OFFS

The model is optimized for real-time performance in the web browser and on a wide variety of mobile devices, and may not provide pixel perfect masks.

ENVIRONMENT

When degrading the environment light, adding noise, or fast motions, or including large occluders, one can expect degradation of quality of the predicted mask.

Ethical Considerations

HUMAN LIFE

The model is not intended for human life-critical decisions. The primary intended application is entertainment.

PRIVACY

This model was trained and evaluated on images, including consented images of people using a mobile AR application captured with smartphone cameras in various “in-the-wild” conditions.

Training Factors and Subgroups

INSTRUMENTATION

- The majority dataset images were captured on a diverse set of front and back-facing smartphone cameras.
- These images were captured in a real-world environment with different light, noise and motion conditions via an AR (Augmented Reality) application.

ENVIRONMENTS

The model is trained on images with various lighting, noise and motion conditions and with diverse augmentations. However, its quality can degrade in extreme conditions.

GROUPS

The 17 groups are based on the United Nations geoscheme with the following amendments: Melanesia, Micronesia, and Polynesia have been united due to their size; Europe excludes EU countries; Middle Africa and Melanesia, Micronesia, and Polynesia regions have fewer evaluation samples; see table below.

Australia and New Zealand
Melanesia, Micronesia, and Polynesia
Europe (excluding EU)
Central Asia
Eastern Asia
Southeastern Asia
Southern Asia
Western Asia
Caribbean
Central America
South America
Northern America
Northern Africa
Eastern Africa
Middle Africa
Southern Africa
Western Africa

Based on Model Cards for Model Reporting, In Proceedings of FAT* Conference (FAT*2019). ACM, New York, NY, USA

Evaluation metrics

Model Performance Measures

IoU, Intersection over Union

We evaluate the performance of our model by computing the ratio of the intersection of the predicted mask with the ground truth mask, and their union for the person class. Typical errors occur along the boundary of the true segmentation mask and may move it by a few pixels or lose thin features.

Evaluation results

Geographical Evaluation Results

DATA

- **1594 images, 100 images from each of 17 the geographical subregions** (except 2 subregions Melanesia + Micronesia + Polynesia, and Middle Africa).
- All samples are picked from the same source as training samples and are characterized as smartphone camera photos taken in real-world environments (see specification in "Factors and Subgroups - Instrumentation").

EVALUATION RESULTS

Detailed evaluation for segmentation across 17 geographical subregions is presented in the table below.

Region	**General model IOU (%) with 95% confidence interval**	**Landscape model IOU (%) with 95% confidence interval**	**Number of images**
Australia and New Zealand	96.80 +/- 0.61%	96.03 +/- 0.84%	100
Central America	96.85 +/- 0.70%	96.44 +/- 0.66%	100
Central Asia	96.39 +/- 0.77%	95.85 +/- 0.79%	100
Caribbean	96.18 +/- 0.78%	95.37 +/- 0.88%	100
Eastern Africa	95.88 +/- 1.25%	95.78 +/- 1.00%	100
Eastern Asia	**97.59** +/- 0.48%	**97.27** +/- 0.49%	100
Europe	96.23 +/- 0.82%	96.18 +/- 0.68%	100
Middle Africa	96.54 +/- 1.19%	96.21 +/- 1.21%	43
Northern Africa	96.42 +/- 0.80%	95.97 +/- 0.90%	100
Northern America	97.14 +/- 0.45%	96.61 +/- 0.56%	100
Melanesia + Micronesia + Polynesia	96.00 +/- 1.00%	95.05 +/- 1.39%	51
Southern Africa	96.02 +/- 1.15%	95.89 +/- 0.78%	100
South America	**95.71** +/- 1.17%	95.59 +/- 0.90%	100

Southern Asia	96.65 +/- 0.56%	96.04 +/- 0.62%	100
Southeastern Asia	96.91 +/- 0.58%	96.45 +/- 0.64%	100
Western Africa	95.75 +/- 1.38%	94.71 +/- 1.57%	100
Western Asia	97.18 +/- 0.58%	96.41 +/- 0.89%	100
Average	**96.48 +/- 0.84**%	**95.99 +/- 0.87%**	

Geographical Fairness Evaluation Results

FAIRNESS CRITERIA

We consider a model to be performing poorly for a particular group if
a) Any region is further away than 3 stdev from the average of the model's performance across regions OR
b) Any region is further away than twice the human annotation from the average of the models performance across regions, in our case 2 * (1-98.74%) = 2.52%

FAIRNESS METRICS & BASELINE

We asked 7 annotators to re-annotate the validation dataset, yielding a person IoU of **98.74%**
This is a high inter-annotator agreement, suggesting that the IoU metric is a strong indicator of the person's segmentation mask.

FAIRNESS RESULTS

General model: Evaluation across 17 regions of the models on selfie datasets yields an average performance of 96.48 +/- 0.84% with a range of [95.71%, 97.59%] across regions.

Landscape model: Evaluation across 17 regions of the models on selfie datasets yields an average performance of 95.99 +/- 0.87% with a range of [94.71%, 97.27%] across regions.

Comparison with our fairness criteria yields a maximum discrepancy between average and worst performing regions of **1.11%** for the general model, and **1.28%** for the landscape model, lower than the criteria.

Skin Tone and Gender

DATA

1594 images, 100 images from each of 17 the geographical subregions (except 2 subregions Melanesia + Micronesia + Polynesia, and Middle Africa) were annotated with perceived gender and skin tone (from 1 to 6) based on the Fitzpatrick scale.

FAIRNESS RESULTS

General model: Evaluation on selfie datasets results in an average performance of 96.57% with a range of [95.64%, 96.74%] across all skin tones. The maximum discrepancy between worst and best performing categories is 1.1%.

Evaluation across gender yields an average performance of 96.57% with a range of [96.25%, 96.74%]. The maximum discrepancy is 0.61%.

Landscape model: Evaluation on selfie datasets results in an average performance of 96.08% with a range of [95.40%, 96.55%] across all skin tones. The maximum discrepancy between worst and best performing categories is 1.15%.

Evaluation across gender yields an average performance of 96.08% with a range of [95.77%, 96.37%]. The maximum discrepancy is 0.6%.

Skin Tone Type	% of dataset	General Model	Landscape Model
1	5.03%	**96.74%**	95.90%
2	15.82%	96.71%	**96.55%**
3	33.57%	96.65%	96.21%
4	27.24%	96.67%	95.97%
5	13.31%	96.28%	95.76%
6	5.03%	**95.64%**	**95.40%**
Average		**96.57%**	**96.08%**
Range		1.1%	1.15%

Gender	% of dataset	General Model	Landscape Model
Female	47.58%	**96.25%**	**95.77%**
Male	52.42%	**96.86%**	**96.37%**
Average		96.57%	**96.08%**
Range		0.61%	0.6%

Definitions

AUGMENTED REALITY (AR)
Augmented reality, a technology that superimposes a computer-generated image on a user's view of the real world, thus providing a composite view.

INTERSECTION OVER UNION
A measure of similarity. In the segmentation case, the ratio between the area of intersection of two masks and the area covered by their union.