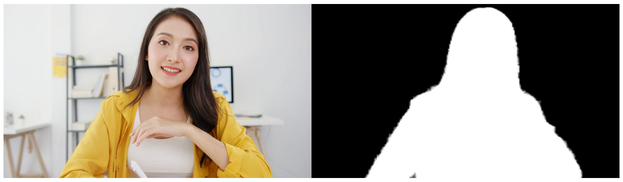
MODEL CARD

MLKit Selfie Segmentation

MODEL DETAILS

A lightweight model (249KB size) to segment the prominent humans[1] in the scene in videos captured by a smartphone. Runs in real-time via XNNPack TFLite backend.

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



Left: Input frame. Right: Output person mask.

MODEL SPECIFICATIONS

Model Type
Convolutional Neural Network

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

Input(s)
A frame of video or an image, represented as a 256 x 256 x 3 tensor. Channels order: RGB with values in [0.0, 1.0].

Output(s)
256 x 256 x 2 tensor for the light model with masks for background (channel 0) and person (channel 1) where values are in range [MIN_FLOAT, MAX_FLOAT]. The user has to apply softmax across both channels to yield foreground probability in [0.0, 1.0].

AUTHORS

Who created this model?
Siargey Pisarchyk, Google
Tingbo Hou, Google
Karthik Raveendran, Google

DATE

Feb 16, 2021

LICENSED UNDER

[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 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

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

- **1700 images, 100 images from each of 17 the geographical subregions** (see specification in Section "Factors and Subgroups").
- 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	IoU(%) (95% confidence interval)	Number of images
australia_nz	96.81	100
c_america	96.19	100
c_asia	96.90	100
caribbean	96.45	100
e_africa	96.00	100
e_asia	**97.64**	100
europe	96.26	100
m_africa	96.06	43
n_africa	96.65	100
n_america	96.49	100

nesias	97.17	51
s_africa	**95.89**	100
s_america	96.93	100
s_asia	96.20	100
se_asia	96.71	100
w_africa	95.91	100
w_asia	97.26	100
average	**96.57**	
range	**1.75**	

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

Evaluation across 17 regions on selfie datasets representative of the primary use case results yields an average performance of 96.57% with a range of [95.89%, 97.64%].

Comparison with our fairness criteria yields a maximum discrepancy between the worst and the best performing regions of 1.75% for the model.

Skin Tone and Gender

DATA

1700 images, 100 images from each of 17 the geographical subregions were annotated with perceived gender and skin tone (from 1 to 6) based on the Fitzpatrick scale.

FAIRNESS RESULTS

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

Evaluation across gender yields an average performance of 96.57% with a range of [96.25%, 96.86%] and maximum discrepancy of 0.61%

Skin Tone Type	% of dataset	IoU
1	5.02%	**96.74**
2	15.81%	96.71
3	33.56%	96.65
4	27.23%	96.67
5	13.30%	96.28
6	5.02%	**95.64**
Average		**96.57**
Range		1.1

Gender	% of dataset	IoU (%)
Female	47.55	**96.25**
Male	52.38	**96.86**
Average		96.57
Range		0.61

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.