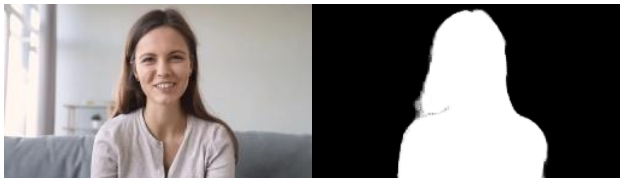
MODEL CARD

MediaPipe Meet Segmentation

MODEL DETAILS

A lightweight model (400KB size) to segment the prominent humans[1] in the scene in videos captured by a smartphone or web camera. Runs in real-time (~120 FPS) on a laptop CPU 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 144 x 3 tensor (for the full model), or 160 x 96 x 3 tensor (for the light model). Channels order: RGB with values in [0.0, 1.0].

Output(s)
256 x 144 x 2 tensor for the full model or 160 x 96 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] and user has to apply softmax across both channels to yield foreground probability in [0.0, 1.0].

AUTHORS

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

DATE

Oct 12, 2020

[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.

Australia and New Zealand
Melanesia, Micronesia, and Polynesia
Europe*
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	Full model (95% confidence interval)	Lite model (95% confidence interval)
australia_nz	0.955 +/- 0.00982	0.947 +/- 0.0112
c_america	0.964 +/- 0.0062	0.952 +/- 0.00835
c_asia	0.966 +/- 0.00839	0.954 +/- 0.0104
caribbean	0.952 +/- 0.00802	0.938 +/- 0.011
e_africa	0.947 +/- 0.0112	0.932 +/- 0.0132
e_asia	0.966 +/- 0.0055	0.954 +/- 0.00834
europe	0.963 +/- 0.00685	0.95 +/- 0.00917
m_africa	0.94 +/- 0.0135	0.93 +/- 0.0146
n_africa	0.962 +/- 0.00964	0.95 +/- 0.0095
n_america	0.957 +/- 0.00861	0.952 +/- 0.00785

nesias	**0.933 +/- 0.018**	**0.922 +/- 0.0206**
s_africa	0.964 +/- 0.00535	0.956 +/- 0.00716
s_america	0.953 +/- 0.0125	0.944 +/- 0.0139
s_asia	**0.967 +/- 0.00503**	**0.956 +/- 0.00544**
se_asia	0.961 +/- 0.00799	0.951 +/- 0.00867
w_africa	0.942 +/- 0.017	0.933 +/- 0.0174
w_asia	0.962 +/- 0.00666	0.953 +/- 0.00746
average	0.956 +/- 0.0103	0.946 +/- 0.0105
range	**0.034**	**0.034**

Geographical Fairness Evaluation Results

FAIRNESS CRITERIA

We consider a model to be performing unfairly across representative groups 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 of full and lite models on selfie datasets representative of Meet primary use case results yields an average performance of 95.6% +/- 1% stdev with a range of [93.3%, 96.7%] across regions for the full model and an average performance of 94.6% +/-1% stdev with a range of [92.2%, 95.6%] across regions for the lite model.

Comparison with our fairness criteria yields a maximum discrepancy between average and worst performing regions of 2.3% for the full and 2.4% for the light 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 Meet primary use case results in an average performance of 95.1% with a range of [94.1%, 96.1%] across all skin tones for the full model and an average performance of 94.1% with a range of [93.0%, 95.3%] across regions for the lite model. The maximum discrepancy between worst and best performing categories is 2% for the full model and 2.3% for the lite model.

Evaluation across gender yields an average performance of 95.6% with a range of [95.5%, 95.8%] for the full model, and an average of 94.6% with a range of [94.4%, 94.8%] for the lite model. The maximum discrepancy is 0.3% for the full model and 0.4% for the lite model.

Skin Tone Type	% of dataset	Full Model	Lite Model
1	0.47%	0.942	0.933
2	14.88%	0.961	**0.953**
3	37.76%	**0.961**	0.950
4	31.00%	0.957	0.946
5	11.53%	**0.941**	**0.930**
6	4.24%	0.944	0.936
Average		0.951	0.941
Range		0.020	0.023

Gender	% of dataset	Full Model	Lite Model
Female	41.65%	**0.955**	**0.944**
Male	58.24%	**0.958**	**0.948**
Average		0.956	0.946
Range		0.004	0.003

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.