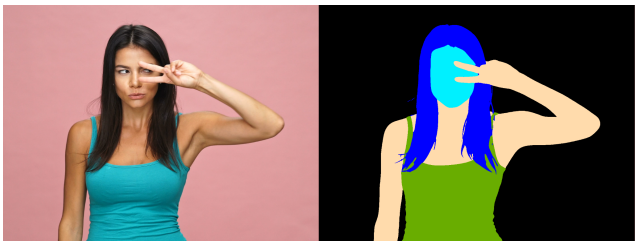
MODEL CARD

# Mediapipe Multiclass Segmentation 15/03/2023

## MODEL DETAILS

Two models to predict segmentation masks for human subjects in real time from images captured by the human subject. Runs via Mediapipe image segmenter API for on-device in real-time. The model is a multiclass segmentation model and classifies each pixel as [background, hair, body, face, clothes, others]. Supports single or multiple people in the frame, selfies and full body images.



*Left: Input frame. Middle: Predicted category mask.*

## AUTHORS

Adel Ahmadyan, Google

## MODEL DATE

May 10, 2023

## MODEL SPECIFICATIONS

**Model Type**

- Vision Transformer Neural Network

**Model Architecture**

- Vision Transformer, with customized bottleneck and decoder architecture for real-time performance.

**Inputs**

- normalized RGB image, at resolution of [256, 256, 3] for smaller model and [512, 512, 3] for larger model. The input does not need to be resized to match the image size, and can take portrait, or landscape or any other aspect ratio images as input.

**Output(s)**

- Segmentation mask, represented as tensor of probabilities as float values in [0, 1] range of the size [256x256x6] and [512, 512, 6] where 6 is the number of predicted classes. The output will be post-processed (maxed) via mediapipe image segmenter API to get category mask.

## LICENSED UNDER

Apache License, Version 2.0

## Intended Uses

### APPLICATIONS

- Human segmentation from images/videos in interactive entertainment, and video conferencing applications.

### DOMAIN & USERS

- Augmented Reality
- Video Conferencing
- Entertainment

### OUT-OF-SCOPE APPLICATIONS

Not appropriate for:

- This model is not intended for human life-critical decisions and applications that require pixel perfect masks
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology.

## Limitations

### 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, backlit subjects, adding noise, or fast motions, or including large occluders, one can expect degradation of quality of the predicted mask.

## Factors and Subgroups

### ENVIRONMENTS

- Model is trained on thousands of samples with diverse identities (i.e. unique subjects), various blendshape combinations representing both common and random expressions and transformations (e.g. rotations).

# Evaluation, Datasets and Results

## Skin Tone and Gender Evaluation

DATASET

**Contains 918 images, totalling 1902 instances of human subjects**. Some images contained groups of people. Each human subject was annotated with perceived gender (male and female) and skin tone (from 1 to 10) based on the monk scale.

FAIRNESS CRITERIA

The SDM of each subgroup should be within one standard deviation from those values of the entire dataset to be considered fair.

EVALUATION RESULTS

Detailed evaluation across genders and skin tones is presented in the tables below.

FAIRNESS RESULTS

Examine whether subgroups are within one standard deviation away from the metrics of the entire dataset:
Across Gender:

- SDM: worst case 80.89, difference is 0.2, within the standard deviation of 10.2

Across skin tone:

- Mean IoU worst case 71.86, difference is 9.24, within the standard deviation of 10.2

Observed discrepancy across different genders and skin tones is less than one defined in our fairness criteria. We therefore consider the model performing well across groups.

512x512 resolution model, The IoU was computed at the input image resolution.

| Gender | Test subset items and % | | mean IOU | STDEV |
|---|---|---|---|---|
| Male | 759 | 39.9% | 81.48 | 10.09 |
| Female | 1143 | 60.1% | 80.89 | 10.50 |
| Total | 1902 | 100% | 81.10 | 10.2 |

*Gender evaluation, Signed Deviation*

| Skintone | Test subset items and % | | mean IOU | STDEV |
|---|---|---|---|---|
| Tone_1 + Tone_2 | 819 | 43% | *81.70* | *10.23* |
| Tone_3 | 396 | 20% | *80.93* | *9.20* |
| Tone_4 | 129 | 6% | *82.17* | *9.64* |
| Tone_5 | 81 | 4% | *81.53* | *9.03* |
| Tone_6 | 168 | 8% | *81.04* | *11.22* |
| Tone_7 | 230 | 12% | *79.98* | *11.17* |
| Tone_8 | 48 | 2% | *80.45* | *9.85* |
| Tone_9 + Tone_10 | 31 | 1.6% | *71.86* | *14.32* |
| | | | | |
| Total | 1902 | 100% | 81.10 | 10.2 |

*Skin tone evaluation, Signed Deviation*

256x256 resolution model, The IoU was computed at the input image resolution.

| Gender | Test subset items and % | | mean IOU | STDEV |
|---|---|---|---|---|
| Male | 759 | 39.9% | 77.10 | 11.46 |
| Female | 1143 | 60.1% | 77.32 | 10.94 |
| Total | 1902 | 100% | 77.23 | 11.14 |

*Gender evaluation, Signed Deviation*

| Skintone | Test subset items and % | | mean IOU | STDEV |
|---|---|---|---|---|
| Tone_1 + Tone_2 | 819 | 43% | 77.22 | 10.99 |
| Tone_3 | 396 | 20% | 77.59 | 11.37 |
| Tone_4 | 129 | 6% | 78.81 | 10.60 |
| Tone_5 | 81 | 4% | 78.16 | 10.27 |
| Tone_6 | 168 | 8% | 79.18 | 10.69 |
| Tone_7 | 230 | 12% | 75.98 | 11.08 |
| Tone_8 | 48 | 2% | 73.46 | 11.93 |
| Tone_9 + Tone_10 | 31 | 1.6% | 68.25 | 12.52 |
| | | | | |
| Total | 1902 | 100% | 77.23 | 11.14 |

*Skin tone evaluation, Signed Deviation*

# 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.

# Appendix

`List of predicted classes`

1 - background
2 - hair
3 - body-skin
4 - face-skin
5 - clothes
6 - others (accessories)