Model Card Version: 1.0_2023
License: Apache 2.0

Guideline Depth Model

Model:
https://github.com/google-research/project-guideline/blob/main/project_guideline/vision/models/depth.tflite

A U-Net based model for predicting depth maps from monocular images in certain outdoor environments (e.g. street, park, trail) with dynamic objects (e.g. human, pet, bicycle). The model is relatively lightweight (10.8MB size), and capable of running super-real-time (~260FPS on Pixel 7 Edge TPU, ~45FPS on Pixel 7 GPU, ~5FPS on Pixel 7 single-core CPU).

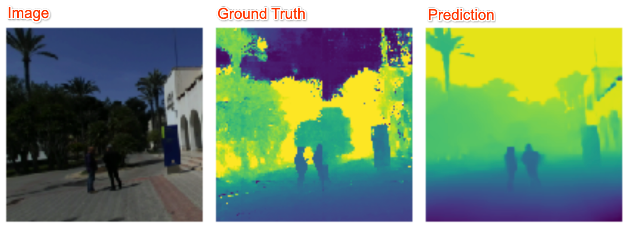
Model Snapshot

Model Overview

MODEL ARCHITECTURE	INPUT(S)	OUTPUT(S)
U-Net model with MobileNetV2 encoder.	3-channel RGB image of size [1, 192, 192, 3] in uint8 ([0, 255]).	Depth map of size [1, 192, 192, 1] in float32. The depth map is inverse of metric depth.

Usage

APPLICATION	BENEFITS	KNOWN CAVEATS
Used in Project Guideline for obstacle detection.	The model is capable of real-time depth estimation from monocular images which is useful in a variety of vision applications, particularly on mobile devices that lack ToF (Time of Flight) and other distance sensors. The model is notably better than other alternatives on outdoor, long-range data. Intel's DPT (Dense Prediction Transformer), a comparable SOTA open-source depth estimation	The model may fail to produce accurate results in environments and conditions deviating from those in the dataset (e.g. indoor scene, inclement weather). The output values are relative to a variety of factors including camera intrinsics (e.g. field of view). Calibration is required to interpret the output as absolute depth (see *Downstream Dependencies*).

	model, is 1.7GB and requires several Transformer operations in series making it unsuitable for real-time, mobile environments. This model is based on MobileNetV2 and several orders of magnitude smaller and faster.	

Model Creators

MODEL CONTACT	MODEL AUTHOR(S)	CITATION
Kimberly Wilber (kwilber@google.com)	Kimberly Wilber, Google; Abhishek Kar, Google; Xuan Yang, Google;	Unavailable

System Type

SYSTEM DESCRIPTION	UPSTREAM DEPENDENCIES	DOWNSTREAM DEPENDENCIES
This can be used either as a standalone model or integrated to other systems (e.g. [Project Guideline])	Resize to input resolution 192×192, using center-crop for best results.	If downstream requires metric depth, `metric_depth = 1.0 / (a * output + b)`. The value of *a* and *b* (scale and shift) need to be calibrated before using and can vary based on camera intrinsics (e.g. field of view) and extrinsics (e.g. view angle).

Implementation Frameworks

HARDWARE & SOFTWARE FOR TRAINING	HARDWARE & SOFTWARE FOR DEPLOYMENT
• GPU (NVidia A100) • TensorFlow v2	• Pixel 6+ device, Edge TPU • MediaPipe v0.10 Image Segmenter • TensorFlow Lite v2

Compute Requirements

COMPUTE REQUIREMENTS FOR FINE-TUNING*		COMPUTE REQUIREMENTS FOR INFERENCE*	
Number of Chips	4x NVidia A100	Number of Chips	1
Training Time (days)	0.5	Training Time (days)	N/A
Total Computation (floating pt operations)	Unavailable	Total Computation (floating pt operations)	Unavailable
Measured Performance (TFLOPS/s)	Unavailable	Measured Performance (TFLOPS/s)	~180
Energy Consumption (MWh)	Unavailable	Energy Consumption (MWh)	Unavailable

**Modeled after Patterson, David, et al. "[Carbon emissions and large neural network training.]" arXiv preprint arXiv:2104.10350 (2021).*

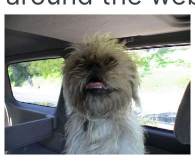
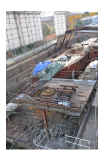
Model Characteristics

MODEL INITIALIZATION		MODEL STATUS		MODEL STATS	
Model training proceeds in two phases: - **Pretrain:** Model is pretrained on OpenImages for 1,000,000 steps (step size = 48 images), or about 10 epochs. - **Fine-tune:** Model is then fine-tuned on SANPO for 100,000 steps.		The model is static and not updated regularly.			
Training Epochs	10	**Dataset Name**	OpenImages with DPT pseudolabels, SANPO	**Size**	10.8MB
Base Learning Rate	0.001	**Version**	1.0	**Weights**	16.2M
Method	Gradient descent (Adam)	**Release Date**	July 2023	**Layers**	~45
Loss	MidasNet scale/shift invariant loss	**Update Cadence**	N/A	**Latency**	~4ms (Edge TPU)

PRUNING		QUANTIZATION		DIFFERENTIAL PRIVACY
No		Yes, Float16		SANPO dataset includes face and license plate blurring.
Methods	N/A	**Methods**	Model is trained at 32-bit precision. As a	

			post-processing step, TFLiteConverter quantizes the model without retraining/fine-tuning.
Structuring	N/A	**Pre-quantized Representation**	fp32
Sparsity Level	N/A	**End Bit Representation**	fp16
Number of Params at Sparsity	N/A	**Hardware**	EdgeTPU, GPU
Accuracy at Final Sparsity after Training	N/A		
Perplexity at Final Sparsity after Training	N/A		

Data Overview

TRAINING DATASET SNAPSHOT	DATASET MAINTENANCE & VERSIONS	INSTRUMENTATION
Two sources of training data: - **DPT/OpenImages:** the popular OpenImages dataset, with pseudolabels from DPT. Around 8M images from around the web.   	The training data is static.	The SANPO dataset was captured using a custom collection rig utilizing Stereolabs ZED stereo cameras. Data is recorded in Stereolab's proprietary SVO file format then pre-processed to extract video frames and metadata.

- **SANPO:** ~100,000 images with depth from stereo cameras   					
Dataset Size	8M (DPT/OpenImages) + 100K (SANPO)	**Current Version**	1.0	**Instrumentation Criteria**	
Number of Instances	N/A	**Update Cadence for Online Data**	No online data is used for either dataset.	**Focal spot size**	Radiation from capture instruments is not focused.
Number of Fields	2 (RGB image and depth map)	**Sampling methods**	Training samples were drawn evenly between DPT and SANPO images.	**Cooling method**	Custom cooling apparatus utilizing a 12V blower fan.
Labeled Classes	N/A	**Validation methods**	Manual	**Avg Adult Effective Dose (mSv)**	ZED cameras and Pixel 6 devices are not a significant source of electromagnetic radiation in the gamma ray band. No lead shielding precautions were taken.
Number of Labels	N/A	**Processing methods**	Depth maps generated from stereo images.	**Operational voltage range**	12V, 5A max, provided by on-board battery
Average labels per instance	N/A	**Annotation methods**	N/A		
Missing Labels	N/A				
DATA PRE-PROCESSING		DEMOGRAPHIC GROUPS		EVALUATION DATA	

- **DPT/OpenImages** images are processed by running them through DPT to get "pseudolabels" (high-quality depth maps). - **SANPO images** are preprocessed in the following way: 1. Each stereo camera stream is segmented into videos of at most 30 seconds long. 2. CREStereo creates a disparity map (an estimate of the depth at each pixel) by comparing the left and right camera view. 3. RGB frames are processed to blur detected faces and license plates.	- Data does not contain labeled groups or demographic attributes. - Both datasets contain blurred, consensual, and/or PII-removed images of incidental people. Demographic group could be plausibly inferred by looking at the images.	25% of dataset sessions are withheld for evaluation.

Evaluation Results

Aggregate Evaluation Results

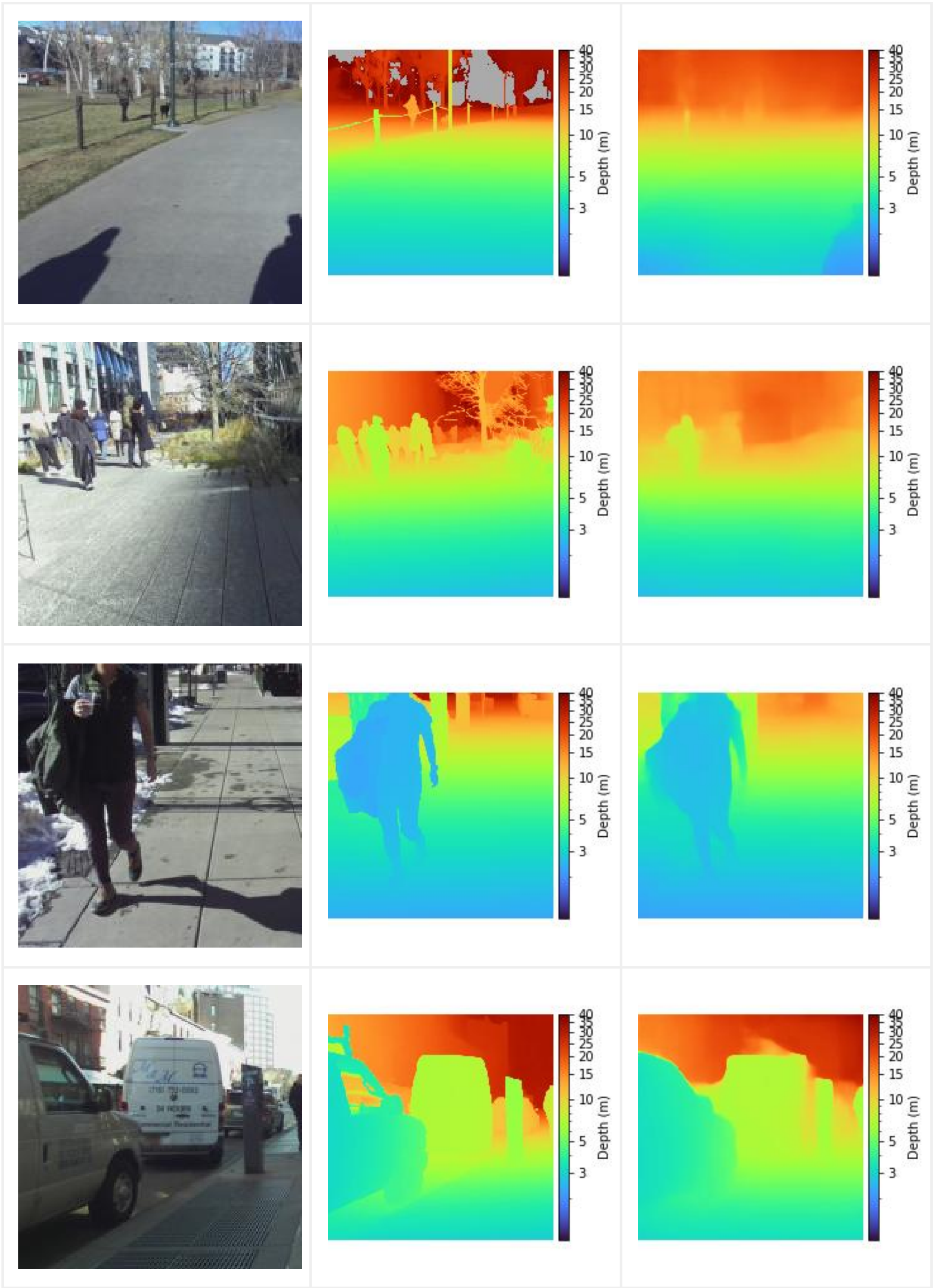
EVALUATION PROCESS	EVALUATION RESULTS
Metrics for depth estimation methods compare the predicted depth and groundtruth depth at each pixel within the output depth map. The aligned $\Delta_{1.25}$ metric is used, which counts the fraction of pixels that are within 125% of the groundtruth; that is, for input image x and groundtruth map y, $$\Delta_{1.25}(x,\ y) = E\left[\max\left(\frac{g(x,y)}{y},\ \frac{y}{g(x,y)}\right) \leq 1.25\right],$$ where $g()$ maps from model output ("disparity") to metric depth, $$g(x,y) = \frac{1}{\alpha f(x) + \beta}.$$ Here, $f(x)$ is the model output, and α and β are scaling coefficients that are chosen for each image to best align the predicted depth with the groundtruth,	This model's performance on this data is $\Delta_{1.25}$=0.805. Qualitative results are given below: Image | Groundtruth | Prediction

Image	Groundtruth	Prediction

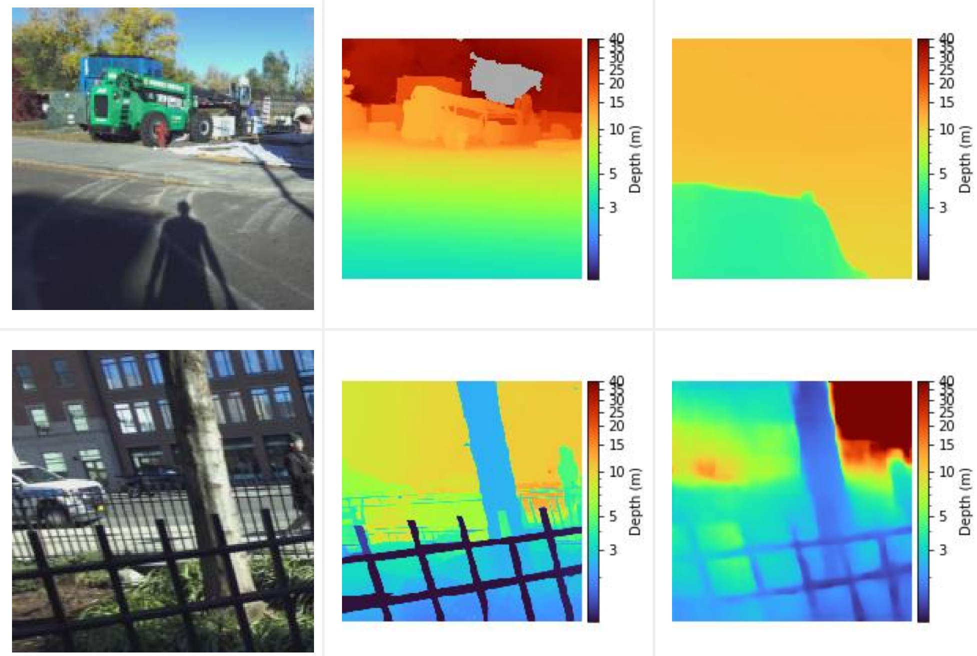
$$\alpha, \beta = \arg\min_{\alpha, \beta} \left\| \frac{1}{\alpha f(x) + \beta} - y \right\|^2 .$$

In practice, a RANSAC (Random Sample Consensus) method has been used to select the best α, β by modeling the relation between the depth of high confidence tracking points provided by ARCore VIO (Visual Inertial Odometry) and the corresponding points in the raw depth map.

Evaluation Set: The model was evaluated on a 25% hold-out of SANPO dataset.



Failure cases



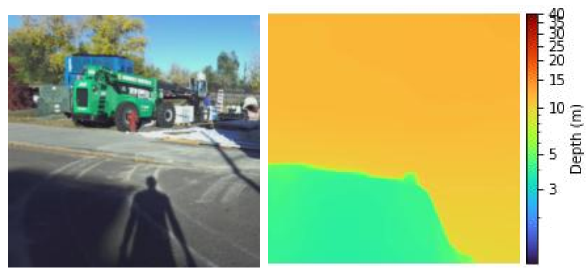
Model Usage & Limitations

SENSITIVE USE	LIMITATIONS	ETHICAL CONSIDERATIONS & RISKS
No sensitive deployment cases identified.	**Input conditions:** The model is tuned for a variety of outdoor scenarios including parks, streets, and trails with dawn to dusk lighting and no inclement weather. The performance may be impacted when deviating from these conditions. The image is expected to be from a typical mobile device rear camera (approx 70-90deg	• The training dataset contains unlabeled images of various demographic groups. The diversity of these demographic groups is limited by the scope of the dataset. • When using the model to detect depth of humans, the performance may vary across skin tones or other demographic characteristics.

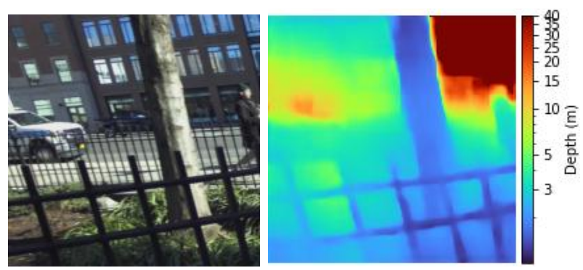
field of view). Accurary will be impacted with wide-angle or telephoto lenses.

Output Caveats:
The output depth value estimations are impacted by camera instrinsics and extrinsics. When estimated absolute depth is required, the output can be calibrated with scale/shift parameters using the formula: `metric_depth = 1.0 / (a * output + b)`.

The output depth may not accurately reflect the scene. The distance to objects may be incorrectly estimated or non-congruent.



Shadow detected as obstacle, other objects missing from depth output.



Inconsistent fence and background depth

- The model is not appropriate for safety-critical applications.