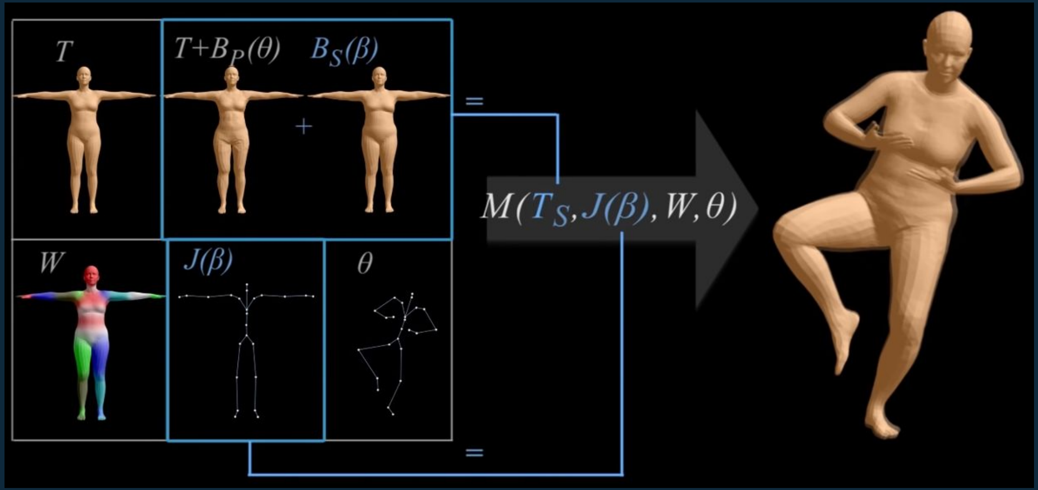
DensePose

Overview

Establish dense correspondences between an RGB image and a surface-based representation of the human body ***in the wild***.

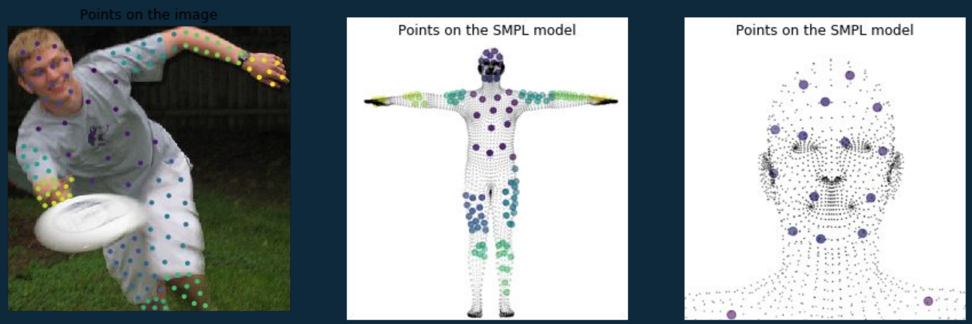
SMPL Model

Skinned Multi-Person Linear model (SMPL) is a realistic 3D model of the human body that is based on skinning and blend shapes and is learned from thousands of 3D body scans.



DensePose-COCO Dataset

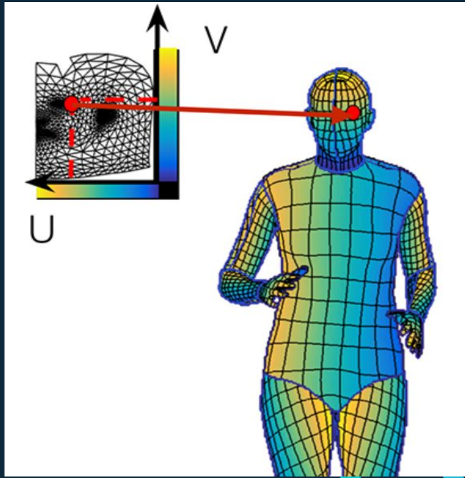
A novel dataset that contains sparse correspondences of sampled pixels in the image to a surface-based representation of the human body, which includes a SMPL model and a UV map that specifies the surface-based location of the pixel in the model's texture space. It contains annotations for **50K** humans, comprising of more than **5 million** manually annotated correspondences.



Learning Dense Human Pose Estimation

The model combines a classification and a regression task. In a first step, we classify a pixel as belonging to either background or one among the surface parts. In a second step, a regression system indicates the exact coordinates of the pixel within the part.

$$c^* = \mathrm{argmax}_c P(c|i), \quad [U,V] = R^{c^*}(i)$$

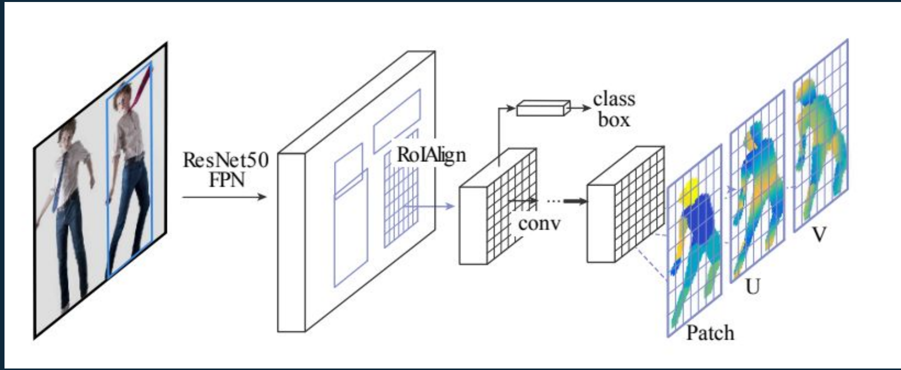


Evaluation Metric

Geodesic Point Similarity (GPS) provides a correspondence matching score and is used as an evaluation metric for the predicted correspondences. It is based on the ***geodesic distance*** between two points which is the length of the shortest path that lies on the surface between them.

$$\mathrm{GPS}_j = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2\kappa^2} \right),$$

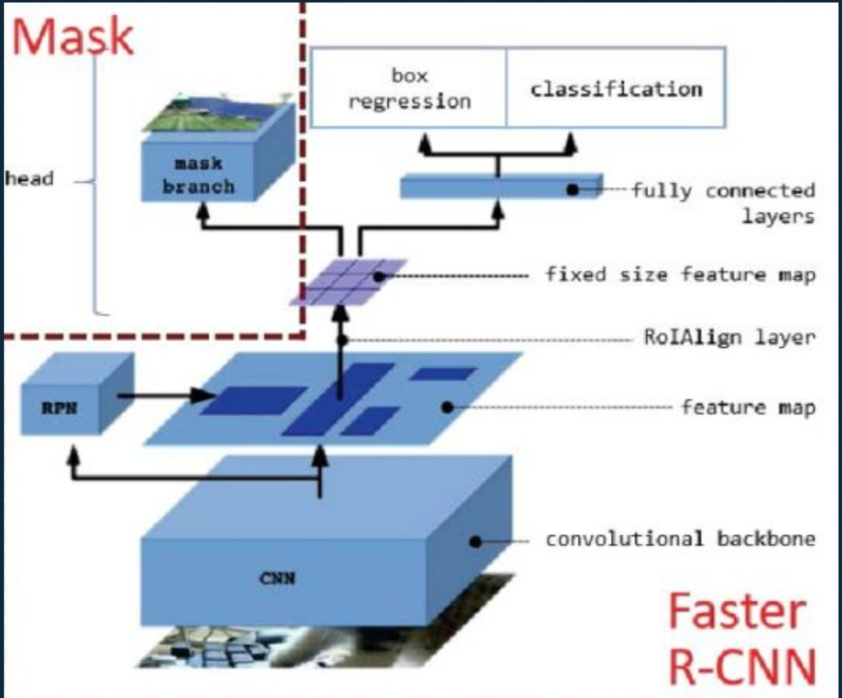
Model Architecture



The architecture resembles that of Mask-RCNN with the Feature Pyramid Network (FPN) features, and ROI-Align pooling so as to obtain dense part labels and coordinates within each of the selected regions.

The model is further tweaked to implement cascading across multiple tasks such as keypoint estimation and instance segmentation.

‘DensePose-RCNN’ -DenseReg approach with the Mask-RCNN architecture



Fully-convolutional dense pose regression

Classification+regression using FCN

Classification-This amounts to a labelling task that is trained using a standard cross-entropy loss.

C can take 25 values (one is background), meaning that Px is a 25-way classification unit, and we train 24 regression functions Rc , each of which provides 2D coordinates within its respective part c. While training, we use a cross-entropy loss for the part classification and a smooth L1 loss for training each regressor.

Region-based Dense Pose Regression

Same architecture used in the keypoint branch of MaskRCNN, consisting of a stack of 8 alternating 3×3 fully convolutional and ReLU layers with 512 channels. At the top of this branch we have the same classification and regression losses as in the FCN baseline, but we now use a supervision signal that is cropped within the proposed region

Multi-task cascaded architectures

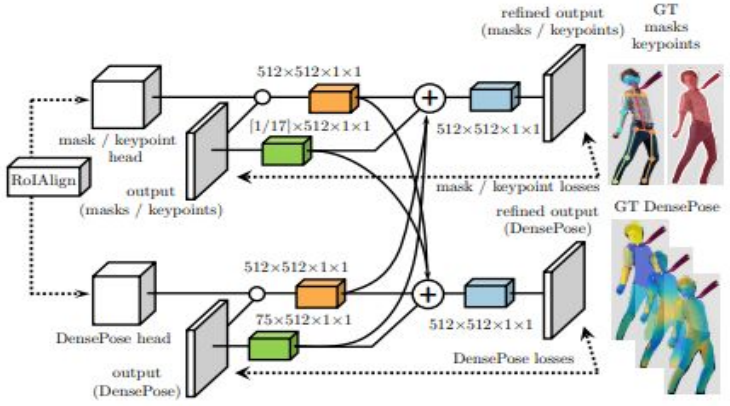
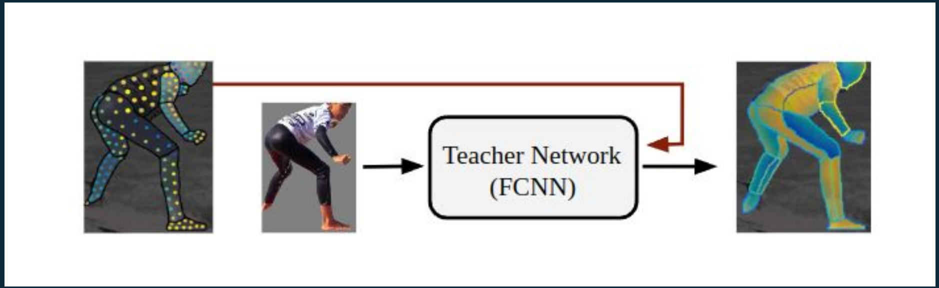


Figure 8: Cross-cascading architecture: The output of the RoIAlign module in Fig. 7 feeds into the DensePose network as well as auxiliary networks for other tasks (masks, keypoints). Once first-stage predictions are obtained from all tasks, they are combined and then fed into a second-stage refinement unit of each branch.

Distillation

The DensePose-RCNN model can be trained directly using the annotated points as supervision. However, quality of the trained model can be enhanced by ***inpainting*** the values of the supervision signal on positions that are not originally annotated using a ***teacher*** network: A fully-convolutional neural network that reconstructs the ground-truth values given images scale-normalized images and the segmentation masks.



Texture Transfer

The estimated dense coordinates can be used to map a texture from the UV space to image pixels. The texture space used for the SMPL model is the ***atlas*** texture space, which allows easy design of custom textures. The texture shown on the right is obtained from the ***SURREAL*** dataset, and is augmented onto a test image.

Scope

Dataset	A down-sample of the DensePose-COCO dataset (10%) inpainted by a F-CNN teacher network
Baseline Model	A Mask-RCNN model with multi-task cascading
Experiments	Perform Dense Pose estimation and compare results with SMPLify model using various evaluation metrics and conduct ablation studies. Build a pipeline to carry out Texture Transfer on any image using custom textures.
Computing Resources	Ada GPU cluster

Computing Resources

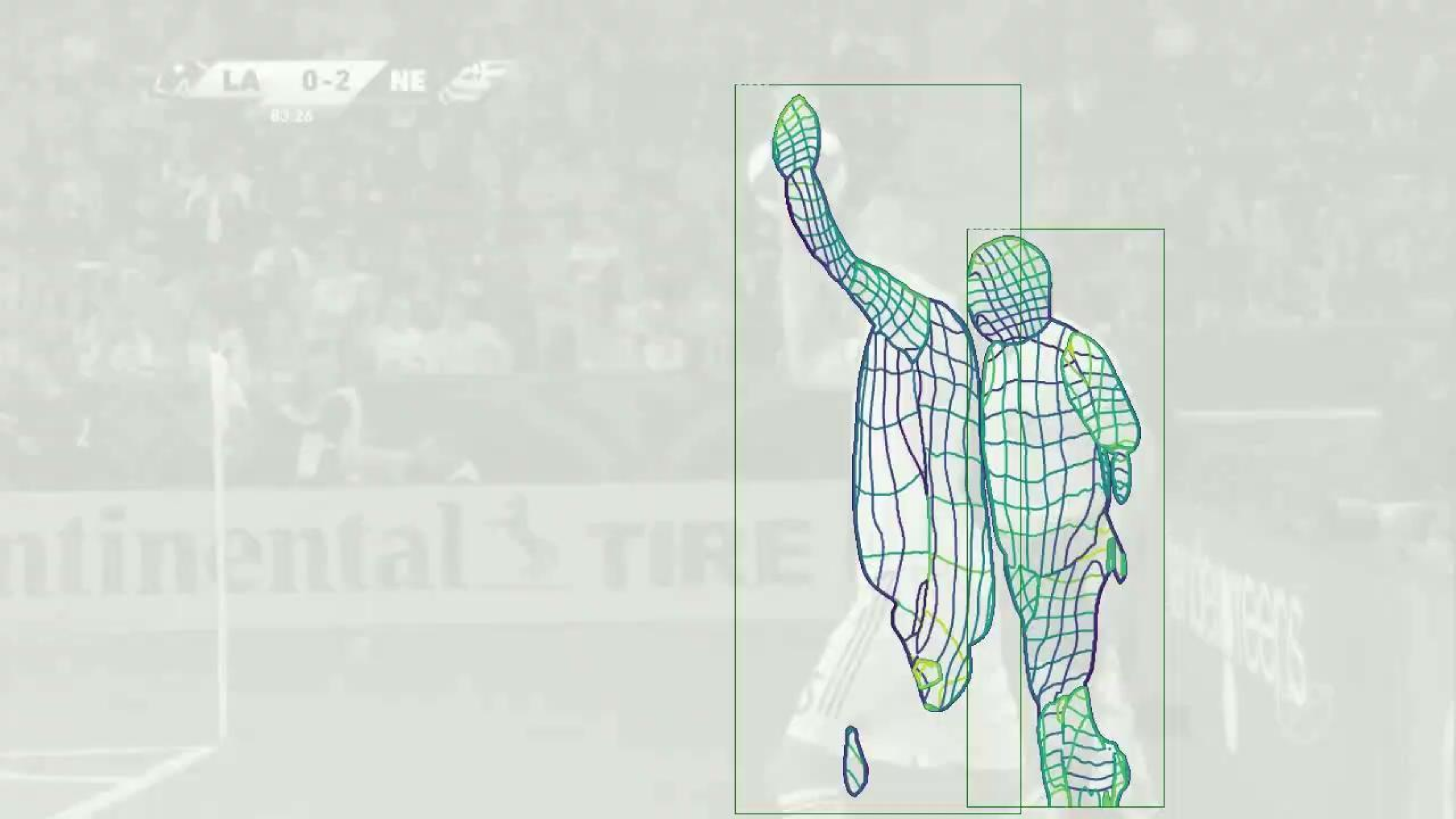
1. GPUs: The authors trained their models on eight NVIDIA Tesla P100 GPUs, which have a peak performance of 9.3 TFLOPS each. Since Ada cluster has four Nvidia GeForce GTX 1080 Ti GPUs or four Nvidia GeForce RTX 2080 Ti GPUs, which have peak performance of 11.34 TFLOPS and 13.45 TFLOPS respectively, they should be sufficient for training the models.
2. Memory: The authors do not mention the exact amount of memory required for training, but it is likely to be several tens of GBs, given the complexity of the model and the size of the COCO dataset. Since each node on the Ada cluster has 128 GB of RAM, this should be sufficient for the training process.
3. Training time: The authors report that the training process took approximately two weeks. With access to 2 GPUs, the training time may be longer, but it should still be feasible to complete the training process on the Ada cluster.
4. Inference time: The authors report that on a single NVIDIA Titan X GPU, their system takes about 0.12 seconds to estimate the dense pose for a single image. With access to four GPUs with higher performance, inference time is likely to be lower on the Ada cluster.

Overall, the Ada cluster is likely to meet the requirements for implementing the DensePose paper, albeit with longer training times due to the smaller number of GPUs.

Densepose Results

Model used: densepose_rcnn_R_50_FPN_s1x

LA 0-2 NE
83:26
ntinental TIRE


LA 0-2 NE
ntinental TIRE

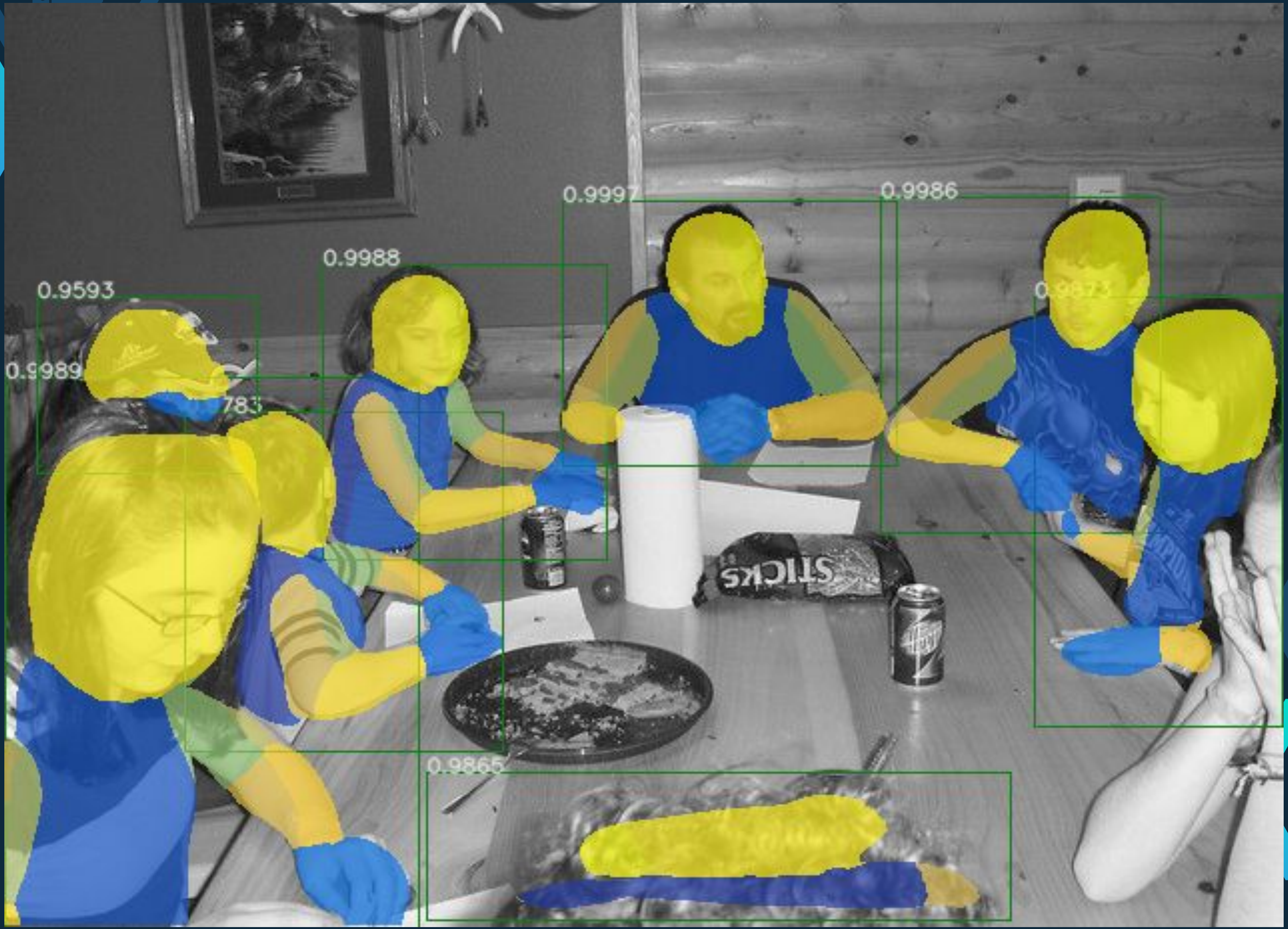

0.9968
0.9432
0.9231
0.9146
0.8663
0.9270
0.9837
0.9611
0.9917
0.9669
0.8039
0.8454
0.9986
0.9419
0.9700
0.9468
0.8134


MorganChase
US OPEN
A USTA EVENT


0.8037
0.8095
0.8390.8475
0.9483
MorganChase
US OPEN
A USTA EVENT
0.9986
0.9953


0.8037
0.8095
0.8475
0.9483
US OPEN
A USTA EVENT
0.9986
0.9953


STICKS

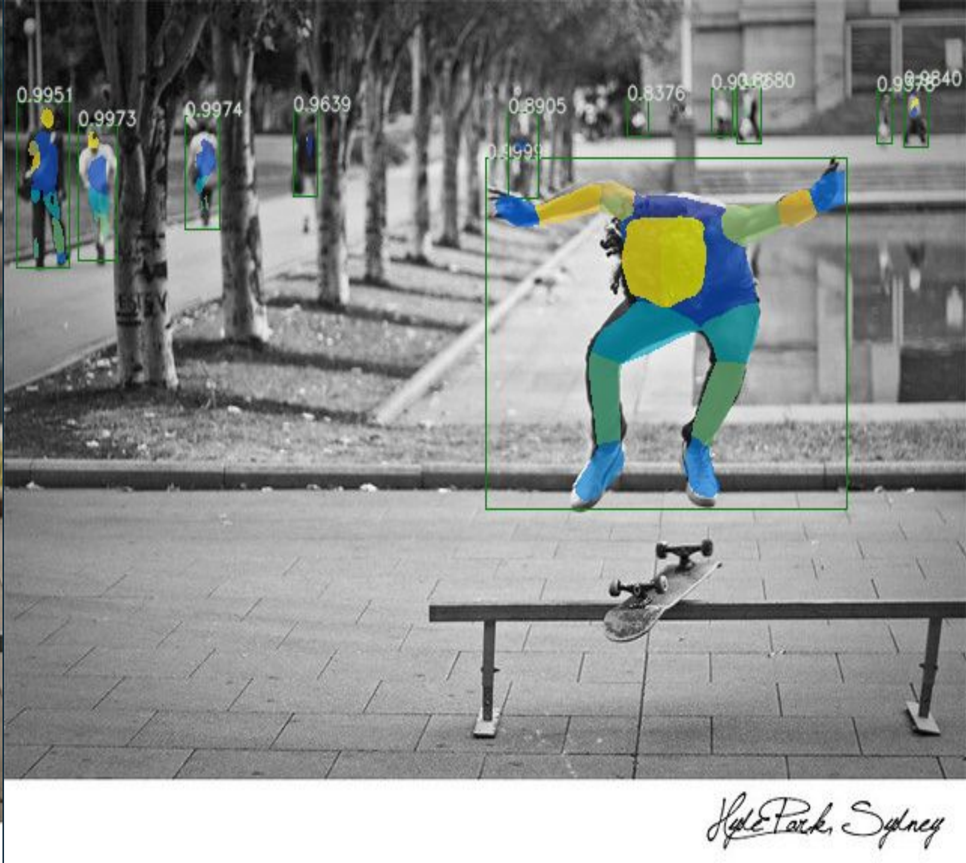

0.9997
0.9986
0.9988
0.9593
0.9989
0.9865

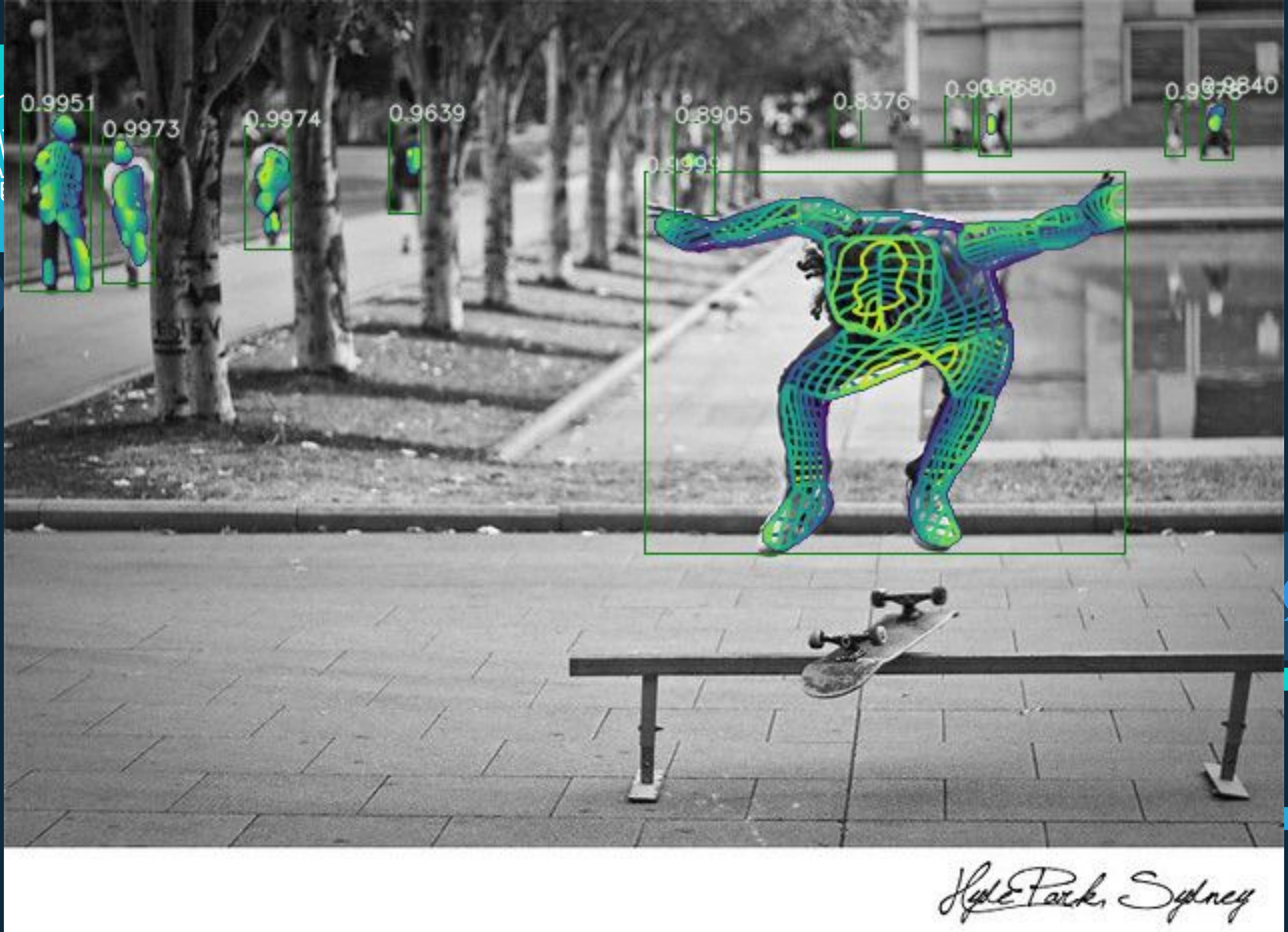

0.9997
0.9986
0.9988
0.9593
0.9573
0.9783
0.9865
STICKS


0.9998

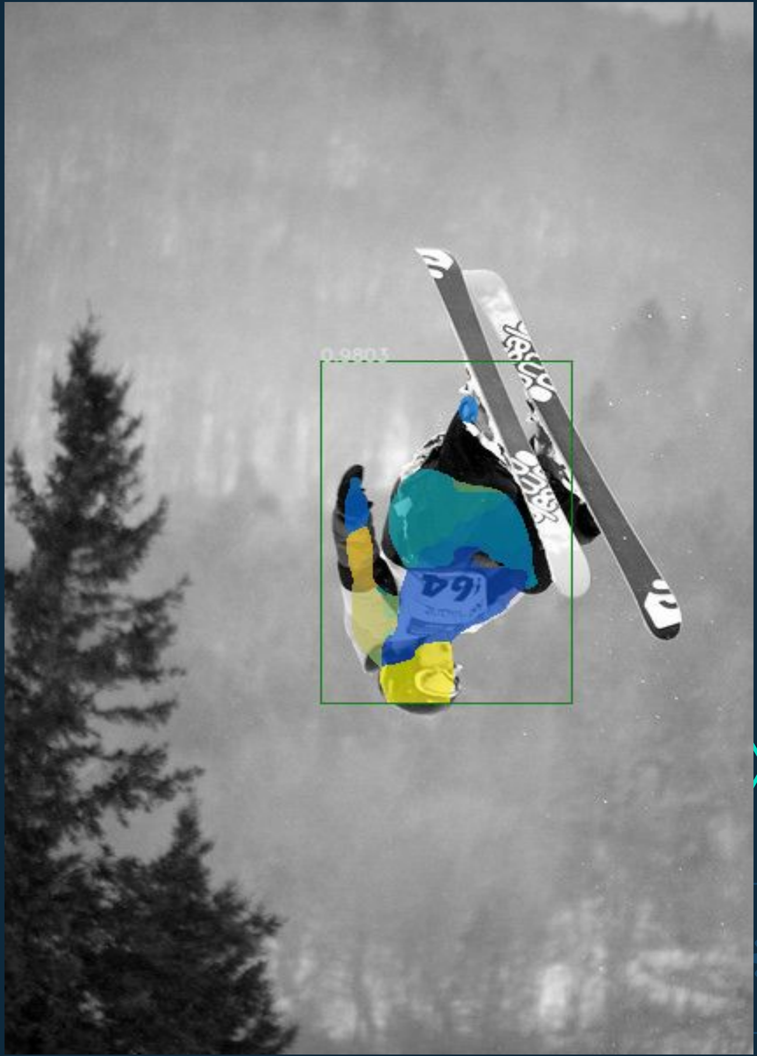

0.9998


Hyde Park, Sydney


0.9951
0.9973
0.9974
0.9639
0.8905
0.8376
Hyde Park, Sydney


0.9951
0.9973
0.9974
0.9639
0.8905
0.8376
Hyde Park, Sydney

Complex poses


0.9803

Resolution


0.9790
0.9644
0.8065
0.9907
0.9996


0.9790
0.9644
0.8065
0.9907
0.9996


0.8722


0.8722


UNITED STATES POST OFFICE
UNITED STATES POST OFFICE
EXIT
EXIT
Food Court


0.9982
0.9474
0.9994
0.8640
0.8702
UNITED STATES POST OFFICE
UNITED STATES POST OFFICE
EXIT
EXIT


UNITED STATES POST OFFICE
UNITED STATES POST OFFICE
EXIT
EXIT
0.9982
0.8640
0.8702
0.9474
0.9994

lighting


0.9310
0.9926


0.9310
0.9926

Occlusions


0.9998

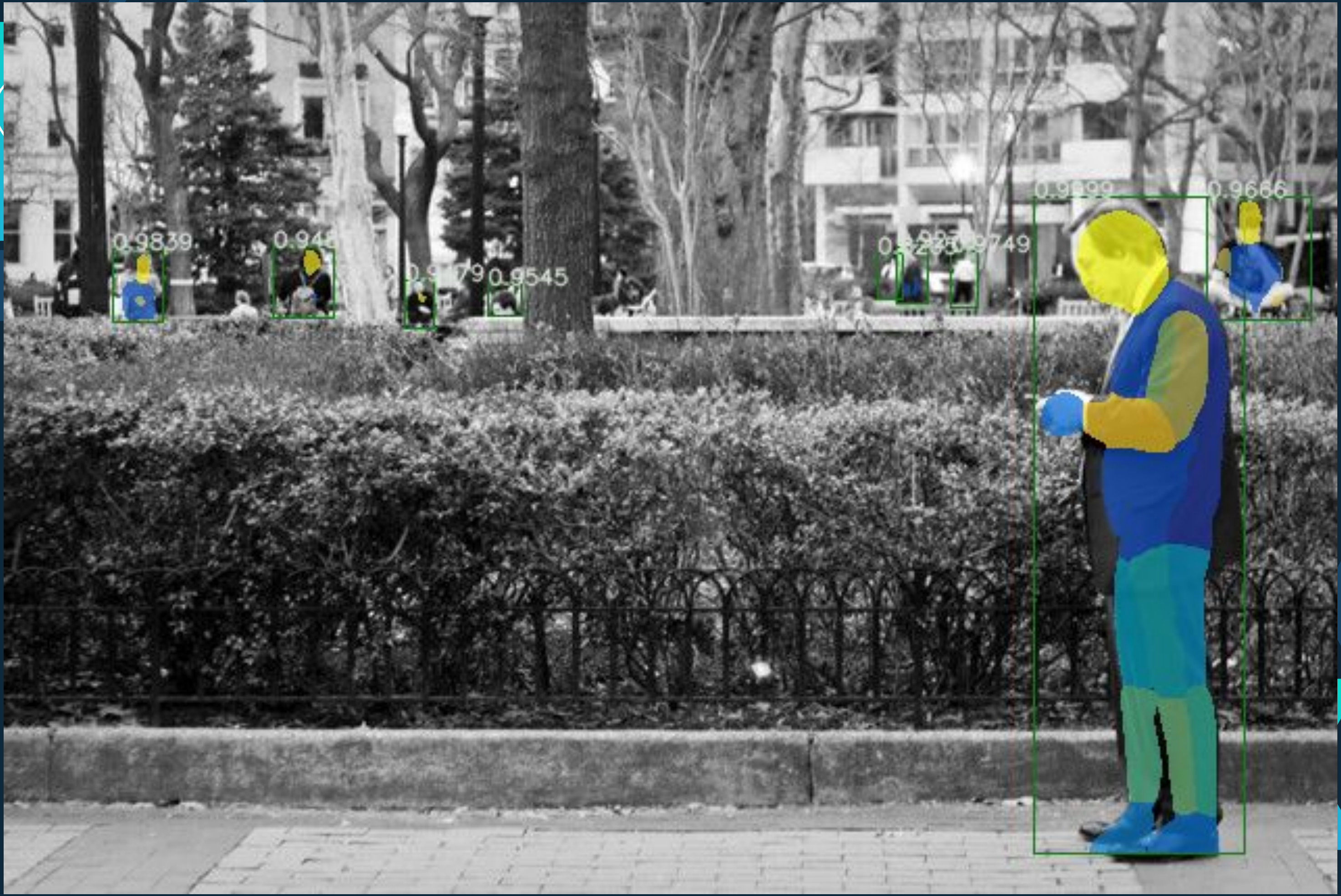

0.9998

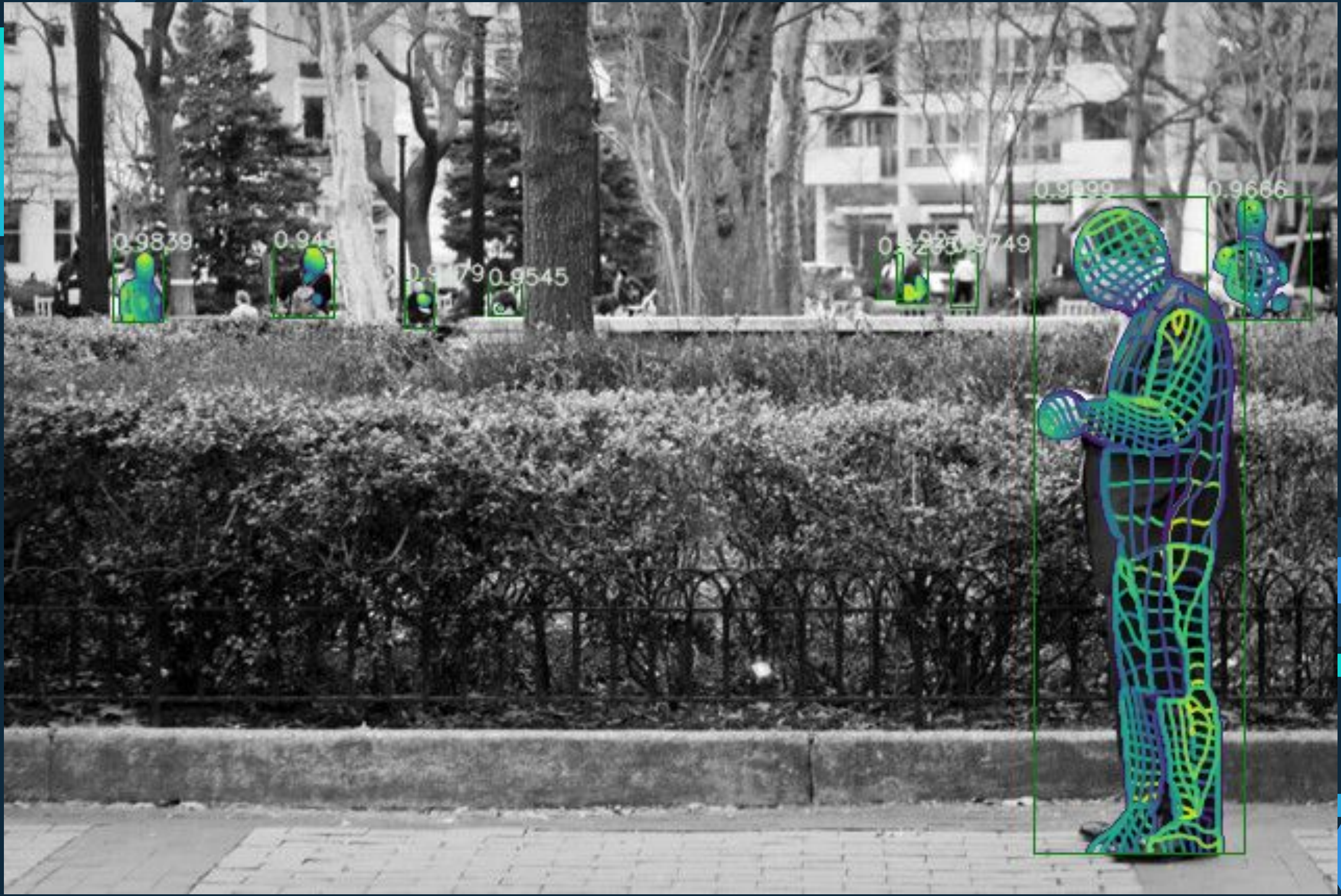

0.9921
0.9996


0.9921
0.9996

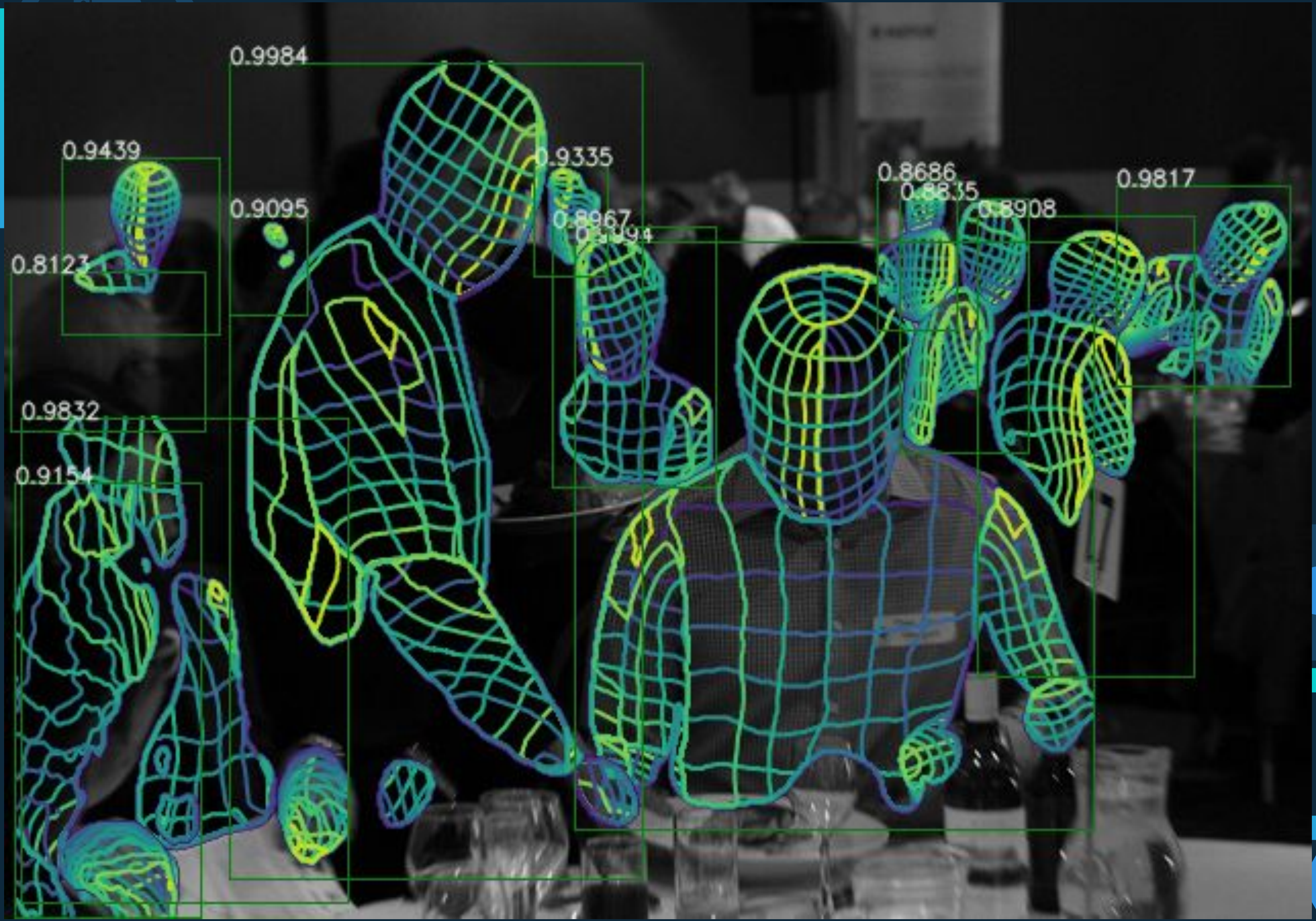

0.9563


0.9563


0.9839
0.9545
0.9666


0.9839
0.9545
0.9999
0.9666


17


0.9984
0.9439
0.9335
0.9095
0.8967
0.8686
0.9817
0.8908
0.8123
0.9832
0.9154


0.9984
0.9439
0.9335
0.9095
0.8123
0.8686
0.8835
0.8908
0.9817
0.9832


0.8684
0.9295
0.8768


0.8684
0.9223
0.9295
0.8768