








PressureVision: Estimating Hand Pressure from a Single RGB Image: Supplementary

Patrick Grady¹, Chengcheng Tang², Samarth Brahmabhatt³, Christopher D. Twigg², Chengde Wan², James Hays¹, and Charles C. Kemp¹

¹ Georgia Institute of Technology

² Meta Reality Labs

³ Intel Labs

1 Introduction

This document provides additional details and qualitative results to supplement the main paper. Section 2 provides more details about the PressureVisionDB dataset. Section 3 gives more details about training PressureVisionNet, and Section 4 provides further results.

2 PressureVisionDB Details

2.1 Table of Actions

During data collection of PressureVisionDB, participants complete a list of actions once for one hand, then repeat for the other hand. These actions are listed in Table 3.

2.2 Data Sampling Rate

Data from four OptiTrack Prime Color cameras is captured at 1080p resolution at 60 Hz, and 185x105 pressure images are collected from the Sensel Morph at approximately 115 Hz. Due to the large size of the dataset, the data is subsampled to 15 Hz for all experiments.

2.3 Lighting Combinations

The capture setup used during the collection of PressureVisionDB includes four RGB cameras and three light sources: “left”, “center”, and “right” (see Figure 1). Figure 2 shows images captured from each camera for each lighting condition.

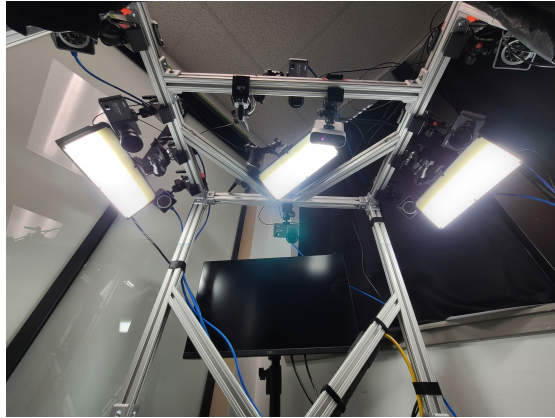


Fig. 1. The top of the capture setup includes four OptiTrack Prime Color cameras and three OptiTrack eStrobe light sources. The light sources can be turned on or off in any combination.

3 PressureVisionNet Details

PressureVisionNet was implemented using the PyTorch library [5] and used the SE-ResNeXt50 encoder [2,3,7] and FPN decoder [4] implementations from the *segmentation-models-pytorch* project [8]. The encoder network weights are initialized from pretraining on ImageNet [1], while the decoder weights are randomly initialized.

PressureVisionNet is optimized with the Adam optimizer with a batch size of 8. The network is trained with a learning rate of $1e - 3$ for 100k iterations, then with a learning rate of $1e - 4$ for 500k iterations. No data augmentation is performed during training.

3.1 Input Images

Images are cropped to include a 50-pixel border around the pressure sensor, and are resized to 480x384 pixels for the network. During experiments with reduced image quality (Section 4.3), images are first downsampled, then upsampled, such that the network architecture may remain constant.

The pressure image from the pressure sensor is warped into image space using a homography transform.

4 Further PressureVisionNet Results

4.1 Qualitative Results

Further results from PressureVisionNet are included in Figures 3-8. These results are *randomly* selected from frames in the test set. As there a large number of

frames where the hand is not present or is not near the object, frames with no ground truth pressure are not shown.

4.2 Temporal PressureVisionNet

Motion can be useful for machine perception of human activities [6]. To evaluate if motion provides salient cues for visual hand pressure estimation, we extend our base model to incorporate temporal information. The PressureVisionNet (PV-Net) structure is modified by concatenating the encoded features from multiple frames before the decoder module. The encoder networks have tied weights. The network was trained and tested on sequences of 4 frames, 0.2 sec apart. As expected, this temporal network outperforms our base model (Table 1). However, due to the ease of use and similar performance of the single-frame model, we focus on this for all experiments. Notably, the single-frame model may still leverage some motion cues by perceiving motion blur present in our RGB images.

Table 1. A version of PressureVisionNet incorporating temporal information slightly outperforms our single-frame model.

Method	Frames	Temporal Acc	Contact IoU	Vol. IoU	MAE
PV-Net	1	96.2%	55.8%	41.3%	39.9 Pa
PV-Net-Temporal	4	96.8%	56.6%	43.1%	39.4 Pa

4.3 Degraded Imagery

We also evaluated the performance of our network trained and tested with degraded images (Table 2). PressureVisionNet uses 480x384 images around the contact surface, captured in a controlled, well-lit setting. We show the performance of a model trained and tested with monochrome images. The drop in Volumetric IoU performance suggests that the model uses color when inferring pressure, but other information is sufficient for contact estimation. We also evaluated models trained and tested with lower resolution images. There was a modest drop in performance until the resolution went below 120x96 pixels. This suggests that our approach may be applicable to images captured in a less controlled setting in which hands occupy a smaller part of the image.

Table 2. Performance decreases slowly as resolution decreases. Evaluating on monochrome images shows a reduction in Volumetric IoU.

Method	Temporal Acc	Contact IoU	Volumetric IoU	MAE
Zero Guesser	53.7%	0.0%	0.0%	51.9 Pa
RGB-15x12	78.9%	17.6%	11.7%	55.0 Pa
RGB-30x24	91.4%	38.5%	28.0%	49.6 Pa
RGB-60x48	94.9%	50.5%	36.2%	45.5 Pa
RGB-120x96	95.8%	53.6%	39.5%	42.8 Pa
RGB-240x192	95.9%	54.2%	39.3%	40.2 Pa
Mono-480x384	96.1%	53.2%	37.7%	40.9 Pa
PressureVisionNet RGB-480x384	96.2%	55.8%	41.3%	39.9 Pa

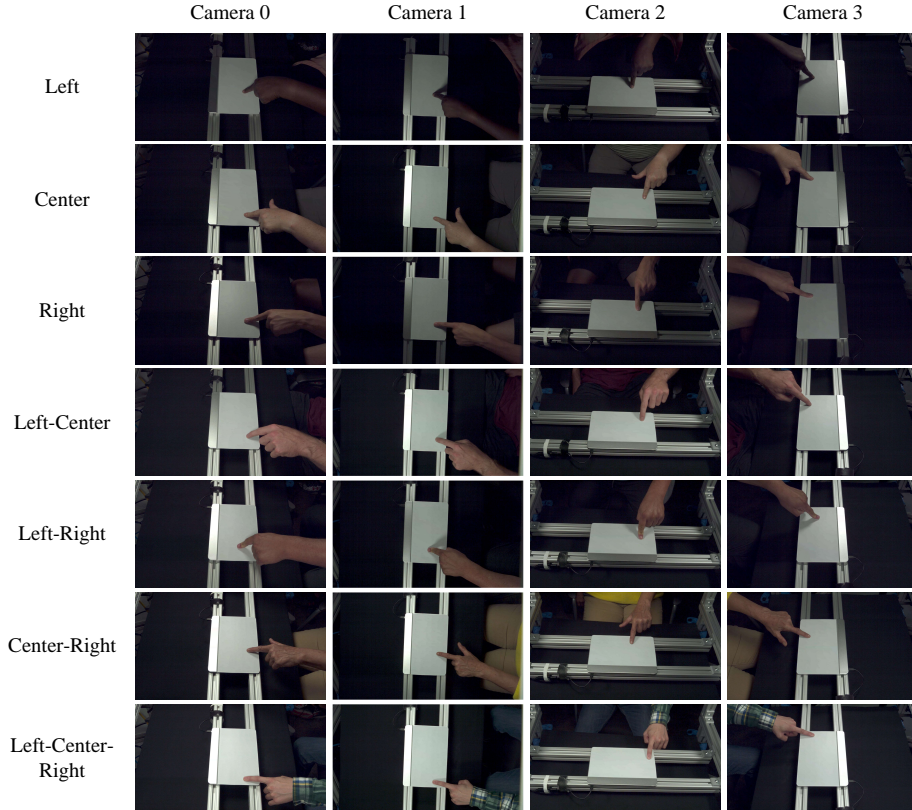


Fig. 2. Data is captured with four cameras. Three lights are turned on and off to create seven lighting conditions.

Table 3. The list of actions executed for each hand during the collection of PressureVisionDB. For actions where the *force level* is given, the participant is prompted to apply the specified amount of force. For actions where it is not specified, the participant may apply any force level.

Action name	Repetitions	Force Level
Calibration routine	1	-
Press index finger	5	Low
Press index finger	5	High
Press index finger	5	No Contact
Index finger, pull towards	5	-
Index finger, push left	5	-
Index finger, push away	5	-
Index finger, push right	5	-
Press palm	5	Low
Press palm	5	High
Press palm	5	No Contact
Press palm and fingers	5	Low
Press palm and fingers	5	High
Press palm and fingers	5	No Contact
Press fingers	5	Low
Press fingers	5	High
Press fingers	5	No Contact
Press finger one by one, flat	3	Low
Press finger one by one, flat	3	High
Press finger one by one, cupped	3	Low
Press finger one by one, cupped	3	High
Fingertips down, push away	5	-
Fingertips down, pull towards	5	-
Grasp edge, thumb down, uncurled fingers	5	-
Grasp edge, thumb up, uncurled fingers	5	-
Grasp edge, thumb down, curled fingers	5	-
Grasp edge, thumb up, curled fingers	5	-
Pinch, thumb down	5	Low
Pinch, thumb down	5	High
Pinch, thumb down	5	No contact
Pinch, thumb up	5	Low
Pinch, thumb up	5	High
Pinch, thumb up	5	No contact
Draw word	5	-
Pinch-zoom	5	-

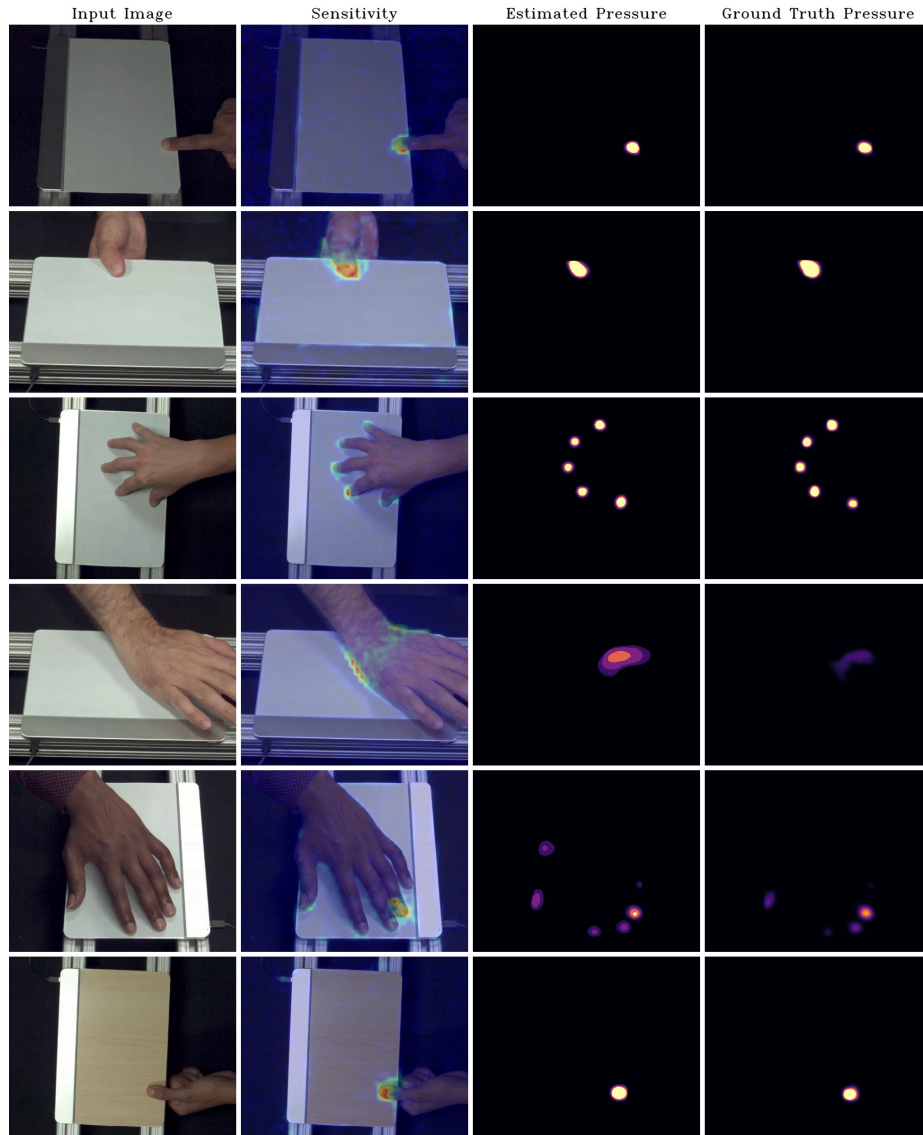


Fig. 3. Qualitative results *randomly* selected from the test set. Only frames with ground truth pressure are shown.

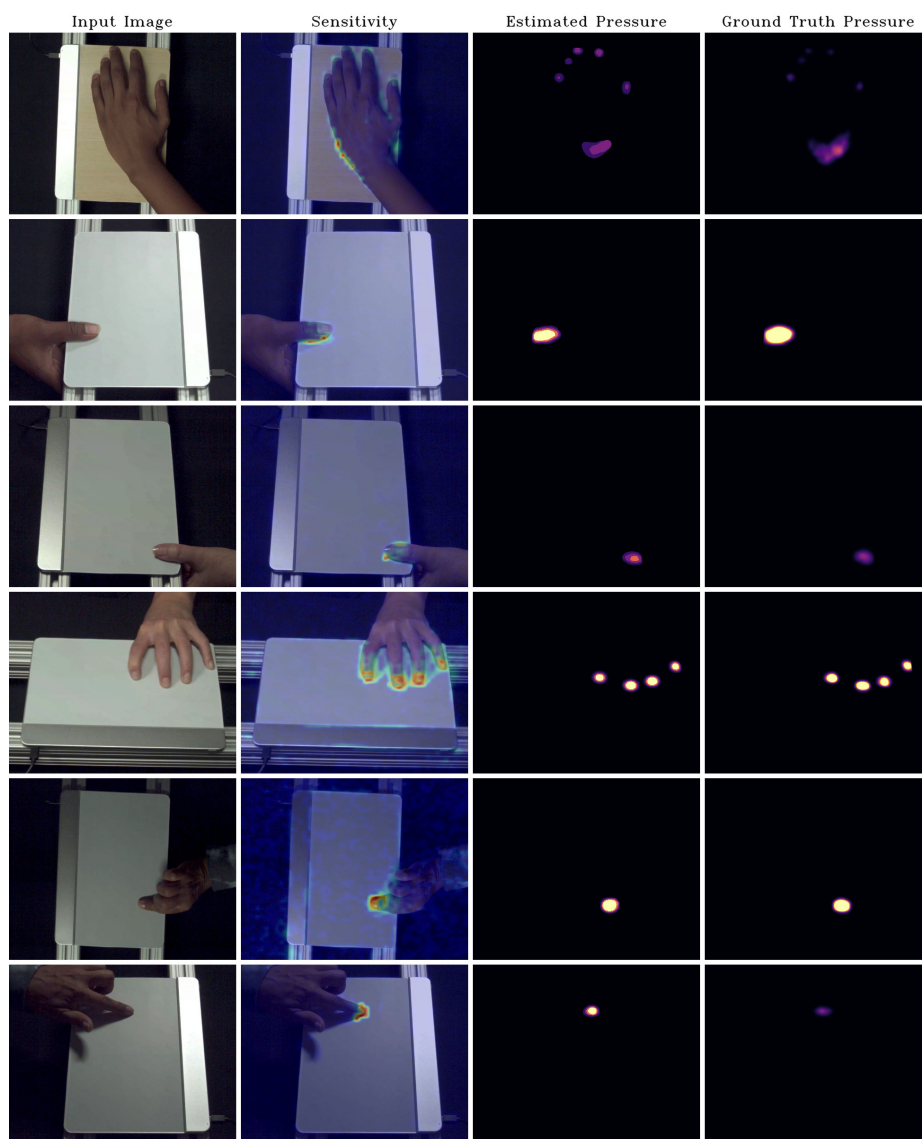


Fig. 4. Qualitative results *randomly* selected from the test set. Only frames with ground truth pressure are shown.

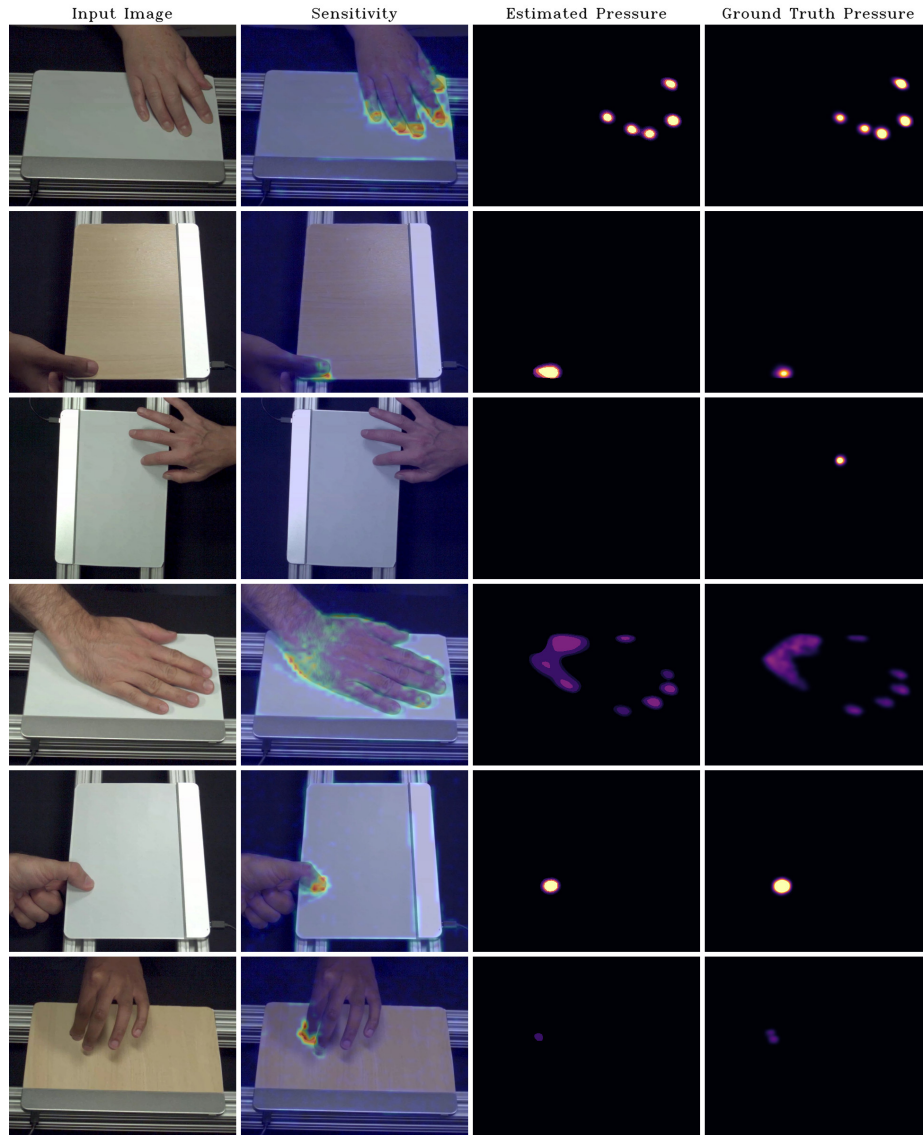


Fig. 5. Qualitative results *randomly* selected from the test set. Only frames with ground truth pressure are shown.

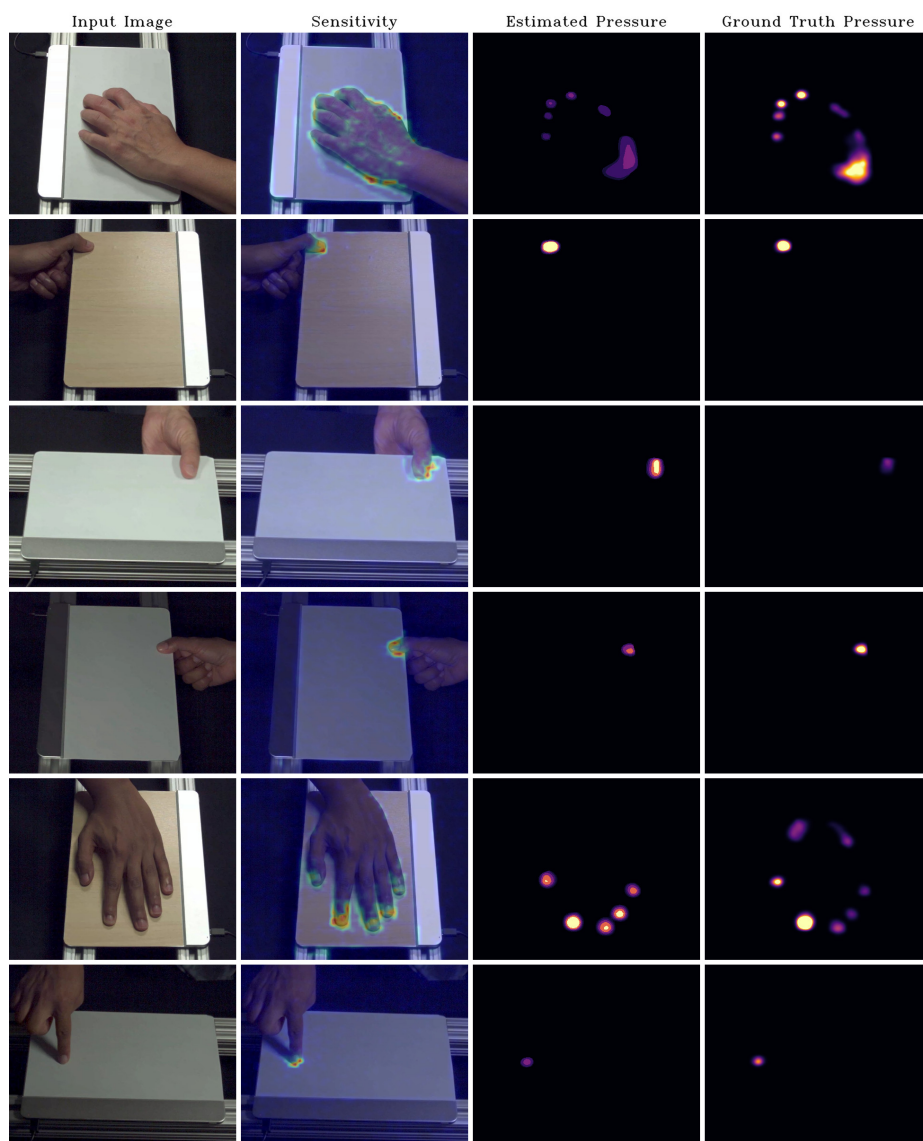


Fig. 6. Qualitative results *randomly* selected from the test set. Only frames with ground truth pressure are shown.

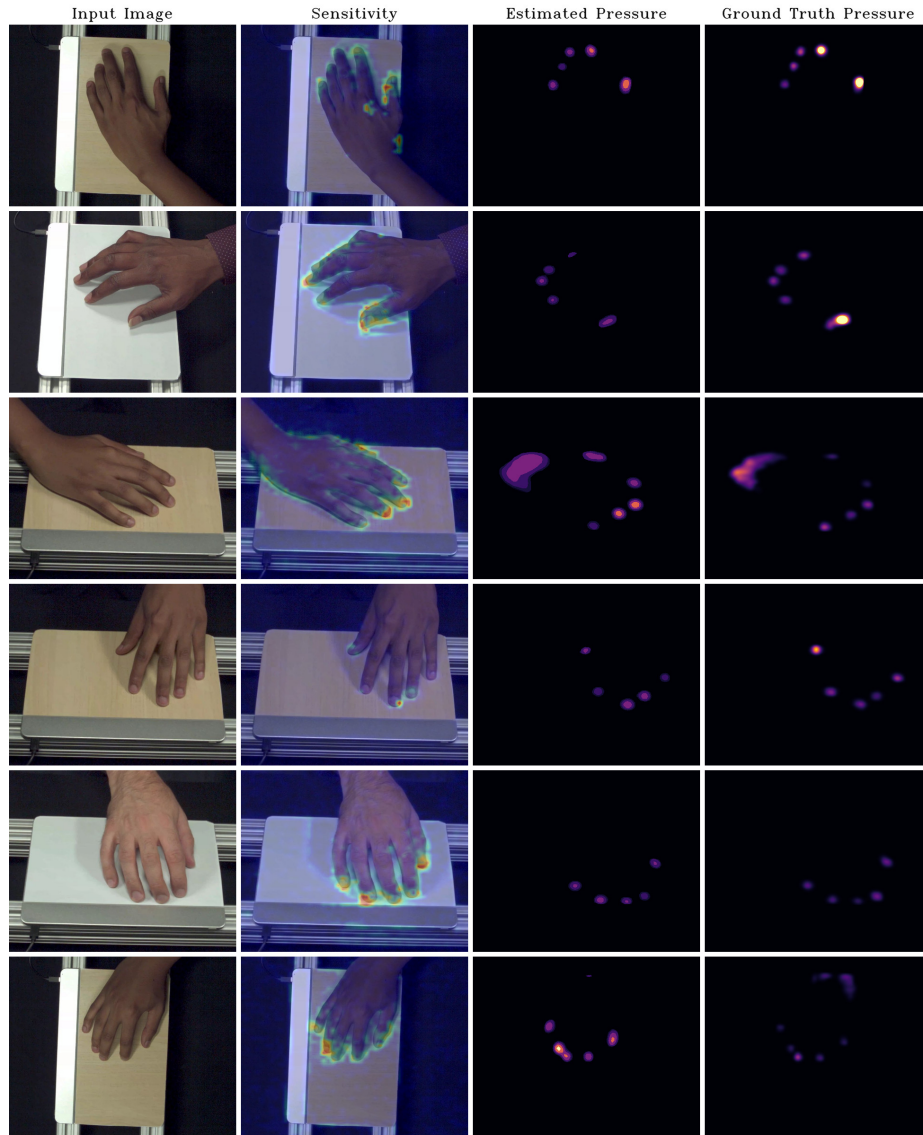


Fig. 7. Qualitative results *randomly* selected from the test set. Only frames with ground truth pressure are shown.

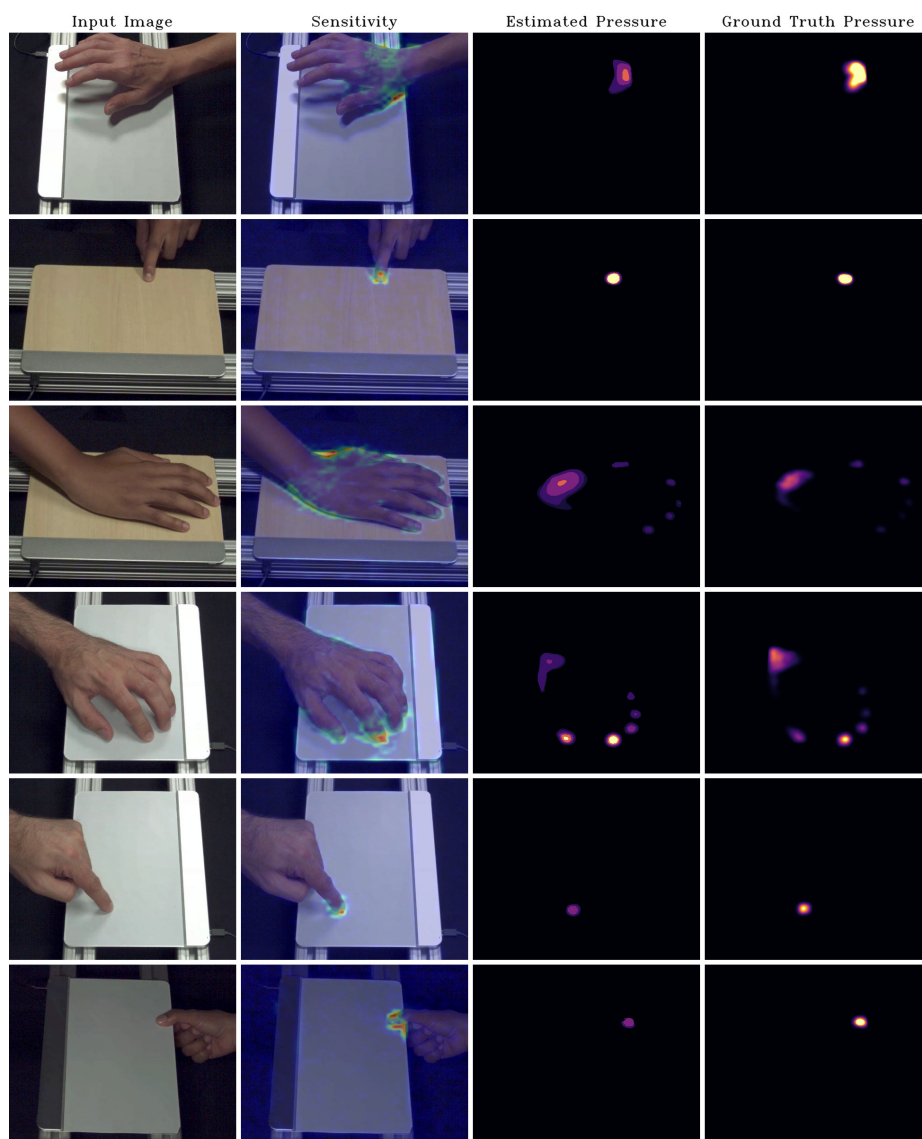


Fig. 8. Qualitative results *randomly* selected from the test set. Only frames with ground truth pressure are shown.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. IEEE (2009)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
3. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141 (2018)
4. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125 (2017)
5. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)
6. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. p. 568–576. NIPS’14 (2014)
7. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5987–5995 (2017)
8. Yakubovskiy, P.: Segmentation models pytorch (2020)