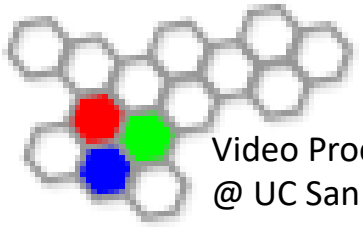
Hand Segmentation for Hand-Object Interaction from Depth Map

Byeongkeun Kang[*], Kar-Han Tan[†], Nan Jiang[*], Hung-Shuo Tai[†], Daniel Tretter[‡], Truong Nguyen[*]

[*]UC San Diego, [†]NovuMind Inc., [‡]Hewlett-Packard, Inc.



Video Processing Lab
@ UC San Diego

UC San Diego

Motivation

Hand segmentation

Hand pose estimation

Hand tracking

Gesture recognition

Sign language recognition

Interface in vehicle

Robot / drone control

Medical application

Augmented / virtual reality

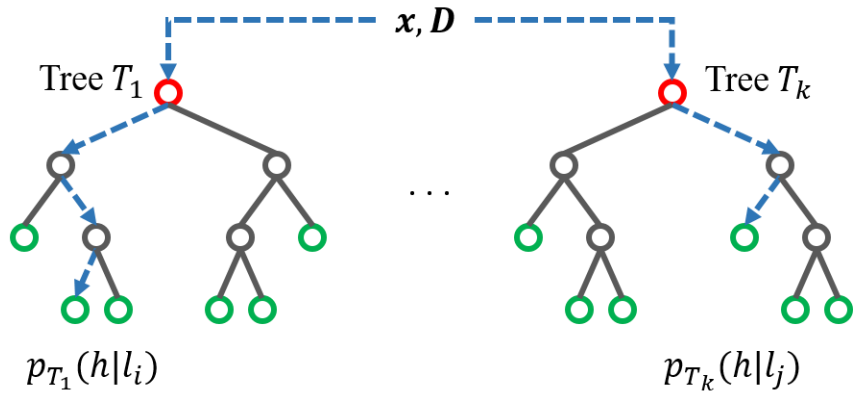
Related Works: Hand Segmentation

- Color image
  - Skin color
    - Threshold in HSV space [1-4]
    - Color histogram-based model [5]
    - Gaussian mixture model [6]
  - Limitation: other body parts, skin color objects, skin pigment difference, light condition variations

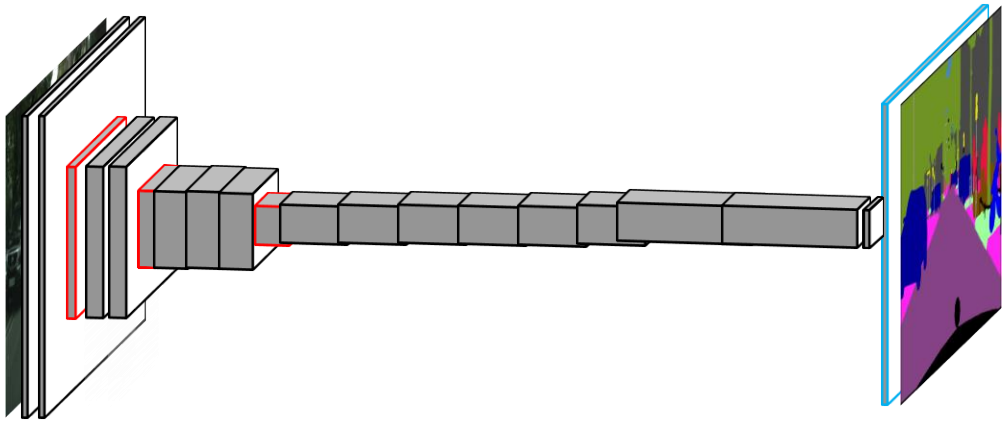
- Depth map
  - Wristband [7-9]
  - Random decision forest (RDF) [10-12]

Related Works: Semantic Segmentation

- Random decision forest (RDF) [10-12]
  - Accurate / robust
  - Short processing time



- Convolutional neural network [13-14]
  - More accurate / robust
  - Longer processing time
    (hard to achieve real-time)

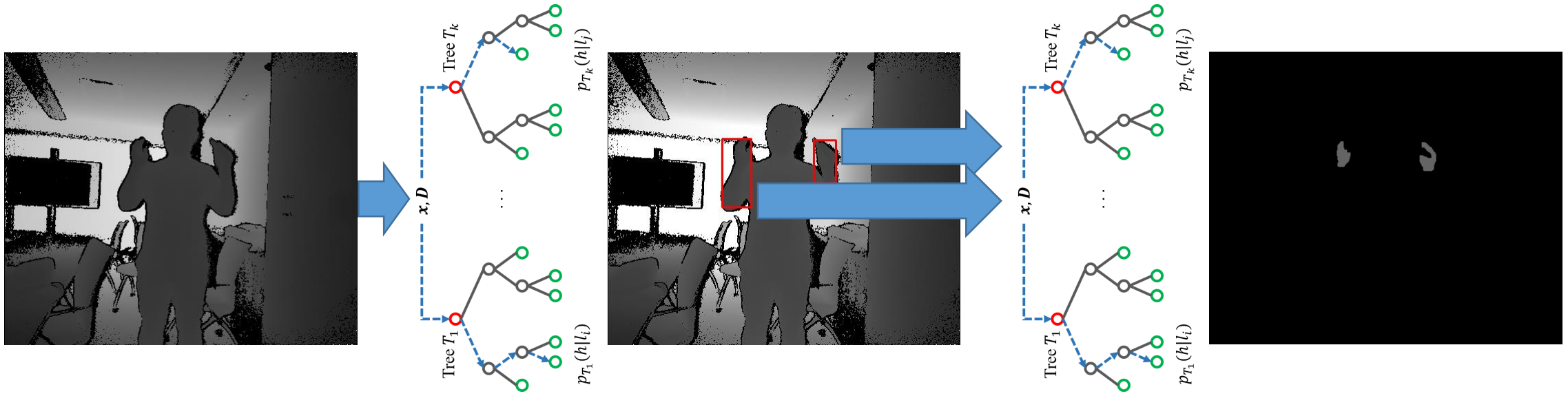


Abstract

- Goal: hand segmentation algorithm for hand-object interaction
- Input: only a depth map
  - To avoid the limitations of skin color-based methods
- Method: two-stage RDF
  - 1st RDF: hand detection
  - 2nd RDF: hand segmentation
- Result: high accuracy in short processing time (~10ms / frame)

Method: Two-stage RDF

- 1st RDF: detection on an entire depth map.
- 2nd RDF: segmentation in the detected region.

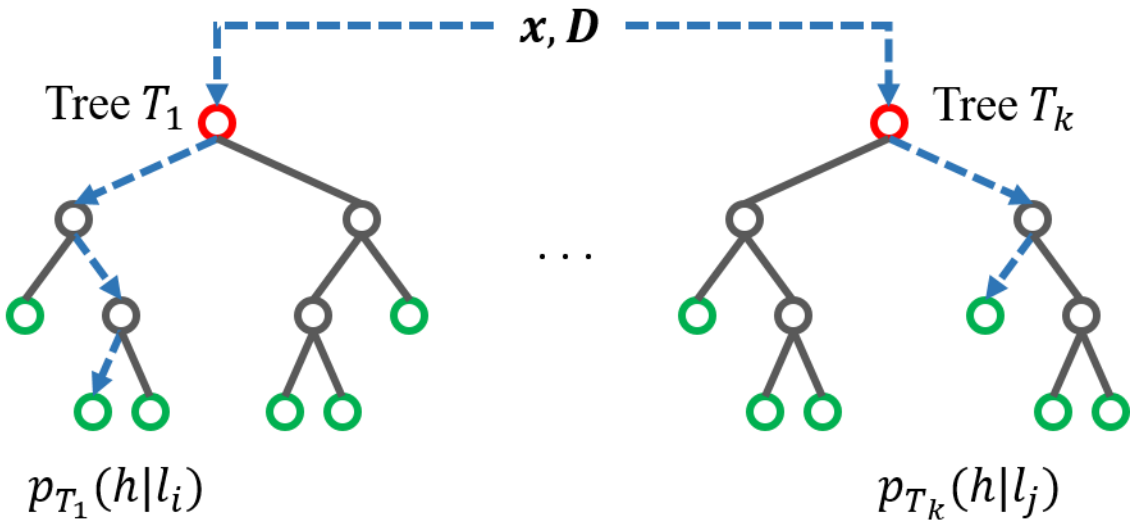


1st RDF: detection

2nd RDF: segmentation

Method: RDF

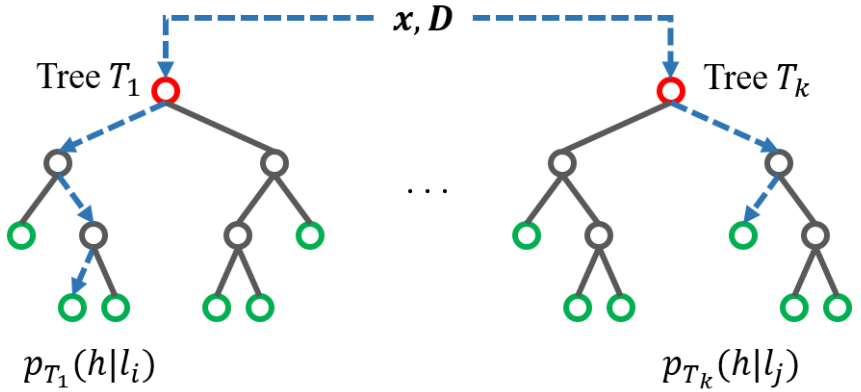
- Consists of a collection of decision trees.
- Each tree is composed of a root node, splitting nodes, and leaf nodes.



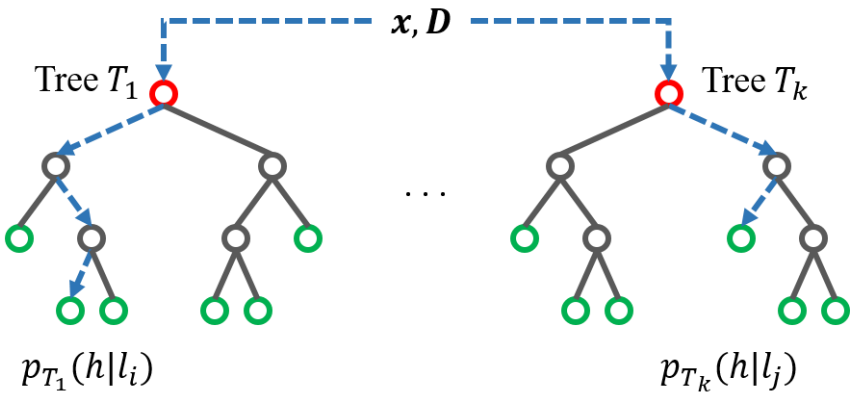
Random decision Forest. Red: root nodes, Black: splitting nodes, Green: leaf nodes.

Method: RDF (Training)

1. Select training data (partial data points in a set of images).
2. Learn a split function at each splitting node.
    1. Generate possible candidates.
    2. Select the candidate with the minimum loss.
3. Repeat (2) until a leaf node.
4. Store conditional probability at each leaf node.



Method: RDF (Testing)



1. Each pixel enters root nodes in the forest.
2. Classify the pixel to child nodes until a leaf node.
3. Load the learned conditional probability.
4. Average the probabilities.

$$p(h|x) = \frac{1}{|\mathrm{T}|}\sum_{T\in\mathrm{T}} p_T(h|l)$$

- h: class, x: data, l: leaf node
- T: learned forest, $|\mathrm{T}|$: the number of trees

Method: Modified Bilateral Filter

- RDF's prediction is independent for each pixel.

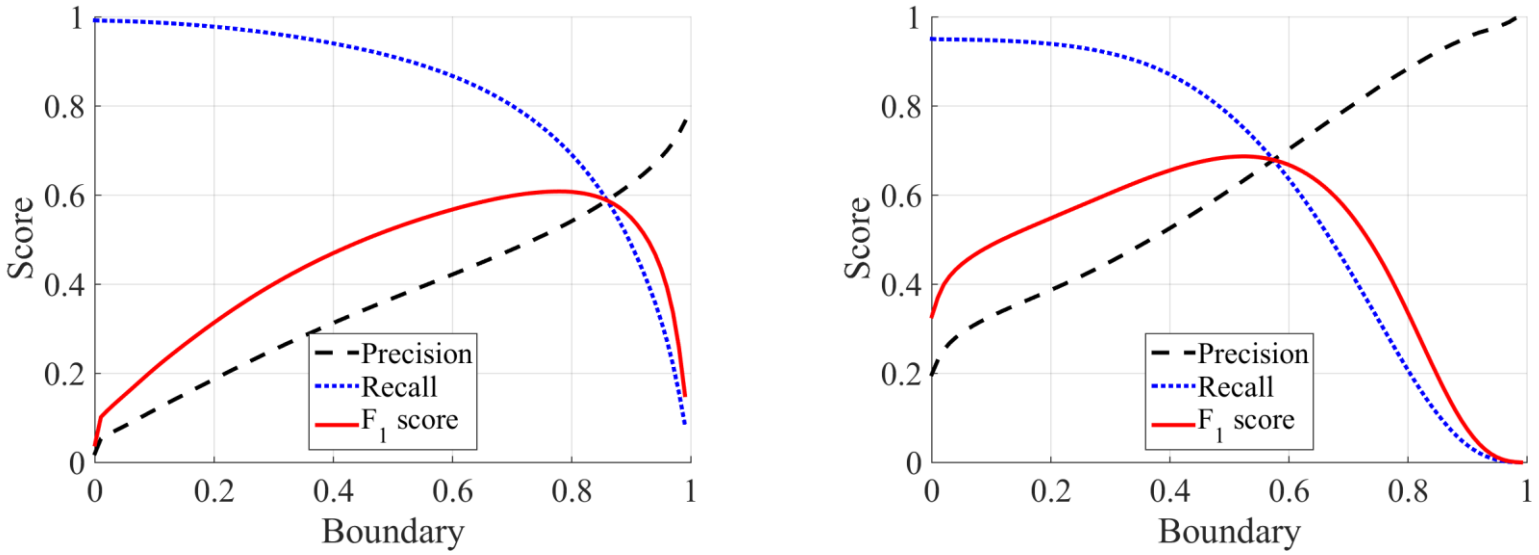
  ⇒ Stabilize by averaging the predictions in close distance.

$$\tilde{p}(h|\boldsymbol{x}) = \frac{1}{W}\sum_{\boldsymbol{x}_i \in \Omega} g_r\left(\left|\boldsymbol{D}_{\boldsymbol{x}_i} - \boldsymbol{D}_{\boldsymbol{x}}\right|\right) g_s(\|\boldsymbol{x}_i - \boldsymbol{x}\|) p(h|\boldsymbol{x}_i)$$

  - $g_r(\cdot)$, $g_s(\cdot)$: Gaussian functions
  - $\left|\boldsymbol{D}_{\boldsymbol{x}_i} - \boldsymbol{D}_{\boldsymbol{x}}\right|$: depth difference
  - $\|\boldsymbol{x}_i - \boldsymbol{x}\|$: spatial distance

Method: Decision Boundary

- Typical boundary is 0.5 ⇒ Not the best parameter.
- Search with the step size of 0.01 exhaustively.



Scores depending on the decision boundary. Left: RDF in the first stage. Right: RDF in the second stage.

Dataset: HOI dataset

- 27,525 pairs of depth maps and ground truth labels (training: 19,470, validation: 2,706, testing: 5,349)
- 6 people
- 21 different objects
- Available at https://github.com/byeongkeun-kang/HOI-dataset



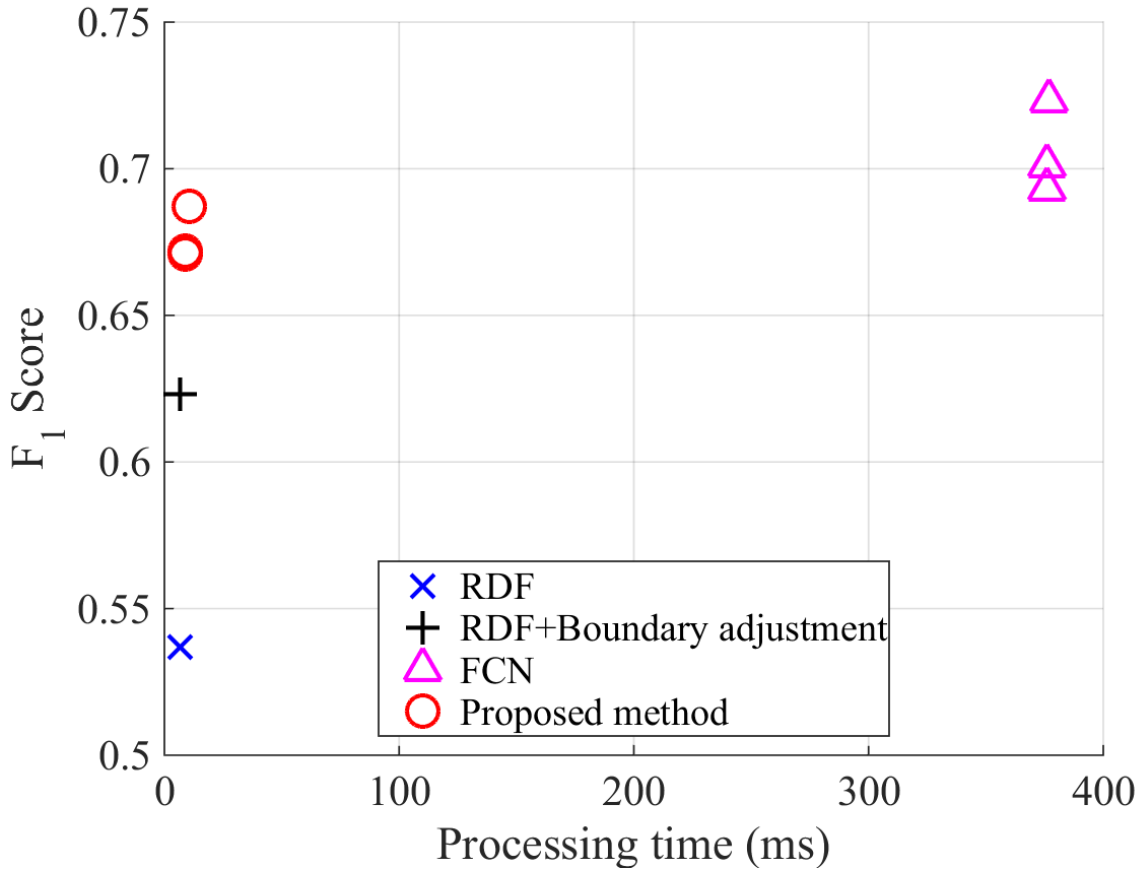
Result: Visual Comparison

Ground truth | RDF [11, 12] | RDF [11, 12] +Adjustment | FCN-8s [13, 14] | Proposed method

Result: Quantitative Comparison

Method			Score			Processing time (ms)
Method	Boundary	Filter	Precision	Recall	F_1 score	
RDF [1, 2]	0.50	-	38.1	91.2	53.7	6.7
	0.78	-	54.5	72.7	62.3	6.7
FCN-32s [3, 4]			70.0	68.6	69.3	376
FCN-16s [3, 4]			68.0	72.2	70.1	376
FCN-8s [3, 4]			70.4	74.4	72.3	377
Proposed method	0.50, 0.50	-	59.2	77.4	67.1	8.9
	0.50, 0.52	-	60.8	75.1	67.2	8.9
	0.50, 0.52	11×11	62.9	75.6	68.7	10.7

Result: Analysis of Accuracy and Efficiency



Summary

- Task: hand segmentation for hand-object interaction
- Input: only a depth map
- Method: two-stage RDF
- Result: high accuracy in short processing time

Questions

References (1/2)

[1] J. Romero, H. Kjellstrom, and D. Kragic, “Hands in action: real-time 3d reconstruction of hands in interaction with objects,” in Robotics and Automation (ICRA), 2010 IEEE International Conference on, May 2010.

[2] J. Romero, H. Kjellstrom, C. H. Ek, and D. Kragic, “Non-parametric hand pose estimation with object context,” Image Vision Comput., vol. 31, no. 8, Aug. 2013.

[3] I. Oikonomidis, N. Kyriazis, and A.A. Argyros, “Full dof tracking of a hand interacting with an object bymodeling occlusions and physical constraints,” in Computer Vision (ICCV), 2011 IEEE International Conference on, 2011.

[4] A. A. Argyros and M. I. A. Lourakis, “Real-time tracking of multiple skin-colored objects with a possibly moving camera,” in Computer Vision - ECCV, 2004.

[5] Y. Wang, J. Min, J. Zhang, Y. Liu, F. Xu, Q. Dai, and J. Chai, “Video-based hand manipulation capture through composite motion control,” ACM Trans. Graph., vol. 32, no. 4, July 2013.

[6] D. Tzionas and J. Gall, “3d object reconstruction from hand-object interactions,” in International Conference on Computer Vision (ICCV), Dec. 2015.

[7] C. Qian, X. Sun, Y.Wei, X. Tang, and J. Sun, “Realtime and robust hand tracking from depth,” in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, June 2014.

[8] B. Kang, Y. Lee, and T. Nguyen, “Efficient hand articulations tracking using adaptive hand model and depth map,” in Advances in Visual Computing, Dec. 2015.

References (2/2)

[9] B. Kang, S. Tripathi, and T. Nguyen, “Real-time sign language fingerspelling recognition using convolutional neural networks from depth map,” in Pattern Recognition, 2015 3rd IAPR Asian Conference on, Nov. 2015.

[10] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi, “Accurate, robust, and flexible real-time hand tracking,” in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015.

[11] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, “Realtime continuous pose recovery of human hands using convolutional networks,” ACM Trans. Graph., 2014.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, June 2011.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[14] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.