First-Person View Hand Segmentation of Multi-Modal Hand Activity Video Dataset

Sangpil Kim[1]
kim2030@purdue.edu
Hyung-gun Chi[1]
chi45@purdue.edu
Xiao Hu[2]
hu440@purdue.edu
Anirudh Vegesana[2]
avegesan@purdue.edu
Karthik Ramani[1]
ramani@purdue.edu

[1] C Design Lab | Purdue University
West Lafayette, Indiana, USA

[2] Electrical and Computer Engineering | Purdue University
West Lafayette, Indiana, USA

Abstract

First-person-view videos of hands interacting with tools are widely used in the computer vision industry. However, creating a dataset with pixel-wise segmentation of hands is challenging since most videos are captured with fingertips occluded by the hand dorsum and grasped tools. Current methods often rely on manually segmenting hands to create annotations, which is inefficient and costly. To relieve this challenge, we create a method that utilizes thermal information of hands for efficient pixel-wise hand segmentation to create a multi-modal activity video dataset. Our method is not affected by fingertip and joint occlusions and does not require hand pose ground truth. We show our method to be 24 times faster than the traditional polygon labeling method while maintaining high quality. With the segmentation method, we propose a multi-modal hand activity video dataset with 790 sequences and 401,765 frames of "hands using tools" videos captured by thermal and RGB-D cameras with hand segmentation data. We analyze multiple models for hand segmentation performance and benchmark four segmentation networks. We show that our multi-modal dataset with fusing Long-Wave InfraRed (LWIR) and RGB-D frames achieves 5% better hand IoU performance than using RGB frames.

1 Introduction

Hands are crucial in many industrial computer vision applications, such as augmented reality, virtual reality, or human-computer interaction. Recognizing hands with vision systems is necessary to interact between people and digital devices. Therefore, understanding hands with computer vision systems has been deeply explored through hand tracking [41, 54], hand pose estimation [18, 21, 22, 27, 29], grasp detection [10, 47], hand gesture recognition [58], multi-view prediction [29], and hand-action classification [50]. These works require segmenting hands from the background to increase the accuracy of performance.

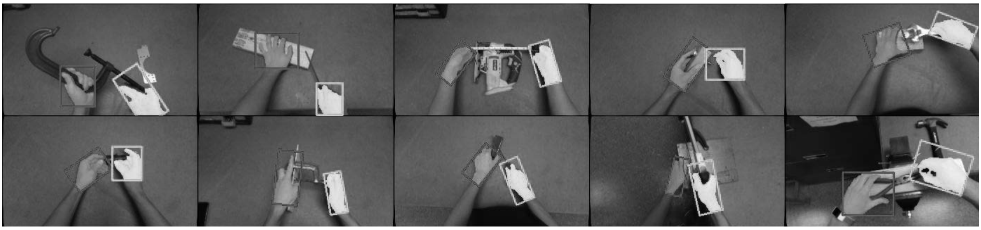

Figure 1: Sample frames from our hand segmentation video dataset. Red and green masks represent left and right hand, respectively.

Most of these applications require first-person-view images of hands in actions with tools. However, segmenting hands interacting with tools from the background is a challenging problem because (a) fingertips are heavily occluded by the hand dorsum and tools in the first-person view, (b) tools are held with various grasps, and (c) shapes of tools or objects are infinite. The traditional approach to create a RGB hand segmentation video dataset is through manual pixel-wise labeling [5, 7, 32, 50]. However, time and cost of person-in-the-loop segmentation grows linearly as the number of frames increases, which reduces the scalability. Therefore, developing an efficient segmentation method is important for creating a segmented hand video dataset.

We utilize hand temperature as prior information for the labeling process. However, pixels from Long-Wave InfraRed (LWIR) thermal images of hands may falsely include pixels from the surroundings that share the same temperature. In our method, we utilize crowd workers to provide an Axis-Aligned Bounding Box (AABB) to localize the detailed boundary of each hand. We further relax the AABB creation task by training a tracker with a small amount of AABB results. The tracker learns the shape features of hands and then uses them to estimate an Oriented Minimum Bounding Box (OMBB) for each hand. Therefore, we use both spatial and thermal features of hands to segment them from the background and tools. Our approach is effective regardless of finger tips or finger joints occlusion and does not require hand pose ground truth. We prove that our method is much more efficient than the traditional pixel-wise labeling tasks (sec. 6) while maintaining a high performance (sec. 7).

Optimizing deep neural networks with a single modality may lead to failures when the network fails to extract distinctive features from the input source. Multiple modalities are used to provide distinctive features to the networks [14, 56, 60] and has been an emerging area [11, 26, 44] due to the enhancement of computation power and sensors. Human body temperature is relatively constant [9] and has been widely used in pedestrian detection [28], biological image processing [19], and gesture recognition [4]. An additional advantage of LWIR is that it is invariant to colors, textures, and lighting conditions. Color information may mislead vision systems distinguishing shape features. Therefore, we create a multi-modal (LWIR, RGB, and depth) hand segmentation video dataset which consists of 790 sequences and 401,765 frames of "hands using tools" videos. Compared to the other existing hand segmentation datasets, our dataset contains three different modalities and head-mounted camera Inertial Measurement Unit (IMU) information. Sample frames from our dataset are shown in Figure 1 and a detailed comparison with other hands datasets is shown in Table 1.

With the video dataset, we analyze fusing three modalities with DeepLabV3+ [13] and benchmark five different state-of-the-art segmentation methods [13, 49, 50, 51, 61]. We observe that the neural networks can automatically learn important cues from three different modalities: LWIR, RGB, and depth. The jointly-learned features of these three modalities prevent confusion between hands and backgrounds.

The main contributions of this paper are as follows:

- We collect large-scale action videos in a first-person view which contain LWIR, RGB, depth, and IMU information. This dataset can be used for hand segmentation research using multiple modalities.
- We develop a framework that can significantly reduce segmentation efforts by leveraging hand temperature for creating pixel-wise hand segmentation ground truth when a person is holding tools. Our method does not require hand pose labels nor a hand mesh model.
- We analyze the effectiveness of multiple modalities for hand segmentation task with deep neural networks and found the optimal combination which is fusing thermal (LWIR), RGB, and depth modalities.

Pixel-wise RGB Hand Seg. Dataset	Egocentric	Both Hands	Depth	LWIR	IMU	#Frames
HandNet [57]			✓			202,928
HOF [50]		✓				300
Hand-CNN [38]		✓				40,531
EgoHands+ [50]	✓	✓				800
EYTH [50]	✓	✓				1,290
EgoHands [5]	✓	✓				4,800
WorkingHands [48]	✓	✓	✓			3,900
Ours	✓	✓	✓	✓	✓	401,765

Table 1: Comparison table of pixel-wise hand segmentation datasets. We exclude *syntactic* frames on WorkingHands. HandNet provides RGB frames, but hands are covered with wires.

2 Related work

Efficient annotation methods are crucial in creating labels for large-scale video datasets. Although complex boundaries can be traced manually, labeling image/video datasets with object masks is extremely time consuming [34]. A successful method is to have a neural network produce polygon annotations of objects interactively using humans-in-the-loop to improve the accuracy [1, 31, 39]. The machine can provide the human with information to manipulate [3] and generate segments [2], matting layers [59], or boundary fragments [43]. Sequences in video datasets consist of similar frames which contain redundant information. For moving objects especially, several notable and established annotation tools with auto-tracking demonstrate great performance on efficiency improvement [6, 46]. Unlike these works, we extract shape features from thermal frames for pixel-wise hand segmentation. Therefore, our annotation pipeline is invariant to colors and textures. The invariance property improves the quality of labels and efficiency.

Pixel-wise hand segmentation dataset with deep neural networks has been widely used for segmenting objects in videos [12, 40, 42, 61]. Deep-learning-based methods use convolutional neural networks to predict each pixel in an image by extracting feature representations. One of the popular structures is an encoder-decoder structure which projects a high-dimensional image into the latent vector and decodes the latent vector into class-wise pixel space [13, 35]. Researchers create hand segmentation datasets by manually drawing polygon or coloring the hand on RGB frames [33, 38, 48, 48, 50]. Other works use hardware sensors to predict joint position and render with mesh models; however, the sensors are visible in the RGB images [7, 57]. Alternatively, studies showed good performance when utilizing hand pose estimation and mesh models to segment hands in frames [45, 62]. Our method doesn't require training a hand pose network [20] to generate high accuracy on hand pose estimation under heavily occluded situations.

Multi-modal data is widely used to capture robust features because of color and texture invariance and its expression of an object in an explicit 3D information [15, 17, 37, 52]. Therefore, LWIR is widely used in detecting objects [52] and controlling unmanned aerial vehicles due to its outstanding ability of identifying objects with a distinct temperature from the surroundings [24, 49, 55]. MFNet [23] showed that fusing LWIR and RGB frames significantly enhances the segmentation performance. Luo *et al.* [36] uses thermal information to segment coarse point clouds scenes. Thus, many RGB pixels can not be labeled from the coarse labeled point clouds. Most works with LWIR sensors are for scene segmentation and autonomous driving system. Unlike other works, we use LWIR sensors to threshold the hand temperature and remove backgrounds omitting a similar wavelength of thermal radiation with spatial features for hand segmentation dataset creation (sec. 8).

3 Efficient multi-modal hand segmentation

We record our videos with a low-cost non-radiometric thermal camera, Flir Boson 320, to capture relative LWIR data. RGB resolution then is rescaled to match resolution of the LWIR sensor with two camera frustums. We use an Intel D435i depth camera for RGB, depth, and IMU information acquisition. These sensors are placed within a 3D printed case and mounted in front of a helmet to make the camera location consistent over sequences.

Mapping LWIR onto RGB-D
To segment hands with LWIR frames, we narrow down the search space by finding a Thermal Mask (TM), denoted I_{tm}, with LWIR frames, denoted I_{lwir}^{raw}, and $I_{lwir} = \mathcal{T}(I_{lwir}^{raw})$ where $\mathcal{T}$ is transformation function that transforms a LWIR plane to a RGB plane.

$$I_{tm} = \omega(I_{lwir}) \tag{1}$$

The bounded value for a pixel in the target frame corresponding to a spatial location (i,j) in I_{tm} is defined as follows:

$$\omega^{(i,j)}(I_{lwir}^{(i,j)}) = \begin{cases} 1 \ \text{ if } a \leq I_{lwir}^{(i,j)} \leq b & (2) \\ 0 \ \text{ otherwise} & (3) \end{cases}$$

,where a and b are upper bound and lower bound of hand temperature. $I_{tm} \in [0,1]^{H \times W}$ and $I_{lwir} \in \mathbb{R}^{H \times W}$. We map the I_{lwir}^{raw} onto the RGB frame, denoted I_{rgb}, with depth maps. To align depth maps and I_{lwir}^{raw}, we find the spatial relationship between the I_{lwir}^{raw} and depth camera. A projection of an object in pixel space is derived by multiplying the camera matrix (K_T for LWIR camera and K_D for depth camera) with an object point.

$$p_D = K_D \cdot P_D \tag{4}$$

$$\lambda \cdot p_T = K_T \cdot P_T \tag{5}$$

,where $p_D = [u_D, v_D, w_D]^T$, $p_T = [u_T, v_T, 1]^T$, a projected point in depth and LWIR camera pixel plane, respectively. P_D and P_T are an object point in depth and LWIR camera coordinate, respectively. λ is a scale factor and w_D is depth value in camera space. The spatial relation between two cameras is defined by equation 6 where R is a 3D rotation matrix and T is a translation matrix.

$$P_T = R \cdot P_D + T \tag{6}$$

By combining equation 4, 5, and 6, we can get an equation:

$$\lambda \cdot p_T = K_T \cdot (R \cdot K_D^{-1} \cdot p_D + T) \tag{7}$$

By solving equation 7, we transform I_{lwir}^{raw} to the depth plane and depth plane to the RGB plane. The detail of solving equation 7 is explained in the supplementary document. The RGB and depth frames are aligned using the Intel RealSense API. the different resolution and field of view between two cameras are adjusted using intrinsic and extrinsic camera parameters. After the alignment, we threshold the hand temperature by setting the lower and upper bounds in human temperature as depicted in Figure 2 b. These bounds are manually captured for every sequence and used as priors that segment hands from surrounding backgrounds and the hand-held objects. To create accurate bounds, we overlapped the I_{lwir}^{raw} on the depth maps as seen in Figure 2 column a. Finally, we get the segmented hands by filtering thermal mask (see Figure 2 column d).

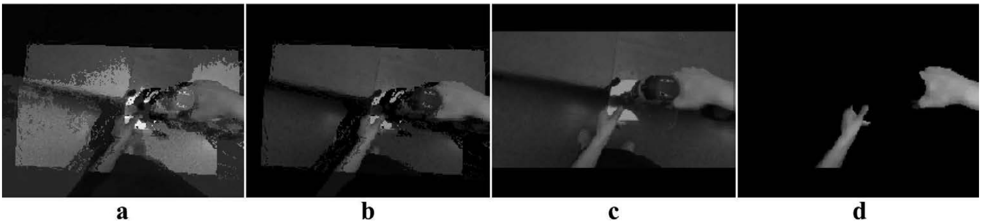

Figure 2: Aligning I_{lwir}^{raw} into the I_{rgb} with depth maps. First, (a) I_{lwir}^{raw} and I_{rgb} are overlapped onto the depth maps. (b) Next, the projected I_{lwir}^{raw} is bounded by hand temperature to capture possible hand regions. (c) Then, the projected I_{rgb} and I_{lwir}^{raw} are transformed on the RGB plane. (d) Finally, hands are cropped from backgrounds. For visualization, I_{lwir} is color mapped by a high value as red and a low value as blue.

Removing mislabeled pixels and identifying orientation

We occasionally observe mislabeled pixels from backgrounds that have similar temperature as hands. To remove these mislabeled pixels, we use a tracking algorithm, named

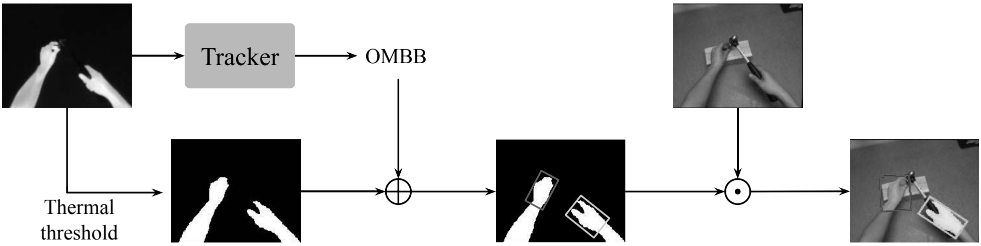

Figure 3: Overview of our proposed segmentation method.

SiamMask [53], to localize hands with OMBBs shown in Figure 3. We train the tracker with I_{tm} and AABBs of the hand. We use Amazon Mechanical Turk (AMT) to crowdsource the creation of hand AABBs. For training the tracker, I_{tm} and corresponding AABBs are used as targets, and I_{lwir} is used as inputs. Therefore, the tracker is color and texture invariant, which improves tracking performance of the tracker (sec. 7). After the training process, given initial AABBs, the tracker predicts OMBBs and classifies them sequentially OMBBs as a left hand or a right hand through the frames. This implies that the tracker learns hand shape features. These OMBBs are used to remove mislabeled pixels by intersecting I_{tm} and OMBBs.

4 Experiments

In this section, we show the efficiency of our proposed segmentation method, the performance of the tracker, and the analysis of multi-modal sources for hand segmentation, and we also benchmark our dataset with five different segmentation methods. We consider the hand segmentation problem as a two-class segmentation task, and we plot the maximum probability between the two classes, background and hands, per pixel for the prediction mask creation. For evaluation metrics, we use the Intersection over Union (IoU) of the hand (hIoU) and the background (bIoU). We define the mean IoU (mIoU) as the mean of these two class IoUs. For the test dataset, we use manually-annotated labels which consist of sequences that are used in neither the training set nor the validation set.

4.1 Dataset overview

To segment hands from objects, we create a pixel-wise hand segmentation dataset with subjects holding objects and tools. The dataset consists of 401,765 frames and 790 sequences. Our dataset has a large number of sequences and frames compared with the other datasets as shown in Table 1. We manually annotated 13,792 frames from 136 sequences to create a test dataset. The video dataset contains five subjects, 15 actions, and 23 tools. The distribution of the dataset across the actions and tools is plotted in the supplementary document. For annotating per pixel label of hands, we use I_{lwir} as prior knowledge and used a tracker to identify orientation of hands as well as whether a hand is a left hand or a right hand.

4.2 Efficiency of dataset creation

We evaluate the accuracy of I_{tm} with manually-labeled frames. I_{tm} is defined by the temperature of the hands. It shows fairly reasonable accuracy of 0.849 in hIoU. We find that

I_{tm} reduces false positive in the sequences where non-hand area has the same temperature as hands. These mislabeled pixels are the main reason of the hIoU degradation. We remove these mislabeled pixels with the tracker given initial AABBs [53]. Utilizing I_{lwir} improves the efficiency of the manually labeling frames. We profile the amount of time that it takes to annotate frames using four different methods for pixel-wise segmentation labeling. First, annotators use PolyRNN++ [1] with I_{rgb}. Second, annotators label hands on a tablet with a tablet pen given I_{rgb}. Third, annotators label hands on a tablet with a tablet pen given masked $I'_{rgb} = I_{rgb} \odot I_{tm}$, where $\odot$ is the Hadamard product. Lastly, annotators draw AABBs on top of I_{rgb}. Ten people annotated a random sample of twenty frames with the four different methods. We then averaged the annotation time, yielding the results in Table 2. We find that drawing AABBs is 24 times faster than the other methods. This implies that our method is 24 times faster than PolyRNN++ since our method intersects AABBs on I_{tm}. Additionally, the third method is two times faster than the second approach and six times faster than using PolyRNN++. Masking hands with I_{tm} significantly narrows down the region for annotators, which reduces the labeling time. To validate the quality of the annotation methods, we randomly sample 136 sequences and manually annotate 13,792 frames. With these manually annotated frames, we evaluate the IoUs of I_{tm} with and without AABBs labels (see Table 3). The total cost of our method is \$2,343 for annotating 401,765 frames, with \$8/hr as the minimum wage of our contractor. To annotate 401,765 frames with a tablet and pen, 76 seconds per frame, it requires \$67,853, which is 28.96 times more expensive than our method. The time cost of PollyRNN++ is 122 seconds per frame, resulting in \$108,923 for the pixel-wise annotation, which is 46.49 times more expensive than our method.

Annotation Methods	Avg. Time (second)
Drawing polygon with PolyRNN++ [1]	122
Painting hands on I_{rgb} with a tablet pen	76
Painting hands on I'_{rgb} with a tablet pen	24
Drawing AABBs on I_{rgb}	**5**

Table 2: Comparison table of annotation average time cost per frame. $I'_{rgb} = I_{rgb} \odot I_{tm}$. Lower average time is better.

4.3 Performance of tracker

We use SiamMask as our tracker [53] and train the tracker with a seeding dataset consisting of 518 sequences that have 11,718 frames labeled by crowd workers. The dataset is divided into 441 sequences that have 7,882 frames for training and 77 sequences that have 3,836 frames for validation. The tracker is trained in two ways: with I_{rgb} frames and with I_{lwir} frames. For evaluation, we use the manually labeled frames and metrics from Section 4.2. From the evaluation, we find that the tracker with I_{lwir} frames outperforms the others as shown in Table 3. The tracker with I_{rgb} tends to detect more forearm than the tracker with I_{lwir} as shown in Figure 4. This implies that the tracker with I_{lwir} is more sensitive in finding convex shape of wrist than the tracker with I_{rgb}, yielding better orientation of the hand and tighter OMBBs. The tracker performs well in most of cases; however, we need to re-initialize it with AABBs when the tracker fails to estimate the next frame. We also find that the tracker fails to track the hands when the two hands are heavily overlapping. In this case, we need to manually draw the OMBBs. Conventionally, creating pixel-wise segmentation is done by manually drawing polygons. The major drawback of this method is that polygon can not define smooth curves, not like our method. Particularly in hand-with-object cases, the boundary of fingers and objects creates many holes and curves. Our method can label pixel by pixel and represents smooth curves. Therefore, mIoU of PolyRNN++ is 0.895 [1] which is not as accurate as our method which is 0.923.

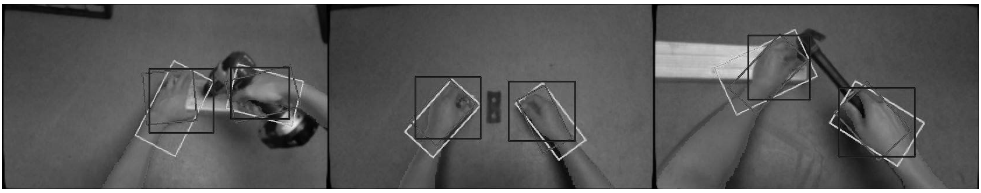

Figure 4: Visualization of three different bounding boxes. I_{tm} is overlapped onto the I_{rgb} as a red mask for visualization. The bounding boxes with red, blue, and green represent AABBs, OMBBs (I_{rgb}), and OMBBs (I_{lwir}), respectively.

Annotation Source	mIoU	hIoU	bIoU
I_{tm}	0.913	0.849	0.977
I_{tm} with AABBs	0.917	0.842	**0.992**
I_{tm} with OMBBs (I_{rgb})	0.921	0.851	0.990
I_{tm} with OMBBs (I_{lwir})	**0.923**	**0.855**	0.990

Table 3: Comparison of the quality of the annotated AABBs and the tracker-generated OMBBs.

4.4 Multi-modal sequence analysis

We analyze the effect of multi-modal sequences, $\{I_{rgb}, I_{lwir}, I_{depth}\}$, for hand segmentation by conducting seven ablation studies using all possible combinations of $\{I_{rgb}, I_{lwir}, I_{depth}\}$ as input modalities. We perform the seven ablation studies to find out how I_{lwir} contributes in training neural networks. For all experiments in this section, we use randomly sampled 50K frames and split into two sets as the following: 40K frames as train dataset and 10K frames as test dataset. The frames in the test dataset are labeled manually. We use DeepLabV3+ [13] as base model and add additional encoders to fuse additional modalities. Our fusing method is detailed on the supplementary document. The rationale of using DeepLabV3+ is that it outperforms other methods [50, 51, 61] in hand segmentation benchmark experiments, only using RGB, as shown in Table 4. It also has the fewest parameters. We use ResNet 101 [25] which is pre-trained on the ImageNet [16] as a backbone network. All experiments use an equal number of encoders as the number of input modalities. From experiments, we found that I_{lwir} guides the network in finding better minima by observing both the loss drops and performance improvement as shown in Figure 5 when the I_{lwir} is used. Including I_{lwir} enhances the performance of hand segmentation by 5% in hIOU score compared to $\{I_{rgb}, I_{depth}\}$. We also find that including IMU information improves 0.006 mIoU improvement. Increments are observed for bIOU and mIOU scores as well. Therefore, I_{lwir} is a robust feature for hand segmentation. The three modalities contain complementary properties, which generates robust features, compensates weak points of each other, and leverages their advantages. The models have been trained using Stochastic Gradient Descent (SGD) [8] and the ADAM optimizer [30] with initial parameters: learning rate as 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We decay the learning rate by 0.1 every 250K steps. Additional configurations of the experiments are listed in the supplementary document. We use a single TITAN RTX GPU and an Intel i7-6850K CPU for the experiments.

Source	mIoU	hIoU	bIoU	#Parameters
D	0.857	0.753	0.960	**59.3 M**
L	0.907	0.840	0.974	**59.3 M**
C	0.884	0.800	0.968	**59.3 M**
D,L	0.897	0.822	0.972	118.0 M
D,C	0.906	0.838	0.975	118.0 M
L,C	0.921	0.862	0.979	118.0 M
L,C,D	**0.931**	**0.880**	**0.982**	176.6 M

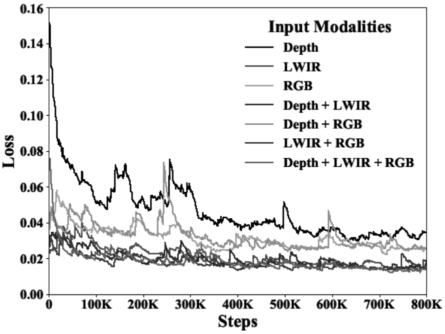

Figure 5: Comparison of segmentation IoUs of seven experiments using different input modalities. The higher values are better in IoU and the lower values are better in #Parameters. C, D, and L stands for RGB, depth, and LWIR frames, respectively. DeepLabV3+ [13] is used for the experiments.

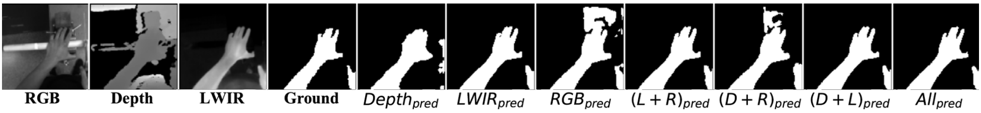

Figure 6: Qualitative results of the seven different ablation experiments. L, R, D, and All stand for I_{lwir}, I_{rgb}, I_{depth}, and $\{I_{rgb}, I_{lwir}, I_{depth}\}$, respectively.

4.5 Hand segmentation benchmark

To validate the performance of using multiple modalities, we compare our method with five state-of-the-art segmentation methods [13, 13, 50, 51, 61] which use I_{rgb} and segmentation networks [49] and jointly use I_{rgb} and I_{lwir} as input modalities. We use the same dataset as Section 4.4 and hyper-parameters listed on the original papers. We notice that RTFNet [49] performs second-best among all methods, indicating I_{lwir} provides the most meaningful prior knowledge for segmenting hands in frames. DeepLabV3+* with three modalities outperforms the second-best method, RTFnet, by 4% in hIoU and 30% fewer parameters, as shown in Table 4.

5 Conclusion and discussion

In this work, we propose a robust and efficient pixel-wise hand segmentation method and a multi-modal dataset. We record rich sequences with three different image modalities and IMU information of first-person-view images with pixel-wise hands and action labels. We found that using multiple modalities achieves 4% better hIoU when compared to the existing

	mIoU	hIoU	bIoU	Model Size
HIW [50]	0.865	0.770	0.865	118.0 M
PSPet [61]	0.897	0.823	0.972	70.4 M
DUC-HDC [51]	0.893	0.815	0.961	69.2 M
RTFNet [49]	0.911	0.846	0.976	254.5 M
DeepLabV3+ [13]	0.907	0.840	0.974	**59.3 M**
DeepLabV3+* [13]	**0.931**	**0.880**	**0.982**	176.6 M

Table 4: Comparison of quantitative results with other segmentation methods. DeepLabV3+* is trained with fused $\{I_{rgb}, I_{lwir}, I_{depth}\}$. The higher values are better in IoU and the lower values are better in size of model parameters.

state-of-the-art methods for hand segmentation. We also show that our multi-modal dataset with fusing LWIR and RGB-D frames achieves 5% better hand IoU performance than using just RGB-D frames. Also, we notice that only using I_{lwir} gives poorer results than using other modalities such as RGB and depth. This could be because thermal signature of hand is shared by other body parts. The proposed method is 24 times faster than PolyRNN++ with similar quality of manually-labeled frames. One limitation we find is that the tracker does not work properly when two hands are heavily overlapped. The future development will be focusing on improving the dataset for more diverse hand-related tasks such as hand-object pose estimation, object reconstruction when a person is holding the object, and hand action recognition.

6 Acknowledgment

We wish to give a special thanks to the reviewers for their invaluable feedback. This work is partially supported by NSF under the grants FW-HTF 1839971 and OIA 1937036. We also acknowledge the Feddersen Chair Funds. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. pages 859–868, 06 2018. doi: 10.1109/CVPR.2018.00096.

[2] Yağız Aksoy, Tunç Ozan Aydın, Aljoša Smolić, and Marc Pollefeys. Unmixing-based soft color segmentation for image manipulation. *ACM Trans. Graph.*, 36(2):19:1–19:19, 2017.

[3] Mykhaylo Andriluka, Jasper Uijlings, and Vittorio Ferrari. Fluid annotation: A human-machine collaboration interface for full image annotation. pages 1957–1966, 10 2018. doi: 10.1145/3240508.3241916.

[4] Jörg Appenrodt, Ayoub Al-Hamadi, and Bernd Michaelis. Data gathering for gesture recognition systems based on single color-, stereo color-and thermal cameras. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 3(1): 37–50, 2010.

[5] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957, 2015.

[6] Tewodros A Biresaw, Tahir Nawaz, James Ferryman, and Anthony I Dell. Vitbat: Video tracking and behavior annotation tool. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 295–301. IEEE, 2016.

[7] A. K. Bojja, F. Mueller, S. R. Malireddi, M. Oberweger, V. Lepetit, C. Theobalt, K. M. Yi, and A. Tagliasacchi. Handseg: An automatically labeled dataset for hand segmentation from depth images. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 151–158, May 2019. doi: 10.1109/CRV.2019.00028.

[8] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[9] AC Burton. The range and variability of the blood flow in the human fingers and the vasomotor regulation of body temperature. *American Journal of Physiology-Legacy Content*, 127(3):437–453, 1939.

[10] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3. Ann Arbor, Michigan;, 2016.

[11] Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multi-modal correlation learning for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2644–2652, 2017.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[14] Chiho Choi, Sangpil Kim, and Karthik Ramani. Learning hand articulations by hallucinating heat distribution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3104–3113, 2017.

[15] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3): 934–948, 2018.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[18] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1-2):52–73, 2007.

[19] Xavier Font-Aragones, Marcos Faundez-Zanuy, and Jiri Mekyska. Thermal hand image segmentation for biometric recognition. *IEEE Aerospace and Electronic Systems Magazine*, 28(6):4–14, 2013.

[20] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[21] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1991–2000, 2017.

[22] Oliver Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and Olga Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (TOG)*, 38(4):41, 2019.

[23] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[24] Wilfried Hartmann, Sebastian Tilch, Henri Eisenbeiss, and Konrad Schindler. Determination of the uav position by automatic processing of thermal images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 39:B6, 2012.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun. Real time hand pose estimation using depth sensors. In *Consumer depth cameras for computer vision*, pages 119–137. Springer, 2013.

[28] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. Domain adaptation for privacy-preserving pedestrian detection in thermal imagery. In *International Conference on Image Analysis and Processing*, pages 203–213. Springer, 2019.

[29] Sangpil Kim, Nick Winovich, Hyung-Gun Chi, Guang Lin, and Karthik Ramani. Latent transformations neural network for object view synthesis. *The Visual Computer*, pages 1–15, 2019.

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Hoang Le, Long Mai, Brian Price, Scott Cohen, Hailin Jin, and Feng Liu. Interactive boundary prediction for object selection. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[32] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.

[33] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[34] Hubert Lin, Paul Upchurch, and Kavita Bala. Block annotation: Better image annotation for semantic segmentation with sub-image decomposition, 02 2020.

[35] Dong Liu, Yue Li, Jianping Lin, and Houqiang Li. Deep learning-based video coding: A review and a case study, 04 2019.

[36] Rachel Luo, Ozan Sener, and Silvio Savarese. Scene semantic reconstruction from egocentric rgb-d-thermal videos. In *2017 International Conference on 3D Vision (3DV)*, pages 593–602. IEEE, 2017.

[37] Konstantinos Makantasis, Antonios Nikitakis, Anastasios D Doulamis, Nikolaos D Doulamis, and Ioannis Papaefstathiou. Data-driven background subtraction algorithm for in-camera acceleration in thermal imagery. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2090–2104, 2017.

[38] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9567–9576, 2019.

[39] Jifeng Ning, Lei Zhang, David Zhang, and Chengke Wu. Interactive image segmentation by maximal similarity based region merging. *Pattern Recognition*, 43(2):445–456, 2010.

[40] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[41] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011.

[42] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[43] X. Qin, S. He, Z. Zhang, M. Dehghan, and M. Jagersand. Bylabel: A boundary based semi-automatic image annotation tool. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1804–1813, 2018.

[44] Sarah Rastegar, Mahdieh Soleymani, Hamid R. Rabiee, and Seyed Mohsen Shojaee. Mdl-cw: A multimodal deep learning framework with cross weights. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[45] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[46] Bryan Russell, Antonio Torralba, Kevin Murphy, and William Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77, 05 2008. doi: 10.1007/s11263-007-0090-8.

[47] Thomas Schlömer, Benjamin Poppinga, Niels Henze, and Susanne Boll. Gesture recognition with a wii controller. In *Proceedings of the 2nd international conference on Tangible and embedded interaction*, pages 11–14. ACM, 2008.

[48] Roy Shilkrot, Supreeth Narasimhaswamy, Saif Vazir, and Minh Hoai. Workinghands: A hand-tool assembly dataset for image segmentation and activity mining. In *BMVC*, 2019.

[49] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3): 2576–2583, 2019.

[50] Aisha Khan Urooj and Ali Borji. Analysis of hand segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2018.

[51] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018.

[52] Peng Wang and Xiangzhi Bai. Thermal infrared pedestrian segmentation based on conditional gan. *IEEE transactions on image processing*, 28(12):6007–6021, 2019.

[53] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1328–1338, 2019.

[54] Robert Y Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, 28(3):63, 2009.

[55] Sean Ward, Jordon Hensler, Bilal Alsalam, and Luis Felipe Gonzalez. Autonomous uavs wildlife detection using thermal imaging, predictive navigation and computer vision. In *2016 IEEE Aerospace Conference*, pages 1–8. IEEE, 2016.

[56] Xiao-Yong Wei, Yu-Gang Jiang, and Chong-Wah Ngo. Concept-driven multi-modality fusion for video search. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(1):62–73, 2011.

[57] Aaron Wetzler, Ron Slossberg, and Ron Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 33.1–33.12. BMVA Press, September 2015. ISBN 1-901725-53-7. doi: 10.5244/C.29.33. URL `https://dx.doi.org/10.5244/C.29.33`.

[58] Felix Zhan. Hand gesture recognition with convolution neural networks. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 295–298, 2019.

[59] Lishi Zhang, Chenghan Fu, and Jia Li. Collaborative annotation of semantic objects in images with multi-granularity supervisions. pages 474–482, 10 2018. doi: 10.1145/3240508.3240540.

[60] Qiang Zhang, Yi Liu, Rick S Blum, Jungong Han, and Dacheng Tao. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information Fusion*, 40:57–75, 2018.

[61] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[62] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 813–822, 2019.