Multispectral Video Dataset for First-person View Hand Segmentation during Activity with Objects and Tools

Anonymous CVPR submission

Paper ID 1857

Abstract

Hands have been the subject of computer vision studies because of its fundamental importance as an enabling tool for human interactions in virtual reality and augmented reality. Hand segmentation is important for understanding human intent and action while hands are interacting with objects and the environment. However, parsing the hands from hand-held objects is a difficult problem because of both object and self-occlusion, the high number of degrees of freedom of the hand, different grasping strategies used by humans for holding objects, and infinitely complex shapes of objects that hands hold. To relieve these challenges, we propose a robust and efficient pixel-wise hand annotation method which requires minimal human labor to create a large dataset. This dataset involves hand interacting with objects. We use human temperature as an additional feature to color and depth features. We recorded 979 sequences and 491,192 frames of "hands using tools" videos with Long-Wave InfraRed (LWIR) and RGB-D cameras and then label hands by identifying human body temperature and geometric constrains of hands. The quality of our annotation method was evaluated by comparing it with manually labeled images. We also analyze the segmentation performance of DeepLabV3+ using a different combination of LWIR and RGB-D images as an input. From the analysis results, we developed MultiSpectral Network (MSNet) which is based on DeepLabV3+ and it achieves 4% better hand IoU performance with 30% fewer parameters than the second best network among 5 state of the art methods on our video dataset.

1. Introduction

Visualization technologies such as Augmented Reality (AR) and Virtual Reality (VR) will spawn the next generation of workspaces, where the digital world can be made to intimately merge with the physical. In this process, it will enable transformative and more intuitive workflows. However for AR and VR to reach that maturity and become truly ubiquitous, one of the sets of problems that need to be robustly solved are problems related to the hands, especially its interactions with objects and the environment. This is because hands are the primary interface that helps human beings interact with virtual objects and innumerable tools and devices. Thus there is a need to understand the hands and its interactions with hand-held tools, objects and the environment.

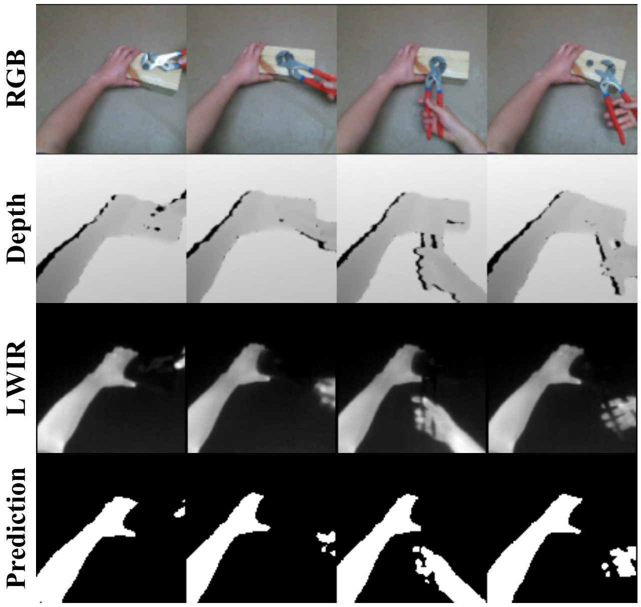

Figure 1: Hand segmentation with multispectral data. LWIR stands for Long-Wave InfraRed.

There are many works in the computer vision area, which are related to hands such as hand tracking [40, 31], hand pose estimation [13, 21, 17, 16, 23], grasp detection [7, 34], hand gesture recognition [42], and hand-action classification [37]. These works require localizing the hands to isolate them from the human body and environment. Thus hand localization is directly related to their final performance.

Localizing hands in the first-person view while hands interact with objects is a challenging problem because (a) objects and hands block each other while a person is interacting with objects and tools, (b) humans hold objects with various grasps and added to this (c) the hands and objects rotate and translate spontaneously.

There are few datasets directed towards hand segmentation [4, 3, 26, 37]. They use human labor to annotate each hand. Therefore, developing an efficient method with almost zero human intervention that labels pixels of hands are important. Human body temperature is relatively constant [6] and has been widely used in pedestrian detection [22], biological image processing [15], and gesture recognition [2]. However, to the best of our knowledge, there are no methods that directly use the thermal information for creating a large-scale pixel-wise hand segmentation dataset. We used both geometric constraints of human hands and human body temperature to isolate hands from the background and objects it is holding. To validate our automated annotation method, we manually annotated hands from some of the frames and validate that our annotation method achieves near human-level performance.

Prior strategies to use the features from deep neural networks for segmentation tasks by decoding the hand image features often lead to failures, such as when the objective function gets stuck in local minima. The neural networks can automatically learn important cues from three different modalities: thermal, color, and depth. The jointly learn features of these three modalities prevents confusion between hands, the surrounding environment, and objects. We also compared our method with the other existing methods and our method outperformed existing state-of-the-art methods [36, 37, 43, 9, 39] in the pixel-wise hand segmentation tasks by a large margin.

Our contributions are the following:

- We developed a framework that can significantly reduce human annotation efforts by leveraging human body temperature for creating pixel-wise hand segmentation ground truth when a person is holding objects and performing tasks.
- We collected large-scale action videos in a first-person view with LWIR and RGB-D cameras and evaluated it against manually annotated images.
- We analyzed the effectiveness of multispectral images for hand segmentation task with deep neural networks and found the optimal combination of fusing long-wave infrared, color, and depth modalities to segment hand pixels.

The rest of the paper is structured as follows. In the next section, we explain related works. Dataset creation without human labor is elaborated in section 3. and in section 4 we analyze how LWIR works with RGB-D images and we propose MultiSpectral Network (MSNet) for pixel-wise hand segmentation. In section 5, we discuss the experiment results and limitations. Finally, we present our conclusions in section 6.

2. Related works

Hand localization In the first-person view, hand localization is an important pre-requisite for many problems such as handled object recognition [14, 33], hand pose estimation [21], and hand gesture recognition [35, 20] as hands are present in the majority of first-person view video frames. Li *et al*.'s work [25] is among the first works in the area of hand segmentation. They propose datasets as well as local features and a global appearance-based mixture model for pixel-wise segmentation of hand. Bambach *et al*. [3] collected first-person view hand interaction videos with Google glasses and manually annotated 4,800 frames with left-right hand instances. The dataset contains interaction with other people doing four activities. Compared to the above manually annotated data-sets we significantly enhanced the annotation process with a low-cost LWIR sensor. Our dataset is more focused on manufacturing based activities with machine tools such as grinding, sawing, and assembling, etc. in the first-person view. Manufacturing area is a useful application for training and skilling using AR. Abhishake *et al*. [4] proposed a method that segments hands from the depth map that are paired with RGB images where a single person wears colored gloves. They also automatically obtain the hand segments but require colored gloves for their process. Unlike ours, their RGB modality doesn't reflect the skin color of the hands. Our proposed method labels hands in diverse images: color, depth, and Long-Wave InfraRed (LWIR). Providing diverse modalities enhances the opportunity to try other methods for the researchers.

Multispectral data is widely used to capture robust features which are beyond the visible spectrum [28, 12, 10, 38]. This is because a depth sensor can only capture geometric information invariant under lightening conditions. Using LWIR helps to identify objects which have distinct temperature from surrounding backgrounds. For this reason, LWIR is widely used in pedestrian detections [38] and controlling unmanned aerial vehicles [41, 19]. MFNet [18] showed that incorporating the thermal information with RGB significantly increased the segmentation performance. Sun *et al*. [36] proposed semantic segmentation networks that fuse RGB and thermal information for autonomous vehicles. However, all these works used thermal information for scene segmentation for autonomous driving. However, in our work we use LWIR as guidance to find the Region Of Interest (ROI) and narrow down the search space for labeling hands at pixel-level to create a segmentation dataset.

Pixel-wise segmentation with deep neural networks have been widely used for segmenting objects [32, 43, 8, 30, 9]. Deep learning-based methods use convolutional neural networks to extract feature representations to predict each pixel. One of the popular structure 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. However, the pre-requisite of deep-learning methods is a large-scale dataset, and thus an appropriate method that can create such a high quality annotated dataset at speed, with lesser manual effort is crucial. Khan *et al*. [37] contributed to hand segmentation analysis by using RefineNet [27] and showed huge improvement. They manually created a hand segmentation dataset consisting of RGB images. Unlike creating the dataset with human labor, we labeled hand videos efficiently by leveraging human body temperatures. Since video data contains a large number of frames, the need for an efficient way of creating ground truth becomes even more pronounced.

3. Automating hand labeling

Our dataset has a fairly large number of sequences and frames compared with the other datasets as shown in Table 1. For labeling a large-scale pixel-wise hand segmentation dataset, we recorded the daily action videos with a low-cost non-radiometric thermal camera, Flir Boson 320, which captures relative thermal distribution but not the exact temperature. We used an Intel D435i depth camera for RGB-D and Inertial Measurement Unit (IMU) information acquisition. These sensors were placed within a 3D printed case and mounted in front of the helmet to make the camera location consistent over sequences as shown in Figure 2.

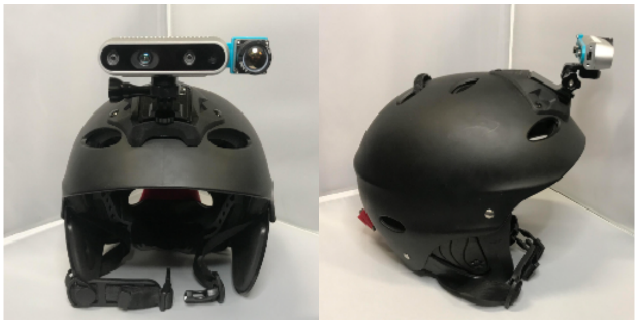

Figure 2: Head-mounted multispectral sensors on the helmet. Multispectral cameras were mounted on the helmet to stabilize the location of them for each person.

3.1. Isolating hands with multispectral imaging

To isolate hands with their LWIR images, we narrow down a search space by finding a Region Of Interest (ROI) with LWIR images. We map the LWIR frames onto the RGB frames with depth maps. To align depth maps and LWIR images, we need to find the spatial relationship between the LWIR and depth cameras. A projection of an object in pixel space is derived by multiplying the camera matrix ($\boldsymbol{K_T}$ for LWIR and $\boldsymbol{K_D}$ for depth camera) to object point.

Dataset	Sensor	Seq.	Frames	Labels
EgoHands [3]	RGB	48	4,800	Seg.
EgoHands+ [37]	RGB	8	800	Seg.+Action
GTEA [26]	RGB	28	663	Seg.+Action
EYTH [37]	RGB	3	1290	Seg.
HOF [37]	RGB	-	300	Seg.
SegHand [4]	Depth	-	210,000	Seg.
Ours	RGB-D-T-IMU	979	491,192	Seg.+Action

Table 1: Comparison table of hand segmentation datasets. T,D, and IMU represent LWIR, depth, and inertial measurement unit, respectively.

$$\boldsymbol{p_D} = \boldsymbol{K_D} \cdot \boldsymbol{P_D} \tag{1}$$

$$\lambda \cdot \boldsymbol{p_T} = \boldsymbol{K_T} \cdot \boldsymbol{P_T} \tag{2}$$

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

$$\boldsymbol{P_T} = \boldsymbol{R} \cdot \boldsymbol{P_D} + \boldsymbol{T} \tag{3}$$

By combining equation (1), (2), and (3), we can get an equation below,

$$\lambda \cdot \boldsymbol{p_T} = \boldsymbol{K_T} \cdot (\boldsymbol{R} \cdot \boldsymbol{K_D^{-1}} \cdot \boldsymbol{p_D} + \boldsymbol{T}) \tag{4}$$

By solving equation (4) in terms of $\boldsymbol{R}$ and $\boldsymbol{T}$ with datum points in both cameras, we can transform LWIR images to the depth plane and depth plane to the LWIR plane. The detail of solving equation (4) is explained in the appendix. The RGB and depth image is already aligned using the Intel RealSense API. As a result, we can transform the LWIR and RGB image, which has the same coordinate as the depth plane, to each other as shown in Figure 3. Then, we threshold hands temperature manually by finding the lower and upper bound as depicted in Figure 3 b. These bounds were manually captured for every sequence and used as priors that isolate hands from surrounding backgrounds and the handheld objects. To create accurate bounds, we overlapped the LWIR image on the depth maps as seen in Figure 3 column a. This process can be done automatically by recording

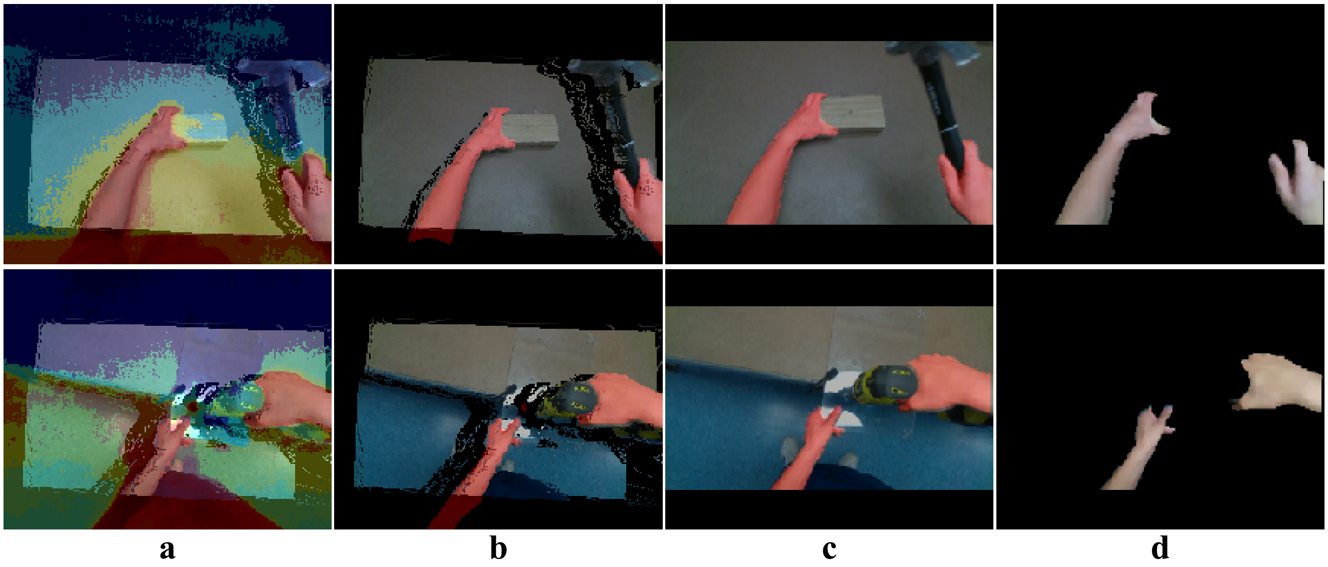

Figure 3: Aligning LWIR values into the RGB images with depth maps. First, (a) LWIR and RGB images are overlapped onto the depth maps. (b) LWIR images are bounded by body temperature to capture possible hand regions. (c) Then images are transformed on RGB plane. Finally, (d) we crop the hands from the image. For visualization, LWIR is color mapped by a high value as red and a low value as blue.

the videos with an expensive radiometric thermal camera as they capture absolute temperature instead of the relative temperature and multiple RGB-D cameras. Finally, we can get the segmented hands by filtering the thermal mask (see Figure 3 column d).

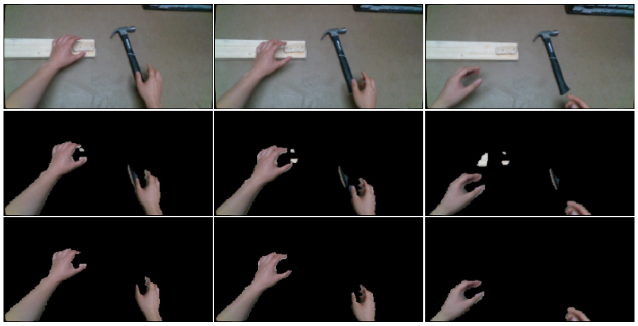

Figure 4: Removing artifacts with geometric constraints of human hands. 3rd row from the top are after applying final refining stages of our method.

3.2. Geometric constraints of human body

LWIR mapping is not sufficient enough to parse hands from the surrounding backgrounds due to the heat diffusion and objects which have similar temperature with the human body. If humans hold an object for a certain amount of time then the temperature of the local object becomes similar to human hands which are depicted in Figure 4. This happens more often with objects which are made of highly conductive materials. Therefore, we used geometric constraints of the wrist structure to remove these heated objects which were being captured by temperature thresholding. To do so, we binarized the LWIR image by thresholding human body temperature in a range, and created blobs with connected components algorithm. Then we applied erosion to remove small blobs and we examine the concavity of the blobs to determine if it is wrist or not. The distance transformations and elliptical shape optimization are used to detect the wrists [29] by finding a sudden increase in the width of the forearm. After determining wrists, we additionally validated remaining blobs if the blobs are touching the frames of the image. This is because we observed that the beginning and ends of the frames only have partial hands without wrists. Therefore, if the blobs are not touching the frames or not containing the wrist then we determine that they are not hands.

4. Experiments

We explored ways to use multispectral data, LWIR and RGB-D, for hand segmentation. First, we performed seven ablation studies to find out how LWIR helped in the training of the neural networks. Through those experiments, we found that the best fusion method is using all three modalities: LWIR, RGB, and depth. For evaluation metrics, we used Intersection over Union (IoU) of hand (hIoU)

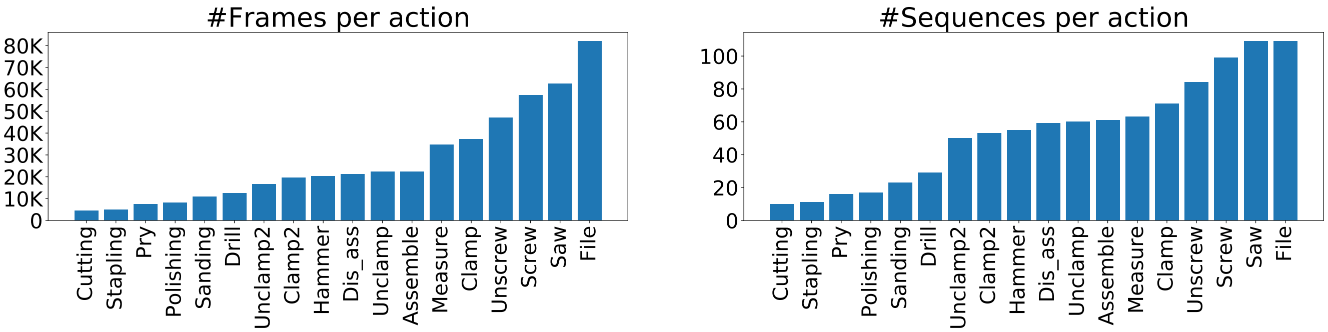

Figure 5: The dataset distribution regarding actions per frames and sequences. “Clamp on table”, “Unclamp on table” and “Disassemble” are mentioned as clamp2, unclamp2 and disass for brevity purposes, respectively.

and background (bIoU). We use mean IoU (mIoU) as mean of each class IoU. We considered the hand segmentation problem as a two-class segmentation task and we plotted the maximum probability between two classes, background, and hands, per pixel for the prediction mask creation. For all experiments in this section, we divided the dataset which is labeled with our method into two sets: 80% of sequences as training and 20% of sequences as validation. For the test set, we used manually annotated dataset which consists of sequences that are not used in both of the training set and the validation set. The models have been trained using Stochastic Gradient Descent (SGD)[5] and the ADAM optimizer [24] with initial parameters: learning rate as 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We decayed the learning rate by 0.1 in every 250K steps. For hardware devices, we used a single TITAN RTX GPU and an Intel i7-6850K CPU for the experiments.

4.1. Dataset creation

To segment hands from objects, we created a pixel-wise hand segmentation dataset with a person holding objects and tools. This dataset consists of multiple sequences of RGB-D, LWIR, and IMU of camera information in a first-person view. The dataset consists of 491,192 frames with 979 sequences which have five individual persons performing 18 actions with 36 objects and 30 tools. The distribution of the dataset along with action is shown as histogram at Figure 5. For annotating human hands, we used LWIR frames as a prior and used geometric constraints of the human body to parse hands out from surrounding backgrounds. For recording sequences, each person was asked to complete 18 tasks with handheld tools and objects. To validate our annotation method, we randomly sampled 20% of sequences and manually annotated every fifth frame of each video, where annotators annotated human hands and forearms. Annotators used tablet and tablet-pen to enhance the accuracy and efficacy of the annotation process.

	mIoU	hIoU	bIoU
Hand ROI	0.845	0.718	0.971
Hand ROI + GC	**0.950**	**0.913**	**0.987**

Table 2: Evaluation of hand ROI and removing mis-labeled hands with Geometric Constraints (GC). Higher value is better.

	Avg. Time (s)
PolyRNN++ [1]	122
Tablet pen w/o hand ROI	76
Tablet pen w/ hand ROI (Ours)	**24**

Table 3: The average time cost for annotating hands per image. Using our hand ROI annotators improve drastically over other methods. Lower time is better.

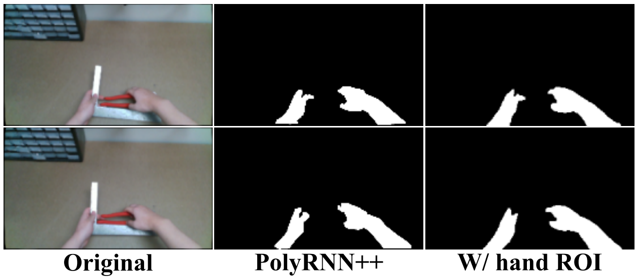

Figure 6: Qualitative comparison of manually annotated images with PolyRNN++ [1] against annotators using tablets and pen with hand ROI masked images generated using our method.

4.2. Labeling hands with LWIR

In this section, we attempt to quantify the quality of our proposed annotation method. For the labeling process, we first detect ROI of hands with LWIR and human body tem-

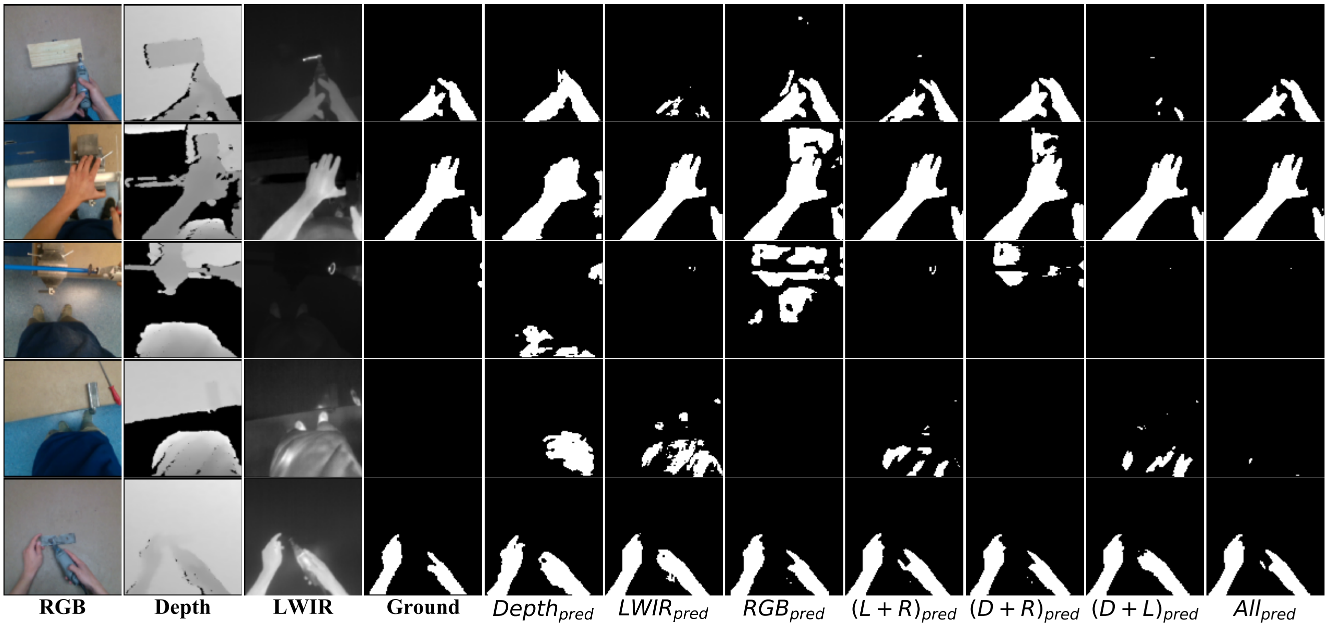

Figure 7: Qualitative results of the ablation study of seven different combinations of LWIR and RGB-D images for hand segmentation. (L + R) denotes Infrared and RGB. Similarly (D + R) denotes depth and RGB and (D + L) denotes depth and Infrared. In Fourth row we see that thermal camera catches heat signatures. However those are not from the hand. Using Infrared for prediction (6th col) gives false positives (F.P.). Only using depth channel also gives F.P. (5th col). However for this row (4th) we see that using RGB (7th, 9th and 11th col.) gives good results. In 3rd row we see that using RGB (7th col.) gives F.P. (the table has skin color). Also using depth only (5th col.) gives F.P. However when we use infrared (6th,8th,10th and 11th col.) we get good results. This exemplifies the importance of using infrared channel.

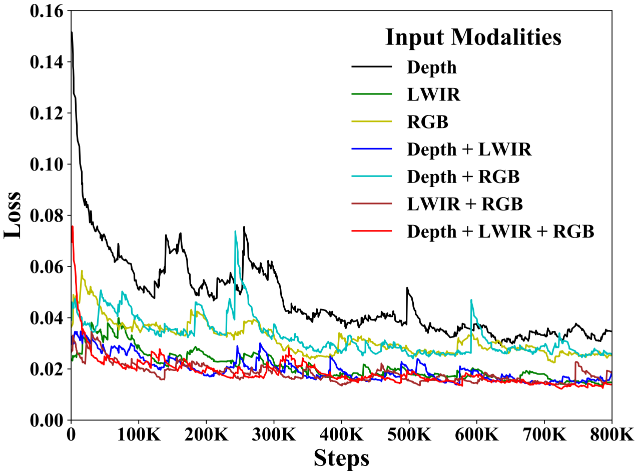

Figure 8: Loss plots of seven different ablation study. Each experiment uses modalities. The DeepLabV3+ [9] was used in this ablation study.

Input modalities	mIoU	hIoU	bIoU	Model Size
Depth	0.857	0.753	0.960	**59.3 M**
LWIR	0.884	0.800	0.968	**59.3 M**
RGB	0.907	0.840	0.974	**59.3 M**
LWIR + RGB	0.921	0.862	0.979	118.0 M
Depth + RGB	0.906	0.838	0.975	118.0 M
Depth + LWIR	0.897	0.822	0.972	118.0 M
All	**0.931**	**0.880**	**0.982**	176.6 M

Table 4: Segmentation results of seven combinations of multispectral data as input modalities. The higher values are better in IoU and the lower values are better in size. DeepLabV3+ [9] was used for the ablation experiments.

perature and then we removed artifacts with geometric constraints of human hand structure. We evaluate the accuracy of the ROI with the human-annotated dataset and it shows fairly reasonable accuracy which is 0.718 in hIoU. We found that ROI imposes false-positive area in the sequences which were created by heated parts of tools such as a head of the grinder or files. These artifacts are the main reason for the degradation of the hIoU and bIoU scores. We reduced these false-positive artifacts by removing blobs when blobs are not satisfying either of two conditions: the blobs silhouette resembles human hands or they are touching the edge of the image frames. After removing artifacts with these geometric constraints, hIoU improved by 27% and bIoU improved by 1.6% as shown in Table 2. As seen in the results, hIoU improved significantly which shows that a lot of false-positive ROI parts are removed. We noticed that using geometric constraint removes a lot of false positive

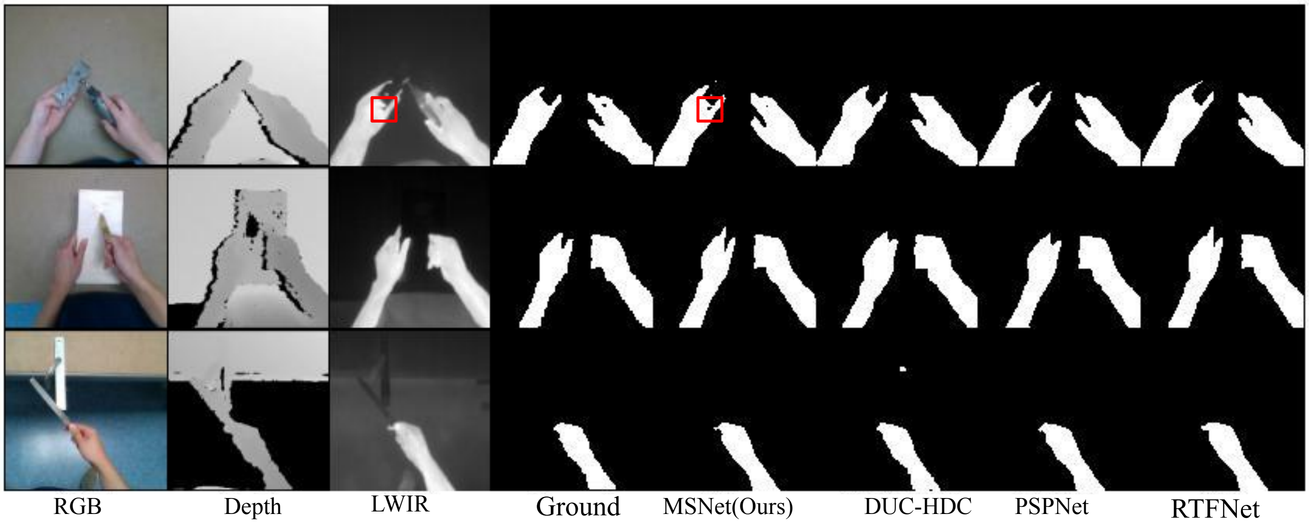

Figure 9: Qualitatively comparing MSNet with DUC-HDC [39], PSPNet [43], and RTFNet [36]. In the first row there is a small kink in the hand segmentation boundary near tip of the thumb (highlighted by the red rectangle). Prediction from MSNet captures that intricate detail (highlighted by the red rectangle) of hands which human annotator has left out.

in the ROI. This improvement shows that geometric constraints which we used for detecting wrists are useful for removing the heated surrounding objects from the ROI.

Utilizing LWIR also improves the efficacy of the human labeling process. We compared the efficacy of annotators who used PolyRNN++ [1] against annotators using a tablet pen with an image seeded with hand ROI masks and also against annotators using tablet pen with an image without seeded hand ROI masks. We asked ten people to annotate twenty images which are randomly selected from the dataset with three different methods. The annotator group using tablet pen with the image that is masked with the hand ROI finished the annotation tasks approximately twice faster than those without ROI masking and six times faster than ones using PolyRNN++, as shown in Table 3 with better quality as shown in Figure 6. Masking hands with hand ROI significantly narrows down the region to annotate which helps in reducing the labeling time.

4.2.1 Multispectral sequence analysis

We analyzed the effect of multispectral sequences, LWIR and RGB-D, for hand segmentation with our dataset by conducting seven ablation studies which are all possible 7 combinations of LWIR, depth, and RGB as input modalities. We used DeepLabV3+ [9] as our base model and added additional encoder to fuse additional modalities. The rationale of using DeepLabV3+ is that it outperformed other methods [37, 43, 39] in the hand segmentation benchmark experiments, only using RGB, as shown in Table 5 and it is the lightest in terms of parameters. We used ResNet 101 which was pre-trained on the ImageNet [11] as our backbone network which extracted robust features in the encoder for this ablation study. We first experimented with the effectiveness of each modality by inputting each modality into the encoder. Then, we experimented what the combination of multiple modalities. For multiple modality experiments, we used the same number of the encoder as the input modalities and concatenated the low-level and high-level features, which are then used as input of the decoder. From these experiments, we found that LWIR guides the network in finding better minima by observing that both the training loss drops and hIoU is enhanced of unseen hands when the LWIR images are used (see Figure 8 and Table 4).

From the experimental results, we developed MultiSpectral Network (MSNet) as shown in Figure 10. MSNet jointly use all modalities as inputs, which compensates weak points of each modality and leverages their advantages. The detail architecture definitions are listed in the appendix.

4.3. Hand segmentation benchmark

To validate the performance of the MSNet which was trained with multispectral information learning we compared it with four state-of-the-art segmentation methods [9, 43, 39, 37] which use RGB images and segmentation networks [36] which jointly use RGB and LWIR as input modalities. We noticed that RTFNet [36], which used LWIR and RGB, performed second-best among all methods. We found in our ablation study, as shown in 4, that including LWIR enhances the performance of hand segmentation by 4.2% for hIOU scores compared to when using only RGB-D. Increments are observed for bIOU and mIOU scores too.

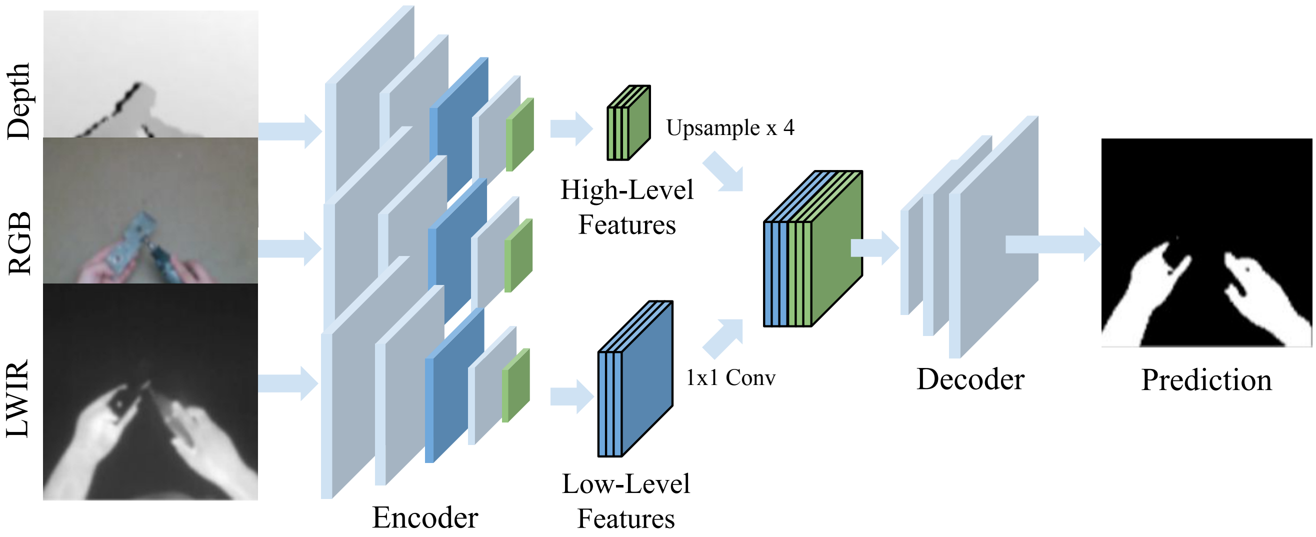

Figure 10: Overview of MSNet for hand segmentation with LWIR and RGB-D data. The network encodes three modalities with three independent backbone networks. The network is based on the DeepLabV3+ [9].

MSNet also outperformed the second-best method, RTFnet, by 4% in hIoU with 30% lesser parameter, as shown in Table 5.

	mIoU	hIoU	bIoU	Model Size
HIW [37]	0.865	0.770	0.865	118.0 M
PSPet [43]	0.897	0.823	0.972	70.4 M
RTFNet [36]	0.911	0.846	0.976	254.5 M
DeepLabV3+ [9]	0.909	0.842	0.975	**59.3 M**
DUC-HDC [39]	0.893	0.815	0.961	69.2 M
MSNet (Ours)	**0.931**	**0.880**	**0.982**	176.6 M

Table 5: Quantitative results of other segmentation methods. The higher values are better in IoU and the lower values are better in size of model parameters.

5. Discussion and limitations

We automated the annotation process of the hands at the pixel-level with a low-cost LWIR and RGB-D sensors. The proposed annotation method requires humans to identify the upper and lower bounds of LWIR values to create an ROI of hands, since we are using the low-cost non-radiometric LWIR sensor which captures relative temperatures. Also, this can be automated with a high-end radiometric LWIR sensor that captures absolute temperatures. Therefore, LWIR sensors that captures the temperature information of the human body is useful for detecting hands as it narrows down the search space. This ROI area helps in not only training segmentation methods but also enhances the efficacy of the human annotators. In Table3 we show the drastic improvements made on annotation time when humans use our ROI as seed for annotation mask. As for the limitations of using LWIR, when objects which have similar temperatures to human body are in the scene, then the networks sometimes detects them as hands. Also, we noticed that only using LWIR image gives poorer results than using other modalities such as RGB-D. This could be because thermal signature of hand is shared by other body parts. Heat diffusion to surroundings also causes nearby objects to get similar thermal signatures as that of the hand.

6. Conclusion

In this work, we propose a robust and efficient annotation method for pixel-wise labeling human hands. Our fast and near automatic pixel-wise hand annotation method can ease dataset creation for more research on vision problems related to hands. We recorded rich sequences that have LWIR, RGB-D, and IMU information in first-person view with pixel-wise hands and action labels. We use our dataset, for analyzing the uses of multispectral, LWIR and RGB-D images for training neural networks. From the experiment results, we propose MSNet which is based on DeepLabV3+[9], which leverages advantage of each of LWIR and RGB-D modalities. The proposed MSNet achieves better hIoU when compared to the existing state-of-the-art-methods for hand segmentation, on our dataset by a margin of 4% compared to the second best method. It also has 30% less parameters than the second-best method, RTFNet [36]. As a future work, the dataset could further used 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.

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *CVPR*, 2018. 5, 7

[2] 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. 2

[3] 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. 2, 3

[4] 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. 2, 3

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 5

[6] 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. 2

[7] 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. 1

[8] 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. 3

[9] 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. 2, 3, 6, 7, 8

[10] 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. 2

[11] 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. 7

[12] 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. 2

[13] 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. 1

[14] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 2

[15] 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. 2

[16] 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. 1

[17] 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. 1

[18] 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. 2

[19] 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. 2

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 2

[21] 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. 1, 2

[22] 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. 2

[23] 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. 1

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Cheng Li and Kris M Kitani. Pixel-level hand detection in ego-centric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2013. 2

[26] 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. 2, 3

[27] 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. 3

[28] 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. 2

[29] Gourav Modanwal and Kishor Sarawadekar. A robust wrist point detection algorithm using geometric features. *Pattern Recognition Letters*, 110:72–78, 2018. 4

[30] 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. 3

[31] 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. 1

[32] 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. 3

[33] Xiaofeng Ren and Chunhui Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3137–3144. IEEE, 2010. 2

[34] 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. 1

[35] Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2620–2628, 2016. 2

[36] 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. 2, 7, 8

[37] 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. 1, 2, 3, 7, 8

[38] Peng Wang and Xiangzhi Bai. Thermal infrared pedestrian segmentation based on conditional gan. *IEEE transactions on image processing*, 28(12):6007–6021, 2019. 2

[39] 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. 2, 7, 8

[40] Robert Y Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, 28(3):63, 2009. 1

[41] 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. 2

[42] 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. 1

[43] 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. 2, 3, 7, 8