Supplementary document

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

1 Derivation of transformation function

In this section, we derive the equation 7 in the main paper, which map the LongWave InfraRed (LWIR) frames onto the RGB frames with depth maps. RGB frames are mapped with depth maps by RealSense API.

1.1 Notations

Principle points : $[u_0, v_0]$
Focal points of camera : $[f_x, f_y]$

Camera intrinsic matrix : $K = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}$

Depth camera intrinsic matrix : K_D
LWIR camera intrinsic matrix : K_T

3D rotation matrix : $R = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$
Translation matrix : $T = [t_1, t_2, t_3]^T$
Scale factor : λ
Projected point in depth image : $p_D = [u_d, v_d, w_d]^T$
Projected pixel in LWIR image : $p_T = [u_t, v_t, 1]^T$
Point in depth camera coordinate : $P_D = [x_d, y_d, z_d]^T$
Point in LWIR camera coordinate : $P_T = [x_t, y_t, z_t]^T$
, where $\{f, t, h, u_0, v_0, x, y, z\} \in \mathbb{R}$ and $\{u, v, w\} \in \mathbb{N}$.

1.2 Derivation of transformation matrix

To align two different images which are from different location of the cameras, we have to know transformation matrix (R, T) between them. To calculate transformation matrix, it requires to use at least six corresponding datum points of each camera. We took a picture from both cameras with marked datum points in different locations and use these points to calculate the rotation matrix R and translation matrix T. We get the camera intrinsic matrix (K_T, K_D) by calibrating both cameras.

The points in metric space are projected to the camera plane by multiplying intrinsic matrix of camera (Equation 1, 3). Therefore, by multiplying the inverse matrix (K_D^{-1}) of intrinsic camera matrix to a projected point (p_D) in depth image, we get a point (P_D) in depth camera coordinate (Equation 1). We then transform P_D to thermal camera domain using R and T as equation 2.

$$P_D = K_D^{-1} \cdot p_D \tag{1}$$

$$P_T = R \cdot P_D + T \tag{2}$$

$$\lambda p_T = K_T \cdot P_T \tag{3}$$

We derive equation 4 by combining equation 1, 2, and 3.

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

By comparing a transformed depth point to the thermal plane and a thermal datum point, we can derive the transformation matrix. By calculating matrix equation 4, we get three equations 5, 6 and 7.

$$\begin{aligned} &f_x(h_{11}x_d + h_{12}y_d + h_{13}z_d + t_1) \\ &\quad + u_0(h_{31}x_d + h_{32}y_d + h_{33}z_d + t_3) - \lambda u_T = 0 \end{aligned} \tag{5}$$

$$\begin{aligned} &f_y(h_{21}x_d + h_{22}y_d + h_{23}z_d + t_2) \\ &\quad + v_0(h_{31}x_d + h_{32}y_d + h_{33}z_d + t_3) - \lambda v_T = 0 \end{aligned} \tag{6}$$

$$h_{31}x_d + h_{32}y_d + h_{33}z_d + t_3 = \lambda \tag{7}$$

By substituting λ in equation 5 and 6 by equation 7, we get

$$\begin{aligned} &f_x(h_{11}x_d + h_{12}y_d + h_{13}z_d + t_1) \\ &\quad + u_0(h_{31}x_d + h_{32}y_d + h_{33}z_d + t_3) \\ &\quad - u_T(h_{31}x_d + h_{32}y_d + h_{33}z_d + t_3) = 0 \end{aligned} \tag{8}$$

$$\begin{aligned} &f_y(h_{21}x_d + h_{22}y_d + h_{23}z_d + t_2) \\ &\quad + v_0(h_{31}x_d + h_{32}y_d + h_{33}z_d + t_3) \\ &\quad - v_T(h_{31}x_d + h_{32}y_d + h_{33}z_d + t_3) = 0 \end{aligned} \tag{9}$$

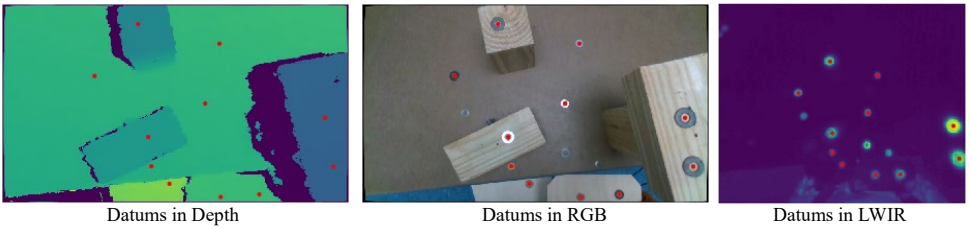
We collected the corresponding datum points (Figure 1) from both the depth and LWIR cameras. These points are the values (x_d, y_d, z_d, u_T, v_T) of the equations 8 and 9. To solve the

equations 8 and 9 in terms of h and t, we need at least six datum points, which is number of unknown. Datum points in three different modalities are shown in Figure 1. The RGB and depth image is already aligned using RealSense API. We align RGB-D and LWIR camera by mapping the datum points in different locations. Red dots on the images represent datum points. By solving nonhomogeneous system, we get rotation (R) and translation (T) matrix between two cameras.

$$p_T = \frac{1}{\lambda}(K_T R K_D^{-1} p_D + T) \tag{10}$$

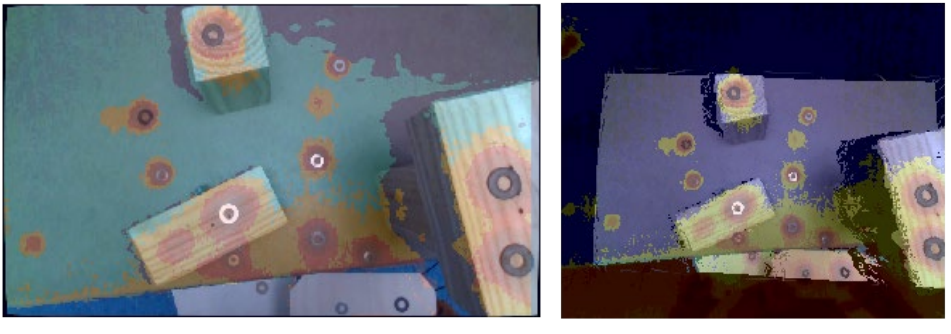
$$p_D = K_D R^{-1} K_T^{-1}(\lambda p_T - T)) \tag{11}$$

The depth map is aligned to the LWIR image in thermal camera plane with equation 10 (left image at Figure 2) and the LWIR image is aligned to the RGB frame in RGB-D camera plane with equation 11 (right image at Figure 2).



Datums in Depth Datums in RGB Datums in LWIR

Figure 1: Datum points in three modalities.



(a) Projection of LWIR onto RGBD plane (b) Projection of RGBD onto LWIR plane

Figure 2: (a) is a projection of LWIR image to RGBD plane and (b) is a projection of RGBD image to LWIR plane.

1.3 Results

We get intrinsic and extrinsic matrix of cameras through this process and the values are listed below,

$$K_T = \begin{bmatrix} 201.115 & 0.000 & 163.209 \\ 0.000 & 201.046 & 122.208 \\ 0.000 & 0.000 & 1.000 \end{bmatrix}$$

$$K_D = \begin{bmatrix} 608.598 & 0.000 & 432.474 \\ 0.000 & 606.760 & 239.820 \\ 0.000 & 0.000 & 1.000 \end{bmatrix}$$

$$R = \begin{bmatrix} -0.495 & 0.017 & -0.344 \\ -0.017 & -0.494 & -0.208 \\ 0.013 & 0.023 & 0.018 \end{bmatrix}$$

$$T = \begin{bmatrix} 0.295 & 0.172 & -0.478 \end{bmatrix}^T$$

2 Experiment details and additional results

The input resolution of all experiments in the main paper is 380×380. The RGB-D and LWIR images were resized with a bi-linear interpolation into 380×380 resolution. The cameras are mounted on helmet with 3D printed case as shown in Figure 3. Depth maps and LWIR images, which is a single channel, were duplicated three times and concatenated in channel-wise to feed into the backbone network, when depth and LWIR were used as input modalities to the networks. The backbone network is pre-trained on ImageNet [2]. Three encoders are used to fuse three modalities as shown in Figure 4 and the prediction along with inputs are shown in Figure 6. The LWIR and depth values were normalized to be within range of [0,1] and each channel of RGB images were normalized by the standard deviations and means of the ImageNet dataset. We provide additional qualitative comparison images of DeepLabV3+ with RGB-D and LWIR modalities as inputs and other state-of-the-art segmentation methods as shown in Figure 7. The dataset statistic is detailed in Figure 9. The qualitative comparison of annotation method is shown in Figure 5.

Figure 3: Helmet and 3D printed case.

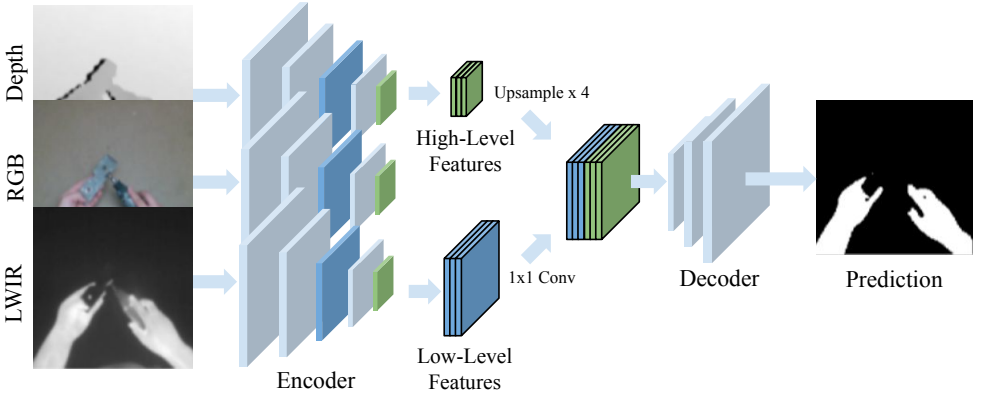


Figure 4: DeepLabV3+ with three encoders that fusing RGB-D and LWIR modalities.

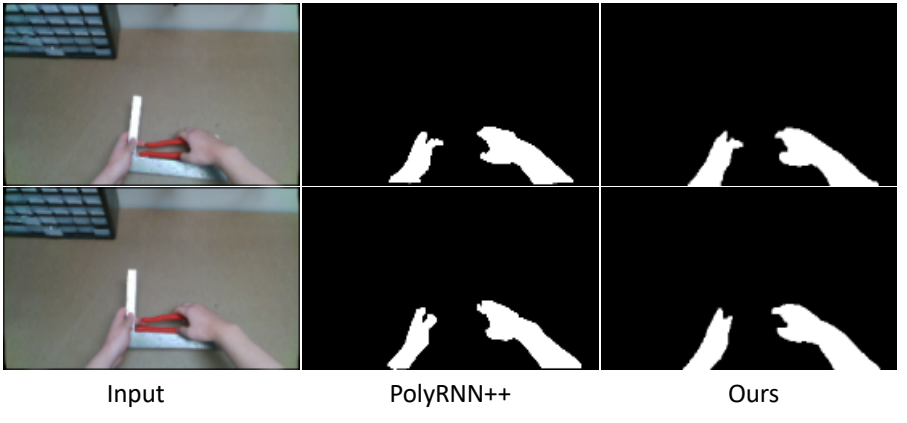


Figure 5: Qualitative comparison of manually annotated images with PolyRNN++ [1] against annotators using tablets and pen with hand ROI masked images generated using our method.

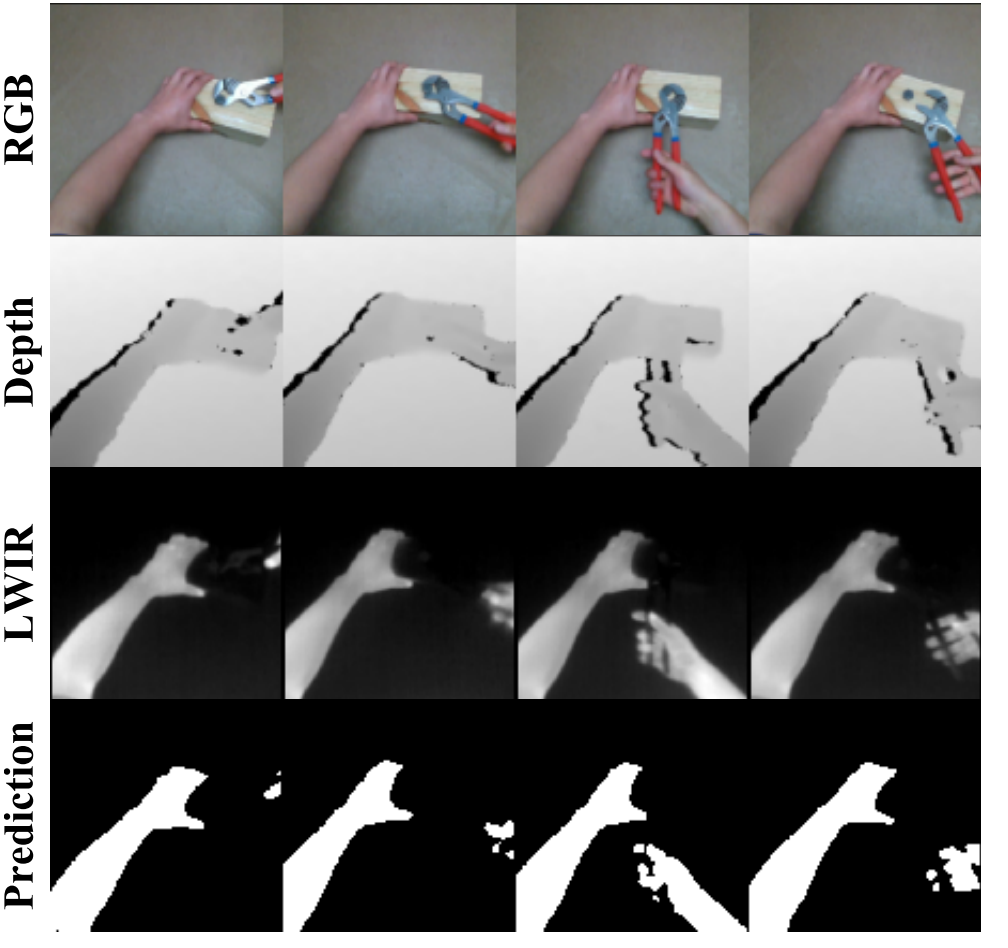


Figure 6: Predictions from DeepLabV3+ with three encoders that fusing RGB-D and LWIR modalities.

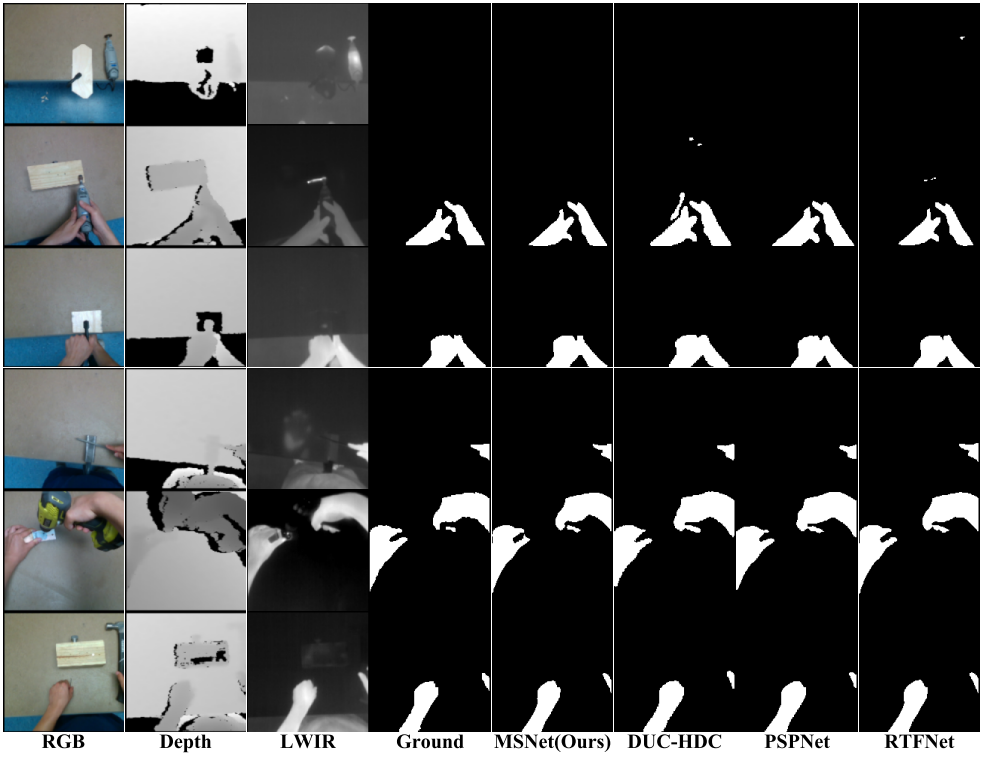


Figure 7: Qualitative results of MSNet and other state-of-the-art methods. MSNet takes LWIR and RGB-D maps as input. RGB and LWIR are used in RTFNet. The other methods used RGB as their input modality.

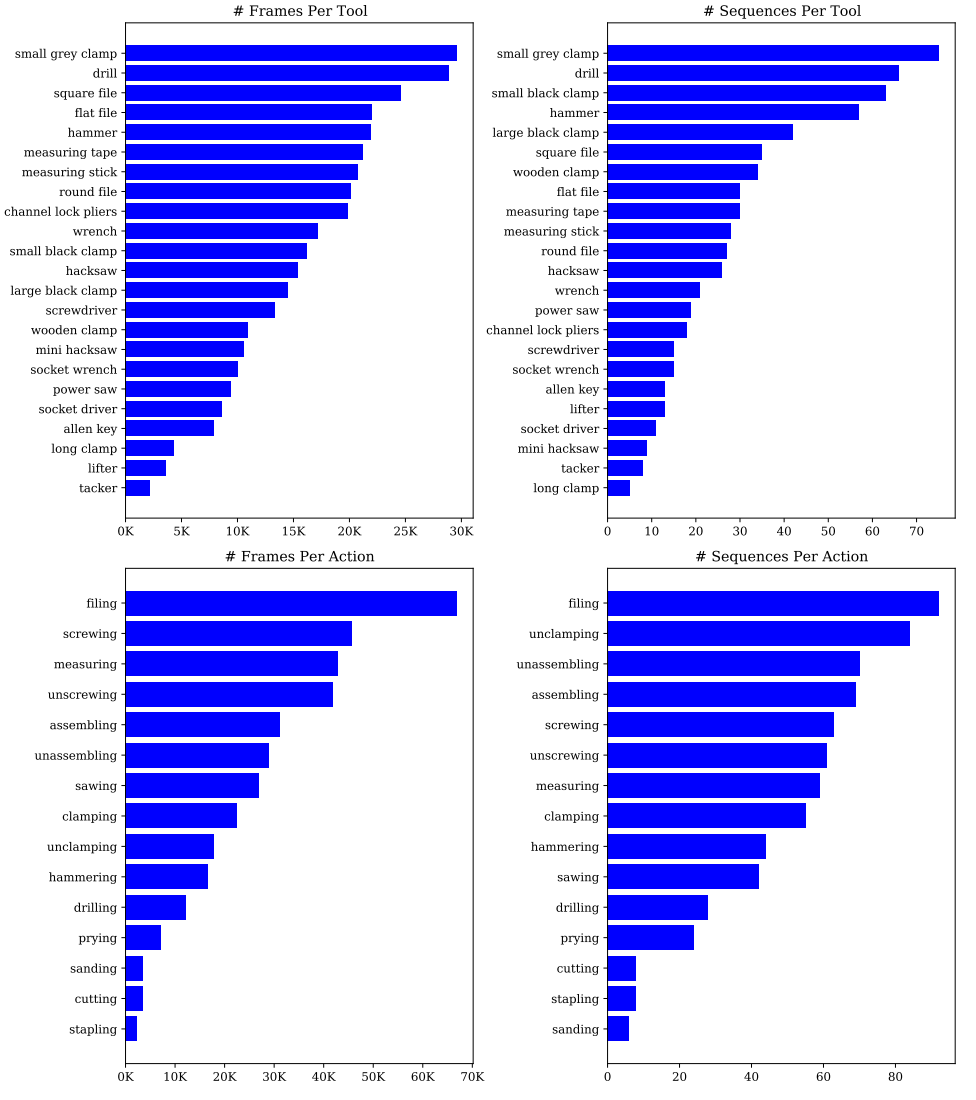


Figure 8: The number of frames and sequences of each tool and each object.

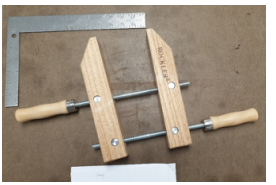
Allen Key

Channel Lock Pliers

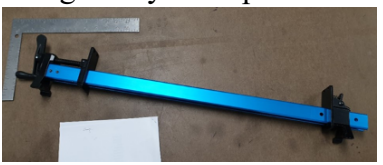
Small Grey Clamp

Large Grey Clamp

Large Black Clamp



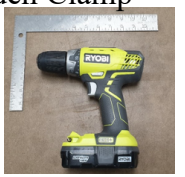
Wooden Clamp



Long Clamp

Lifter

Tacker



Drill

Square File

Flat File

Round File

Measuring Tape

Hammer

Power Saw

Hacksaw

Screwdriver

Wrench

References

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

[2] 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.