Homework 3: Pose Estimation on Unsegmented RGB-D Data

Kaiyuan Wang

December 8, 2022

1 Method

Two approaches have been attempted, the first one being 2D segmentation network + ICP, and the second one being PVN3D. Various difficulties have been encounter while experimenting with PVN3D, which given the tight schedule of this course, could not be resolved in time. This report therefore focuses on the first approach: **2D segmentation + ICP**. The ICP pipeline is identical to that used for homework2, so its details are omitted in this report.

1.1 Network Architecture

The 2D segmentation network uses the UNet architecture [1], which consists of four down-sampling layers and four up-sampling layers. Each down-sampling layer consists of a 3×3 convolution, a ReLU activation, and a 2×2 max-pooling, applied in succession. Each up-sampling layer consists of a 2×2 up-convolution, two 3 convolution, a ReLU activation, applied in succession.

Each down-sampling layer expand number of feature channels by a factor of two, while each up-sampling layer reduce number of feature channels by half. The four down-sampling layers each have channels 128, 256, 512, 1024; correspondingly, the four up-sampling layers each have channels 1024, 512, 256, 128. Each up-sampling layer takes two feature maps as inputs: one from the corresponding down-sampling layer, another from the previous up-sampling layer. The two feature maps are concatenated channel-wise, then feed into the convolution layer. The structure is also shown in figure 1.

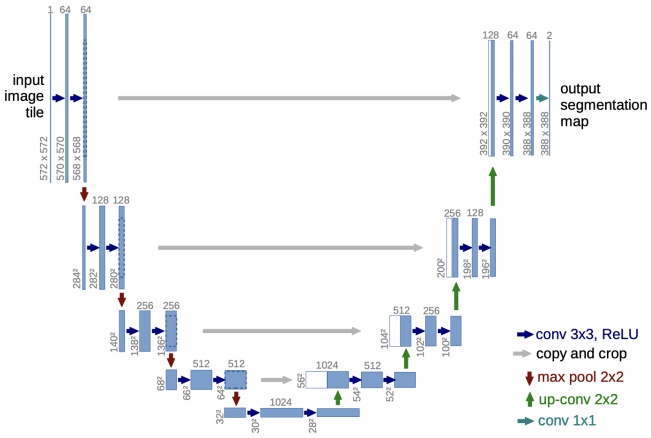

Figure 1: UNet architecture from [1].

1.2 Losses

A combination of Multi-class cross entropy loss and focal loss is used. The Multi-class cross entropy loss is used to regularize pixel-level classification:

$$L_{\text{entropy}}(x, y) = -\sum_{c=1}^{C} \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} y_{n,c}$$

where $C = 82$ denotes number of classes, subscript n denotes the n-th sample in the batch, x, y denote model prediction and ground truth segmentation tensors respectively.

Focal loss is used to address class imbalance issue:

$$\begin{aligned} L_{\text{focal}} &= -\alpha (1 - q_i)^{\gamma} \log(q_i) \\ &\text{where } q_i = c_i \cdot l_i \end{aligned}$$

where $c_i \in \mathbb{R}^C$ is the predicted confidence of the i-th pixel belonging to each class, and $l_i \in \mathbb{R}^C$ is the ground truth one-hot encoded label of the i-th pixel.

A visualization of segmented output is shown in figure 2.

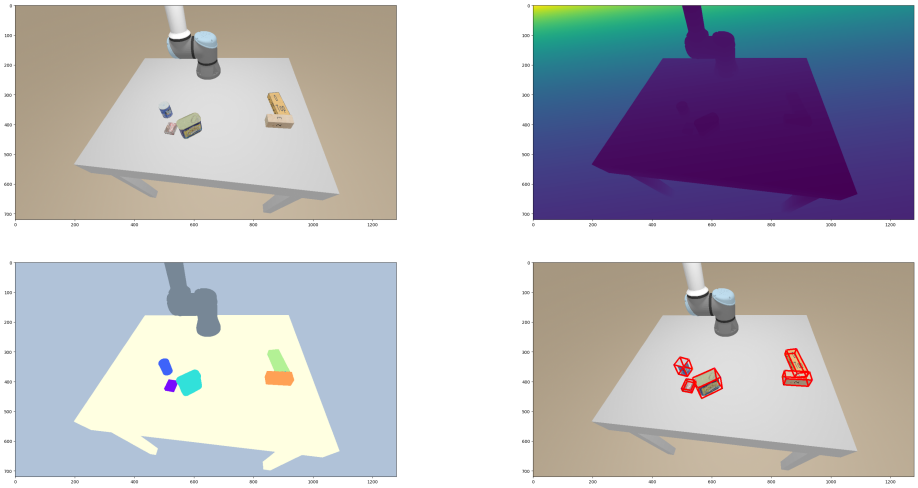

Figure 2: Visualization of model output on a test scene. Top left: input RGB image. Top right: input depth map. Bottom left: 2D segmentation mask output by UNet. Bottom right: pose estimation.

2 Experiments

2.1 Data

3964 images with ground truth label are used for training, 236 images with ground truth label are used for validation, and 400 images are used for testing. No preprocessing is applied. One reason is that the training set has ample images for the UNet to learn a good segmentation model (hence no random rotation or horizontal flip are needed to increase training set size), another reason is that the original resolution image yield better segmentation result than resized images (more details about this is discussed in subsection 2.3).

2.2 Training Details

RMSProp algorithm is used to optimize loss function, with learning rate $1e-5$, batch size 1, and 2 epochs. A linear decay learning rate scheduler is alos applied. The batch size was limited to one because of limited GPU memory - only one GPU with 10GB VRAM.

2.3 Ablation Studies and Visualization

First, several different loss function choices are investigated: 1) focal loss only, 2) dice loss only (which was adopted by the original UNet paper [1]), 3) cross entropy loss only, 4) cross entropy and focal loss. The segmentation mask output by UNet trained using each of the above loss choices is shown in figure 3. Empirically, the combination of cross entropy and focal losses yield the best result.

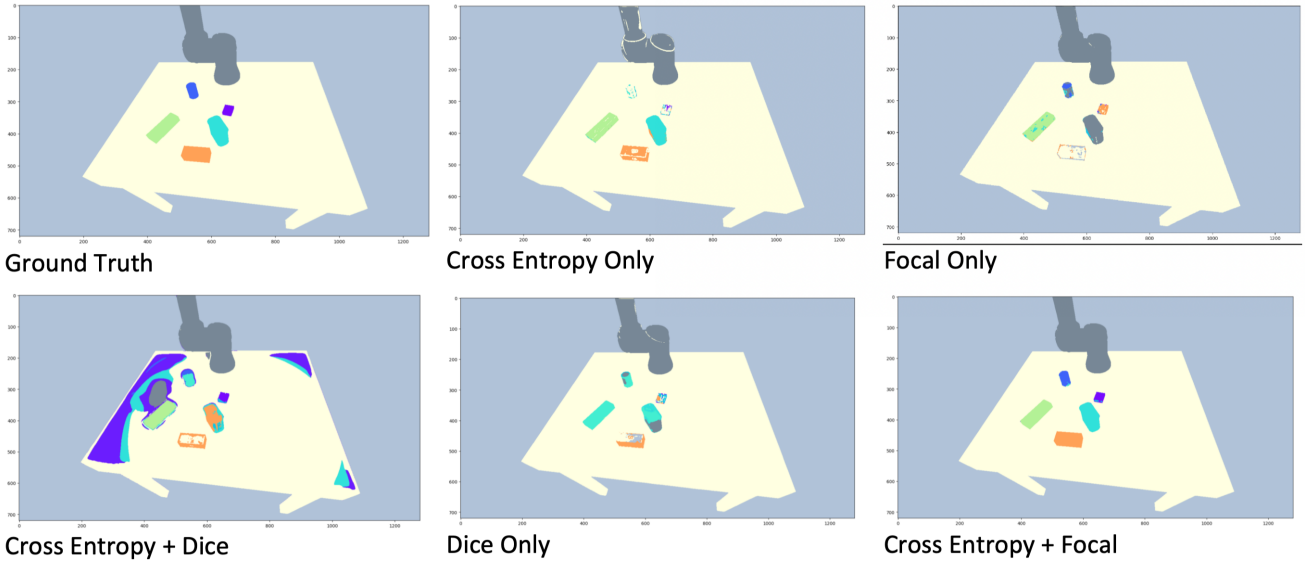

Figure 3: UNet segmentation masks (trained using difference losses).

Secondly, the effect of inappropriate learning rate and or over-fitting is investigated by adopting an unreasonably large learning rate (i.e. 1e-2) and epochs (i.e. 50), while using the optimal loss combination. As shown in figure 4, overfitting leads to a degenerative model that's incapable of producing fine-grained segmentation masks for small objects.

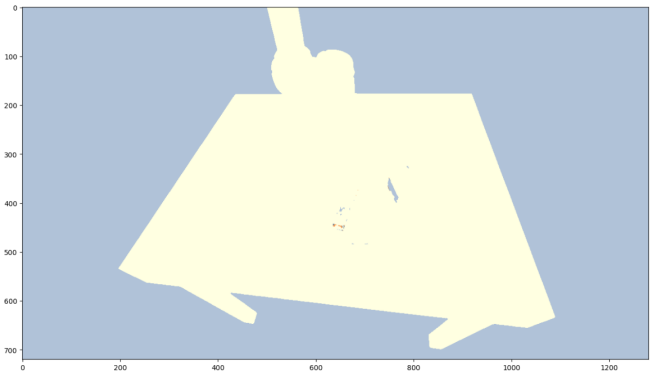

Figure 4: Segmentation mask produced by overfitted UNet.

Thirdly, to further justify the loss choice, the effect of segmentation mask quality on pose estimation result is investigated. As shown in figure 5, a low-quality segmentation mask is sparse for some objects, thereby dramatically reducing the target points available for the IPC algorithm, resulting in inferior results.

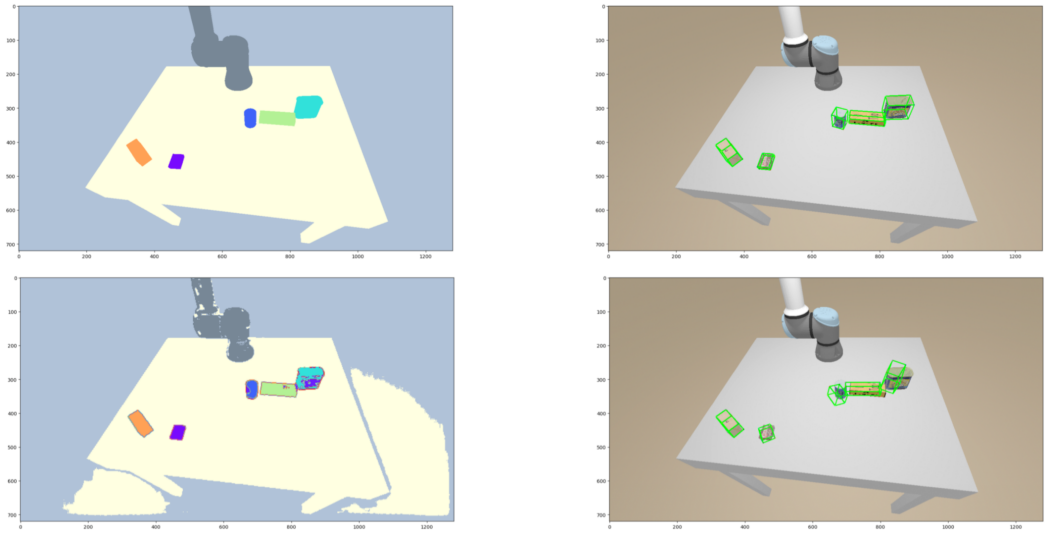

Figure 5: Top row: Pose estimation result using a high-quality segmentation mask. Bottom row: Pose estimation result using a low-quality segmentation mask.

2.4 Performance

The model achieves **66%, 71%** accuracy for 5 degree and 10 degree pose respectively. The submission is available through this link (user name is **k5wang**).

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. DOI: 10.48550/ARXIV.1505.04597. URL: https://arxiv.org/abs/1505.04597.