Toward Human-Like Grasp: Dexterous Grasping via Semantic Representation of Object-Hand
(Supplementary Material)

Anonymous ICCV submission

Paper ID 2918

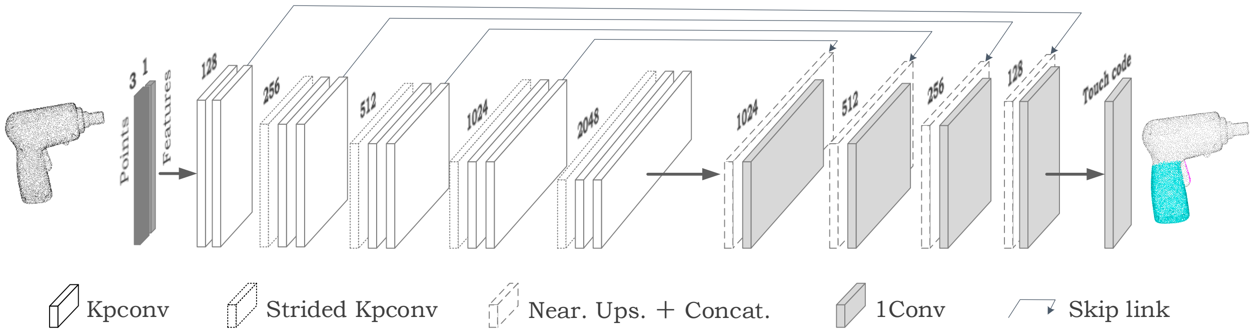

Figure 1. **The overall architecture of semantic segmentation framework. The original point cloud of the object with feature '1' is fed into the network, which is a fully convolutional network based on KPConv [1]. The output of the network is the predicted 'touch code' of each point.**

This supplementary document is organized as follows:

- Sec. A describes semantic segmentation network architecture.
- Sec. B presents semantic segmentation results.
- Sec. C details grasp synthesis network parameter.

A. Detail of Semantic Segmentation Network Architecture

As mentioned in the main body, the primary purpose of this paper is to synthesize a functional grasp. Due to the space limitation, we introduce the semantic segmentation network here. The semantic segmentation network is important, the accuracy of which directly affects the success rate of subsequent functional grasp synthesis in practical application. We are here to provide a baseline network for the following works, and the reference is KPConv [1].

Before inputting data into the network, we carry out online data augmentation on the original point cloud of object, including point cloud rotation, deformation, adding noise and so on. Then, as shown in Fig. 1, the point cloud of the object with one-dimensional feature '1' is fed into the network, which is a fully convolutional network based on KPConv [1]. The first half is the point cloud feature extractor built by kpconv and strided kpconv, and the second half is the deconvolution layers constructed by neighbor upsampling and one-dimensional convolution. The remaining network details such as feature connections and feature dimensional changes are indicated in Fig. 1. The initial parameter settings of kpconv layers in this network are: the size of the first subsampling grid is $2mm$; the number of kernel points is 15; the radius of first kpconv is $5mm$; the radius of the influence area of each kernel point in first kpconv is $2.4mm$.

B. Semantic Segmentation Results

In this section, we demonstrate experimentally that the object semantic segmentation of human grasping experience can be predicted by network, and this predictive ability of network can be generalized to the object not in the training set.

Fig. 2 show some representative segmentation results on the test set. From the perspective of visual perception, most objects can be well segmented. This show that the segmentation ability of the network can be generalized to the samples that the network has not seen. However, there are still

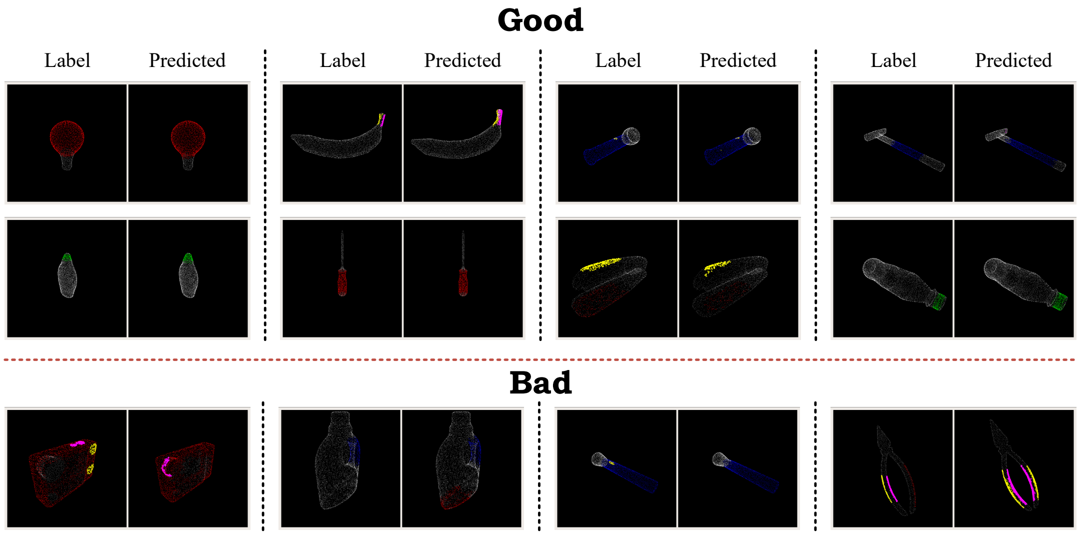

Figure 2. Segmentation results of function area of network on test set.

Test set	
mean acc	mean IoU
88.0%	61.2%

Table 1. The accuracy(acc) and the mean Intersection-over-Union (mean IoU) of the segmentation results on the test set. The higher the value is, the better the segmentation effects.

some bad segmentation results. For example, for those objects which differ greatly from the training set or have less of the same category in the training set, the segmentation result is poor, such as cramer, pitcher. Some objects with small functional areas that are difficult to segment, such as flashlights. And some objects are segmented incorrectly because of symmetry, such as pliers. These problems can be solved by expanding the dataset, designing special network to improve the fine-grained segmentation ability, or designing special loss functions to strengthen the constraint.

Tab. 1 show two numerical metrics. The first is the accuracy (acc), which shows the percentage of correct prediction points in the total input points. The second is mean Intersection-over-Union (mIoU) scores of point cloud. As we can see, for such a fine-grained segmentation problem, our baseline network can achieve accuracy of 88% and mIoU of 61.2%. It is normal that the mIoU is lower than accuracy, because there are few points in many functional areas, where the wrong prediction will greatly affect the IoU value. For example, there are 20000 points for the whole flashlight, but only about 100 points for the flashlight switch.

C. Detail of Grasp Synthesis Network Parameter

The main text mainly describes the overall structure of the grasp synthesis network, here we will supplement the detailed parameters of the network. The parameters of each layer are shown in Fig. 3, and the initial parameter settings for kpconv are the same as described in Sec. A.

For $L_{attraction}$:

$$L_{attraction} = \sum_{i=1}^{N} \sum_{j=1}^{16} \alpha_j \cdot dis\left(k_j, o_{ij}\right)) \tag{1}$$

where the weights α_j of the distal thumb, the distal index finger and the distal middle finger are 10, 5 and 2, remaining 1 for the rest.

For $L_{repulsion}$:

$$L_{repulsion} = \sum_{i=1}^{N} \sum_{j=1}^{16} \gamma_j \cdot max\left(log \frac{\beta_j}{dis\left(k_j, \widetilde{o_{ij}}\right) + \varepsilon}, 0\right) \tag{2}$$

where the weights γ_j for 16 links are “1, 1, 2, 1, 1, 2, 1, 1, 2, 1, 1, 2, 1, 1, 2, 5”, and the thresholds β_j are “10, 15, 20, 10, 15, 20, 10, 15, 20, 10, 15, 20, 10, 15, 20, 60” (mm).

For L_{angle}:

$$L_{angle} = \sum_{n=1}^{N} max\left(\theta_n - \theta_n^{max}, 0\right) + max\left(\theta_n^{min} - \theta_n, 0\right) \tag{3}$$

please refer to the factory settings of the specific robotic hand.

For $L_{self-collision}$:

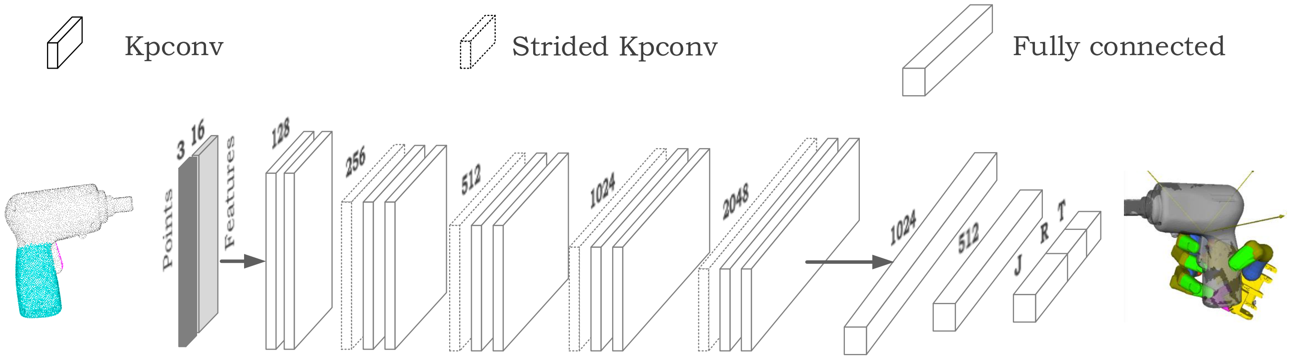

Figure 3. Some detailed parameters of the grasp synthesis network. The loss functions are omitted.

$$L_{self\_collision} = \sum_{i=1}^{16}\sum_{j=1}^{16}\mu_{ij}\cdot max\left(\delta_{ij} - dis\left(k_i, k_j\right), 0\right) \quad (4)$$

we only set a distance threshold δ_{ij} of $30mm$ for fingertips that are prone to collision, and 0 for the rest. The weights μ_{ij} are all set to 1.

References

[1] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6411–6420, 2019. 1