Robust Loss Functions for Object Grasping under Limited Ground Truth

Abstract

Object grasping is a crucial technology enabling robots to perceive and interact with the environment sufficiently. However, in practical applications, researchers are faced with missing or noisy ground truth while training the convolutional neural network, which decreases the accuracy of the model. Therefore, different loss functions are proposed to deal with these problems to improve the accuracy of the neural network. For missing ground truth, a new predicted category probability method is defined for unlabeled samples, which works effectively in conjunction with the pseudo-labeling method. Furthermore, for noisy ground truth, a symmetric loss function is introduced to resist the corruption of label noises. The proposed loss functions are powerful, robust, and easy to use. Experimental results based on the typical grasping neural network show that our method can improve performance by 2 to 13 percent.

Keywords: object grasping, missing ground truth, noisy ground truth, robust loss function

1 Introduction

Although manipulating objects is a simple task for humans, it is still a challenging problem for robots to achieve effective grasping of any object. It is widely accepted that object grasping is one of the basic operations to achieve robot control. Solving this problem will promote the use of robotics in industrial cases such as part assembly and binning.

In the last decade, convolutional neural network has achieved significant success on detection, classification and regression tasks. It is reasonable for researchers to introduce CNN into object grasping. Typical steps utilizing deep learning for grasp generation are shown in Figure 1. The input scenes are cluttered with multiple target objects occluding each other. Therefore, the researchers need to segment the target objects firstly and utilize 6D pose estimation techniques to accurately estimate the positions of the objects. Then the grasp generation network can generate grasp candidates for the target objects. Finally, grasps for the target objects need to be filtered by the evaluation network including collision detection, robustness testing and success rate filtering and etc.

However, object grasping problem still remains complicated to solve. The first challenge is the cluttering scenes existed in the environments where other similar objects may greatly decrease the successful possibility of grasping the target object. The second challenge is the generalization ability of the model. In the industrial environment, the object grasping algorithm should not only achieve accurate grasp of objects in the training dataset, but also generate effective grasp for the unseen objects. Furthermore, the most serious and important challenge is the validity and quality of data. Training data obtained in an industrial environment is probably facing with the problems of missing or noisy ground truth, which will reduce the performance of the neural network. Missing ground truth probably result in the risk of overfitting and lead to training instability. Simultaneously, due to the lack of the ground truth, the grasp evaluation network is unable to evaluate the performance of grasps accurately. As for noisy ground truth, it potentially misleads the training model into incorrect training, and correcting these erroneous labels requires a huge amount of time and resources.

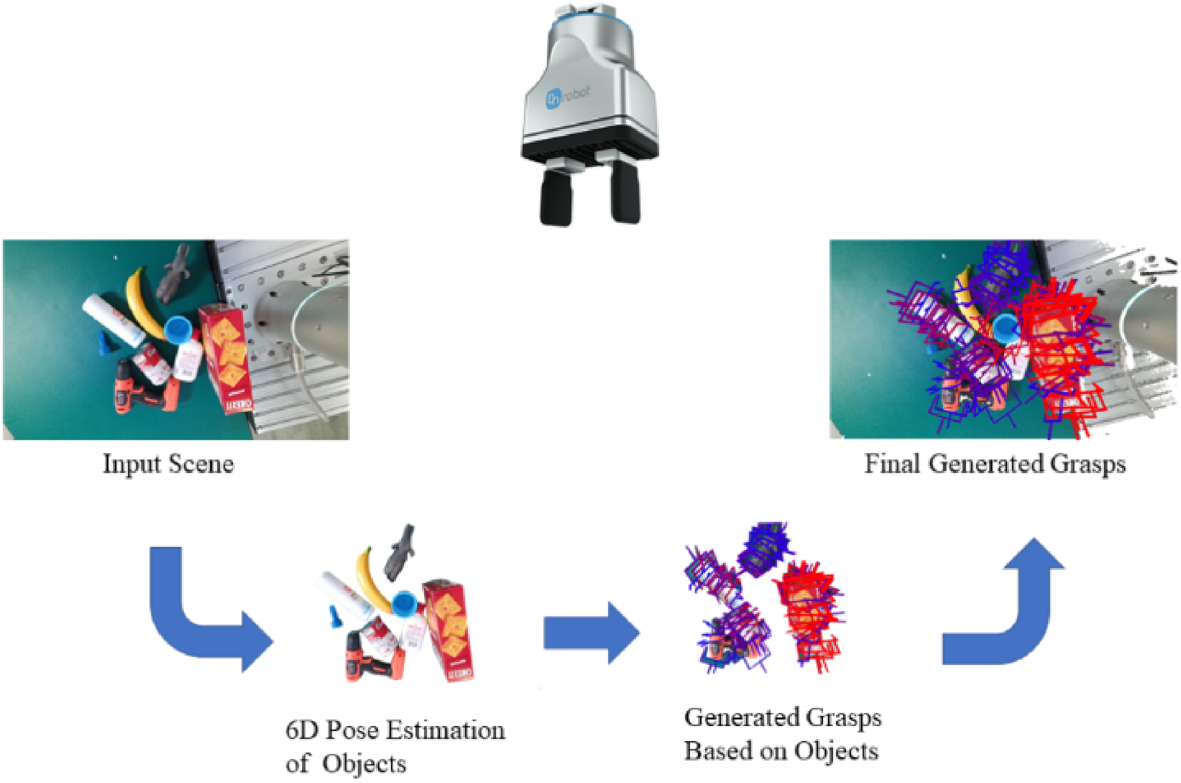
Input Scene
Final Generated Grasps
6D Pose Estimation of Objects
Generated Grasps Based on Objects

Figure 1: The most classical steps for generating grasps. After receiving the scene data as input, the neural network will segment the objects firstly and the position of the target objects will be determined by 6D pose estimation technique. Subsequently, grasps based on the target objects can be generated. Finally, all generated grasps candidates need to be evaluated in order to find the best ones, and the evaluation methods includes collision detection, success rate of grasping and etc. The gripper of the robot is capable of performing the grasp following the generated grasps.

Our proposed loss functions contribute to solve the problems of missing or noisy ground truth. Firstly, to deal with missing ground truth, a new predicted category probability method is introduced for unlabeled samples to generate better pseudo-labels. In particular, this new formulation conveys more information of missing labels. Secondly, this idea is extended to noisy ground truth, and further construct a symmetric loss function to reduce the corruption of label noises. A large number of experiments show that the performance of the algorithm can be greatly improved. To the best of our knowledge, it is the first work to address missing or noisy ground truth in object grasping. We develop new loss functions, which are very effective and can be incorporated into existing grasping neural networks easily.

2 Related work

In this part, object grasping algorithms based on RGB or RGB-D data are reviewed firstly. Then several representative grasp algorithms related to point cloud are introduced. Subsequently, typical methods for missing ground truth and noisy ground truth are discussed.

Grasping algorithms utilizing 2D and 2.5D data Since Ian Lenz systematically introduced deep learning techniques into object grasp algorithms [1, 2], more and more researchers were inspired to work on relevant research. There are primarily two directions: improving the accuracy of the grasp algorithm or the ability to handle with the cluttered environment. Researchers typically enhanced the accuracy by introducing advanced deep learning network into object grasp algorithms [3, 4, 5]. GR-ConvNet [3] utilized ResNet in the network, and FCGDN [4] introduced oriented anchor box to generate grasp candidates. GraspNet [5] based on the AutoEncoder could extract more accurate feature maps from the input images. As for handling with the clutter, many researchers interested in applying robotic grasping into industrial scenarios had made significant contributions to this field [6, 7, 8, 9, 10]. Sergey et al. [6] developed a grasp success rate prediction model and integrated it with Cross-Entropy Method algorithm to achieve the continuous control. Gualtieri et al. [7] constructed an evaluation model that evaluates a vast number of generated grasp candidates, enabling the selection of proper grasps. Mahler et al. [8] used reinforcement learning to solve the problem of clutter. Following that, they continued their work based on vacuum end effectors [9]. A grasp-first-then-recognize work-flow for grasping was proposed by Andy et al. [10], which performs well in the stowing task.

Grasping algorithms utilizing point cloud data Since PointNet [11] and PointNet++ [12] have demonstrated exceptional performance in the detection, segmentation and classification tasks based on point cloud data, the types of data used in object grasping algorithms had generally transformed from RGB or RGB-D images into point cloud [13, 14, 15, 16, 17]. Simultaneously, the performance of 6D pose estimation [18, 19, 20] was also greatly improved with the invention of PointNet++, which resulted in more researchers to choose PointNet++ as their backbone of the network. 6-DOF GraspNet [21] utilized PointNet++ as its backbone to estimate the 6D poses of the target object in order to obtain its accurate position, which could increase the successful possibility of the final grasp. Similar to PointNet++, other advanced deep learning techniques were also incorporated into object grasping algorithms, such as Transformers [22], self-Supervised learning [23, 24, 25] and transfer learning [26, 27, 28]. In addition to incorporate deep learning techniques, researchers also begun to investigate how to generate grasps on objects made of different materials. To grasp transparent objects, [29, 30] altered the structure of the gripper in order to solve the problem of inaccurate depth information for transparent objects captured by cameras. As for flexible objects, [31] constructed the prediction model to estimate the position of the objects in different time series.

Missing and noisy ground truth Many typical methods from deep learning actually also play an important role in the area of the representation learning for missing labels (also called semi-supervised learning). CCGAN [32] demonstrates that the surrounding parts of the input image is able to provide the context information, which can contribute the generator to generate pixels for the missing parts. As for pseudo-label methods [33], Entropy Minimization [34], a strategy to prevent stop the boundary from passing through the dense data points region, provides the foundational theoretical knowledge. For example, Noisy Student [35] put forward a semi-supervised method which proposed two network models, one is called student and another is teacher, to incorporate during training. [36] is commonly regarded as one of the most seminar work in the research area of the label-noise representation learning, above which an additional constrained linear "noise" layer has been introduced. This layer adjusts the output of the network to simulate a noisy label distribution. Since then, this specific area became flourished. The experiment results from [37] strongly demonstrate that the robustness against label destruction, especially for large-scale noisy datasets, could be promoted by pre-training. Dividemix [38] came up with the idea that learning all the samples is not essential for the neural network. Network can pick up the non-confusing samples as their datasets. Furthermore, the experiments demonstrate that a staged training approach can effectively alleviate the inference caused by noisy labels [39, 40]. In addition, many scholars established high-quality datasets for limited ground truth conditions [41, 42, 43]. Despite the variety of algorithms for missing and noisy ground truth, we have not found one designed for robot grasping tasks. The propose of this paper fills a gap in this field.

3 Problem statement

Grasp under the certain framework can be represented as three elements, the orientation, the translation and the width of the gripper. Therefore, we define the grasp $\boldsymbol{G}$ as:

$$\boldsymbol{G} = [\boldsymbol{R}\,\boldsymbol{t}\,w],$$

where $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ represents the orientation of the gripper, $\boldsymbol{t} \in \mathbb{R}^{3\times1}$ represents the center of the grasp and $w \in \mathbb{R}$ represents the appropriate grasping width of the target object. The determinant of the orientation matrix $\boldsymbol{R}$ must equal one and the inverse of it is its transpose, which is almost impossible for the network to learn. The classic solution is to decouple the orientation matrix as viewpoint classification and in-plane rotation. Then, just as shown in Figure 2, we can formulate the final grasp $\boldsymbol{G}$ as:

$$\boldsymbol{G} = [\boldsymbol{v}\,d\,r\,\boldsymbol{t}\,w],$$

where $\boldsymbol{v}$ represents the approaching vector, d represents the distance between the center of the grasp and the center of the gripper and r represents the in-plane rotation around the approaching axis. Besides, in order to make our grasp representation more visually understandable, we choose the most popular gripper, two-finger parallel gripper, as the example in our Figure 2. Grasp representation for other kinds of grippers can be defined in the same way.

4 Robust loss functions

4.1 Loss function for missing ground truth

Given a classification task, C is defined as the number of categories in the samples. $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$ represent the neural network predictions for labeled and unlabeled samples respectively. As for samples, $\boldsymbol{A}$ represents the labels of labeled samples and $\hat{\boldsymbol{A}}$ represents the predicted labels of unlabeled samples. N_l and N_u are defined as the number of labeled and unlabeled samples. Similarly, b_n^m and $\hat{b}_n^m$ represent the m^{th} component of neural network predictions of the n^{th} sample in the labeled and unlabeled samples. a_n^m represents the m^{th} component of label of the n^{th} sample in the labeled samples and $\hat{a}_n^m$ represents the m^{th} component of predicted label of the n^{th} sample in the unlabeled samples. Besides, $p(m|n) = softmax(b_n^m)$ is defined as the predicted probability of b_n^m and $\hat{p}(m|n) = softmax(\hat{b}_n^m)$ as the predicted probability of $\hat{b}_n^m$. $\hat{\boldsymbol{p}}(n) = [\hat{p}(1|n), \ldots, \hat{p}(m|n), \ldots, \hat{p}(C|n)]$ represents the predicted probability vector of the n^{th} unlabeled sample.

To deal with the labeled samples, we adopt the cross entropy loss function:

$$L_w(\boldsymbol{B}, \boldsymbol{A}) = -\frac{1}{N_l} \sum_{n=1}^{N_l} \sum_{m=1}^{C} a_n^m \cdot \log(p(m|n)),$$

While dealing with the situation of missing ground truth, the approach of pseudo-labels is selected. The core idea of pseudo-labels is to make use of the model itself to generate labels for unlabeled data. In particular, we evaluate the possibility of the artificial labels based on the argmax of the model's output. A predefined threshold is set up, which can filter out interferences from low possibilities. Therefore, we can propose our preliminary loss function:

$$L_p(\hat{\boldsymbol{B}}, \hat{\boldsymbol{A}}) = -\frac{1}{N_u} \sum_{n=1}^{N_u} \mu(\max(\hat{\boldsymbol{p}}(n)) > \gamma) \cdot \sum_{m=1}^{C} \hat{a}_n^m \cdot \log(\hat{p}(m|n)),$$

where $\max(\hat{\boldsymbol{p}}(n))$ denotes the highest value in $\hat{\boldsymbol{p}}(n)$, γ denotes the threshold and function $\mu(x) = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases}$. While $\max(\hat{\boldsymbol{p}}(n)) > \gamma$, we believe that the confidence of $\hat{\boldsymbol{p}}(n)$ is high and the corresponding loss function term will be reserved. Otherwise, the corresponding term will be discarded. However, we find that if $\hat{a}_n^m$ is a binary variable (either 0 or 1), it obtained less information in the predictions of the network. In order to enhance the generalization capabilities of the network, our new predicted probability as follows:

$$\hat{s}_n^m = \xi \hat{p}(m|n) + \frac{1-\xi}{C-1}\left(1 - \hat{p}(m|n)\right), \quad (1)$$

where ξ is a constant from 0 to 1. By substituting $\hat{s}_n^m$ for $\hat{a}_n^m$, it allows the loss function term for unlabeled samples to incorporate more network prediction information, leading to increased robustness during training. Therefore, our new loss function for unlabeled samples is:

$$L_u(\hat{\boldsymbol{B}}, \hat{\boldsymbol{A}}) = -\frac{1}{N_u}\sum_{n=1}^{N_u} \mu(\max(\hat{\boldsymbol{p}}(n)) > \gamma) \cdot \sum_{m=1}^{C} \hat{s}_n^m \cdot \log(\hat{p}(m|n)). \tag{2}$$

By combining the loss function of labeled and unlabeled samples, we obtain the final loss function:

$$L_m = \lambda_1 L_w(\boldsymbol{B}, \boldsymbol{A}) + \lambda_2 L_u(\hat{\boldsymbol{B}}, \hat{\boldsymbol{A}}),$$

where λ_1 and λ_2 are weights.

4.2 Loss function for missing ground truth

Given a classification task, C is defined as the number of categories in the samples. D represents samples of the dataset. Then, we define the probability of each label $c \in \{1, \dots, C\}$ as $p(c|d) = \frac{e^{g_c}}{\sum_{j=1}^{c} e^{g_j}}$, where g_j are the logits and $d \in D$.

As for noisy ground truth, symmetric cross entropy learning algorithm [44] which is able to strike a balance between sufficient learning and robustness to noisy labels plays a pivotal role. According to this algorithm, the loss function for noisy ground truth should be defined as:

$$L_t = -\sum_{c=1}^{C} q(c|d) \log p(c|d) - \sum_{c=1}^{C} p(c|d) \log q(c|d),$$

where $q(c|d)$ should be a binary variable (either 0 or 1) and qualify $\sum_{c=1}^{C} q(c|d) = 1$. However, as mentioned in the loss function for missing ground truth, the loss function cannot bring all useful information in the noisy ground truth. Therefore, we adopt the similar idea of the construction of $\hat{s}_n^m$. Then $s(c|d)$ is defined as:

$$s(c|d) = \delta p(c|d) + \frac{1-\delta}{C-1}(1 - p(c|d)),$$

where δ is a constant from 0 to 1. So $s(c|d)$ is used to define our new loss function for noisy ground truth as:

$$L_n = -\alpha_1 \sum_{c=1}^{C} s(c|d) \log p(c|d) - \alpha_2 \sum_{c=1}^{C} p(c|d) \log s(c|d),$$

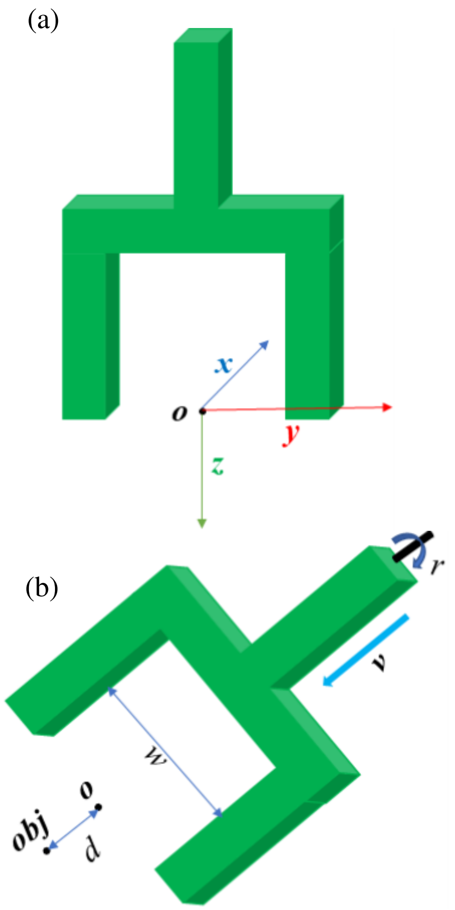

Figure 2: The representation of the final grasp. (a) The coordinate system of the gripper. (b) Our final representation of the grasp. ***obj*** denotes the center of the object. In practical grasping scenarios, the gripper will follow the direction of $\boldsymbol{v}$ to move forward for the distance of d and grasp the target object with the width w.

where α_1 and α_2 are weights. In order to express it clearly, we propose the following definition:

$$L_{ce} = -\sum_{c=1}^{C} s(c|d) \log p(c|d),$$

$$L_{rce} = -\sum_{c=1}^{C} p(c|d) \log s(c|d).$$

Therefore, the final loss function for the noisy ground truth can be rewritten as:

$$L_n = \alpha_1 L_{ce} + \alpha_2 L_{rce} \tag{3}$$

5 Experiments

In this section, we adopt the network and dataset from GraspNet-1Billion [45] to demonstrate that our loss functions can improve the performance while facing with the problem of missing or noisy ground truth. The dataset of GraspNet-1Billion contains 97,280 RGB-D images, which consists of 190 cluttered scenes with 88 different objects. Particularly, 512 RGB-D images are provided with each scene. Objects in each scene contain various grasp poses. The basic network of GrasspNet-1Billion [45] consists of three parts: ApproachNet, OperationNet, and ToleranceNet. ApproachNet is used to extract features and approaching vectors from point cloud input data, after which OperationNet utilizes the extracted features and approaching vectors to generate grasp candidates. ToleranceNet simultaneously provide the robustness and feasibility of the grasp candidates. Considering that ToleranceNet is only used for evaluating the robustness and feasibility of the grasp and is not involved in the calculation of the values of grasp candidates, we only modified the loss functions of ApproachNet and OperationNet when dealing with the conditions under limited data.

5.1 Missing ground truth

The ground truth of one grasp is consisted of the in-plane rotation, width of the gripper, and grasping confidence score. We adopted a strategy of Missing Completely at Random (MCAR) for the ground truth, setting up constant κ_1 as the proportion of data removed in the original dataset. During training, if encountering complete ground truth, the normal loss function mentioned in the original paper [45] was utilized for training. If the ground truth is missing, the loss function proposed in section 4.1 was used.

The grasping confidence score is related to the loss function of the ApproachNet. Following the proposed loss function (as defined in Equation (2)) and the loss function from [45], we can modify the loss function of ApproachNet in GraspNet-1Billion into the following formula:

$$\begin{aligned} L^A(\{c_i\},\{s_{ij}\}) =& \frac{1}{N_{cls}}\sum_i L_{cls}(c_i,c_i^*) + \beta_1\frac{1}{N_{reg}}\cdot \\ & \sum_i\sum_j c_i^*\mathbf{1}(|v_{ij},v_{ij}^*| < 5^\circ)\cdot F_A, \end{aligned}$$

$$F_A = \begin{cases} L_{reg}(s_{ij},s_{ij}^*), & \text{with ground truth} \\ L_u(s_{ij},\hat{s}_{ij}), & \text{without ground truth} \end{cases}$$

where c_i represents the binary value for whether it is graspable or not for each point i, c_i^* is regarded as 1 if point i is positive and 0 if negative, s_{ij} represents the predicted confidence score and j means the viewpoint, $\hat{s}_{ij}$ is the predicted value based on the Equation (1), s_{ij}^* is the corresponding ground truth of s_{ij} and $|v_{ij},v_{ij}^*|$ represents the angle difference between these two approaching vectors. β_1 is the constant which is usually set as 0.5. L_{cls} we use here denotes a two class softmax loss and L_{reg} denotes the smooth L_1 loss.

The in-plane rotation and the width of the gripper are associated with the loss function of the OperationNet. Following the same idea to deal with the ApproachNet, we can also modify the loss function of OperationNet in GraspNet-1Billion into the following formula:

$$\begin{aligned} L^R(R_{ij},S_{ij},W_{ij}) =& \sum_{c=1}^{C}(\frac{1}{N_{cls}}\sum_{ij}F_{O1} \\ & + \beta_2\frac{1}{N_{reg}}\sum_{ij}F_{O2} \\ & + \beta_3\frac{1}{N_{reg}}\sum_{ij}F_{O3}), \end{aligned}$$

$$F_{O1} = \begin{cases} L_{cls}^d(R_{ij},R_{ij}^*), & \text{with ground truth} \\ L_u(R_{ij},\hat{R}_{ij}), & \text{without ground truth,} \end{cases}$$

$$F_{O2} = \begin{cases} L_{reg}^d(S_{ij},S_{ij}^*), & \text{with ground truth} \\ L_u(S_{ij},\hat{S}_{ij}), & \text{without ground truth,} \end{cases}$$

$$F_{O3} = \begin{cases} L_{reg}^d(W_{ij},W_{ij}^*), & \text{with ground truth} \\ L_u(W_{ij},\hat{W}_{ij}), & \text{without ground truth,} \end{cases}$$

where R_{ij} represents the rotation degree, S_{ij} denotes the grasp confidence score, W_{ij} are regarded as gripper width. $\hat{R}_{ij}$, $\hat{S}_{ij}$ and $\hat{W}_{ij}$ represent the predicted values of R_{ij}, S_{ij}, and W_{ij}

Table 1: Evaluation based on ground truth of different missing ratios.

κ_1	Methods	Seen			Unseen			Novel		
		AP	$\text{AP}_{0.8}$	$\text{AP}_{0.4}$	AP	$\text{AP}_{0.8}$	$\text{AP}_{0.4}$	AP	$\text{AP}_{0.8}$	$\text{AP}_{0.4}$
50%	[45]	23.35	28.09	14.10	22.98	29.05	12.80	8.23	7.87	3.38
	Ours	**23.86**	27.12	**14.45**	22.95	**30.21**	**13.82**	**9.14**	**8.66**	**3.79**
60%	[45]	21.98	26.42	13.25	22.89	25.63	11.95	7.28	6.75	3.10
	Ours	**22.48**	**26.99**	13.16	**23.94**	**26.86**	**12.93**	**8.01**	**7.49**	**3.48**
70%	[45]	20.05	22.07	12.58	18.80	22.22	11.38	6.33	5.85	2.95
	Ours	**20.60**	22.02	**13.19**	**19.72**	**23.38**	11.34	**6.90**	**6.55**	**3.34**

respectively. R^*_{ij}, S^*_{ij} and W^*_{ij} represent the corresponding ground truth of R_{ij}, S_{ij}, and W_{ij} respectively. L^d represents the mean loss for the d^{th} binned distance. β_2 and β_3 are constants. In particular, L_{cls} represents the sigmoid cross entropy loss function.

Table I demonstrates the evaluation results utilizing various ratios of ground truth. The percentages in the first column, κ_1, represent the proportion of missing ground truth in the dataset. By comparing to its original values under complete ground truth, the average AP value of GraspNet-1Billlion [45] decrease around 20%, which proves that the performance of the grasping algorithm is highly dependent on the ground truth. In situations where ground truth is missing, training the grasping algorithm on seen and unseen datasets does not significantly deteriorate, as the target objects in these test datasets are highly similar to the objects provided in the training dataset. However, its generalization ability is greatly affected, which is shown from the evaluation results of novel dataset. By employing our loss function, the model not only achieves an average increase of 3% on seen and unseen datasets but also significantly enhances its generalization capability, with its performance on novel dataset increasing averagely by 15%.

5.2 *Noisy ground truth*

Following the similar idea from missing ground truth condition, the noisy ground truth was generated through changing the values of the in-plane rotation and the width of the gripper. In particular, as we did not modify the grasping confidence scores, the loss function of the ApproachNet remains unchanged. Firstly, we set up two constants, κ_2 and ϵ. κ_2 represents the proportion of data modified in the original dataset. ϵ represents the extent to which the ground truth is modified. Both κ_2 and ϵ are in the range from 0 to 1. In this part of experiment, we randomly selected a propotion, κ_2, of the ground truth and multiplied them all by ϵ to achieve the effect of noisy ground truth.

The in-plane rotation and the width of the gripper are in connection with the loss function of the OperationNet. Following the proposed loss function (as defined in Equation (3)) and the loss function from [45], the loss function of OperationNet for the noisy ground truth in GraspNet-1Billion can be constructed as the following formula:

$$\begin{aligned} L^R(R_{ij}, S_{ij}, W_{ij}) = &\sum_{c=1}^{C}(\frac{1}{N_{cls}}\sum_{ij}(L_{ce}(R_{ij}, R^*_{ij}) \\ &+ L_{rce}(R_{ij}, R^*_{ij})) \\ &+ \eta_2 \frac{1}{N_{reg}}\sum_{ij}(L_{ce}(S_{ij}, S^*_{ij}) \\ &+ L_{rce}(S_{ij}, S^*_{ij})) \\ &+ \eta_3 \frac{1}{N_{reg}}\sum_{ij}(L_{ce}(W_{ij}, W^*_{ij}) \\ &+ L_{rce}(W_{ij}, W^*_{ij}))), \end{aligned}$$

the parameters setting is the same as Section 5.1. η_2 and η_3 are constants.

Table II shows the evaluation results using various ratios and noisy factors of ground truth.

Table 2: Evaluation based on ground truth of different ratios and noisy factors.

κ_2	ϵ	Me-thods	Seen			Unseen			Novel		
			AP	AP$_{\mathbf{0.8}}$	AP$_{\mathbf{0.4}}$	**AP**	AP$_{\mathbf{0.8}}$	AP$_{\mathbf{0.4}}$	**AP**	AP$_{\mathbf{0.8}}$	AP$_{\mathbf{0.4}}$
50%	0.5	[45]	25.01	30.69	14.64	23.02	30.10	12.23	6.49	6.56	2.62
		Ours	**25.28**	30.65	**14.94**	**23.98**	30.08	**12.71**	**7.43**	**7.75**	**2.95**
50%	0.6	[45]	24.73	30.39	14.53	22.97	29.93	12.11	6.44	6.45	2.66
		Ours	**25.05**	**30.71**	14.46	**23.95**	**31.79**	**12.56**	**7.47**	**7.64**	**3.00**
60%	0.5	[45]	24.18	29.69	14.17	22.79	29.90	12.08	6.39	6.33	2.58
		Ours	**24.58**	**30.10**	14.16	**23.79**	**32.00**	**12.50**	**7.52**	**7.42**	**2.92**
60%	0.6	[45]	23.63	28.96	13.81	24.80	29.45	11.92	5.97	6.20	2.50
		Ours	**24.14**	**29.46**	**14.19**	**26.09**	28.58	**12.46**	**7.14**	**7.37**	**2.84**

The first column and the second column represent the ratios of noisy ground truth κ_2 and the noisy factors ϵ respectively. By constructing a symmetric loss function, the influence of noisy ground truth was reduced on the final calculation of the loss function. This approach enhances the effectiveness of training process. The improvement in values of Table II demonstrates that the accuracy and generalization ability have both been further enhanced under conditions of noisy ground truth. On seen and unseen datasets, the performances contain an overall increase of 2%, and on the novel datasets, the performances obtain an overall increase of 10%.

6 Conclusion

In this paper, we proposed two loss functions to address the issues of missing ground truth and noisy ground truth in grasping algorithms. For missing ground truth, the pseudo-labeling approach is utilized to mitigate the problem of insufficient data where our proposed predicted category possibility method plays an essential role. For noisy ground truth, a symmetric loss function is constructed to reduce the impact of label noises on training. Through comparative experiments, we demonstrated that these two loss functions can not only enhance the accuracy of algorithm but also improve its generalization ability under the condition of limited data. For future work, we will attempt to build a dataset specifically for training grasping algorithms under limited conditions, filling a gap in this field.

References

[1] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1316–1322. IEEE, 2015.

[2] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[3] Sulabh Kumra, Shirin Joshi, and Ferat Sahin. Antipodal robotic grasping using generative residual convolutional neural network. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9626–9633. IEEE, 2020.

[4] Xinwen Zhou, Xuguang Lan, Hanbo Zhang, Zhiqiang Tian, Yang Zhang, and Narming Zheng. Fully convolutional grasp detection network with oriented anchor box. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7223–7230. IEEE, 2018.

[5] Umar Asif, Jianbin Tang, and Stefan Harrer. Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices. In *IJCAI*, volume 7, pages 4875–4882, 2018.

[6] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[7] Marcus Gualtieri, Andreas Ten Pas, Kate Saenko, and Robert Platt. High precision grasp pose detection in dense clutter. in 2016 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 598–605.

[8] Jeffrey Mahler and Ken Goldberg. Learning deep policies for robot bin picking by simulating robust grasping sequences. In *Conference on robot learning*, pages 515–524. PMLR, 2017.

[9] Jeffrey Mahler, Matthew Matl, Xinyu Liu, Albert Li, David Gealy, and Ken Goldberg. Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In *2018 IEEE International Conference on robotics and automation (ICRA)*, pages 5620–5627. IEEE, 2018.

[10] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705, 2022.

[11] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[12] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[13] Mingze Wei, Yaomin Huang, Zhiyuan Xu, Ning Liu, Zhengping Che, Xinyu Zhang, Chaomin Shen, Feifei Feng, Chun Shan, and Jian Tang. Cmg-net: An end-to-end contact-based multi-finger dexterous grasping network. *arXiv preprint arXiv:2303.13182*, 2023.

[14] Isabella Huang, Yashraj Narang, Ruzena Bajcsy, Fabio Ramos, Tucker Hermans, and Dieter Fox. Defgraspnets: Grasp planning on 3d fields with graph neural nets. *arXiv preprint arXiv:2303.16138*, 2023.

[15] Jacob Varley, Jonathan Weisz, Jared Weiss, and Peter Allen. Generating multi-fingered robotic grasps via deep learning. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4415–4420. IEEE, 2015.

[16] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2901–2910, 2019.

[17] Hongzhuo Liang, Xiaojian Ma, Shuang Li, Michael Görner, Song Tang, Bin Fang, Fuchun Sun, and Jianwei Zhang. Pointnetgpd: Detecting grasp configurations from point sets. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3629–3635. IEEE, 2019.

[18] Haojie Huang, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Edge grasp network: A graph-based se (3)-invariant approach to grasp detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3882–3888. IEEE, 2023.

[19] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp

detection for transparent and specular objects using generalizable nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1757–1763. IEEE, 2023.

[20] Yiye Chen, Yunzhi Lin, Ruinian Xu, and Patricio A Vela. Keypoint-graspnet: Keypoint-based 6-dof grasp generation from the monocular rgb-d input. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7988–7995. IEEE, 2023.

[21] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3):4978–4985, 2020.

[22] Zhixuan Liu, Zibo Chen, Shangjin Xie, and Wei-Shi Zheng. Transgrasp: A multi-scale hierarchical point transformer for 7-dof grasp detection. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1533–1539. IEEE, 2022.

[23] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8973–8979. IEEE, 2019.

[24] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8973–8979. IEEE, 2019.

[25] Menglong Zhu, Konstantinos G Derpanis, Yinfei Yang, Samarth Brahmbhatt, Mabel Zhang, Cody Phillips, Matthieu Lecce, and Kostas Daniilidis. Single image 3d object detection and pose estimation for grasping. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3936–3943. IEEE, 2014.

[26] Andrei A Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *Conference on robot learning*, pages 262–270. PMLR, 2017.

[27] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*, 2018.

[28] Abhishek Gupta, Clemens Eppner, Sergey Levine, and Pieter Abbeel. Learning dexterous manipulation for a soft robotic hand from human demonstrations. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3786–3793. IEEE, 2016.

[29] Jaesik Chang, Minju Kim, Seongmin Kang, Heungwoo Han, Sunpyo Hong, Kyunghun Jang, and Sungchul Kang. Ghostpose: Multi-view pose estimation of transparent objects for robot hand grasping. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5749–5755. IEEE, 2021.

[30] Yingjie Tang, Junhong Chen, Zhenguo Yang, Zehang Lin, Qing Li, and Wenyin Liu. Depthgrasp: depth completion of transparent objects using self-attentive adversarial network with spectral residual for grasping. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5710–5716. IEEE, 2021.

[31] Hideyuki Ichiwara, Hiroshi Ito, Kenjiro Yamamoto, Hiroki Mori, and Tetsuya Ogata. Contact-rich manipulation of a flexible object based on deep predictive learning using vision and tactility. In *2022 International Conference on*

Robotics and Automation (ICRA), pages 5375–5381. IEEE, 2022.

[32] Emily Denton, Sam Gross, and Rob Fergus. Semi-supervised learning with context-conditional generative adversarial networks. *arXiv preprint arXiv:1611.06430*, 2016.

[33] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[34] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In François Denis, editor, *Actes de CAP 05, Conférence francophone sur l'apprentissage automatique - 2005, Nice, France, du 31 mai au 3 juin 2005*, pages 281–296. PUG, 2005.

[35] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. Computer Vision Foundation / IEEE, 2020.

[36] Bekhzod Olimov, Jeonghong Kim, and Anand Paul. Dcbt-net: Training deep convolutional neural networks with extremely noisy labels. *IEEE Access*, 8:220482–220495, 2020.

[37] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2712–2721. PMLR, 2019.

[38] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff A. Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6234–6243. PMLR, 2019.

[39] Yueming Lyu and Ivor W. Tsang. Curriculum loss: Robust learning and generalization against label corruption. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[40] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor W. Tsang, and Masashi Sugiyama. SIGUA: forgetting may make learning with noisy labels more robust. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4006–4016. PMLR, 2020.

[41] Kyungjune Baek, Seungho Lee, and Hyunjung Shim. Learning from better supervision: Self-distillation for learning with noisy labels. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 1829–1835. IEEE, 2022.

[42] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective, 2020.

[43] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2309–2318. PMLR, 2018.

[44] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 322–330. IEEE, 2019.

[45] Haoshu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11441–11450. Computer Vision Foundation / IEEE, 2020.