FS-Net: Fast Shape-based Network for Category-Level 6D Object Pose Estimation with Decoupled Rotation Mechanism

Wei Chen [1] Xi Jia[1] Hyung Jin Chang[1] Jinming Duan[1] Linlin Shen[2] Aleš Leonardis[1]

[1] School of Computer Science, University of Birmingham
[2] Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University
{✉wxc795, X.Jia.1, a.leonardis}@cs.bham.ac.uk, {H.J.Chang,J.Duan}@bham.ac.uk,llshen@szu.edu.cn

Abstract

In this paper, we focus on category-level 6D pose and size estimation from a monocular RGB-D image. Previous methods suffer from inefficient category-level pose feature extraction, which leads to low accuracy and inference speed. To tackle this problem, we propose a fast shape-based network (FS-Net) with efficient category-level feature extraction for 6D pose estimation. First, we design an orientation aware autoencoder with 3D graph convolution for latent feature extraction. Thanks to the shift and scale-invariance properties of 3D graph convolution, the learned latent feature is insensitive to point shift and object size. Then, to efficiently decode category-level rotation information from the latent feature, we propose a novel decoupled rotation mechanism that employs two decoders to complementarily access the rotation information. For translation and size, we estimate them by two residuals: the difference between the mean of object points and ground truth translation, and the difference between the mean size of the category and ground truth size, respectively. Finally, to increase the generalization ability of the FS-Net, we propose an online box-cage based 3D deformation mechanism to augment the training data. Extensive experiments on two benchmark datasets show that the proposed method achieves state-of-the-art performance in both category- and instance-level 6D object pose estimation. Especially in category-level pose estimation, without extra synthetic data, our method outperforms existing methods by 6.3% on the NOCS-REAL dataset [1].

1. Introduction

Estimating 6D object pose plays an essential role in many computer vision tasks such as augmented reality [19, 20], virtual reality [2], and smart robotic arm [46, 35]. For instance-level 6D pose estimation, in which training set and test set contain the same objects, huge progress has been made in recent years [41, 28, 21, 15, 10]. However, category-level 6D pose estimation remains challenging as the object shape and color are various in the same category. Existing methods addressed this problem by mapping the different objects in the same category into a uniform model via RGB feature or RGB-D fusion feature. For example, Wang *et al.* [40] trained a modified Mask R-CNN [9] to predict the normalized object coordinate space (NOCS) map of different objects based on RGB feature, and then computed the pose with observed depth and NOCS map by Umeyama algorithm [36]. Chen *et al.* [4] proposed to learn a canonical shape space (CASS) to tackle intra-class shape variations with RGB-D fusion feature [39]. Tian *et al.* [34] trained a network to predict the NOCS map of different objects, with the uniform shape prior learned from a shape collection, and RGB-D fusion feature [39].

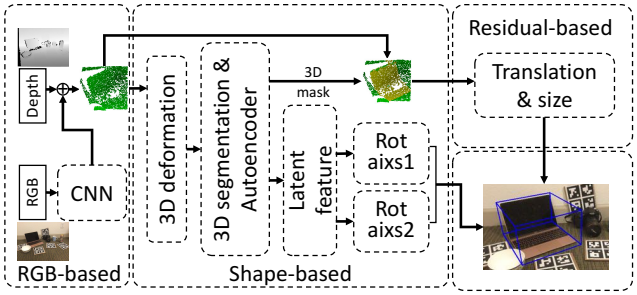

Figure 1. **FS-Net** comprises different networks for different tasks. The RGB-based network is used for 2D object detection, and the shape-based network is used for 3D segmentation and rotation estimation. The residual-based network is used for translation and size estimation with segmented points.

Although these methods achieved state-of-the-art performance, there are still two remaining issues. Firstly, the

[1]The code is at https://github.com/DC1991/FS-Net

benefits of using RGB feature or RGB-D fusion feature for category-level pose estimation are still questionable. Vlach et al. [37] showed that people focus more on shape than color when categorizing objects, as different objects in the same category have very different colors but stable shapes (shown in Figure 3). Thereby the use of RGB feature for category-level pose estimation may lead to low performance due to huge color variation in the test scene. For this issue, to alleviate the color variation, we merely use the RGB feature for 2D detection while using the shape feature learned with point cloud extracted from depth image for category-level pose estimation.

Secondly, learning a representative uniform shape requires a large amount of training data. Therefore, the performance of these methods is not guaranteed with limited training examples. To overcome this issue, we propose a 3D graph convolution (3DGC) autoencoder [18] to effectively learn the category-level pose feature via observed points reconstruction of different objects instead of uniform shape mapping. We further propose an online box-cage based 3D data augmentation mechanism to reduce the dependencies of labeled data.

In this paper, the newly proposed FS-Net consists of three parts: 2D detection, 3D segmentation & rotation estimation, and translation & size estimation. In 2D detection part, we use the YOLOv3 [30] to detect the object bounding box for coarse object points obtainment [6]. As to the 3D segmentation & rotation estimation part, we design a 3DGC autoencoder to perform segmentation and observed points reconstruction jointly. The autoencoder encodes orientation information in the latent feature. Then we propose the decoupled rotation mechanism that uses two decoders to decode the category-level rotation information. For translation and size estimation, since they are all point coordinates related, we design a coordinate residual estimation network based on PointNet [26] to estimate the translation residual and size residuals. To further increase the generalization ability of FS-Net, we use the proposed online 3D deformation for data augmentation. To summarize, the main contributions of this paper are as follows:

- We propose a fast shape-based network to estimate category-level 6D object size and pose. Due to the efficient category-level pose feature extraction, the framework runs at 20 FPS on a GTX 1080 Ti GPU.
- We propose a 3DGC autoencoder to reconstruct the observed points for latent orientation feature learning. Then we design a decoupled rotation mechanism to fully decode the orientation information. This decoupled mechanism allows us to naturally handle the circle symmetry object (in Section 3.3).
- Based-on the shape similarity, we propose a novel box-cage based 3D deformation mechanism to augment the training data. With this mechanism, the pose accuracy of FS-Net is improved by 7.7%.

2. Related Works

2.1. Instance-Level Pose Estimation

In instance-level pose estimation, a known 3D object model is usually available for training and testing. Based on the 3D model, instance-level pose estimation can be roughly divided into three types: template matching based, correspondences-based, and voting-based methods. Template matching methods [11, 29, 21] aligned the template to the observed image or depth map via hand-crafted or deep learning feature descriptors. As they need the 3D object model to generate the template pool, their applications in category-level 6D pose estimation are limited. Correspondences-based methods trained their model to establish 2D-3D correspondences [28, 29, 23] or 3D-3D correspondences [6, 5]. Then they solved perspective-n-point and SVD problem with 2D-3D and 3D-3D correspondences [13], respectively. Some methods [5, 1] also used these correspondences to generate voting candidates, and then used RANSAC [8] algorithm for selecting the best candidate. However, the generation of canonical 3D keypoints is based on the known 3D object model that is not available when predicting the category-level pose.

2.2. Category-Level Pose Estimation

Compared to instance-level, the major challenge of category-level pose estimation is the intra-class object variation, including shape and color variation. To handle the object variation problem, [40] proposed to map the different objects in the same category to a NOCS map. Then they used semantic segmentation to access the observed points cloud with known camera parameters. The 6D pose and size are calculated by the Umeyama algorithm [36] with the NOCS map and the observed points. Shape-Prior [34] adopted similar method with [40], but both extra shape prior knowledge and dense-fusion feature [39], instead of RGB feature, are used. CASS [4] estimated the 6D pose via the learning of a canonical shape space with dense-fusion feature [39]. Since the RGB feature is sensitive to color variation, the performance of their methods in category-level pose estimation is limited. In contrast, our method is shape feature-based which is robust for this task.

2.3. 3D Data Augmentation

In 3D object detection tasks [6, 25, 31, 5], online data augmentation techniques such as translation, random flipping, shifting, scaling, and rotation are applied to original point clouds for training data augmentation. However, these operations cannot change the shape property of the object.

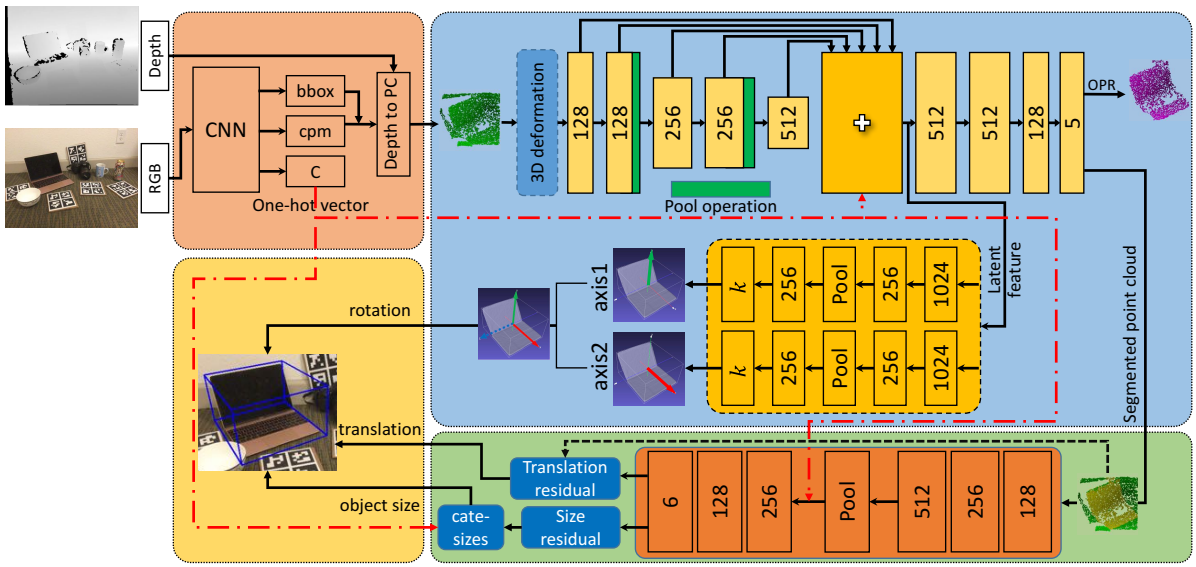

Figure 2. **Architecture of FS-Net.** The input of FS-Net is an RGB-D image. For RGB channels, we use a 2D detector to detect the object 2D location, category label 'C' (used as a one-hot feature for next tasks), and class probability map (cpm) (generate the 3D sphere center via maximum probability location and camera parameters). With this information and depth channel, the points in a compact 3D sphere are generated. Given the points in the 3D sphere, we first use the proposed 3D deformation mechanism for data augmentation. After that, we use a shape-based 3DGC autoencoder to perform observed points reconstruction (OPR), as well as point cloud segmentation, for orientation latent feature learning. Then, we decode the rotation information into two perpendicular vectors from the latent feature. Finally, we use a residual estimation network to predict the translation and size residuals. 'cate-sizes' denotes the pre-calculated average sizes of different categories, 'k' is the rotation vector dimension, and the hollow '+' means feature concatenation.

Simply adopting these operations on point clouds is not able to handle the shape variation problem in the 3D task. To address this, [7] proposed part-aware augmentation which operates on the semantic parts of the 3D object with five manipulations: dropout, swap, mix, sparing, and noise injection. However, how to decide the semantic parts are ambiguous. In contrast, we propose a box-cage based 3D data augmentation mechanism which can generate the various shape variants (shown in Figure 5) and avoid semantic parts decision procedure.

3. Proposed Method

In this section, we describe the detailed architecture of FS-Net shown in Figure 2. Firstly, we use the YOLOv3 to detect the object location with RGB input. Secondly, we use 3DGC autoencoder to perform 3D segmentation and observed points reconstruction; the latent feature can learn orientation information through the process. Then we propose a novel decoupled rotation mechanism for decoding orientation information. Thirdly, we use PointNet [26] to estimate the translation and object size. Finally, to increase the generalization ability of FS-Net, we propose the box-cage based 3D deformation mechanism.

Figure 3. **Stable shape and various color.** Top row: three bowl instances randomly chosen from the NOCS-REAL dataset. Bottom row: three bowl instances randomly cropped from the internet image search results (using the keyword 'bowl'). The color is varied, while the shape is relatively stable.

3.1. Object Detection

Following [6], we train a YOLOv3 [30] to fast detect the object bounding box in RGB images, and output class (category) labels. Then we adopt the 3D sphere to locate the point cloud of the target object quickly. With these techniques, the 2D detection part provides a compact 3D learning space for the following tasks. Different from other category-level 6D object pose estimation methods that need

semantic segmentation masks, we only need object bounding boxes. Since object detection is faster than semantic segmentation [30, 9], the detection speed of our method is faster than previous methods.

3.2. Shape-Based Network

The output points of object detection contain both object and background points. To access the points that belong to the target object and calculate the rotation of the object, we need a network that performs two tasks: 3D segmentation and rotation estimation.

Although there are many network architectures that directly process point cloud [26, 27, 45], most of the architectures calculate on point coordinates, which means their networks are sensitive to point clouds shift and size variation [18]. This decreases the pose estimation accuracy.

To tackle the point clouds shift, Frustum-P [25] and G2L-Net [6] employed the estimated translation to align the segmented point clouds to local coordinate space. However, their methods cannot handle the intra-class size variation.

To solve the point clouds shift and size variation problem, we propose a 3DGC autoencoder to extract the point cloud shape feature for segmentation and rotation estimation. 3DGC is designed for point cloud classification and object part segmentation; our work shows that 3DGC can also be used for category-level 6D pose estimation task.

3.2.1 3D Graph Convolution

3DGC kernel consists of m unit vectors. The m kernel vectors are applied to the n vectors generated by the center point with its n-nearest neighbors. Then, the convolution value is the sum of cosine similarity between kernel vectors and the n-nearest vectors. In a 2D convolution network, the trained network learned a weighted kernel, which has a higher response with a matched RGB value, while the 3DGC network learned the orientations of the m vectors in the kernel. The weighted 3DGC kernel has a higher response with a matched 3D pattern which is defined by the center point with its n-nearest neighbors. Please refer to [18] for more details.

3.2.2 Rotation-Aware Autoencoder

Based on the 3DGC, we design an autoencoder for the estimation of category-level object rotation. To extract the latent rotation feature, we train the autoencoder to reconstruct the observed points transformed from the observed depth map of the object. There are several advantages to this strategy: 1) the reconstruction of observed points is view-based and symmetry invariant [32, 33], 2) the reconstruction of observed points is easier than that of a complete object model (shown in Table 2), and 3) more representative orientation feature can be learned (shown in Table 1).

In [32, 33], the authors also reconstructed the input images to observed views. However, the input and output of their models are 2D images that are different from our 3D point cloud input and output. Furthermore, our network architecture is also different from theirs.

We utilize Chamfer Distance to train the autoencoder, the reconstruction loss function $\mathcal{L}_{rec}$ is defined as

$$\mathcal{L}_{rec} = \sum_{x_i \in M_c} \min_{\hat{x}_i \in \hat{M}_c} \|x_i - \hat{x}_i\|_2^2 + \sum_{\hat{x}_i \in \hat{M}_c} \min_{x_i \in M_c} \|x_i - \hat{x}_i\|_2^2, \tag{1}$$

where M_c and $\hat{M}_c$ denote the ground truth point cloud and reconstructed point cloud, respectively. x_i and $\hat{x}_i$ are the points in M_c and $\hat{M}_c$. With the help of 3D segmentation mask, we only use the features extracted from the observed object points for reconstruction.

After the network convergence, the encoder learned the rotation-aware latent feature. Since the 3DGC is scale and shift-invariant, the observed points reconstruction enforces the autoencoder to learn the scale and shift-invariant orientation feature under corresponding rotation. In the next subsection, we will describe how we decode rotation information from this latent feature.

3.3. Decoupled Rotation Estimation

Given the latent feature which contains rotation information, our task is to decode the category-level rotation feature. To achieve this, we utilize two decoders to extract the rotation information in a decoupled fashion. The two decoders decode the rotation information into two perpendicular vectors under corresponding rotation. These two vectors can represent rotation information completely (shown in Figure 4).

Since the two vectors are mutually perpendicular, the decoded rotation information related to them is independent; we can use one of them to recover part rotation information of the object. For example, in Figure 8, we use the green vector axis to recover the pose. We can see that the green boxes and blue boxes are aligned well in the recovered axis.

Each decoder only needs to extract the orientation information along corresponding vector which is easier than the estimation of the complete rotation. The loss function is based on cosine similarity and defined as

$$\mathcal{L}_{rot} = \frac{\langle \hat{\mathbf{v}}_1, \mathbf{v}_1 \rangle}{\|\hat{\mathbf{v}}_1\| \|\mathbf{v}_1\|} + \lambda_r \frac{\langle \hat{\mathbf{v}}_2, \mathbf{v}_2 \rangle}{\|\hat{\mathbf{v}}_2\| \|\mathbf{v}_2\|}, \tag{2}$$

where $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$ are the predicted vectors. $\mathbf{v}_1$ and $\mathbf{v}_2$ are the ground truth, and λ_r is the balance parameter.

The balance parameter λ_r makes our network easy to handle circular symmetry object such as bottle, and for such circular symmetry object, the red vector is not necessary (shown in Figure 4). Without loss of generality, we assume that the green vector is along the symmetry axis; then, we

set λ_r as zero to handle the circular symmetry objects. For other types of symmetric objects, we can employ the rotation mapping function used in [24, 34] to map the relevant rotation matrices to a unique one.

Please note that our decoupled rotation is different to the rotation representation proposed in [44]. They took the first two columns from a rotation matrix as the new representation, which has no geometric meaning. In contrast, our representation is defined based on the shape of the target object, and our representation can avoid the discontinuity issue mentioned in [44, 24].

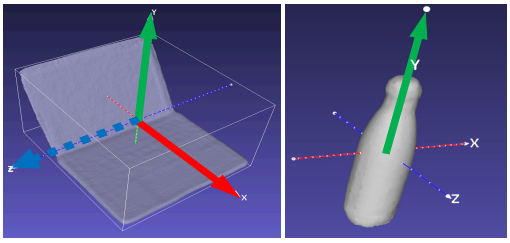

Figure 4. **Rotation represented by vectors.** Left: The object rotation can be represented by two perpendicular vectors (green vector and red vector); Right: For circular symmetry object like the bottle, only the green vector matters.

3.4. Residual Prediction Network

As both translation and object size are related to points coordinates, inspired by [25, 6], we train a tiny PointNet [26] that takes segmented point cloud as input. More concretely, the PointNet performs two related tasks: 1) estimating the residual between the translation ground truth and the mean value of the segmented point cloud; 2) estimating the residual between object size and the mean category size.

For size residual, we pre-calculate the mean size $[\overline{x}, \overline{y}, \overline{z}]^T$ of each category by

$$\begin{bmatrix} \overline{x} \\ \overline{y} \\ \overline{z} \end{bmatrix} = \frac{1}{N} \sum_{i=1}^{N} [x_i, y_i, z_i]^T, \tag{3}$$

where N is the amount of the object in that category. Then for object o in that category the ground truth $[\delta_x^o, \delta_y^o, \delta_z^o]^T$ of the size residual estimation is calculated as: $[x_o, y_o, z_o]^T - [\overline{x}, \overline{y}, \overline{z}]^T$.

We use mean square error (MSE) loss to predict both the translation and size residual. The total loss function $\mathcal{L}_{res}$ is defined as: $\mathcal{L}_{res} = \mathcal{L}_{tra} + \mathcal{L}_{size}$, where $\mathcal{L}_{tra}$ and $\mathcal{L}_{size}$ are losses for translation and size residual, respectively.

3.5. 3D Deformation Mechanism

One major problem in category-level 6D pose estimation is the intra-class shape variation. The existing methods employed two large synthetic datasets, i.e. CAMERA [40] and 3D model dataset [3] to learn this variation. However, this strategy not only needs extra hardware resources to store these big synthetic datasets but also increases the (pre-)training time.

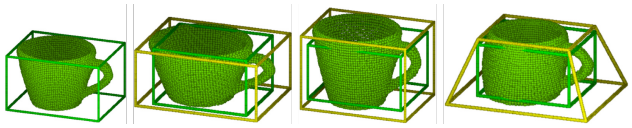

Figure 5. **3D deformed examples**. The new training examples can be generated by enlarging, shrinking, or changing the area of some surfaces of the box-cages. The left one is the original point could with original 3D box-cage, i.e. 3D bounding box. The right three ones are the deformed point clouds with deformed box-cages (shown in yellow color). The green boxes are the original 3D bounding boxes before deformation.

To alleviate the shape variation issue, based on the fact that the shapes of most objects in the same category are similar [37] (shown in Figure 3), we propose an online box-cage based 3D deformation mechanism for training data augmentation. We pre-define a box-cage for each rigid object (shown in Figure 5). Each point is assigned to its nearest surface of the cage; when we deform the surface, the corresponding points move as well.

Though box-cage can be designed more refined, in experiments, we find that with a simple box cage, i.e. 3D bounding box of the object, the generalization ability of FS-Net is considerably improved (Table 1). Different from [42], we do not require an extra training process to obtain the box-cage of the object, and we do not need the target shape to learn the deformation operation either. Our mechanism is entirely online, which saves training time and storage space.

To make the deformation operation easier, we first transfer the points to the canonical coordinate system and then perform 3D deformation. Finally we transform them to global scene:

$$\{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_n\} = R(\mathbb{F}_{3D}(R^T(\mathcal{P} - T))) + T, \tag{4}$$

where $\mathcal{P}$ is the points generated after the 2D detection step. R, T are the pose ground truth. $\{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_n\}$ are the new generated training examples. $\mathbb{F}_{3D}$ is 3D deformation which includes cage enlarging, shrinking, changing the area of some surfaces.

4. Experiments

4.1. Datasets

NOCS [40] is benchmark dataset for category-level 6D object pose estimation. It consists of real-world dataset (8K RGB-D images: 4300 for training, 950 for validation and

Table 1. **Ablation studies on NOCS-REAL dataset**. We use two different metrics to measure performance. '3DGC' means the 3D graph convolution. 'OPR' means observed points reconstruction. 'DR' represents the decoupled rotation mechanism. 'DEF' denotes the online 3D deformation. In the last row, the values in the bracket are the performance for the reconstruction of the complete object model transformed by the corresponding pose. Please note that we provide the ground truth 2D bounding box for different methods for the sake of the ablation study.

Method	3DGC	DEF	OPR	DR	IoU_{50}	$10°10$ **cm**
G2L [6]	×	✓	×	×	94.65%	31.0%
G2L+DR	×	✓	×	✓	96.21%	47.81%
Med1	✓	✓	×	×	97.98%	46.4%
Med2	✓	✓	✓	×	95.61%	46.8%
Med3	✓	✓	×	✓	97.34%	61.1%
Med4	✓	×	✓	✓	97.30%	58.2%
Med5	✓	✓	✓	✓	98.04% (94.44%)	65.9% (58.0%)

2750 for testing) and synthetic dataset (300K composited images: 25K are set for validation).

LINEMOD [12] is a widely used instance-level 6D object pose estimation dataset which consists of 13 different objects with significant shape variation.

We use the automatic point-wise labeling techniques proposed in [5] to access the label of each point in both training sets.

4.2. Implementation Details

We use Pytorch [22] to implement our pipeline. All experiments are deployed on a PC with i7-4930K 3.4GHz CPU and GTX 1080Ti GPU.

First, to locate the object in RGB images, we fine-tune the YOLOv3 pre-trained on COCO dataset [17] with the training dataset. Then we jointly train the 3DGC autoencoder and residual estimation network. The total loss function is defined as

$$\mathcal{L}_{Shape} = \lambda_{seg}\mathcal{L}_{seg}+\lambda_{rec}\mathcal{L}_{rec}+\lambda_{rot}\mathcal{L}_{rot}+\lambda_{res}\mathcal{L}_{res}, \quad (5)$$

where λs are the balance parameters. We empirically set them as 0.001, 1, 0.001, and 1 to keep different loss values at the same magnitude. We use cross entropy for 3D segmentation loss function $\mathcal{L}_{seg}$.

We adopt Adam [14] to optimize the FS-Net. The initial learning rate is 0.001, and we halve it every 10 epochs. The maximum epoch is 50.

4.3. Evaluation Metrics

For category-level pose estimation, we adopt the same metrics, 3D $IoU_{25,50,75}$ and $n°m$ **cm** used in [40, 4, 34]. For instance-level pose estimation, we compare the performance of FS-Net with other state-of-the-art instance-level methods using the ADD-(S) metric [12].

4.4. Ablation Studies

We use the G2L-Net [6] as the baseline method, which extracted the latent feature for rotation estimation via point-wise orientated vector regression, and the ground truth of rotation is the eight corners of the 3D bounding box with corresponding rotation. The loss function for rotation estimation is the mean square error between predicted 3D coordinates and ground truth. Compared to baseline, our proposed work has three novelties: a) view-based 3DGC autoencoder for observed point cloud reconstruction; b) rotation decoupled mechanism; c) online 3D deformation mechanism.

In Table 1, we report the experimental results of three novelties on the NOCS-REAL dataset. Comparing Med3 and Med5, we find that reconstruction of the observed point cloud can learn better pose feature. The performance of Med2(Med1, G2L) and Med5(Med3, G2L+DR) shows that the proposed decoupled rotation mechanism can effectively extract the rotation information. The results of Med4 and Med5 demonstrate the effectiveness of the 3D deformation mechanism, which increases the pose accuracy by 7.7% in terms of $10°10$ **cm** metric. We also compare the different reconstruction choices: the reconstruction of observed points and the complete object model with corresponding rotation. From the last row of Table 1, we can see that the observed points reconstruction can learn better rotation feature. Overall, Table 1 shows that the proposed novelties can improve the accuracy significantly.

4.5. Generalization Performance

NOCS-REAL dataset provides 4.3k real images that covers various poses of different objects in different categories for training. That means the category-level pose information is rich in the training set. Thanks to the effectively pose feature extraction, FS-Net achieves state-of-the-art performance even with part of the real-world training data. We randomly choose different percentages of the training set to train FS-Net and test it on the whole testing set. Figure 6 shows that: 1) FS-Net is robust to the size of the training dataset and has good category-level feature extraction ability. Even with 20% of the training dataset, the FS-Net can still achieve state-of-the-art performance; 2) the 3D deformation mechanism significantly improves the robustness and performance of FS-Net.

4.6. Evaluation of Reconstruction

Point cloud reconstruction has a close relationship with pose estimation performance. We computed the Chamfer Distance of the reconstructed point cloud with the ground truth point cloud and compared it with other reconstruction types used by other methods. From Table 2, we can see that the average reconstruction error of our method is 0.86, which is 72.9% and 18.9% lower than that of Shape-Prior [34] and CASS [4], respectively. It shows that our method achieves better pose estimation results via a simpler reconstruction task.

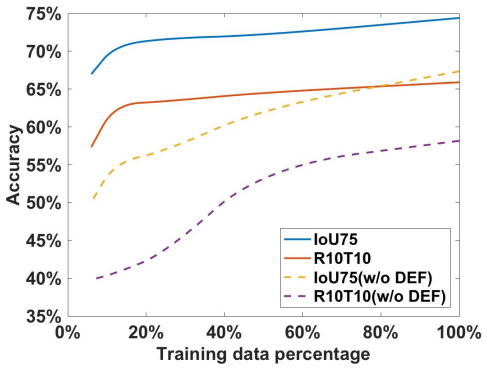

Figure 6. **Generalization performance.** With the given 2D bounding box and a randomly chosen 3D sphere center, we show how the training set size affects the pose estimation performance. 'w/o DEF' means no 3D deformation mechanism is adopted during training.

4.7. Comparison with State-of-the-Arts

4.7.1 Category-Level Pose Estimation

We compare FS-Net with NOCS [40], CASS [4], Shape-Prior [34], and 6D-PACK [38] on NOCS-REAL dataset in Table 4. It demonstrates our proposed method outperforms the other state-of-the-art methods on both accuracy and speed. Specifically, on 3D detection metric IOU_{50}, our FS-Net outperforms the previous best method, NOCS, by 11.7%, and the running speed is 4 times faster. In terms of 6D pose metric $5°$**5cm** and $10°10$ **cm**, FS-Net outperforms the CASS by the margins of 4.7% and 6.3%, respectively. FS-Net even outperforms 6D-PACK under 3D detection metric IOU_{50}, which is a 6D tracker and needs an initial 6D pose and object size to start. See Figure 7 for more quantitative details. The qualitative results are shown in Figure 8. Please note that we only use real-world data (NOCS-REAL) to train our pose estimation part. Other methods use both synthetic dataset (CAMERA) [40] and real-world data for training. The number of training examples in CAMERA is 275K, which is more than 60 times that of NOCS-REAL (4.3K). It shows that FS-Net can efficiently extract the category-level pose feature with fewer data.

4.7.2 Instance-Level Pose Estimation

We compare the instance-level pose estimation results of FS-Net on the LINEMOD dataset with other state-of-the-arts instance-level methods. From Table 3, we can see that FS-Net achieves comparable results on both accuracy and speed. It shows that our method can effectively extract both category-level and instance-level pose features.

Table 2. **Reconstruction type comparison.** The comparison is on NOCS-REAL dataset with the Chamfer Distance metric ($\times 10^{-3}$). 'Complete' means the reconstruction of the complete 3D model. 'Observed' denotes the reconstruction of the observed points.

Methods	CASS [4]	Shape-Prior [34]	Ours
	Complete	Complete	Observed
Bottle	0.75	3.44	1.2
Bowl	0.38	1.21	0.39
Camera	0.77	8.89	**0.44**
Can	0.42	1.56	0.62
Laptop	3.73	2.91	**2.23**
Mug	0.32	1.02	**0.29**
Average	1.06	3.17	**0.86**

Table 3. **Instance-level comparison on LINEMOD dataset.** Our method achieves a comparable performance with the state-of-the-art in both speed and accuracy.

Method	Input	ADD-(S)	Speed(FPS)
PVNet [23]	RGB	86.3%	25
CDPN [16]	RGB	89.9%	33
DPOD [43]	RGB	95.2%	33
G2L-Net [6]	RGBD	98.7%	23
Densefusion[39]	RGBD	94.3%	16
PVN3D [10]	RGBD	99.4%	5
Ours	RGBD	97.6%	20

4.8. Running Time

Given a 640× 480 RGB-D image, our method runs at 20 FPS with Intel i7-4930K CPU and 1080Ti GPU, which is 2 times faster than the previous fastest method 6-PACK [38]. Specifically, the 2D detection takes about 10ms to proceed. The pose and size estimation takes about 40ms.

5. Conclusion

In this paper, we propose a fast category-level pose estimation method that runs at 20 FPS which is fast enough for real-time applications. The proposed method first extracts the latent feature by the observed points reconstruction with a shape-based 3DGC autoencoder. Then the category-level orientation feature is decoded by the effective decoupled rotation mechanism. Finally, for translation and object size estimation, we use the residual network to estimate them based on residuals estimation. In addition, to increase the generalization ability of FS-Net and save the hardware source, we design an online 3D deformation mechanism for training set augmentation. Extensive experimental results demonstrate that FS-Net is robust to dataset size and can achieve state-of-the-art performance on category- and instance-level pose estimation in both accuracy and speed. As our 3D deformation mechanism and decoupled rotation scheme are model-free, they can be directly applied to other

Table 4. **Category-level performance on NOCS-REAL dataset with different metrics.** We summarize the pose estimation results reported in the origin papers on the NOCS-REAL dataset. '-' means no results are reported under this metric. The values in the bracket are the performance for synthetic NOCS dataset.

Method	IoU_{25}	IoU_{50}	IoU_{75}	5°5**cm**	10°5 **cm**	10°10 **cm**	Speed(FPS)
NOCS [40]	84.9%	80.5%	30.1%(69.5%)	9.5 %(40.9%)	26.7%	26.7%	5
CASS [4]	84.2%	77.7%	-	23.5 %	58.0%	58.3%	-
Shape-Prior [34]	83.4%	77.3%	53.2%(83.1%)	21.4%(59.0%)	54.1%	-	4
6D-PACK [38]	94.2%	-	-	**33.3 %**	-	-	10
Ours	**95.1%**	**92.2%**	**63.5%**(85.17%)	28.2 %(62.01%)	**60.8%**	**64.6%**	**20**

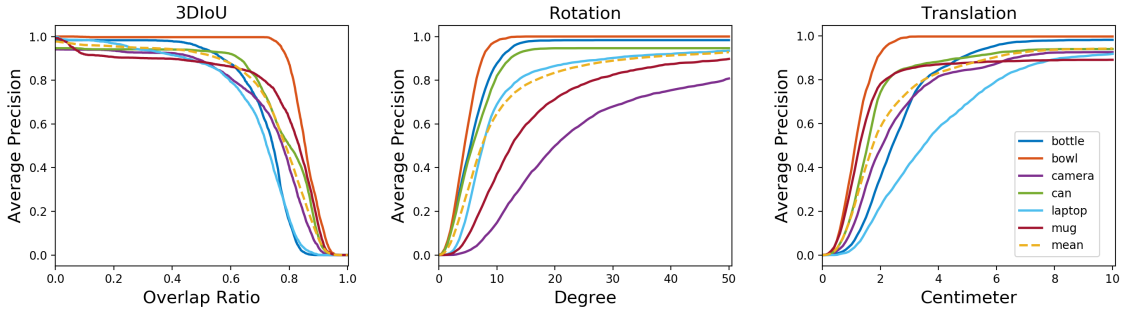

Figure 7. **Result on NOCS-REAL**. The average precision of different thresholds tested on NOCS-REAL dataset with 3D IoU, rotation, and translation error.

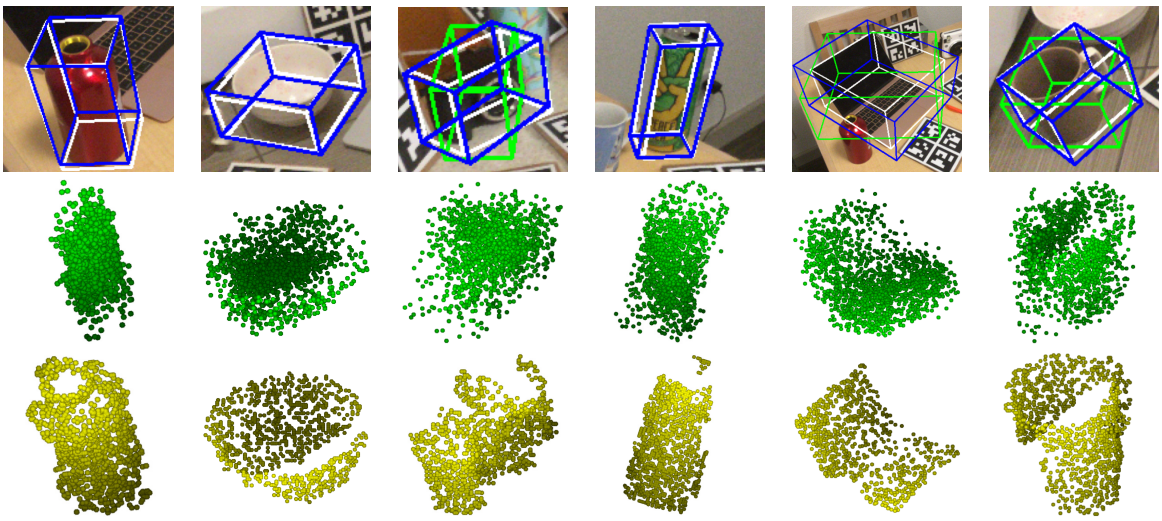

Figure 8. **Qualitative results on NOCS-REAL dataset**. The first row is the pose and size estimation results. White 3D bounding boxes denote ground truth. Blue boxes are the poses recovered from two estimated rotation vectors. The green boxes are the poses recovered from one estimated rotation vector. Our results match ground truth well in both pose and size. The second row is the reconstructed observed points under corresponding poses, although the reconstructed points are not perfectly in line with the target points, the basic orientation information is kept. The third row is the ground truth of the observed points transformed from the observed depth map.

pose estimation methods to boost the performance.

Although FS-Net achieves state-of-the-art performance, it relies on a robust 2D detector to detect the region of interest. In future work, we plan to adopt 3D object detection techniques to directly detect the objects from point clouds.

Acknowledgement

This work was in part supported by the following grants: EPSRC (EP/N019415/1) MURI, EPSRC (EP/S032487/1) CHIST-ERA, and IITP (IITP-2020-2020-0-01789) funded by the Korean government (MSIT). The authors also thank Tze Ho Elden Tse for his valuable comments.

References

[1] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3364–3372, 2016. 2

[2] Grigore C Burdea and Philippe Coiffet. *Virtual Reality Technology*. John Wiley & Sons, 2003. 1

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. In *arXiv preprint arXiv:1512.03012*, 2015. 5

[4] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning Canonical Shape Space for Category-Level 6D Object Pose and Size Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11973–11982, 2020. 1, 2, 6, 7, 8

[5] Wei Chen, Jinming Duan, Hector Basevi, Hyung Jin Chang, and Ales Leonardis. PonitPoseNet: Point Pose Network for Robust 6D Object Pose Estimation. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2, 6

[6] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, and Ales Leonardis. G2L-Net: Global to Local Network for Real-Time 6D Pose Estimation With Embedding Vector Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 4, 5, 6, 7

[7] Jaeseok Choi, Yeji Song, and Nojun Kwak. Part-Aware Data Augmentation for 3D Object Detection in Point Cloud. *arXiv preprint arXiv:2007.13373*, 2020. 3

[8] Martin A Fischler and Robert C Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of The IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 1, 4

[10] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. PVN3D: A Deep Point-wise 3D Keypoints Voting Network for 6DoF Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11632–11641, 2020. 1, 7

[11] Stefan Hinterstoisser, Cedric Cagniart, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit. Gradient Response Maps for Real-Time Detection of Textureless Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):876–888, 2012. 2

[12] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, pages 548–562. Springer, 2012. 6

[13] Wolfgang Kabsch. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Chi Li, Jin Bai, and Gregory D. Hager. A Unified Framework for Multi-View Multi-Class Object Pose Estimation. In *Proceedings of The European Conference on Computer Vision (ECCV)*, September 2018. 1

[16] Zhigang Li, Gu Wang, and Xiangyang Ji. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In *Proceeding of The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 7

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Proceeding of European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 6

[18] Zhi-Hao Lin, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Convolution in the Cloud: Learning Deformable Kernels in 3D Graph Convolution Networks for Point Cloud Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1809, 2020. 2, 4

[19] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose Estimation for Augmented Reality: A Hands-On Survey. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2633–2651, 2016. 1

[20] Eitan Marder-Eppstein. Project Tango. In *ACM SIGGRAPH 2016 Real-Time Live!*, page 40. ACM, 2016. 1

[21] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation. In *Proceedings of The European Conference on Computer Vision (ECCV)*, pages 119–134, 2018. 1, 2

[22] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. PyTorch, 2017. 6

[23] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 2, 7

[24] Giorgia Pitteri, Michaël Ramamonjisoa, Slobodan Ilic, and Vincent Lepetit. On Object Symmetries and 6D Pose Estimation from Images. In *Proceedings of International Conference on 3D Vision (3DV)*, pages 614–622. IEEE, 2019. 5

[25] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum PointNets for 3D Object Detection From RGB-D Data. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 4, 5

[26] 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 (CVPR)*, July 2017. 2, 3, 4, 5

[27] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 5099–5108, 2017. 4

[28] Mahdi Rad and Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3828–3836, 2017. 1, 2

[29] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Feature Mapping for Learning Fast and Accurate 3D Pose Inference from Synthetic Images. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4663–4672, 2018. 2

[30] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2, 3, 4

[31] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 2

[32] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. Multi-path Learning for Object Pose Estimation Across Domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13916–13925, 2020. 4

[33] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018. 4

[34] Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation. *arXiv preprint arXiv:2007.08454*, 2020. 1, 2, 5, 6, 7, 8

[35] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. *arXiv preprint arXiv:1809.10790*, 2018. 1

[36] Shinji Umeyama. Least-Squares Estimation of Transformation Parameters between Two Point Patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, (4):376–380, 1991. 1, 2

[37] Haley A Vlach. How We Categorize Objects is Related to How We Remember Them: the Shape Bias as A Memory Bias. *Journal of experimental child psychology*, 152:12–30, 2016. 2, 5

[38] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-Pack: Category-Level 6D Pose Tracker with Anchor-Based Keypoints. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066, 2020. 7, 8

[39] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 7

[40] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2019. 1, 2, 5, 6, 7, 8

[41] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1

[42] Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural Cages for Detail-Preserving 3D Deformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 75–83, 2020. 5

[43] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD: 6D Pose Object Detector and Refiner. In *Proceedings of The IEEE International Conference on Computer Vision (ICCV)*, pages 1941–1950, 2019. 7

[44] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 5

[45] Yin Zhou and Oncel Tuzel. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, 2018. 4

[46] 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 *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 3936–3943, 2014. 1