

SSP-Pose: Symmetry-Aware Shape Prior Deformation for Direct Category-Level Object Pose Estimation

Ruida Zhang^{1*}, Yan Di^{2*}, Fabian Manhardt³, Federico Tombari^{2,3}, and Xiangyang Ji⁴

Abstract—Category-level pose estimation is a challenging problem due to intra-class shape variations. Recent methods deform pre-computed shape priors to map the observed point cloud into the normalized object coordinate space and then retrieve the pose via post-processing, *i.e.*, Umeyama’s Algorithm. The shortcomings of this two-stage strategy lie in two aspects: 1) The surrogate supervision on the intermediate results can not directly guide the learning of pose, resulting in large pose error after post-processing. 2) The inference speed is limited by the post-processing step. In this paper, to handle these shortcomings, we propose an end-to-end trainable network SSP-Pose for category-level pose estimation, which integrates shape priors into a direct pose regression network. SSP-Pose stacks four individual branches on a shared feature extractor, where two branches are designed to deform and match the prior model with the observed instance, and the other two branches are applied for directly regressing the totally 9 degrees-of-freedom pose and performing symmetry reconstruction and point-wise inlier mask prediction respectively. Consistency loss terms are then naturally exploited to align the outputs of different branches and promote the performance. During inference, only the direct pose regression branch is needed. In this manner, SSP-Pose not only learns category-level pose-sensitive characteristics to boost performance but also keeps a real-time inference speed. Moreover, we utilize the symmetry information of each category to guide the shape prior deformation, and propose a novel symmetry-aware loss to mitigate the matching ambiguity. Extensive experiments on public datasets demonstrate that SSP-Pose produces superior performance compared with competitors with a real-time inference speed at about 25Hz. The codes will be released soon.

I. INTRODUCTION

Object pose estimation is a fundamental task in computer vision society, due to its wide applications in AR/VR [1], robotics [2] and autonomous driving [3], [4], [5]. Researchers have made great progress in recent years and produced several reliable and efficient methods [6], [7], [8]. However, most of them can only deal with a handful of instances and require per-object CAD models, which limits its real-world applications to some extent. Recently, category-level pose estimation is proposed to predict full configurations of rotation, translation and 3D dimensions for previously unseen instances from a known set of categories. It’s a much more challenging problem due to the lack of CAD models and the intra-class shape variations.

Wang *et al.* [10] propose the Normalized Object Coordinate Space (NOCS) and utilize a CNN network to establish

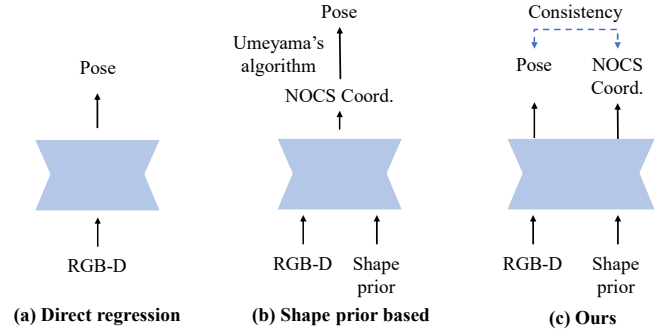


Fig. 1. **Direct regression network vs Shape prior-guided network vs Ours.** Most shape prior guided methods are computational inefficient due to post-processing optimization for pose with Umeyama’s algorithm [9], while the direct regression methods suffer from the lack of prior shape knowledge. Our method takes advantage of both methods and leverages an auxiliary shape prior adaptation branch to aid the direct pose regression branch.

correspondences between the observed point cloud and the NOCS coordinates. Finally, the object pose is recovered with Umeyama’s algorithm [9]. To tackle the intra-class variations, SPD [11] derives a point cloud based mean shape for each category and utilizes it as prior information. It deforms the shape prior to the observed instance, and then maps the observed point cloud to the deformed shape to obtain the NOCS coordinates. SGPA [12], DO-Net [13] and CR-Net [14] follow SPD and utilize shape prior deformation to guide the learning. These methods show powerful strength in extracting shape-sensitive features, which further improves the NOCS coordinate prediction and the pose estimation. However, this strategy suffers from two shortcomings. First, the accuracy of pose estimation is expected to monotonically improve with the quality of intermediate results, which is proved to be infeasible in [7]. The error on the intermediate results does not necessarily reflect the actual error of pose, especially for symmetric objects. Moreover, the inference speed is limited by the time-consuming post-processing step, *i.e.*, Umeyama’s algorithm [9].

Meanwhile, another line of methods that predict the pose directly has aroused wide research interest. FS-Net [15] leverages disentangled representation of rotation to tackle ambiguity caused by symmetry and proposes the 3D bounding box deformation based data augmentation. DualPoseNet [16] employs an explicit decoder to regress the pose directly and an implicit decoder to reconstruct the object shape. Direct methods don’t depend on the coordinate mapping process in comparison with two-stage methods, which makes them inherently faster and end-to-end trainable.

*Authors with equal contributions.

¹ Tsinghua University, zhangrd21@mails.tsinghua.edu.cn

² Technical University of Munich.

³ Google

⁴ Tsinghua University

However, they suffer from the lack of category-level prior knowledge and typically their performance is inferior to shape prior guided competitors.

In this paper, to handle these shortcomings, we propose SSP-Pose, an end-to-end trainable network for category-level pose estimation, which takes advantage of both prior-guided methods and direct regression methods for high accuracy and efficiency. Given an RGB-D image, SSP-Pose first detects the target object and then extracts point-wise features from the point cloud back-projected from the corresponding depth map of the object as well as the category shape prior. Four sub-branches are attached to the features to perform individual prediction tasks. The first two branches predict auxiliary information that deforms and associates the shape prior to the observed instance so that each observed point is connected with its corresponding point in NOCS. The third branch, following FS-Net [17], directly regresses all 9 degrees-of-freedom (DoF) pose parameters in the disentangled representation. The fourth branch performs symmetry reconstruction to jointly reconstruct the object shape utilizing symmetry prior and implicitly guide the learning of pose. Moreover, the fourth branch predicts a point-wise inlier mask to distinguish outliers and improve the robustness of our method. A geometry-guided consistency loss term is naturally introduced to align the outputs from different branches and boost the performance. Note that during inference, only the pose regression branch is needed thus the inference speed is guaranteed to be efficient. Meanwhile, we leverage the symmetric characteristics of several categories *e.g. bottle, can, laptop* to guide the learning of shape prior deformation and introduce a symmetry-aware coordinate loss to mitigate the matching ambiguity in the matching step.

In summary, our main contributions are as follows:

- We propose an end-to-end category-level pose estimation network SSP-Pose, which deforms and associates the shape prior model to the observed object to establish consistency with the directly regressed pose, boosting performance whilst keeping a fast inference speed.
- We utilize the symmetric prior of some categories, *e.g. bottle, laptop*, to guide shape prior deformation. Moreover, we propose a symmetry-aware matching loss to mitigate the ambiguity in the shape prior matching step, stabilizing and promoting the learning of category-level pose-sensitive features.
- SSP-Pose produces state-of-the-art performance on common public benchmarks, whilst achieving real-time inference speed at about 25 FPS.

II. RELATED WORKS

A. Instance-level Object Pose Estimation

In instance-level object pose estimation, the network is trained and tested on several known objects with CAD models. Recently, many efforts have been made to estimate 6DoF pose from RGB images or RGB-D data. As for RGB-only methods, end-to-end methods [18], [19], [20], [21], [22], [23], [6] regress the pose parameters directly, while

two-stage methods [24], [25], [26], [27], [28] first establish 2D-3D correspondences by predicting the 3D coordinate for each pixel or detecting pre-defined keypoints, and then utilize PnP/RANSAC algorithm to solve the pose from intermediate results. Another line of methods [29], [7], [30], [8] jointly take advantage of direct pose regression and correspondence-based pose recovery to boost performance. As for methods based on RGB-D data, FFB6D [31] and DenseFusion [32] combine appearance and geometry information by iterative dense feature fusion, while PVN3D [33] proposes a 3D voting scheme to detect keypoints. Instance-level methods are limited in practice since they can deal with merely a handful of objects and require precise CAD models.

B. Category-level Object Pose Estimation

Category-level pose estimation aims to predict the 6DoF pose along with the 3D metric size of previously unseen objects from a known set of categories. Recently, Wang *et al.* [10] proposes the *Normalized Object Coordinate Space* (NOCS). They represent the observed point cloud with NOCS and recover the pose by the Umeyama's algorithm [9]. CASS [34] learns a canonical shape space by VAE [35] to obtain a view-factorized RGB-D embedding. CenterSnap [36] presents a one-stage pipeline to reduce the computational cost. 6-PACK [37] recovers the pose by tracking inter-frame motion of the object. CAPTRA [38] tracks the object pose by canonicalizing the observed point cloud using the previous predictions, while BundleTrack [39] employs pose graph optimization to enforce spatiotemporal consistency.

To tackle the intra-class shape variation, which is considered as the main bottleneck in category-level pose estimation [34], [11], shape priors are widely utilized [11], [12], [13]. Most of them first deform the shape prior to the observed instance and match the observed point cloud with the deformed prior model to retrieve the NOCS coordinates. Then Umeyama's algorithm is adopted to recover the pose from the NOCS coordinates. Although they have achieved large improvements on several benchmarks, the performance is still far from excellent.

Meanwhile, other methods directly regress the pose parameters. FS-Net [15] proposes a decoupled mechanism for rotation, while DualPoseNet [16] employs an explicit decoder to regress the pose directly and an implicit decoder to reconstruct the object shape. It refines the pose estimation by enforcing consistency between the two decoders. However, they are unable to incorporate prior information. In this paper, we integrate shape priors into the direct regression network, which ensures efficient and accurate inference.

III. SSP-POSE

A schematic overview of SSP-Pose is presented in Fig 2. We first segment the region of objects of interest using an off-the-shelf object detector (*e.g.* MaskRCNN [40]). We randomly sample 1024 points from the back-projected point cloud and feed them into our network. A 3D graph convolutional network (3DGC [41]) based feature extractor is applied to extract point-wise features from the observed point cloud

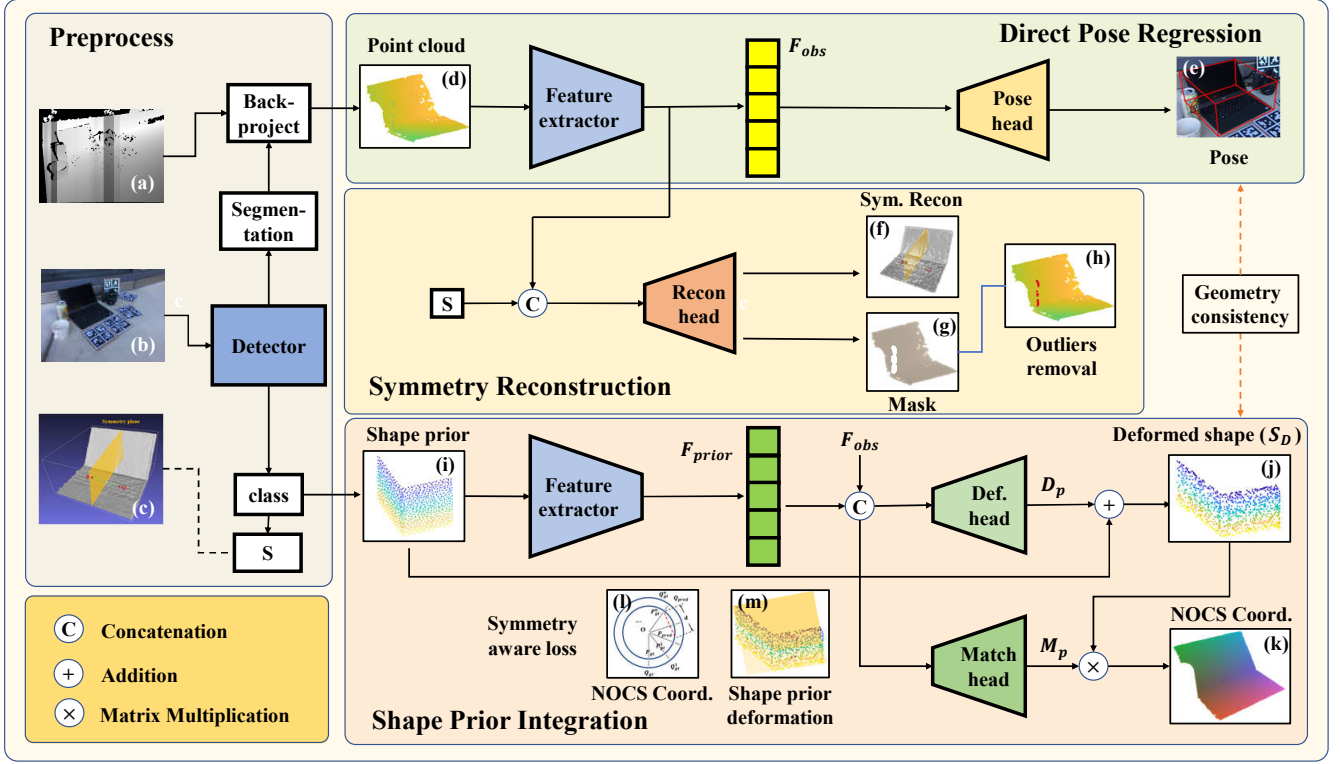


Fig. 2. **Overview of SSP-Pose.** We first use an off-the-shelf object detector (e.g. MaskRCNN [40]) to segment out the object of interest (b) and generate the back-projected point cloud from the depth map (a). Then we adopt 3DGCN[41] as the encoder to extract features of the observed point cloud \mathcal{F}_{obs} . The pose regression head takes \mathcal{F}_{obs} as input and predicts the pose R, t, s (e). We feed \mathcal{F}_{obs} and the symmetry type S (c) of the category into the reconstruction head (Recon. head). It performs symmetry reconstruction (f) and predicts point-wise mask for outlier removal (g, h). Meanwhile, we use the same encoder to extract the feature of the mean shape (i) \mathcal{F}_{prior} . The deformation head (Def. head) and matching head take both \mathcal{F}_{obs} and \mathcal{F}_{prior} as input. The deformation head deforms the mean shape to the observed instance (j), and the matching head matches each observed point with the deformed model and obtains the NOCS coordinates (k). We design a geometry-guided consistency term between the predicted pose and the NOCS coordinates to boost performance. Moreover, we propose a symmetry-aware loss to eliminate ambiguity for coordinate prediction and shape deformation (l, m).

and shape prior respectively. Then four parallel branches are attached for independent tasks. The deformation head and matching head are designed to deform and associate the prior model with the observed instance. The pose regression head is adopted for direct regression of the pose parameters. And the reconstruction head performs symmetry reconstruction as well as point-wise inlier mask prediction. Geometry-guided consistency terms among these branches are finally enforced to align the outputs and boost the performance.

A. Network Architecture

Feature Extractor. Inspired by FS-Net [15], we choose 3DGC [41], which is insensitive to shift and scale, to extract point-wise features. Specifically, we adopt the same 3DGC network to extract shape-sensitive features \mathcal{F}_{obs} , \mathcal{F}_{prior} from both the observed point cloud P_{obs} and the shape prior P_{prior} , allowing for shape prior guided shape reconstruction and pose estimation.

Pose regression head. Previous methods typically harness RANSAC-based Umeyama’s algorithm [9] to filter outliers and recover the pose. However, RANSAC is vulnerable to systematic erroneous point matches, e.g. the mistakenly established point matches caused by symmetry, which leads to performance deterioration. Thus we design an end-to-

end pose regression head and adopt three other auxiliary heads for cross-task consistency. We feed \mathcal{F}_{obs} to our pose head and regress the 9DoF pose directly, ensuring fast inference speed and robust pose estimation. We follow FS-Net [15] to parameterize the translation and size with the residual representation, and parameterize the rotation in a geometry-aware manner. Specifically, for translation t , given the predicted residual translation t_r and the mean of the observed point cloud P_M , we recover $t = t_r + P_M$. Similarly, given the estimated residual size s_r and the pre-computed category mean size S_M , we obtain the scale $s = s_r + S_M$, where $s = \{s_x, s_y, s_z\}$. For rotation R , we predict two rotation vectors along x and y axis r_x, r_y respectively, which are also the first two columns of the rotation matrix. For rotational symmetric objects which are symmetric around an axis, r_x contains ambiguity and only r_y is used to solve the rotation information, as shown in Fig. 4.

Reconstruction head. We adopt a reconstruction head for symmetry-aware reconstruction and outlier removal, which are introduced in Sec. III-D, III-C respectively. The reconstruction head takes \mathcal{F}_{obs} and the category-level symmetry type S as input, and outputs the mirrored input points (Sec. III-D) and the point-wise inlier mask (Sec. III-C).

Deformation and matching heads. These two heads take

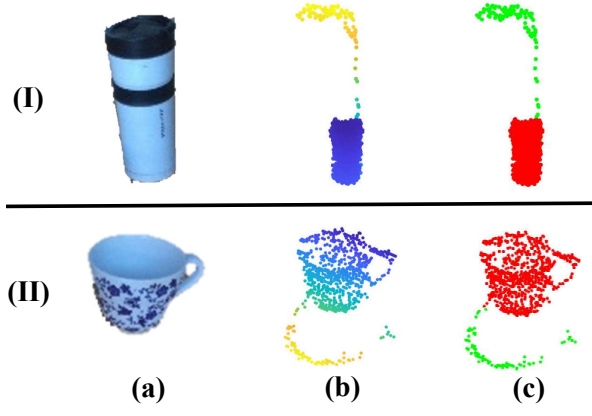


Fig. 3. **Illustration of outlier removal.** (I), (II) are two examples for outlier removal. Using the segmentation results from RGB images (a), we back-project the depth map to retrieve point cloud (b). Due to the limitation of depth sensor, the point cloud contains a non-negligible number of outliers. We compute the ground truth location p_{gt} for each point p using the ground truth NOCS coordinate (red points in (c)) and compare p_{gt} with the observed location p_{obs} to detect outliers (green points in (c)).

both F_{obs} and F_{prior} as input, and output the deformation field D_p and the matching matrix M_p respectively. D_p deforms the shape prior model to the observed instance, and M_p matches each observed point with the points in the deformed model. Similar to SPD [11], the NOCS coordinate C of the target object is estimated as follows,

$$C = M_p(P_{prior} + D_p) \quad (1)$$

B. Geometry-Guided Consistency

We exploit the underlying geometry consistency between the predicted NOCS coordinate C and the predicted pose, and transform it into a loss term to align both outputs and further improve the performance. Given the predicted pose R, t, s , the NOCS coordinate of the point p should be

$$c_p = R^T(p - t)/L \quad (2)$$

where L is the diagonal length of the bounding box.

Thus our consistency loss term is defined as follows,

$$\mathcal{L}_{con} = |c_p - \bar{c}| \quad (3)$$

where \bar{c} is the predicted NOCS coordinate.

C. Outlier Removal

We use an off-the-shelf object detector (e.g. MaskRCNN [40]) to segment the object of interest and back-project the corresponding depth map to obtain the point cloud as the input for the network. Point cloud generated in this way contains a non-negligible number of outliers (as shown in Fig. 3 (b)), which may mislead the network in pose estimation. In order to enhance the robustness of the training process, an outlier removal process is needed.

In Fig. 3, we demonstrate our outlier removal process. We use the ground truth NOCS coordinate and pose to calculate the ground truth location p_{gt} for each observed point p (red points in Fig. 3 (c)).

$$p_{gt} = R(Lc) + t \quad (4)$$

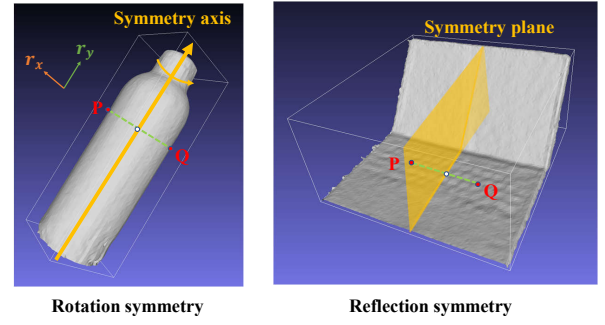


Fig. 4. **Demonstration of two types of symmetry.** Objects with **rotational symmetry** is symmetric along an axis, while objects with **reflection symmetry** is symmetric w.r.t. a plane. Point Q is the mirrored point of point P .

where $L = \sqrt{s_x^2 + s_y^2 + s_z^2}$ stands for the diagonal length of the object bounding box. We detect the outliers by comparing p_{gt} with p (green points in Fig. 3 (c)). Specifically, the point is considered to be an outlier if $|p - p_{gt}| > \lambda_{pt}$, where λ_{pt} is the distance threshold. We set $\lambda_{pt} = 0.1$ in this paper.

The reconstruction head predicts a point-wise mask indicating whether a point is an inlier, which guides the network to identify outliers and improves the robustness of our method. Please note that we do not adopt outlier removal as a pre-processing step like in FS-Net [15]. The advantages lie in two points. First, we enforce the network to distinguish outliers out of the point cloud itself, enabling robust performance when the outlier removal process fails in inference. Second, only the simple direct pose regression network is needed during inference, guaranteeing fast running speed.

D. Symmetry Analysis

We focus on two common types of symmetry: **Reflection Symmetry** and **Rotational Symmetry** [13], which is demonstrated in Fig. 4. Objects with rotational symmetry is symmetric along an axis, while objects with reflection symmetry is symmetric w.r.t. a plane. To simplify the illustration, we assume that the symmetry plane for reflection symmetry is on the xy -plane, the symmetry axis is y axis for rotational symmetry. For arbitrary category, we can adjust the canonical pose to guarantee the assumption.

Symmetry reconstruction. Previous methods ([15], [34]) adopt the point cloud reconstruction as an auxiliary task to improve the generalization ability of the network. Inspired by DO-Net [13], we reconstruct the mirrored point for each observed point according to the symmetry type of the category (Figure 4). For category with no symmetry, e.g. camera, we directly reconstruct the input point cloud. For category with rotational symmetry, e.g. bowl, bottle, can, the mirrored point is the corresponding point of the observed point w.r.t. the symmetry axis. For category with only reflection symmetry, e.g. laptop, the mirrored point can be obtained by flipping w.r.t. the reflection plane. In summary, for a point p , its corresponding mirrored point

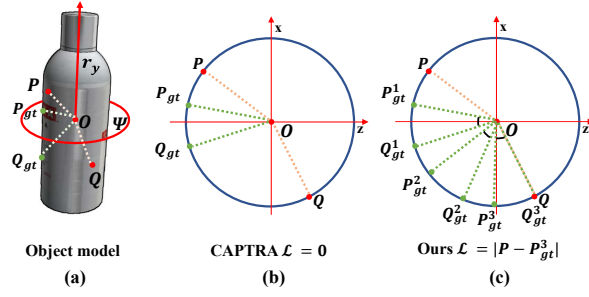


Fig. 5. **Demonstration of symmetry-aware coordinate loss.** We take the symmetric category, *i.e.*, *bottle* as an example (a). Since the y axis in the object coordinate coincides with the symmetry axis r_y , we only need to consider the x, z axis. Thus we project the predicted NOCS coordinates on the plane Ψ to simplify the illustration. In (b) and (c), P and Q are the predicted NOCS coordinates of two observed points and P_{gt}, Q_{gt} are the corresponding ground truth. O denotes the projection of r_y . Since CAPTRA[38] only considers the radius distance, as shown in (b), although the loss is zero, the predicted NOCS coordinates are still not fully constrained. Unlike CAPTRA, in (c), we first generate a set of candidate rotation matrix $R_{gt}^i, i = 1, \dots, n$ by uniformly sampling on the circle. Then for each candidate pose $\{R_{gt}^i, t_{gt}\}$, we generate the candidate NOCS coordinates N_c for all observed points, and compute the loss as in Eq. 8. This loss alleviates the ambiguity caused by symmetry and guides the network to learn symmetry-aware geometric characteristics of each category.

$\mathcal{F}_{mir}(p)$ under the canonical view is defined as follows,

$$\mathcal{F}_{mir}(p) = \begin{cases} [p_x, p_y, p_z] & \text{no symmetry} \\ [-p_x, p_y, -p_z] & \text{rotational symmetry} \\ [p_x, p_y, -p_z] & \text{reflection symmetry} \end{cases} \quad (5)$$

The loss is defined as

$$\mathcal{L}_{recon} = \text{avg}_{p \in \mathcal{P}} |R(\mathcal{F}_{mir}(R^{-1}(p - t))) + t - \bar{p}| \quad (6)$$

where \mathcal{P} is the input point cloud and \bar{p} is the predicted mirrored point of p . In this way, the network not only reconstructs the observed point cloud, but also implicitly learns the symmetry-aware geometric characteristics for pose estimation.

Symmetry-aware shape deformation loss. We propose a symmetry-aware shape deformation loss \mathcal{L}_{sym}^{SP} in order to harness symmetry as prior information for shape deformation (Fig. 2 (j)). Since the category symmetry type remains unchanged after deformation, the mirrored points should also lie on the surface of the object. We employ the Chamfer Distance between the mirrored point cloud \mathcal{S}_M and the deformed shape \mathcal{S}_D as the loss function. Specifically, the symmetry-aware shape recovery loss is defined as follows

$$\mathcal{L}_{sym}^{SP} = \text{avg}_{p \in \mathcal{S}_D} \min_{q \in \mathcal{S}_M} |p - q| + \text{avg}_{q \in \mathcal{S}_M} \min_{p \in \mathcal{S}_D} |p - q| \quad (7)$$

Symmetry-aware coordinate loss. The pose of objects with rotational symmetry (or with more than one reflection planes) is ambiguous since multiple candidate poses can be obtained by rotating it around the symmetry axis. Therefore, the pose-dependent NOCS coordinates (Fig. 2 (k)) also contain ambiguity. Forcing the network to predict the coordinates of symmetric objects under a specific pose will confuse the network. To address this issue, CAPTRA[38] utilizes

$\sqrt{c_x^2 + c_z^2}$ for supervision, where NOCS coordinate $c = \{c_x, c_y, c_z\}$. Since the y axis coincides with the symmetric axis r_y , c_y can be omitted. However, such a supervision manner may fail in several cases. In Fig 5, we provide an example to explain where the technique fails in detail.

In order to tackle this problem, we propose a symmetry-aware NOCS coordinates loss, which alleviates the ambiguity caused by symmetry and guides the predicted coordinates to correspond to a unique pose. We first uniformly sample a set of candidate rotation matrix $R_{gt}^i, i = 1, \dots, n$ according to the symmetry type, as shown in Fig. 5 (c). Then for each candidate pose $\{R_{gt}^i, t_{gt}\}$, we generate the candidate NOCS coordinates N_c for all points, and compute the loss as the minimum L_1 distance between the predicted NOCS coordinates C and all the candidates in N_c ,

$$\mathcal{L}_{coord}^{SP} = \min_{C^* \in N_c} \frac{1}{|C|} \sum_{p \in C, q \in C^*} |p - q| \quad (8)$$

where $|C|$ denotes the point number in C .

E. Overall training objective

The overall training objective is defined as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{pose} + \lambda_2 \mathcal{L}_{SP} + \lambda_3 \mathcal{L}_{mask} + \lambda_4 \mathcal{L}_{recon} + \lambda_5 \mathcal{L}_{con} \quad (9)$$

For \mathcal{L}_{pose} , we follow FS-Net [15] to supervise R, t and s with the ground truth. We replace the L2 loss with L1 loss for faster convergence. \mathcal{L}_{SP} includes the loss terms introduced in Equation 7, 8, along with loss terms for D_p and M_p from SPD [11]. We use L1 loss function for \mathcal{L}_{mask} to supervise the point-wise mask prediction. \mathcal{L}_{recon} and \mathcal{L}_{con} is introduced in Eq. 3 and 6. Please note that we only calculate the loss terms \mathcal{L}_{SP} , \mathcal{L}_{con} and \mathcal{L}_{recon} on the inliers.

IV. EXPERIMENTS

Datasets. We use NOCS-REAL275 and NOCS-CAMERA25 [10] to evaluate SSP-Pose. NOCS-REAL275 is a real-world dataset containing 4.3k images captured from 7 scenes for training and 2.75k images from 6 scenes for testing. It covers objects of 6 known categories, including *bottle*, *bowl*, *camera*, *can*, *laptop* and *mug*. CAMERA25 is a synthetic dataset which is generated by rendering virtual object on the real background and shares the same categories with NOCS-REAL275. It provides 275k images for training and 25k for testing.

Implementation Details. We follow [11], [12] to generate segmentation masks with an off-the-shelf object detector, *i.e.*, Mask-RCNN [40] for fair comparison. We employ data augmentation during training including random uniform noise and random rotational and translational perturbations similar to [15]. All the experiments are conducted on a single NVIDIA A100 GPU and a AMD EPYC 7H12 64-Core CPU. The hyper parameters for all loss terms are kept unchanged during experimentation unless specified, with $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\} = \{8.0, 10.0, 1.0, 1.0, 1.0\}$. We adopt the Ranger optimizer [42], [43], [44] and set the batch size to be 32 and the base learning rate to be 1e-4. The learning

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON NOCS-REAL275 DATASET.

Method	Direct	Prior	IoU_{25}	IoU_{50}	IoU_{75}	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 5cm$	$10^\circ 10cm$	Speed(FPS)
NOCS [10]			84.9	80.5	30.1	7.2	10.0	25.2	26.7	5
CASS [34]			84.2	77.7	-	-	23.5	58.0	58.3	-
SPD [11]		✓	83.4	77.3	53.2	19.3	21.4	54.1	-	4
CR-Net [14]		✓	-	79.3	55.9	27.8	34.3	60.8	-	-
SGPA [12]		✓	-	80.1	61.9	35.9	39.6	70.7	-	-
DO-Net [13]		✓	-	80.4	63.7	24.1	34.8	67.4	-	10
DualPoseNet [16]	✓		-	79.8	62.2	29.3	35.9	66.8	-	2
FS-Net [15]	✓		-	-	-	-	28.2	60.8	64.6	20
FS-Net(Ours)	✓		84.0	81.1	52.0	19.9	33.9	69.1	71.0	20
Ours	✓	✓	84.0	82.3	66.3	34.7	44.6	77.8	79.7	25

Overall best results are in bold. We reimplement FS-Net as **FS-Net(Ours)** for fair comparison because FS-Net uses different detection results.

TABLE II
ABLATION STUDIES ON NOCS-REAL275 DATASETS.

	Direct	Prior	Sym. \mathcal{L}	Sym. Recon.	Removal	IoU_{50}	IoU_{75}	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 5cm$	$10^\circ 10cm$
A1	✓					82.9	62.3	26.9	34.6	72.7	74.3
A2		✓				81.1	48.3	10.6	19.8	55.9	59.3
A3	✓	✓				83.3	63.7	27.5	38.3	72.0	73.6
B1		✓	✓			83.2	58.7	30.5	41.0	71.2	73.3
B2	✓	✓	✓			78.3	64.0	31.1	40.5	74.7	76.6
C	✓	✓	✓	✓		83.1	66.0	33.3	42.1	76.1	78.0
D	✓	✓	✓	✓	✓	82.3	66.3	34.7	44.6	77.8	79.7

Direct means using direct regression for pose, otherwise the pose is solved from predicted NOCS coordinates by Umeyama's algorithm. **Prior** means using shape priors as input and performing shape prior deformation. **Sym. \mathcal{L}** includes the symmetry-aware loss terms for NOCS coordinate prediction and shape deformation, otherwise the traditional loss terms are used. **Sym. Recon.** stands for symmetry reconstruction and **Removal** is the outlier removal process.

rate is annealed at 72% of the training phase using a cosine schedule. We only use the real data during training in the evaluation of dataset NOCS-REAL275. We train a single model for all categories. The total training epoch is 150.

Evaluation Metrics. We follow [10], [12], [16] to adopt 3D IOU that reports the mean precision of 3D intersection over union (IoU) at thresholds of 25%, 50%, 75% to jointly evaluate rotation, translation and size. To directly compare errors in rotation and translation, we also adopt the $5^\circ 2cm$, $5^\circ 5cm$, $10^\circ 5cm$, $10^\circ 10cm$ metrics. A pose is thereby considered correct if the translation and rotation errors are both below the given thresholds.

A. Comparison with State-of-the-Art Methods

In table I, we compare SSP-Pose with state-of-the-art methods on NOCS-REAL275 dataset. Our method is superior to other methods under IoU_{50} , IoU_{75} , $5^\circ 5cm$, $10^\circ 5cm$, $10^\circ 10cm$ metrics, including all shape prior based methods and direct regression methods. In term of IoU_{75} , $5^\circ 5cm$ and $10^\circ 5cm$, our method surpasses the second best method by a large margin. Specifically, we achieve 82.5% mAP under IoU_{50} , 64.2% mAP under IoU_{75} , 44.5% mAP under $5^\circ 5cm$ and 75.7% mAP under $10^\circ 5cm$ respectively, where our performance is 2.4%, 2.3%, 4.9% and 5.0% better than the best shape prior based method SGPA[12], and 2.1%, 2.0%, 8.6% and 8.9% better than the best direct regression method DualPoseNet [16].

In table III, we compare our method with state-of-the-art methods On NOCS-CAMERA25 dataset. In IoU_{75} , $5^\circ 5cm$ and $10^\circ 5cm$ metrics, our result is superior to DualPoseNet

and comparable to CR-Net and SGPA. Although CR-Net [14] achieves better performance under $5^\circ 2cm$ on the synthetic dataset CAMERA25, its accuracy on real dataset REAL275 is inferior in comparison with ours. Similarly, the margin in $5^\circ 2cm$ between SPGA and ours is much smaller in the real world dataset. The reason is that the noise in synthetic depth map is relatively small and does not contain outliers as discussed in Sec. III-C, which does not reflect the real-world situation.

Moreover, our method is superior in running speed since only the pose regression head is needed during inference, as shown in the last column of Tab. I. Regardless of the object detection time, the 9DoF pose estimation process can run at >50 FPS, which enables real-world applications with time requirements. The framerate of our method reaches 25 FPS when using YOLOv3 [45] and ATSA [46] to segment the objects of interest.

We provide a detailed comparison with DualPoseNet [16] on REAL275 dataset in Fig 6. Our method produces significantly superior results in the accuracy of rotation and 3D IoU, especially for the category *camera*. This is also verified by the qualitative comparison in Fig 7. With the guidance of shape priors, SSP-Pose consistently outperforms DualPoseNet in handling complex objects such as *camera*

B. Ablation Studies

Effect of shape prior deformation and direct regression. In Tab. II (A) we conduct evaluation of the effect of shape prior deformation and direct regression. As can be observed, combining direct regression and shape prior de-

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON
NOCS-CAMERA25 DATASET.

Method	IoU_{75}	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 5cm$
NOCS [10]	69.5	32.3	40.9	64.6
DualPoseNet [16]	86.4	64.7	77.2	84.7
SPD [11]	83.1	54.3	59.0	81.5
CR-Net [14]	88.0	72.0	76.4	87.7
SGPA [12]	88.1	70.7	74.5	88.4
Ours	86.8	64.7	75.5	87.4

Overall best results are in bold.

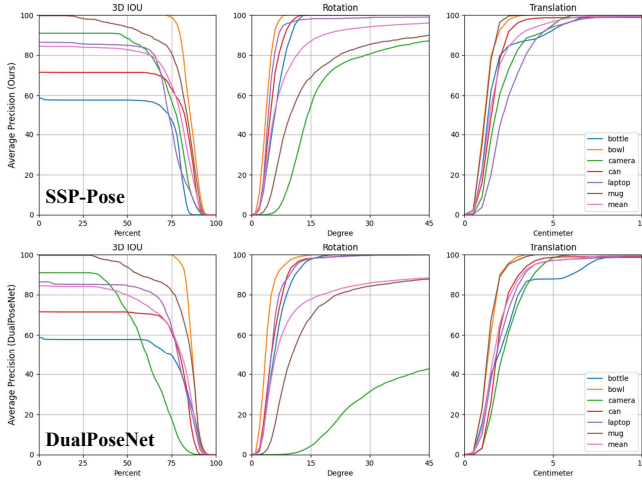


Fig. 6. Quantitative comparison of SSP-Pose and DualPoseNet[16]. We illustrate the precision under different error thresholds on NOCS-REAL275, including 3D IoU, rotation and translation.

formation (A3) is superior to solely using each part (A1,A2) in all metrics. Specifically, comparing A1 and A3, integrating category-level prior information and shape prior deformation process improves accuracy of IoU_{75} and $5^\circ 5cm$ by 1.4% and 3.7% respectively. From A2 to A3, by using direct regression for pose estimation instead of Umeyama’s algorithm, the accuracy of IoU_{75} , $5^\circ 2cm$, $5^\circ 5cm$ and $10^\circ 5cm$ improves more than 15%, which indicates the effectiveness and robustness of direct pose regression.

Effect of symmetry-aware loss. We evaluate the effect of the symmetry-aware loss in Tab. II (B). By comparing B2 and A3, the mAP under $5^\circ 2cm$ and $5^\circ 5cm$ improves 3.6% and 2.2% respectively, which proves that mitigating ambiguity caused by symmetry and utilizing symmetry prior bring significant improvement to pose estimation. Interestingly, when recovering the pose using the Umeyama’s algorithm (B1 and A2), using the symmetry-aware loss improves accuracy of IoU_{75} and $5^\circ 5cm$ by 10.4% and 19.9%, indicating that our symmetry-aware coordinate loss boosts the accuracy of coordinate prediction by a large margin.

Effect of symmetry reconstruction. In Tab. II (C), we add symmetry reconstruction process to B1. The symmetry reconstruction improves IoU_{75} , $5^\circ 2cm$ and $5^\circ 5cm$ by 2.0%, 2.2% and 1.6% respectively. The main reason is that symmetry reconstruction guides the network to learn the symmetry-

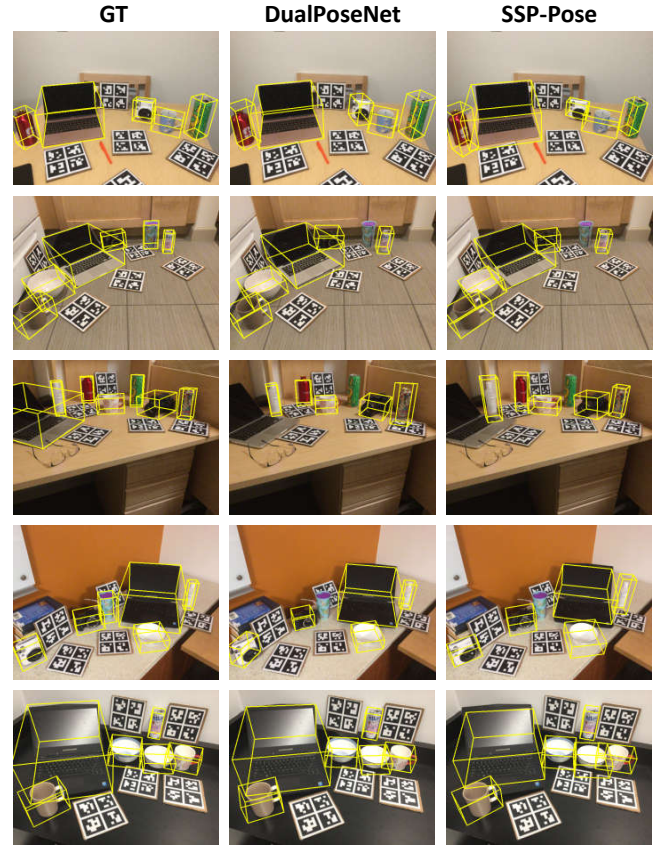


Fig. 7. Qualitative comparison of SSP-Pose and DualPoseNet[16].

aware geometric characteristics of each category for pose estimation.

Effect of outlier removal. We verify the effectiveness of the outlier removal step in comparison of C and D. By only supervising point-wise loss terms on inliers and predicting point-wise mask, the accuracy of $5^\circ 2cm$ and $5^\circ 5cm$ improves by 1.4%, 2.5%, which illustrates the effectiveness of outlier removal technique. Notably, this technique could be easily adopted for other methods in a plug-and-play manner.

V. CONCLUSION

We propose an end-to-end category-level pose estimation network SSP-Pose, which deforms and associates the shape prior model to the observed object to establish consistency with the directly regressed pose, boosting performance whilst keeping a fast inference speed. We utilize the category-level symmetry prior to guide the learning of shape prior deformation and pose regression. Extensive experiments on public datasets demonstrate that SSP-Pose surpasses all competitors whilst keeping a real-time inference speed at 25Hz. In the future, we plan to extend SSP-Pose to more complex scenes, e.g. traffic scenes and robotic manipulation.

REFERENCES

- [1] Y. Su, J. Rambach, N. Minaskan, P. Lesur, A. Pagani, and D. Stricker, “Deep multi-state object pose estimation for augmented reality assembly,” in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2019, pp. 222–227.

- [2] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, "Self-supervised 6d object pose estimation for robot manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3665–3671.
- [3] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [4] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *arXiv preprint arXiv:2111.14382*, 2021.
- [5] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, "Geometry uncertainty projection network for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3111–3121.
- [6] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 574–591.
- [7] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *CVPR*, June 2021.
- [8] Y. Di, F. Manhardt, G. Wang, X. Ji, N. Navab, and F. Tombari, "So-pose: Exploiting self-occlusion for direct 6d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 396–12 405.
- [9] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 376–380, 1991.
- [10] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.
- [11] M. Tian, M. H. Ang, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 530–546.
- [12] K. Chen and Q. Dou, "Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2773–2782.
- [13] H. Lin, Z. Liu, C. Cheang, L. Zhang, Y. Fu, and X. Xue, "Donet: Learning category-level 6d object pose and size estimation from depth observation," *arXiv preprint arXiv:2106.14193*, 2021.
- [14] J. Wang, K. Chen, and Q. Dou, "Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.
- [15] W. Chen, X. Jia, H. J. Chang, J. Duan, S. Linlin, and A. Leonardis, "Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 1581–1590.
- [16] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency," *arXiv preprint arXiv:2103.06526*, 2021.
- [17] D. Yan, Z. Ruida, L. Zhiqiang, M. Fabian, N. Nassir, J. Xiangyang, and F. Tombari, "Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting," in *arXiv preprint*, 2022.
- [18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.
- [19] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.
- [20] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," *RSS*, 2018.
- [21] F. Manhardt, W. Kehl, N. Navab, and F. Tombari, "Deep model-based 6d pose refinement in rgb," in *ECCV*, 2018.
- [22] F. Manhardt, D. M. Arroyo, C. Rupprecht, B. Busam, T. Birdal, N. Navab, and F. Tombari, "Explaining the ambiguity of object detection and 6d pose from visual data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6841–6850.
- [23] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep iterative matching for 6d pose estimation," *IJCV*, pp. 1–22, 2019.
- [24] C. Song, J. Song, and Q. Huang, "Hybridpose: 6d object pose estimation under hybrid representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 431–440.
- [25] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *CVPR*, 2019.
- [26] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: Dense 6d pose object detector in rgb images," in *ICCV*, 2019.
- [27] T. Hodan, D. Barath, and J. Matas, "Epos: Estimating 6d pose of objects with symmetries," in *CVPR*, 2020, pp. 11 703–11 712.
- [28] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation," in *ICCV*, 2019.
- [29] Z. Li, G. Wang, and X. Ji, "CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation," in *ICCV*, 2019, pp. 7678–7687.
- [30] Y. Hu, P. Fua, W. Wang, and M. Salzmann, "Single-stage 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2930–2939.
- [31] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 3003–3013.
- [32] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.
- [33] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. 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*, 2020, pp. 11 632–11 641.
- [34] D. Chen, J. Li, Z. Wang, and K. 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*, 2020, pp. 11 973–11 982.
- [35] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [36] M. Z. Irshad, T. Kollar, M. Laskey, K. Stone, and Z. Kira, "Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation," 2022.
- [37] C. Wang, R. Martín-Martín, D. Xu, J. Lv, C. Lu, L. Fei-Fei, S. Savarese, and Y. Zhu, "6-pack: Category-level 6d pose tracker with anchor-based keypoints," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 059–10 066.
- [38] Y. Weng, H. Wang, Q. Zhou, Y. Qin, Y. Duan, Q. Fan, B. Chen, H. Su, and L. J. Guibas, "Captra: Category-level pose tracking for rigid and articulated objects from point clouds," *arXiv preprint arXiv:2104.03437*, 2021.
- [39] B. Wen and K. E. Bekris, "Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.
- [40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [41] Z.-H. Lin, S.-Y. Huang, and Y.-C. F. 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*, 2020, pp. 1800–1809.
- [42] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *International Conference on Learning Representations*, 2019.
- [43] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, "Lookahead optimizer: k steps forward, 1 step back," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.
- [44] H. Yong, J. Huang, X. Hua, and L. Zhang, "Gradient centralization: A new optimization technique for deep neural networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 635–652.
- [45] B. Tekin, S. N. Sinha, and P. Fua, "Real-Time Seamless Single Shot 6D Object Pose Prediction," in *CVPR*, 2018, pp. 292–301.
- [46] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate rgb-d saliency detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 374–390.