

RBP-Pose: Residual Bounding Box Projection for Category-Level Pose Estimation

Ruida Zhang^{1*}, Yan Di^{2*}, Zhiqiang Lou¹, Fabian Manhardt³, Federico Tombari^{2,3}, and Xiangyang Ji¹

¹ Tsinghua University

² Technical University of Munich

³ Google

{zhangrd21@mails. lzq20@mails. xyji@}tsinghua.edu.cn
shangbuhuan13@gmail.com, fabianmanhardt@google.com, tombari@in.tum.de

Abstract. Category-level object pose estimation aims to predict the 6D pose as well as the 3D metric size of arbitrary objects from a known set of categories. Recent methods harness shape prior adaptation to map the observed point cloud into the canonical space and apply Umeyama algorithm to recover the pose and size. However, their shape prior integration strategy boosts pose estimation indirectly, which leads to insufficient pose-sensitive feature extraction and slow inference speed. To tackle this problem, in this paper, we propose a novel geometry-guided **Residual Object Bounding Box Projection** network **RBP-Pose** that jointly predicts object pose and residual vectors describing the displacements from the shape-prior-indicated object surface projections on the bounding box towards the real surface projections. Such definition of residual vectors is inherently zero-mean and relatively small, and explicitly encapsulates spatial cues of the 3D object for robust and accurate pose regression. We enforce geometry-aware consistency terms to align the predicted pose and residual vectors to further boost performance. Finally, to avoid overfitting and enhance the generalization ability of RBP-Pose, we propose an online non-linear shape augmentation scheme to promote shape diversity during training. Extensive experiments on NOCS datasets demonstrate that RBP-Pose surpasses all existing methods by a large margin, whilst achieving a real-time inference speed.

Keywords: Category-Level Pose Estimation, 3D Object Detection, Scene Understanding

1 Introduction

Category-level object pose estimation describes the task of estimating the full 9 degrees-of-freedom (DoF) object pose (consisting of the 3D rotation, 3D translation and 3D metric size) for objects from a given set of categories. The problem

* Authors with equal contributions.

Codes are released at https://github.com/lolrudy/RBP_Pose.

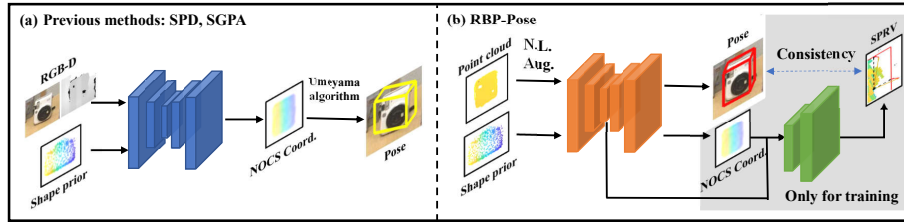


Fig. 1. Comparison of RBP-Pose and previous shape prior adaptation methods. Previous methods [38,2] predict NOCS coordinates (NOCS Coord.) and recover the pose using the Umeyama algorithm [39]. In comparison, we use NOCS coordinates within our point-wise bounding box projection and predict Shape Prior Guided Residual Vectors (SPRV), which encapsulates the pose explicitly. Moreover, we propose Non-Linear Shape Augmentation (N.L. Aug.) to increase shape diversity during training.

has gained wide interest in research due to its essential role in many applications, such as augmented reality [34], robotic manipulation [5] and scene understanding [29,50]. In comparison to conventional instance-level pose estimation [14,13], which assumes the availability of a 3D CAD model for each object of interest, the category-level task puts forward a higher requirement for adaptability to various shapes and textures within each category.

Noteworthy, category-level pose estimation has recently experienced a large leap forward in recent past, thanks to novel deep learning architecture that can directly operate on point clouds [43,3,44,1]. Thereby, most of these works try to establish 3D-3D correspondences between the input point cloud and either a pre-defined normalized object space [43] or a deformed shape prior to better address intra-class shape variability [38,2]. Eventually, the 9DoF pose is commonly recovered using the Umeyama algorithm [39]. Nonetheless, despite achieving great performance, these methods typically still suffer from two shortcomings. First, their shape prior integration only boosts pose estimation indirectly, which leads to insufficient pose-sensitive feature extraction and slow inference speed. Second, due to the relatively small amount of available real-world data [43], these works tend to overfit as they are directly trained on these limited datasets.

As for the lack of real labeled data, we further propose an online non-linear shape augmentation scheme for training to avoid overfitting and enhance the generalization ability of RBP-Pose. In FS-Net [3], the authors propose to stretch or compress the object bounding box to generate new instances. However, the proportion between different parts of the object basically remains unchanged, as shown in Fig. 4. Therefore, we propose a category-specific non-linear shape augmentation technique. In particular, we deform the object shape by adjusting its scale via a truncated parabolic function along the direction of a selected axis. To this end, we either choose the symmetry axis for symmetric objects or select the axis corresponding to the facing direction to avoid unrealistic distortions for

non-symmetric objects. In this way we are able to increase the dataset size while preserving the representative shape characteristic of each category.

Interestingly, to tackle the former limitation, the authors of GPV-Pose [7] have proposed to leverage **D**isplacement **V**ectors from the observed points to the corresponding **P**rojections on the **B**ounding box (DVPB), in an effort to explicitly encapsulate the spatial cues of the 3D object and, thus, improve direct pose regression. While this performs overall well, the representation still exhibits weaknesses. In particular, DVPB is not necessarily a small vector with zero-mean. In fact, the respective values can become very large (as for large objects like laptops), which can make it very difficult for standard networks to predict them accurately. Based on these grounds, in this paper we propose to overcome this limitation by means of integrating shape priors into DVPB. We essentially describe the displacement field from the shape-prior-indicated projections towards the real projections onto the object bounding box. We dub the residual vectors in this displacement field as SPRV for **S**hape **P**rior Guided **R**esidual **V**ectors. SPRV is inherently zero-centered and relatively small, allowing robust estimation with a deep neural network. In practice, we adopt a fully convolutional decoder to directly regress SPRV and then establish geometry-aware consistency with the predicted pose to enhance feature extraction. We experimentally show that our novel geometry-guided **R**esidual **B**ounding Box **P**rojection network RBP-Pose provides state-of-the-art results and clearly outperforms the DVPB representation. Overall, our main contributions are summarized as follows,

1. We propose a **R**esidual **B**ounding Box **P**rojection network (RBP-Pose) that jointly predicts 9DoF pose and shape prior guided residual vectors. We demonstrate that these nearly zero-mean residual vectors can be effectively predicted from our network and well encapsulate the spatial cues of the pose whilst enabling geometry-guided consistency terms.
2. To enhance the robustness of our method, we additionally propose a non-linear shape augmentation scheme to improve shape diversity during training whilst effectively preserving the commonality of geometric characteristics within categories.
3. RBP-Pose runs at inference speed of 25Hz and achieves state-of-the-art performance on both synthetic and real-world datasets.

2 Related Works

Instance-level 6D Pose Estimation. Instance-level pose estimation tries to estimate the 6DoF object pose, composed of the 3D rotation and 3D translation, for a known set of objects with associated 3D CAD models. The majority of monocular methods falls into three groups. The first group of methods [47,27,26,20,18,16] regresses the pose directly, whereas the second group instead establishes 2D-3D correspondences via keypoint detection or dense pixel-wise prediction of 3D coordinates [21,33,31,49,12,30]. The pose can be then obtained by adopting the PnP algorithm. Noteworthy, a few methods [15,6,42]

adopt a neural network to learn the optimization step instead of relying on PnP. The last group of methods [36,35] attempt to learn a pose-sensitive latent embedding for subsequent pose retrieval. As for RGB-D based methods, most works [19,41,11,10] again regress the pose directly, while a few methods [46,17] resort to latent embedding similar to [36,35]. In spite of great advance in recent years, the practical use of instance-level methods is limited as they can typically only deal with a handful of objects and additionally require CAD models.

Category-level Pose Estimation. In the category-level setting, the goal is to predict the 9DoF pose for previously seen or unseen objects from a known set of categories [28,43]. The setting is fairly more challenging due to the large intra-class variations of shape and texture within categories. To tackle this issue, Wang *et al.* [43] derive the *Normalized Object Coordinate Space* (NOCS) as a unified representation. They map the observed point cloud into NOCS and then apply the Umeyama algorithm [39] for pose recovery. CASS [1] introduces a learned canonical shape space instead. FS-Net [3] proposes a decoupled representation for rotation and directly regresses the pose. DualPoseNet [23] adopts two networks for explicit and implicit pose prediction and enforces consistency between them for pose refinement. While 6-PACK [40] tracks the object’s pose by means of semantic keypoints, CAPTRA [45] instead combines coordinate prediction with direct regression. GPV-Pose [7] harnesses geometric insights into bounding box projection to enhance the learning of category-level pose-sensitive features. To explicitly address intra-class shape variation, a certain line of works make use of shape priors [38,2,22,8]. Thereby, SPD [38] extracts the prior point cloud for each category as the mean of all shapes adopting a PointNet [32] auto-encoder. SPD further deforms the shape prior to fit the observed instance and assigns the observed point cloud to the reconstructed shape model. SGPA [2] dynamically adapts the shape prior to the observed instance in accordance with its structural similarity. DO-Net [22] also utilizes shape prior, yet, additionally harnesses the geometric object properties to enhance performance. ACR-Pose [8] adopts a shape prior guided reconstruction network and a discriminator network to learn high-quality canonical representations. Noteworthy, as shape prior integration only improves pose estimation indirectly, all these methods commonly suffer from insufficient pose-sensitive feature extraction and slow inference speed.

3 Methodology

In this paper, we aim at tackling the problem of category-level 9DoF pose estimation. In particular, given an RGB-D image with objects from a set of known categories, our objective is to detect all present object instances in the scene and recover their 9DoF object poses, including the 3DoF rotation as rotation matrix $R \in \mathbb{R}^{3 \times 3}$, the 3DoF translation $t \in \mathbb{R}^3$ and 3DoF metric size $s \in \mathbb{R}^3$.

3.1 Overview

As illustrated in Fig. 2, RBP-Pose consists of 5 modules, responsible for i) input preprocessing, ii) feature extraction from the input and prior point cloud, iii)

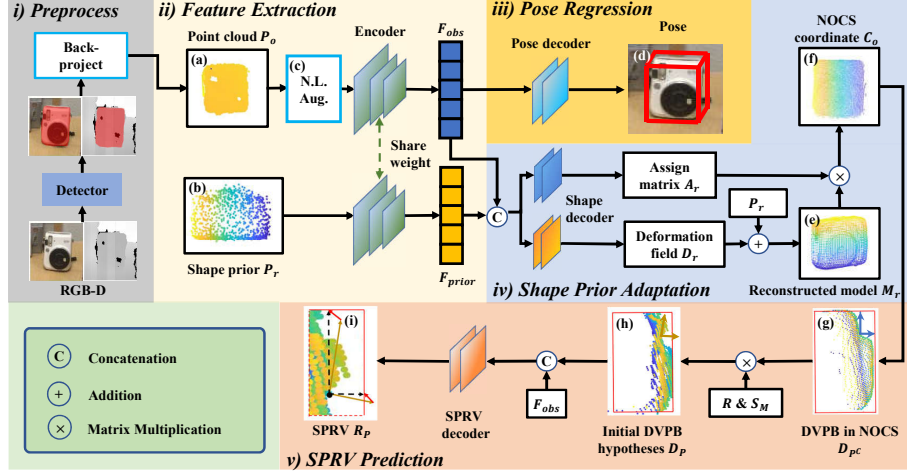


Fig. 2. An overview of RBP-Pose. RBP-Pose takes RGB-D image and shape prior as inputs. We perform a non-linear shape augmentation (c) after extracting point cloud of the object of interest (a). It deforms shape prior (e) and predicts NOCS coordinates (f) to retrieve DVPB in NOCS (g). Integrating the predicted rotation R , the pre-computed category mean size S_M , we compute the initial DVPB hypotheses (h) as the input of SPRV decoder. Finally, RBP-Pose predicts pose (d) and SPRV (i) and enforces consistency between them. During inference, only the Preprocessing, Feature Extraction and Pose Regression modules are needed.

9DoF pose regression, iv) adaptation of the shape prior given the extracted features, and, finally, v) Shape Prior Guided Residual Vectors (SPRV) prediction.

Preprocessing. Given an RGB-D image, we first leverage an off-the-shelf object detector (e.g. Mask-RCNN [9]) to segment objects of interest and then back-project their corresponding depth values to generate the associated object point clouds. We then uniformly sample $N = 1024$ points from each detected object and feed it as the input P_o to the following modules, as shown in Fig. 2 (a).

Feature Extraction. Since 3DGC [24] is insensitive to shift and scale of the given point cloud, we adopt it as our feature extractor to respectively obtain pose-sensitive features F_{obs} and F_{prior} from P_o and a pre-computed mean shape prior P_r (with $M = 1024$ points) as in SPD [38]. We introduce a non-linear shape augmentation scheme to increase the diversity of shapes and promote robustness, which will be discussed in detail in Sec. 3.5. Finally, F_{obs} is fed to the Pose Regression module for direct pose estimation and to the Shape Prior Adaptation module after concatenation with F_{prior} .

9DoF Object Pose Regression. To represent the 3DoF rotation, we follow GPV-Pose [7] and decompose the rotation matrix R into 3 columns $r_x, r_y, r_z \in \mathbb{R}^3$, each representing a plane normal of the object bounding box. We predict the first two columns r_x, r_y along with their uncertainties u_x, u_y , and calculate the calibrated plane normals r'_x, r'_y via uncertainty-aware averaging [7]. Eventually,

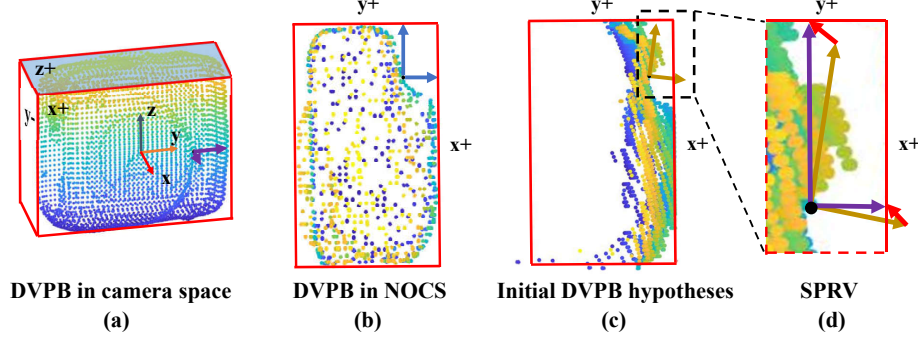


Fig. 3. Illustration of DVPB and SPRV. We show DVPB and SPRV of a point to $x+, y+$ plane. The target is to predict DVPB in the camera space (purple vectors in (a)). For better demonstration, we project the point cloud to $z+$ plane (the blue plane in (a)) in (b)-(d). Using the predicted coordinates in (b), we recover DVPB in NOCS (blue vectors in (b)). We then transform it into the camera space and compute the initial DVPB hypotheses (brown vectors in (c)). RBP-Pose predicts the residual vector from hypotheses to ground truth DVPB (SPRV, red vectors in (d)).

the predicted rotation matrix is recovered as $R' = [r'_x, r'_y, r'_x \times r'_y]$. For translation and size, we follow FS-Net, adopting their residual representation [3]. Specifically, for translation $t = \{t_x, t_y, t_z\}$, given the output residual translation $t_r \in \mathbb{R}^3$ and the mean P_M of the observed visible point cloud P_o , t is recovered as $t = t_r + P_M$. Similarly, given the estimated residual size $s_r \in \mathbb{R}^3$ and the pre-computed category mean size S_M , we have $s = s_r + S_M$, where $s = \{s_x, s_y, s_z\}$.

Shape Prior Adaptation. We first concatenate the feature maps F_{obs} from the observed point cloud P_o and F_{prior} from shape prior P_r in a channel-wise manner. Subsequently, we use two sub-decoders from SPD [38] to predict the row-normalized assignment matrix $A_r \in \mathbb{R}^{N \times M}$ and the deformation field $D_r \in \mathbb{R}^{M \times 3}$, respectively. Given the shape prior \mathcal{P}_r , D_r deforms \mathcal{P}_r to reconstruct the normalized object shape model M_r with $M_r = \mathcal{P}_r + D_r$ (Fig. 2 (e)). Further, A_r associates each point in P_o with M_r . Thereby, the NOCS coordinate $C_o \in \mathbb{R}^{N \times 3}$ of the input point cloud P_o is computed as $C_o = A_r M_r$ (Fig. 2 (f)).

Shape Prior Guided Residual Vectors (SPRV) Prediction. The main contribution of our work resides in the use of Shape Prior Guided Residual Vectors (SPRV) to integrate shape priors into the direct pose regression network, enhancing the performance whilst keeping a fast inference speed. In the following section we will now introduce this module in detail.

3.2 Residual Bounding Box Projection

Preliminaries. In GPV-Pose [7], the authors propose a novel confidence-aware point-wise voting method to recover the bounding box. For each observed object point P , GPV-Pose thereby predicts its **Displacement Vector** towards its **Projections** on each of the 6 **Bounding Box** faces (DVPB), as shown in Fig. 3

(a). Exemplary, when considering the $x+$ plane, the DVPB of the observed point P onto the $x+$ face of the bounding box is defined as,

$$D_{P,x+} = (s_x/2 - \langle r_x, P \rangle + \langle r_x, t_x \rangle) r_x, \quad (1)$$

where $\langle *, * \rangle$ denotes the inner product and, as before, r_x denotes the first column of the rotation matrix R . Thus, each point P provides 6 DVPBs with respect to all 6 bounding box faces $\mathcal{B} = \{x\pm, y\pm, z\pm\}$. Notice that, since symmetries lead to ambiguity in bounding box faces around the corresponding symmetry axis, GPV-Pose only compute the DVPB on the ambiguity-free faces.

Although GPV-Pose reports great results when leveraging DVPB, it still suffers from two important shortcomings. First, DVPB is not necessarily a small vector with zero-mean. In fact, the respective values can become very large (as for large objects like laptops), which can make it very difficult for standard networks to predict them accurately. Second, DVPB is not capable of conducting automatic outlier filtering, hence, noisy point cloud observations may significantly deteriorate the predictions of DVPB. On that account, we propose to incorporate shape prior into DVPB in the form of Shape Prior Guided Residual Vectors (SPRV) to properly address the aforementioned shortcomings.

Shape Prior Guided Residual Vectors (SPRV). As illustrated in Fig. 2 (e), we predict the deformation field D_r that deforms the shape prior P_r to the shape of the observed instance with $M_r = P_r + D_r$. Thereby, during experimentation we made two observations. First, as P_r is outlier-free and D_r is regularized to be small, similar to SPD [38], we can safely assume that M_r contains no outliers, allowing us to accurately recover its bounding box in NOCS by selecting the outermost points along the x , y and z axis, respectively. Second, since A_r is the row-normalized assignment matrix, we know that M_r shares the same bounding box with $C_o = A_r M_r$, which is assumed to be inherently outlier-free and accurate. Based on the above two observations, we can utilize M_r and C_o to provide initial hypotheses for DVPB with respect to each point in P_o .

Specifically, for a point P^C in C_o , as it is in NOCS, its DVPB $D_{P^C,x+}$ (Fig. 3 (b)) can be represented as,

$$D_{P^C,x+} = (s_x^M/2 - P^C) n_x, \quad (2)$$

where s_x^M denotes the size of M_r along the x axis and $n_x = [1, 0, 0]^T$ is the normalized bounding box face normal. We then transform $D_{P^C,x+}$ from NOCS to the camera coordinate to obtain the initial DVPB hypotheses for the corresponding point P in P_o (Fig. 3 (c)) as,

$$D_{P,x+} = L D_{P^C,x+} r_x \quad (3)$$

where $L = \sqrt{s_x^2 + s_y^2 + s_z^2}$ is the diagonal length of the bounding box. Note that L and r_x are calculated from the category mean size S_M and the rotation prediction of our Pose Regression Module respectively.

Given the ground truth DVPB $D_{P,x+}^{gt}$ and initial DVPB hypotheses $D_{P,x+}$, the SPRV of P to the $x+$ bounding box face (Fig. 3 (d)) is calculated as,

$$R_{P,x+} = D_{P,x+}^{gt} - D_{P,x+}. \quad (4)$$

The calculation of SPRV with respect to the other bounding box faces in \mathcal{B} follows the same principal. By this means, SPRV can be approximately modelled with zero-mean Laplacian distribution, which enables effective prediction with a simple network. In the SPRV Prediction module, we feed the estimated initial DVPB hypotheses together with the feature map F_{obs} into a fully convolutional decoder to directly regress SPRV. As this boils down to a multi-task prediction problem, we employ the Laplacian aleatoric uncertainty loss from [4] to weight the different contributions within SPRV according to

$$\begin{aligned}\mathcal{L}_{SPRV}^{data} &= \sum_{P \in P_o} \sum_{j \in \mathcal{B}} \frac{\sqrt{2}}{\sigma_{s_j}} |R_{P_j} - R_{P_j}^{gt}| + \log(\sigma_{s_j}) \\ \mathcal{L}_{SPRV}^{reg} &= \sum_{P \in P_o} \sum_{j \in \mathcal{B}} \frac{\sqrt{2}}{\sigma'_{s_j}} |R_{P_j}| + \log(\sigma'_{s_j}) \quad . \\ \mathcal{L}_{SPRV} &= \mathcal{L}_{SPRV}^{data} + \lambda_0 \mathcal{L}_{SPRV}^{reg}\end{aligned}\tag{5}$$

Thereby, $R_{P_j}^{gt}$ refers to the ground truth SPRV as calculated by the provided ground truth NOCS coordinates and respective pose annotations. Further, σ_{s_j} , σ'_{s_j} denote the standard variation of Laplacian distribution that are utilized to model the uncertainties. Note that the first term $\mathcal{L}_{SPRV}^{data}$ is fully supervised using the respective ground truth, while \mathcal{L}_{SPRV}^{reg} is a regularization term that enforces the SPRV network to predict small displacements. In addition, λ_0 is a weighting parameter to balance the two terms. Note that we do not apply Gaussian-distribution-based losses. We follow GPV-Pose [7] to supervise other branches with \mathcal{L}_1 loss for stability. Thus we adopt Eq. 5 for convenient adjustment of the weight of each term.

3.3 SPRV for Pose Consistency

Since SPRV explicitly encapsulates pose-related cues, we utilize it to enforce geometric consistency between the SPRV prediction and the pose regression. To this end, we first employ the predicted pose to estimate DVPB D^{Pose} according to Eq. 1. We then recover D^{SPRV} via adding the predicted SPRV to the initial hypotheses. Finally, our consistency loss term is defined as follows,

$$\mathcal{L}_{con} = \sum_{P \in P_o} \sum_{j \in \mathcal{B}} |D_{P,b}^{Pose} - D_{P,b}^{SPRV}|,\tag{6}$$

where $|\cdot|$ denotes the \mathcal{L}_1 distance.

3.4 Overall Loss Function

The overall loss function is defined as follows,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{pose} + \lambda_2 \mathcal{L}_{shape} + \lambda_3 \mathcal{L}_{SPRV} + \lambda_4 \mathcal{L}_{con}\tag{7}$$

For \mathcal{L}_{pose} , we utilize the loss terms from GPV-Pose [7] to supervise R, t, s with the ground truth. For \mathcal{L}_{shape} , we adopt the loss terms from SPD [38] to supervise the prediction of the deformation field D_r and the assignment matrix A_r . Further, \mathcal{L}_{SPRV} and \mathcal{L}_{con} are defined in Eq. 5 and Eq. 6. Finally, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ denote the utilized weights to balance the individual loss contributions, and are chosen empirically.

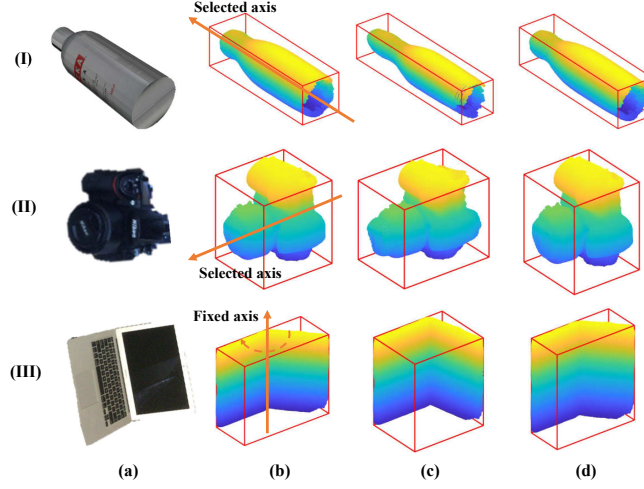


Fig. 4. Demonstration of non-linear data augmentation. (I), (II) and (III) show non-linear data augmentation for *bottle*, *camera* and *laptop*. For the instances (a), we augment the object shape in our non-linear manner from (b) to (c). FS-Net [3] adopts the linear bounding box deformation in data augmentation (d), which can be regarded as a special case of our non-linear shape augmentation.

3.5 Non-Linear Shape Augmentation

To tackle the intra-class shape variations and improve the robustness and generalizability of RBP-Pose, we propose a category-specific non-linear shape augmentation scheme (Fig. 4). In FS-Net [3], the authors augment the shape by stretching or compression of the object bounding box. Their augmentation is linear and unable to cover the large shape variations within a category, since the proportions between different parts of the object basically remain unchanged (Fig. 4 (d)). In contrast, we propose a novel non-linear shape augmentation method which is designed to generate diverse unseen instances, whilst preserving the representative shape features of each category (Fig. 4 (c)).

In particular, we propose two types of augmentation strategies for categories provided by the REAL275 dataset [43]: axis-based non-linear scaling transformation (A1) for *camera*, *bottle*, *can*, *bowl*, *mug* (Fig 4 (I, II)) and plane-based rotation transformation (A2) for *laptop* (Fig 4 (III)).

As for A1, we deform the object shape by adjusting its scale along the direction of a selected axis. For each point P in the canonical object space, its deformation scale $\mathcal{S}_{A1}(P)$ is obtained by $\mathcal{S}_{A1}(P) = \xi(P_*)$, where $\xi(P_*)$ is a random non-linear function and P_* is the projection of P on the selected axis. In this paper, we choose ξ as the parabolic function, thus, we have

$$\mathcal{S}_{A1}(P) = \xi(P_*) = \gamma_{min} + 4(\gamma_{max} - \gamma_{min})(P_*)^2, \quad (8)$$

where $\gamma_{max}, \gamma_{min}$ are uniformly sampled random variables that control the upper and lower bounds of $\mathcal{S}_{A1}(P)$. Exemplary, when selecting y as our augmentation

axis, the respective transformation function is defined as,

$$\mathcal{T}_{A1}(P) = \{\gamma P_x, \mathcal{S}_{A1}(P)P_y, \gamma P_z\}, \quad (9)$$

where γ is the random variable that controls the scaling transformation along x and z axis. In practice, we select the symmetry axis (y -axis) for *bottle*, *can*, *bowl* and *mug* as the transformation axis. Moreover, for *camera*, we select the axis that passes through the camera lens (x -axis), to keep its roundish shape after augmentation. The corresponding transformation function is then defined as in Eq. 8 and Eq. 9, yet, $\mathcal{S}_{A1}(P)$ is only applied to P_x and γ is applied to P_y , P_z .

As for $A2$, since *laptop* is an articulated object consisting of two movable planes, we conduct shape augmentation by modifying the angle between the upper and lower plane (Figure 4 (III)). Thereby, we rotate the upper plane by a certain angle along the fixed axis, while the lower plane remains static. Please refer to the Supplementary Material for details of $A2$ transformation.

4 Experiments

Datasets. We employ the common REAL275 and CAMERA25 [43] benchmark datasets for evaluation. Thereby, REAL275 is a real-world dataset consisting of 7 scenes with 4.3K images for training and 6 scenes with 2.75K images for testing. It covers 6 categories, including *bottle*, *bowl*, *camera*, *can*, *laptop* and *mug*. Each category contains 3 unique instances in both training and test set. On the other hand, CAMERA25 is a synthetic dataset generated by rendering virtual objects on real background. CAMERA25 contains 275k images for training and 25k for testing. Note that CAMERA25 shares the same six categories with REAL275.

Implementation Details. Following [23, 43], we use Mask-RCNN [9] to generate 2D segmentation masks for a fair comparison. As for our category-specific non-linear shape augmentation, we uniformly sample $\gamma_{max} \sim \mathcal{U}(1, 1.3)$, $\gamma_{min} \sim \mathcal{U}(0.7, 1)$ and $\gamma \sim \mathcal{U}(0.8, 1.2)$. Besides our non-linear shape augmentation, we add random Gaussian noise to the input point cloud, and employ random rotational and translational perturbations as well as random scaling of the object. Unless specified, we set the employed balancing factors $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ to $\{8.0, 10.0, 3.0, 1.0\}$. Finally, the parameter λ_0 in Eq. 5 is set to 0.01. We train RBP-Pose in a two-stage manner to stabilize the training process. In the first stage, we only train the pose decoder and the shape decoder employing only \mathcal{L}_{pose} and \mathcal{L}_{shape} . In the second stage we train all the modules except the Pre-processing as explained in Eq. 7. This strategy ensures that our two assumptions in Sec. 3.2 are reasonable and enables smooth training. Notice that similar to other works [43, 38, 2], we train a single model for all categories. Unlike [23, 2] that train with both synthetic and real data for evaluation on REAL275, we only use the real data for training. We train RBP-Pose for 150 epochs in each stage and employ a batch size of 32. We further employ the Ranger optimizer [25, 52, 48] with a base learning rate of 1e-4, annealed at 72% of the training phase using a cosine schedule. Our experiments are conducted on a single NVIDIA-A100 GPU.

Table 1. Comparison with state-of-the-art methods on REAL275 dataset.

Method	Prior	IoU_{75}	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$	Speed(FPS)
NOCS [43]		30.1	7.2	10.0	13.8	25.2	5
CASS [1]		-	-	23.5	-	58.0	-
DualPoseNet [23]		62.2	29.3	35.9	50.0	66.8	2
FS-Net [3]		-	-	28.2	-	60.8	20
FS-Net(Ours)		52.0	19.9	33.9	46.5	69.1	20
GPV-Pose [7]		<u>64.4</u>	32.0	<u>42.9</u>	-	<u>73.3</u>	20
SPD [38]	✓	53.2	19.3	21.4	43.2	54.1	4
CR-Net [44]	✓	55.9	27.8	34.3	47.2	60.8	-
DO-Net [22]	✓	63.7	24.1	34.8	45.3	67.4	10
SGPA [2]	✓	61.9	<u>35.9</u>	39.6	<u>61.3</u>	70.7	-
Ours	✓	67.8	38.2	48.1	63.1	79.2	25

Overall best results are in bold and the second best results are underlined. **Prior** denotes whether the method makes use of shape priors. We reimplement FS-Net as **FS-Net(Ours)** for a fair comparison since FS-Net uses different detection results.

Table 2. Comparison with state-of-the-art methods on CAMERA25 dataset.

Method	Prior	IoU_{50}	IoU_{75}	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$
NOCS [43]		83.9	69.5	32.3	40.9	48.2	64.6
DualPoseNet [23]		92.4	86.4	64.7	70.7	77.2	84.7
GPV-Pose [7]		<u>93.4</u>	<u>88.3</u>	<u>72.1</u>	<u>79.1</u>	-	<u>89.0</u>
SPD [38]	✓	93.2	83.1	54.3	59.0	73.3	81.5
CR-Net [44]	✓	93.8	88.0	72.0	76.4	81.0	87.7
SGPA [2]	✓	93.2	88.1	70.7	74.5	82.7	88.4
Ours	✓	93.1	89.0	73.5	79.6	<u>82.1</u>	89.5

Overall best results are in bold and the second best results are underlined. **Prior** denotes whether the method utilizes shape priors.

Evaluation metrics. Following the widely adopted evaluation scheme [43,2,23], we utilize the two standard metrics for quantitative evaluation of the performance. In particular, we report the mean precision of 3D IoU, which computes intersection over union for two bounding boxes under the predicted and the ground truth pose. Thereby, a prediction is considered correct if the IoU is larger than the employed threshold. On the other hand, to directly evaluate rotation and translation errors, we use the $5^\circ 2cm$, $5^\circ 5cm$, $10^\circ 2cm$ and $10^\circ 5cm$ metrics. A pose is hereby considered correct if the translational and rotational errors are less than the respective thresholds.

4.1 Comparison with State-of-the-art

Performance on NOCS-REAL275. In Tab 1, we compare RBP-Pose with 9 state-of-the-art methods, among which 4 methods utilize shape priors. It can be easily observed that our method outperforms all other competitors by a large

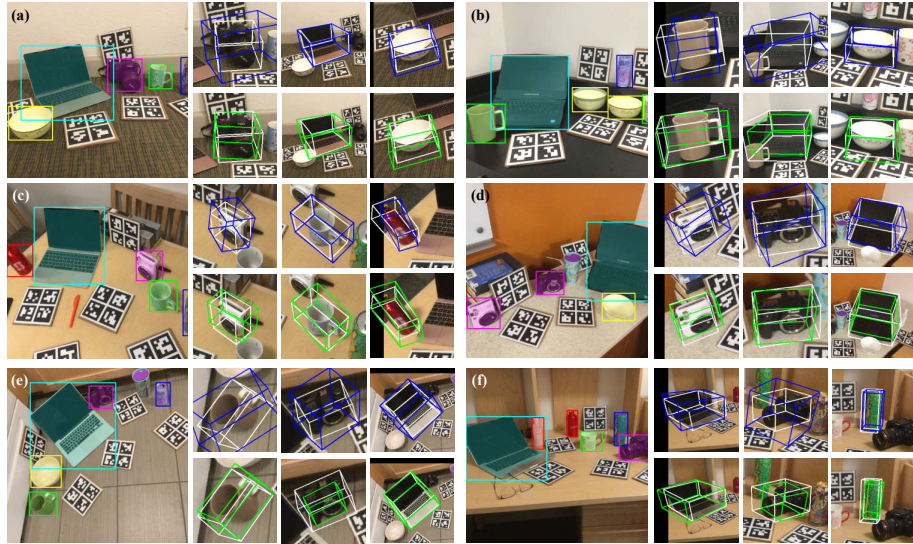


Fig. 5. Qualitative results of our method (green line) and SGPA [2] (blue line). Images (a)-(f) demonstrate 2D segmentation results.

margin. Specifically, under IoU_{75} , we achieve a mAP of 67.8%, which exceeds the second best method DO-Net [22] by 4.1%. Regarding the rotation and translation accuracy, RBP-Pose outperforms SGPA [2] by 2.3% in $5^\circ 2cm$, 8.5% in $5^\circ 5cm$, 1.8% in $10^\circ 2cm$ and 8.5% in $10^\circ 5cm$. Moreover, when comparing with GPV-Pose [7], we can outperform them by 6.2% in $5^\circ 2cm$, 5.2% in $5^\circ 5cm$ and 5.9% in $10^\circ 5cm$. Noteworthy, despite achieving significant accuracy improvements, RBP-Pose still obtains a real-time frame rate of 25Hz when using YOLOv3 [37] and ATSA [51] for object detection. Moreover, we present a detailed per-category comparison for 3D IoU, rotation and translation accuracy of RBP-Pose and SGPA [2] in Fig. 6. It can be deduced that our method obtains superior results over SGPA in terms of mean precision for all metrics, especially in rotation. Moreover, our method is superior in dealing with complex categories with significant intra-class shape variations, *e.g.* *camera* (green line in Fig. 6).

Performance on NOCS-CAMERA25. The results for CAMERA25 are shown in Tab. 2. Our method outperforms all competitors for stricter metrics IoU_{75} , $5^\circ 2cm$, $5^\circ 5cm$ and $10^\circ 5cm$, and is on par with the best methods for IoU_{50} and $10^\circ 2cm$. Specifically, our method exceeds the second best methods for IoU_{75} , $5^\circ 2cm$, $5^\circ 5cm$ and $10^\circ 5cm$ by 0.7%, 1.4%, 0.5% and 0.5%, respectively.

4.2 Ablation Study

Effect of Shape Prior Guided Residual Vectors. In Tab. 3, we evaluate the performance of our method under different configurations. From E1 to E3,

Table 3. Ablation study on DVPB and data augmentation.

	DVPB	SPRV	Con.	L. Aug.	N.L. Aug.	IoU_{75}	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$
E1					✓	65.8	32.6	43.8	57.9	75.6
E2	✓				✓	66.6	33.4	45.9	59.1	77.5
E3		✓			✓	67.7	36.0	44.5	61.6	77.1
E4	✓		✓		✓	66.2	34.4	44.8	61.3	77.5
E5		✓	✓			61.3	23.8	29.7	53.8	66.4
E6		✓	✓	✓		66.2	36.1	47.0	62.2	78.8
E7		✓	✓		✓	67.8	38.2	48.1	63.1	79.2

DVPB denotes predicting DVPB directly, *i.e.*, the decoder only takes F_{obs} as input and outputs DVPB instead of SPRV. **SPRV** denotes SPRV prediction introduced in Sec. 3.2. **Con.** denotes the loss term \mathcal{L}_{con} . **L. Aug.** denotes the linear bounding-box-based shape augmentation from FS-Net [3] and **N.L. Aug.** denotes our non-linear shape augmentation.

we compare three variants of RBP-Pose w.r.t the integration of DVPB: removing the DVPB related modules, predicting DVPB directly and predicting SPRV. By directly predicting DVPB like in GPV-Pose [7], the mAP improves by 0.8% under $5^\circ 2cm$ and 2.1% in $5^\circ 5cm$, which indicates that DVPB explicitly encapsulates pose information, helping the network to extract pose-sensitive features. By utilizing shape priors to generate initial hypothesis of DVPB and additionally predicting SPRV, the performance improves 2.6% under $5^\circ 2cm$ and 2.5% under $10^\circ 2cm$, while the mAP of $5^\circ 5cm$ and $10^\circ 5cm$ decreases. In general, by solely adopting the auxiliary task of predicting SPRV, the translation accuracy rises while the rotation accuracy falls. This, however, can be solved using our consistency loss between SPRV and pose. E7 adopts the consistency term in Eq. 6 based on E3, and boosts the performance by a large margin under all metrics. This shows that the consistency term is able to guide the network to align predictions from different decoders by jointly optimizing them. E4 enforces the consistency term on DVPB without residual reasoning. Performance deteriorates since initial DVPB hypotheses are typically inaccurate. SPRV decoder refines the hypotheses by predicting residuals, and thus enhances overall performance.

Effect of non-linear shape augmentation. In Tab. 3 E5, we remove the non-linear shape augmentation and preserve all other components. Comparing E5 and E7, it can be deduced that the performance degrades dramatically without non-linear shape augmentation, where the mAP of $5^\circ 2cm$ and $5^\circ 5cm$ drops by 15.6% and 18.4%, respectively. The main reason is that we only train the network on real-world data containing only 3 objects for each category with 4k images, leading to severe overfitting. The non-linear data augmentation mitigates this problem and enhances the diversity of shapes in the training data.

Non-linear vs linear shape augmentation. We compare our non-linear shape augmentation with the linear bounding-box-based augmentation from FS-Net [3] in Tab. 3 E6 and E7. Our non-linear shape augmentation boosts the mAP w.r.t. all metrics. Specifically, the accuracy improves by 1.6% for IoU_{75} , 2.1% for $5^\circ 2cm$, 1.1% for $5^\circ 5cm$ and 0.9% for $10^\circ 2cm$. The main reason is that our

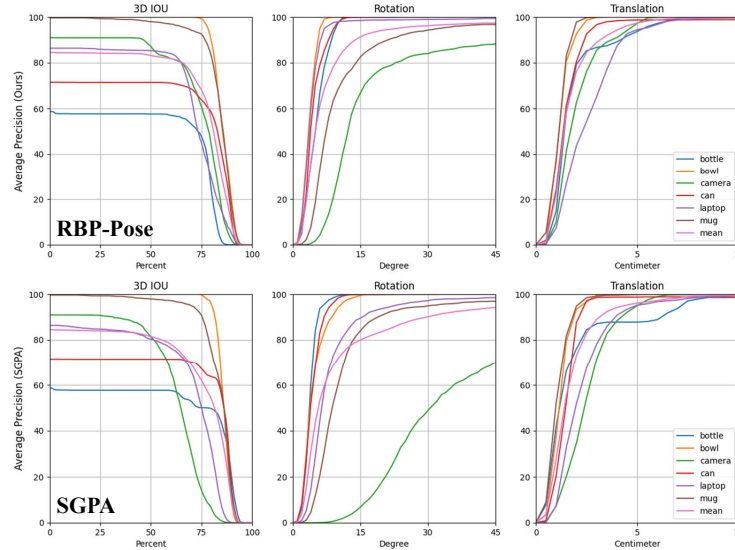


Fig. 6. Quantitative comparison with SGPA [2] on REAL275 in terms of average precision in 3D IoU, Rotation and Translation.

non-linear shape augmentation covers more kinds of shape variations than the linear counterpart, which improves the diversity of training data and mitigates the problem of overfitting.

4.3 Qualitative Results

We provide a qualitative comparison between RBP-Pose and SGPA [2] in Fig. 5. Comparative advantage of our method over SGPA is significant, especially in the accuracy of the rotation estimation. Moreover, our method consistently outperforms SGPA when estimating the pose for the *camera* category, which supports our claim that we can better handle categories with large intra-class variations. We discuss **Failure Cases** and **Limitations** in the supplemental material.

5 Conclusion

In this paper, we propose RBP-Pose, a novel method that leverages Residual Bounding Box Projection for category-level object pose estimation. RBP-Pose jointly predicts 9DoF pose and shape prior guided residual vectors. We illustrate that these nearly zero-mean residual vectors encapsulate the spatial cues of the pose and enable geometry-guided consistency terms. We also propose a non-linear data augmentation scheme to improve shape diversity of the training data. Extensive experiments on the common public benchmark demonstrate the effectiveness of our design and the potential of our method for future real-time applications such as robotic manipulation and augmented reality.

References

1. Chen, D., Li, J., Wang, Z., Xu, K.: 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. pp. 11973–11982 (2020) 2, 4, 11
2. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2773–2782 (2021) 2, 4, 10, 11, 12, 14
3. Chen, W., Jia, X., Chang, H.J., Duan, J., Linlin, S., Leonardis, A.: 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). pp. 1581–1590 (June 2021) 2, 4, 6, 9, 11, 13
4. Chen, Y., Tai, L., Sun, K., Li, M.: Monopair: Monocular 3d object detection using pairwise spatial relationships. In: CVPR. pp. 12093–12102 (2020) 8
5. Deng, X., Xiang, Y., Mousavian, A., Eppner, C., Bretl, T., Fox, D.: Self-supervised 6d object pose estimation for robot manipulation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 3665–3671. IEEE (2020) 2
6. Di, Y., Manhardt, F., Wang, G., Ji, X., Navab, N., Tombari, F.: So-pose: Exploiting self-occlusion for direct 6d pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12396–12405 (2021) 3
7. Di, Y., Zhang, R., Lou, Z., Manhardt, F., Ji, X., Navab, N., Tombari, F.: Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. arXiv preprint (2022) 3, 4, 5, 6, 8, 11, 12, 13
8. Fan, Z., Song, Z., Xu, J., Wang, Z., Wu, K., Liu, H., He, J.: Acr-pose: Adversarial canonical representation reconstruction network for category level 6d object pose estimation. arXiv preprint arXiv:2111.10524 (2021) 4
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 5, 10
10. He, Y., Huang, H., Fan, H., Chen, Q., Sun, J.: 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). pp. 3003–3013 (June 2021) 4
11. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: 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. pp. 11632–11641 (2020) 4
12. Hodan, T., Barath, D., Matas, J.: Epos: Estimating 6d pose of objects with symmetries. In: CVPR. pp. 11703–11712 (2020) 3
13. Hodaň, T., Michel, F., Brachmann, E., Kehl, W., Glent Buch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T.K., Matas, J., Rother, C.: BOP: Benchmark for 6D object pose estimation. European Conference on Computer Vision (ECCV) (2018) 2
14. Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: BOP challenge 2020 on 6D object localization. European Conference on Computer Vision Workshops (ECCVW) (2020) 2
15. Hu, Y., Fua, P., Wang, W., Salzmann, M.: Single-stage 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2930–2939 (2020) 3

16. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017) [3](#)
17. Kehl, W., Milletari, F., Tombari, F., Ilic, S., Navab, N.: Deep Learning of Local RGB-D Patches for 3D Object Detection and 6D Pose Estimation. In: ECCV (2016) [4](#)
18. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: Cosypose: Consistent multi-view multi-object 6d pose estimation. In: European Conference on Computer Vision. pp. 574–591. Springer (2020) [3](#)
19. Li, C., Bai, J., Hager, G.D.: A unified framework for multi-view multi-class object pose estimation. In: Proceedings of the european conference on computer vision (eccv). pp. 254–269 (2018) [4](#)
20. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: DeepIM: Deep iterative matching for 6d pose estimation. IJCV pp. 1–22 (2019) [3](#)
21. Li, Z., Wang, G., Ji, X.: CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In: ICCV. pp. 7678–7687 (2019) [3](#)
22. Lin, H., Liu, Z., Cheang, C., Zhang, L., Fu, Y., Xue, X.: Donet: Learning category-level 6d object pose and size estimation from depth observation. arXiv preprint arXiv:2106.14193 (2021) [4](#), [11](#), [12](#)
23. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: 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) [4](#), [10](#), [11](#)
24. Lin, Z.H., Huang, S.Y., Wang, Y.C.F.: 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. pp. 1800–1809 (2020) [5](#)
25. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: International Conference on Learning Representations (2019) [10](#)
26. Manhardt, F., Arroyo, D.M., Rupperecht, C., Busam, B., Birdal, T., Navab, N., Tombari, F.: Explaining the ambiguity of object detection and 6d pose from visual data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6841–6850 (2019) [3](#)
27. Manhardt, F., Kehl, W., Navab, N., Tombari, F.: Deep model-based 6d pose refinement in rgb. In: ECCV (2018) [3](#)
28. Manhardt, F., Wang, G., Busam, B., Nickel, M., Meier, S., Minciullo, L., Ji, X., Navab, N.: Cps++: Improving class-level 6d pose and shape estimation from monocular images with self-supervised learning. arXiv preprint arXiv:2003.05848v3 (2020) [4](#)
29. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: CVPR. pp. 55–64 (2020) [2](#)
30. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In: ICCV (2019) [3](#)
31. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: CVPR (2019) [3](#)
32. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: 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) [4](#)

33. Song, C., Song, J., Huang, Q.: Hybridpose: 6d object pose estimation under hybrid representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 431–440 (2020) [3](#)
34. Su, Y., Rambach, J., Minaskan, N., Lesur, P., Pagani, A., Stricker, D.: Deep multi-state object pose estimation for augmented reality assembly. In: 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). pp. 222–227. IEEE (2019) [2](#)
35. Sundermeyer, M., Durner, M., Puang, E.Y., Marton, Z.C., Vaskevicius, N., Arras, K.O., Triebel, R.: Multi-path learning for object pose estimation across domains. In: CVPR. pp. 13916–13925 (2020) [4](#)
36. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. In: ECCV (2018) [4](#)
37. Tekin, B., Sinha, S.N., Fua, P.: Real-Time Seamless Single Shot 6D Object Pose Prediction. In: CVPR. pp. 292–301 (2018) [12](#)
38. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: European Conference on Computer Vision. pp. 530–546. Springer (2020) [2](#), [4](#), [5](#), [6](#), [7](#), [8](#), [10](#), [11](#)
39. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence **13**(04), 376–380 (1991). <https://doi.org/10.1109/34.88573> [2](#), [4](#)
40. Wang, C., Martín-Martín, R., Xu, D., Lv, J., Lu, C., Fei-Fei, L., Savarese, S., Zhu, Y.: 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 10059–10066. IEEE (2020) [4](#)
41. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: DenseFusion: 6d object pose estimation by iterative dense fusion. In: CVPR. pp. 3343–3352 (2019) [4](#)
42. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: CVPR (June 2021) [3](#)
43. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: 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. pp. 2642–2651 (2019) [2](#), [4](#), [9](#), [10](#), [11](#)
44. Wang, J., Chen, K., Dou, Q.: Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2021) [2](#), [11](#)
45. Weng, Y., Wang, H., Zhou, Q., Qin, Y., Duan, Y., Fan, Q., Chen, B., Su, H., Guibas, L.J.: Captra: Category-level pose tracking for rigid and articulated objects from point clouds. arXiv preprint arXiv:2104.03437 (2021) [4](#)
46. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: CVPR (2015) [4](#)
47. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. RSS (2018) [3](#)
48. Yong, H., Huang, J., Hua, X., Zhang, L.: Gradient centralization: A new optimization technique for deep neural networks. In: European Conference on Computer Vision. pp. 635–652. Springer (2020) [10](#)
49. Zakharov, S., Shugurov, I., Ilic, S.: Dpod: Dense 6d pose object detector in rgb images. In: ICCV (2019) [3](#)

50. Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., Liu, S.: Holistic 3d scene understanding from a single image with implicit representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8833–8842 (2021) [2](#)
51. Zhang, M., Fei, S.X., Liu, J., Xu, S., Piao, Y., Lu, H.: Asymmetric two-stream architecture for accurate rgb-d saliency detection. In: European Conference on Computer Vision. pp. 374–390. Springer (2020) [12](#)
52. Zhang, M., Lucas, J., Ba, J., Hinton, G.E.: Lookahead optimizer: k steps forward, 1 step back. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019) [10](#)