SGPA: Structure-Guided Prior Adaptation for Category-Level 6D Object Pose Estimation

Kai Chen and Qi Dou
Department of Computer Science and Engineering, The Chinese University of Hong Kong
{kaichen, qdou}@cse.cuhk.edu.hk

Abstract

Category-level 6D object pose estimation aims to predict the position and orientation for unseen objects, which plays a pillar role in many scenarios such as robotics and augmented reality. The significant intra-class variation is the bottleneck challenge in this task yet remains unsolved so far. In this paper, we take advantage of category prior to overcome this problem by innovating a structure-guided prior adaptation scheme to accurately estimate 6D pose for individual objects. Different from existing prior based methods, given one object and its corresponding category prior, we propose to leverage their structure similarity to dynamically adapt the prior to the observed object. The prior adaptation intrinsically associates the adopted prior with different objects, from which we can accurately reconstruct the 3D canonical model of the specific object for pose estimation. To further enhance the structure characteristic of objects, we extract low-rank structure points from the dense object point cloud, therefore more efficiently incorporating sparse structural information during prior adaptation. Extensive experiments on CAMERA25 and REAL275 benchmarks demonstrate significant performance improvement. Project homepage: https://www.cse.cuhk.edu.hk/~kaichen/projects/sgpa/sgpa.html.

1. Introduction

Category-level 6D object pose estimation is increasingly studied and plays a pillar role in many real-world applications such as robotic manipulation [9], augmented reality [24], and 3D scene understanding [7, 18]. The goal is to predict position and orientation for novel objects of the same category, so as to achieve robust applicability. Different from conventional instance-level [12, 20, 30, 35] object pose estimation, which gives instance CAD models and predicts poses for the instances that have been seen during training, category-level task requires capturing the general properties while accounting for the large variation of different instances within a category.

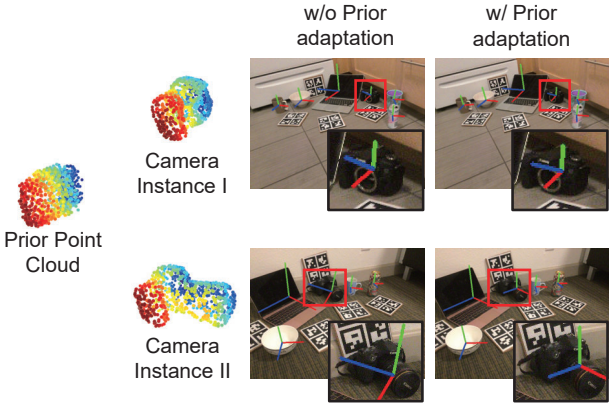

Figure 1. Pose estimation results of two *camera* instances with different structures. (i) *Instance I* is similar to the categorical prior, method w/o prior adaptation can handle pose estimation for such cases. (ii) *Instance II* is very different from the prior in structure, method w/o prior adaptation fails to associate the prior with the instance, leading to inaccurate pose estimation. Our proposed prior adaptation method can overcome this challenge with accurate pose estimation for various novel objects (noting the orientation axis).

Current methods for this challenging problem are still limited so far. First of all, to address intra-class variation of objects, a canonical object space has been resorted as a unified coordinate system [4, 26, 32, 33]. In this normalized space, a 3D structural model is reconstructed for each object with the same size and orientation. However, such normalization lacks explicit representation of shape variations across different objects of the same category, therefore limiting the accuracy of 6D pose estimation. Later on, to overcome this problem, SPD [26] is proposed to reconstruct canonical object models with category-level shape priors. A point cloud prior is built for each category, and further deformed to reconstruct the canonical 3D model for a new object.

However, such category-level prior is static and therefore not adaptable to individual instances, i.e., the same prior is applied to all instances of the same category. This greatly hinders the generalization ability of the method, especially

on those objects with significant difference to the reconstructed prior. For example, as illustrated in Figure 1, when we apply SPD to two different camera instances with different shapes, the performance can be distinct. For camera instance I which has a similar point cloud with the prior, the 6D pose can be well estimated. Unfortunately, for camera instance II which instead has a longer lens, the shape prior is no longer representative for the specific case, thus severely degrading the pose estimation performance.

In this paper, we propose a novel **S**tructure-**G**uided **P**rior **A**daptation network (SGPA), which can dynamically adapt the category-level prior to each particular instance. It adapts the static prior to the observed object according to the structure similarity between the prior and the object. Given the geometry features of prior and object point cloud, our SGPA uses a transformer network to build a long-term dependence between them to model the structure similarity. Based on this similarity, SGPA then dynamically adjusts the prior feature by injecting instance information into the prior feature. Specifically, we propagate the instance semantic feature along the extracted structure similarity to corresponding prior features. We argue that the structure similarity that is overlooked in existing methods can effectively bridge the prior with the instance. In addition, adaptively injecting instance semantic features into the prior can effectively mitigate the gap between the prior and the instance.

Furthermore, densely propagating the semantic feature point by point is prone to introduce noises into the prior feature, because not all points are representative enough to be used to propagate semantic features from instance to prior. To further leverage the inherent structure characteristic of instances for prior adaptation, we design an auxiliary network to extract sparse key-points from the dense input point cloud. Based on the extracted key-point information, we develop a structure regularized low-rank transformer, in which the extracted key-points are assembled with our SGPA for efficient structure guided prior adaptation. The adapted prior feature finally is used in a deformation based framework to reconstruct a canonical model for the instance, and match it with the instance point cloud for 6D pose estimation. We summarize our main contributions as follows:

- We propose a novel prior based category-level 6D object pose estimation framework, in which we dynamically adapt the categorical prior to each particular instance for object pose estimation.
- We propose SGPA, a novel structure-guided prior adaptation network. It uses a transformer network to model the global structure similarity between prior and object, based on which the object semantic information is injected into the prior feature for prior adaptation.
- We propose a structure regularized low-rank transformer. By regularizing the low-rank projection with the projection of point cloud key-points, the derived low-rank transformer manages to leverage the feature on distinctive key-point positions for a more effective prior adaptation.
- We conduct extensive experiments on well-acknowledged CAMERA25 and REAL275 benchmarks. Our method achieves dramatic performance improvements over other existing methods for category-level 6D object pose estimation.

2. Related Works

Instance-Level 6D Object Pose Estimation. In the instance-level setting, the network is trained and tested on the same object instance. Methods [20, 25, 2, 13, 17, 16] mainly focus on learning a robust embedding that is conditioned on the object pose. After that, methods fall into three groups depending on how to use the embedding for pose estimation. The first group of methods [35, 30, 15] directly use the embedding to regress pose parameters. The second group of methods [20, 12, 25, 13] assume the object 3D CAD model is available. They rely on the embedding to match the object observation with the CAD model on predefined landmark positions. Then, correspondence-based optimization techniques [27] are adopted for pose estimation. The third group of methods [19, 31, 36] make use of the object embedding to represent the object in a latent space, based on which differentiable rendering is used for object pose estimation. In general, due to the simplified setting of instance-level pose estimation, the prior information usually is not required for instance-level object pose estimation.

Category-Level 6D Object Pose Estimation. In the category-level setting, methods [22, 29, 14, 33] aim to predict poses for novel objects. Sahin et al. [23] derive a shape-invariant representation for objects and utilize a part-based random forest for pose estimation. Chen et al. [6] adopt neural rendering to synthesize image patches in different poses, which then are used to verify the probability of each possible pose candidate for pose estimation. In order to better overcome the intra-class variation of objects, a more typical way is to perform pose estimation in the canonical object space. Wang et al. [32] directly regress the canonical coordinates for each object on the RGB image. The pose then is estimated based on the dense correspondence between the instance point cloud and the regressed canonical coordinates. Chen et al. [4] develop a variational auto-encoder (VAE) for reconstructing the object model in the canonical space. Subsequently, the pose parameters are directly regressed with a fully connected network. A lack of an explicit model for the deformation of different instances limits the overall performance of these non-prior based methods. Recently, Tian et al. [26] present a prior based method. They first build a category-level prior point

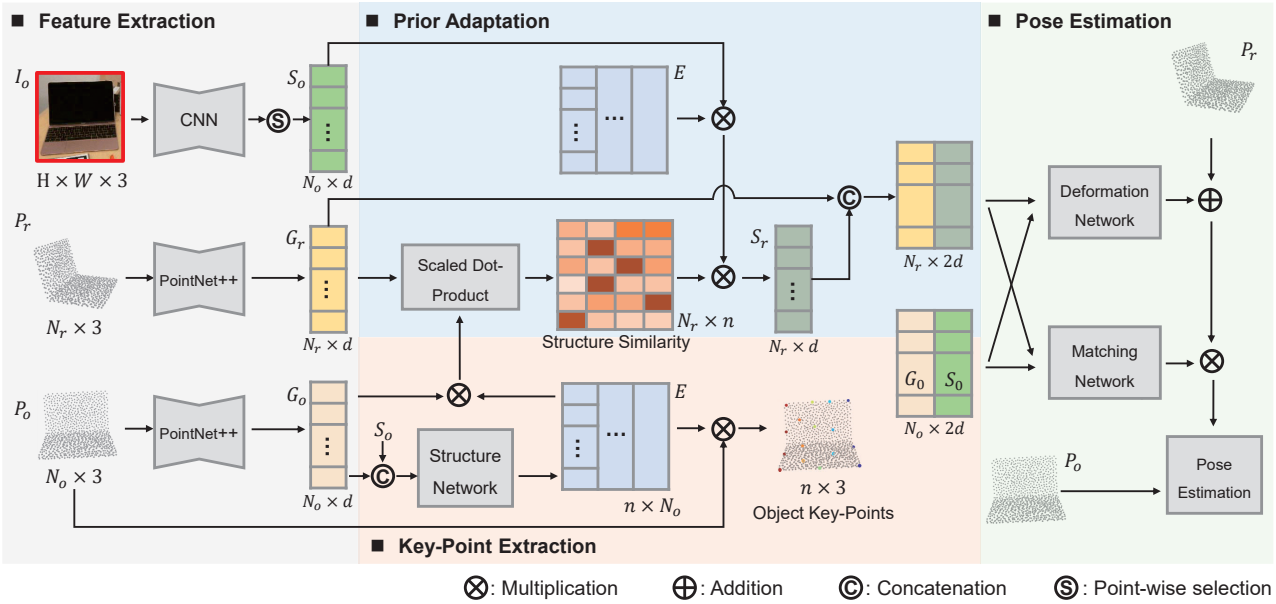

Figure 2. An overview of our proposed structure-guided prior adaptation (SGPA) network for category-level 6D object pose estimation.

cloud for each category of objects in the canonical space. Then, they deform the prior point cloud to reconstruct the canonical model for each instance. Wang et al. [33] further propose CR-Net, a cascaded relation and recurrent reconstruction network to leverage complementary advantages of multi-source inputs for categorical object pose estimation. The introduction of prior information significantly improves the overall performance. In this paper, we present a stronger prior based method by the proposed structure guided prior adaptation network.

3. Methodology

For a target object, let $P_o \in \mathbb{R}^{N_o \times 3}$ and $I_o \in \mathbb{R}^{h \times w \times 3}$ denote its observed point cloud and RGB image patch, where N_o and (h, w) denote the number of object points and the size of image patch. $P_r \in \mathbb{R}^{N_r \times 3}$ is the prior point cloud with N_r points that has the same category with the target object. Taking them as inputs, we present our prior based method for category-level 6D object pose estimation.

3.1. Overview

As illustrated in Figure 2, we propose the SGPA network for category-level object pose estimation. SGPA first uses a feature extraction module to extract object geometry feature $G_o \in \mathbb{R}^{N_o \times d}$, object semantic feature[1] $S_o \in \mathbb{R}^{N_o \times d}$ and prior geometry feature $G_r \in \mathbb{R}^{N_r \times d}$ from P_o, I_o, and P_r respectively. The learned features then would be fed into a structure guided prior adaptation module (Sec. 3.2). In this paper, we implement this prior adaptation module with a transformer-based architecture. It models the global structure similarity between P_o and P_r by correlating their geometry features G_o and G_r. Based on the extracted structure similarity, object semantic feature S_o then is adaptively propagated along the structure similarity from P_o to P_r for prior adaptation. Moreover, instead of densely correlating G_o and G_r on all point positions for prior adaptation, we further design an auxiliary network that predicts n object key-points from the input N_o points. A structure regularized low-rank transformer (Sec. 3.3) then is derived to extract structure similarity and perform prior adaptation based on features located on distinctive object key-points. With the adapted prior feature and the original object feature, a deformation network is utilized to reconstruct the canonical object model by deforming the prior P_r, and a matching network (Sec. 3.4) is adopted to match the reconstructed model with the object point cloud P_o. Finally, a correspondence based algorithm is applied to estimate pose parameters.

[1]Generated by point-wise selection on semantic feature maps.

3.2. Prior Adaptation through Structure Guidance

Given G_o and G_r, which are point-wise geometry features of a target object and its category prior, our SGPA globally correlates them to model the structure similarity between P_o and P_r. The learned structure similarity can be represented into $\mathcal{S} \in \mathbb{R}^{N_r \times N_o}$, in which each element s_{ij} corresponds to a structure similarity value between the point $p_i \in P_r$ and the point $p_j \in P_o$. Intuitively, the larger of the structure similarity value of s_{ij}, the more corresponding semantic feature from S_o should be propagated from p_j to p_i. In other words, we take the structure similarity as a guidance to conduct prior adaptation.

Specifically, we select to apply a transformer network [28] to implement the above scheme. The transformer

architecture recently is proved to be capable of capturing long-term dependencies from sequence / disordered data. The associated multi-head attention mechanism has a strong expressive ability. We leverage this advantage to model the high-level similarity between two disordered point clouds, and based on the similarity to conduct prior adaptation for category-level object pose estimation. In details, we take G_r, G_o and S_o as the *query*, *key*, and *value* of the multi-head attention module:

$$Y^{(m)} = \sigma(Q^{(m)}(K^{(m)})^{\mathsf{T}}/\sqrt{d})V^{(m)}, \quad (1)$$

where $Q^{(m)} = G_r W_Q^{(m)}$, $K^{(m)} = G_o W_K^{(m)}$, and $V^{(m)} = S_o W_V^{(m)}$, where $W_Q^{(m)}$, $W_K^{(m)}$ and $W_V^{(m)}$ all $\in \mathbb{R}^{d\times d}$. They are learnable projection matrices for query, key and value respectively. $\sigma(\cdot)$ denotes the standard *softmax* normalization function, which normalize the similarity value by *row*. $m = 1, 2, ..., M$ indexes the multi-head attention.

In each head, correlating $Q^{(m)}$ and $K^{(m)}$ computes a similarity between P_r and P_o in the projected embedding space. Multiplying the similarity and $K^{(m)}$ gets the semantic feature for the prior, which stems from the semantic feature of the observed object. By using in total M attention blocks, we can model the structure similarity between P_r and P_o comprehensively and fully transfer semantic features to the prior. We concatenate the output features of M attention blocks as:

$$Y = \text{Concat}(Y^{(1)}, Y^{(2)}, ..., Y^{(M)}). \quad (2)$$

Then, we feed Y into the feed forward network to get the adapted semantic feature for the prior $S_r = \text{FFN}(Y)$.

As shown in Figure 2, we subsequently concatenate G_r and S_r to get the complete prior feature after adaptation. In our transformer-based implementation, we resort to multi-attention to extract robust similarity of two geometry features, and then adaptively injecting the semantic feature into the prior. Through this prior feature adaptation, we not only enhance the prior feature with rich semantic features, but also adapt the fixed prior to the varying object instances.

3.3. Enhancing Prior Adaptation by Structure Regularized Low-Rank Transformer

The conventional transformer with the vanilla version of multi-head self-attention densely correlates two geometry features and propagates semantic features point by point. For 6D object pose estimation, this scheme might be neither efficient nor effective enough. On the one hand, the vanilla self-attention incurs a complexity of $O(n^2)$ with respect to the number of object points. Estimating pose for an object with more than thousands of points would be inefficient. On the other hand, for the captured object point cloud, not all positions are representative for prior adaptation, because the object point cloud would be noisy, incomplete and non-uniform in practical environments. To address these issues, in this section, we further describe a structure regularized low-rank transformer for prior adaptation.

Recently, with the rapid development of transformer networks, a lot of research works [1, 8, 34] focus on reducing the overhead of the self-attention. The low-rank transformer [34] is one typical solution, in which the conventional self-attention is replaced with a low-rank attention as:

$$Y^{(m)} = \sigma(Q^{(m)}(E_K^{(m)}K^{(m)})^{\mathsf{T}}/\sqrt{d})(E_V^{(m)}V^{(m)}), \quad (3)$$

where $E_K^{(m)}$ and $E_V^{(m)} \in \mathbb{R}^{n\times N_o}$, and $n \ll N_o$. They are two linear projection matrices that map *key* and *value* into a low-dimension space, so the self-attention map can be computed in a low-rank manner. The problem of this low-rank transformer is that there are not explicit regularization on $E_K^{(m)}$ and $E_V^{(m)}$. As discussed in [34], it reduces the overhead of self-attention with a cost of performance drop.

We argue that when applying the transformer network to object pose estimation, a proper regularization on $E_K^{(m)}$ and $E_V^{(m)}$ is capable of reducing the overhead and preserving (even improving) the pose accuracy simultaneously. We therefore design a novel structure regularized low-rank transformer network. As shown in Figure 2, we constrain the low-rank transformer with object key-points. Specifically, we introduce an auxiliary network to convert the original P_o with N_o points into n object key-points. Inspired by recent intrinsic point estimation [5], we associate P_o and n key-points by a projection matrix $E \in \mathbb{R}^{n\times N_o}$, which is estimated from the concatenated object features $[G_o, S_o]$. Once E is estimated, we set $E_K^{(m)} = E_V^{(m)} = E$ for the low-rank self-attention formulated in Eq. (3). Since E is learned to transform P_o into n key-points, projecting $K^{(m)}$ and $V^{(m)}$ with E can be seen as an approximation of object features on key-points, which effectively regularize the projected feature space to contain as informative as possible features for pose estimation with a reduced complexity.

3.4. Prior-based Object Pose Estimation

After prior adaptation, we get $F_o = [G_o, S_o]$ and $F_r = [G_r, S_r]$, which correspond to object feature and prior feature respectively. SGPA then utilizes two head networks for pose estimation. The first network is a deformation network. It aims to reconstruct the 3D canonical model for the target object, which is achieved by deforming P_r with a point-wise deformation field $D_r \in \mathbb{R}^{N_r\times 3}$:

$$P_r' = P_r + D_r = P_r + \mathcal{F}_d(F_o, F_r), \quad (4)$$

where $\mathcal{F}_d(\cdot)$ denotes the deformation network, and P_r' is the deformed prior point cloud (*a.k.a*, a reconstructed canonical model for the target object). The second network is a matching network. It softly associates P_r' with P_o by estimating a correspondence matrix M_r from P_r' to P_o as:

$$P_o' = M_r \times P_r' = \mathcal{F}_m(F_o, F_r) \times P_r', \quad (5)$$

where $\mathcal{F}_m(\cdot)$ denotes the matching network. $M_r \in \mathbb{R}^{N_o\times N_r}$ is a normalized matrix that compute N_o matched

Table 1. Comparison of our method with four RGBD based state-of-the-art methods on CAMERA25 and REAL275 benchmarks. SPD* denotes our own re-implementation result for SPD [26].

Method	CAMERA25						REAL275					
	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
NOCS [32]	83.9	69.5	32.3	40.9	48.2	64.6	78.0	30.1	7.2	10.0	13.8	25.2
CASS [4]	-	-	-	-	-	-	77.7	-	-	23.5	-	58.0
SPD [26]	93.2	83.1	54.3	59.0	73.3	81.5	77.3	53.2	19.3	21.4	43.2	54.1
SPD* [26]	93.0	85.5	58.1	62.9	75.9	83.8	80.0	56.7	20.0	22.3	45.3	57.9
CR-Net [33]	**93.8**	88.0	**72.0**	**76.4**	81.0	87.7	79.3	55.9	27.8	34.3	47.2	60.8
Ours	93.2	**88.1**	70.7	74.5	**82.7**	**88.4**	**80.1**	**61.9**	**35.9**	**39.6**	**61.3**	**70.7**

points around P'_r for pose estimation. P'_o denotes the predicted matched points that have a point-to-point correspondence relationship with P_o. Given P_o and P'_o, a correspondence based method [27] finally is adopted to jointly estimate object pose and size simultaneously.

3.5. Overall Loss Function

Overall, our SGPA has three estimation targets for 6D object pose estimation: the key-point transformation E, the point-wise deformation field D_r, and the correspondence matrix M_r. In order to train SGPA, we use the following loss function:

$$L = \lambda_1 L_{pose} + \lambda_2 L_{kp}. \tag{6}$$

For L_{pose}, we use the same loss function with SPD [26] for estimating D_r and M_r. The L_{pose} composes of four terms in total. Two of them are used to supervise the predicted D_r and M_r with the ground-truth object model and 6D pose, and the remaining two terms are used to further regularize the value range of D_r and M_r. Please refer to [26] for more detailed formulation of L_{pose}. The L_{kp} is defined as:

$$L_{kp} = \sum_{x_i \in P_o} \min_{y_j \in P_k} \|x_i - y_j\|_2^2 + \sum_{y_i \in P_k} \min_{x_i \in P_o} \|x_i - y_j\|_2^2, \tag{7}$$

where $P_k = E \times P_o$ is the extracted n object key-points. L_{kp} is formulated with the Chamfer Distance (CD) between P_o and P_k, which encourages E to represent a N_o-point model with n key-points where $n \ll N_o$. λ_1 and λ_2 are two balancing weights, which are set to 1.0 in our experiments.

3.6. Implementation Details

For the feature extraction module of SGPA, we use a pointnet++ [21] with four abstraction levels to extract geometry features, which have 512, 256, 128 and 64 centroids respectively. On each abstraction level, multi-scale grouping (MSG) is used to assemble multi-scale features. The MSG block in four levels have scales $(0.01, 0.02)$, $(0.02, 0.04)$, $(0.04, 0.08)$ and $(0.08, 0.16)$ respectively. A four-level PSP [37] network with ResNet-18 [11] as the backbone is used to extract object semantic features from the image patch. For the adaptation module, we use a single-layer transformer network with four self-attention heads as the basic prior adaptation network. The auxiliary key-point extraction network adopts a two-layer perceptron (MLP) block followed by a softmax activation layer. $\mathcal{F}_d$ and $\mathcal{F}_m$ use the same structure as [26] for regressing the deformation field and the correspondence matrix.

4. Experiments

4.1. Datasets

We evaluated our SGPA on both benchmarks of the virtual dataset CAMERA25 and the real dataset REAL275 [32]. Specifically, CAMERA25 has 300K synthetic RGB-D images which are generated by compositing virtual objects with real backgrounds. Among the 300K images, 25K images are used for testing. REAL275 contains 8K RGB-D images that are collected in 18 different real scenes, among which 7 scenes (4300 images) are used for training, 5 scenes (950 images) are used for validation, and the remaining 6 scenes (2750 images) are used for testing. Two datasets cover the same 6 object categories, i.e., *bottle, bowl, camera, can, laptop* and *mug*.

4.2. Experiment Settings

In most object pose estimation methods, to focus on the pose estimation algorithm, the instance segmentation and the subsequent pose estimation are decoupled. We also follow this scheme and generated the instance segmentation result[2] offline with an off-the-shelf network (e.g., Mask-RCNN [10]). After that, we crop the target object from the RGB-D image based on the segmentation result, and recover the instance point cloud by using camera intrinsic parameters. For the prior point cloud, we train an auto-encoder network on the ShapeNet dataset [3], and then we feed the average embedding of all instances that belong to the same category into the trained decoder to get the prior point cloud for this category. For both instance point cloud and prior point cloud, we uniformly sample them into 1024 points. In other words, $N_o = N_r = 1024$ in our experiments. Additionally, we fix $n = 256$ in our experiments, indicating that we use a fixed rank of 256 in our low-rank transformer when adopting it for prior adaptation. A detailed study on parameter n is given in our ablation study.

[2]We use the same segmentation result when conducting experiments.

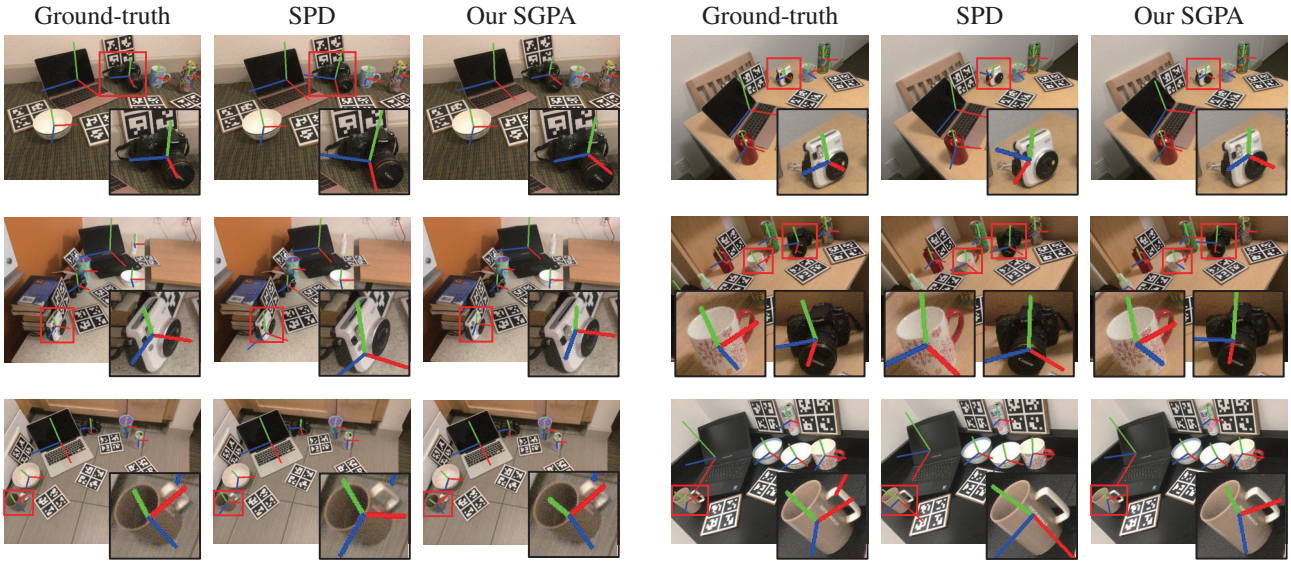

Figure 3. Qualitative comparison of SGPA with SPD on REAL275 dataset.

Table 2. Effects of each component of the proposed method on CAMERA25 and REAL275 benchmarks.

	Adaptation	Low-rank transformer	CAMERA25						REAL275					
			$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
1	-	-	93.0	85.6	61.5	65.3	79.0	85.7	79.8	59.2	23.8	28.4	50.7	62.5
2	✓	-	92.7	87.7	68.2	72.0	82.6	88.2	**80.4**	59.9	33.3	37.2	58.8	69.0
3	✓	✓	**93.2**	**88.1**	**70.7**	**74.5**	**82.7**	**88.4**	80.1	**61.9**	**35.9**	**39.6**	**61.3**	**70.7**

Table 3. Evaluation of the proposed prior adaptation method on CAMERA25 and REAL275 benchmarks when using different approaches to generate the prior point cloud.

Prior	CAMERA25						REAL275					
	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
Random	93.0	87.2	69.5	73.0	81.9	88.0	**81.3**	59.5	33.9	36.6	60.4	68.5
Nearest Neighbor	92.8	**88.2**	69.9	73.7	**83.2**	**88.4**	79.4	59.4	34.8	37.5	59.6	69.9
Embedding	**93.2**	88.1	**70.7**	**74.5**	82.7	**88.4**	80.1	**61.9**	**35.9**	**39.6**	**61.3**	**70.7**

4.3. Evaluation Metrics

Following the widely adopted evaluation scheme [4, 26, 32], we use two aspects of metrics to quantitatively evaluate the pose estimation performance:

- **3D IoU**. It computes the overlap of two 3D bounding boxes under the predicted pose and the ground truth pose respectively. If the ratio of overlapping is larger than a specified ratio, the prediction is judged to be correct. We use IoU_{50} and IoU_{75} for this metric.

- **Rotation and translation errors**. This metric directly computes the rotation and translation errors between the predicted pose and the ground truth pose. If the rotation error is smaller than an angle threshold and the translation is smaller than a distance threshold, the prediction is judged to be correct. We use $5°2cm$, $5°5cm$, $10°2cm$ and $10°5cm$ for this metric.

Given the above two metrics, we report the overall mAP across 6 object categories to compare the performance of different methods.

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

We compared our proposed method with four RGB-D based methods: NOCS [32], CASS [4], SPD [26], and CR-Net [33]. Table 1 gives the comparative results. On both datasets, our proposed method significantly outperforms other existing methods. In terms of IoU_{75}, $5°2cm$, and $10°5cm$, SGPA outperforms NOCS by 18.6%, 38.4%, and 23.8%, and outperforms SPD by 2.6%, 12.6% and 4.6% on the CAMERA25 dataset. The superiority of our method is more obvious on the REAL275 dataset. Specifically, SGPA achieves 61.9% mAP on IoU_{75}, 35.9% mAP on $5°2cm$, and 70.7% mAP on $10°5cm$, which are 31.8%, 28.7%, and 45.5% higher than NOCS, 5.2%, 15.9% and 12.8% higher than SPD, and 6.0%, 8.1% and 9.9% higher than recent CR-Net. These experimental results demonstrate the effectiveness of the proposed SGPA network. Figure 3 further presents qualitative comparison of SPD and our SGPA on the REAL275 dataset. SGPA outperforms SPD in handling geometrically complex objects, such as cameras and mugs. Moreover, Figure 4 presents a more detailed error evaluation result on two datasets. Especially, our SGPA is much more accurate than SPD in terms of rotation.

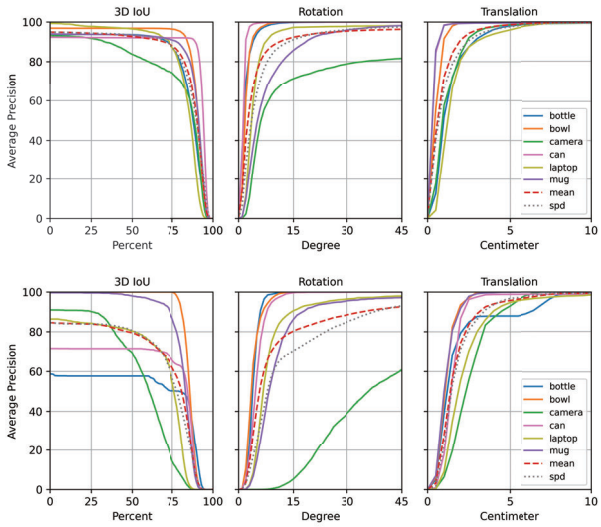

Figure 4. Average precision vs. error thresholds on CAMERA25 (top row) and REAL275 (bottom row).

Table 4. Evaluation of SGPA on CAMERA25 and REAL275 benchmarks when n is set to different values. 'full'=1024.

n	CAMERA25					
	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
16	92.6	86.9	68.2	71.9	**83.5**	86.5
32	92.5	86.6	68.7	72.3	82.1	87.2
64	92.9	87.2	67.3	71.3	82.1	87.6
128	92.5	86.7	68.5	72.3	82.2	87.5
256	**93.2**	**88.1**	**70.7**	**74.5**	82.7	**88.4**
512	92.6	87.8	69.9	73.5	82.7	87.8
full	92.7	87.7	68.2	72.0	82.6	88.2
n	REAL275					
	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
16	79.0	58.0	27.7	30.6	55.0	64.9
32	79.6	59.8	27.9	32.3	52.5	66.7
64	78.7	60.9	30.2	34.3	57.8	69.0
128	79.6	**62.5**	33.6	36.2	58.5	68.8
256	80.1	61.9	**35.9**	**39.6**	**61.3**	**70.7**
512	79.2	61.9	33.9	36.9	60.2	69.2
full	**80.4**	59.9	33.3	37.2	58.8	69.0

4.5. Ablation Studies

To justify the design choice of our method, we conducted the following ablation studies to our method:

- baseline. Not use any prior adaptation. Directly regressing the deformation field and the correspondence matrix from the extracted geometry and semantic features for pose estimation.
- w/ prior adaptation (Sec. 3.2). Conduct prior adaptation with the vanilla transformer network.
- w/ low-rank transformer (Sec. 3.3). Conduct prior adaptation with the proposed structure regularized low-rank transformer. $n = 256$ in this experiment.

Table 2 presents the result of our ablation study on both CAMERA25 and REAL275 datasets. Compared with the baseline method, prior adaptation can significantly improve the pose accuracy with a large margin. Specifically, a vanilla transformer based prior adaptation network improves the mAP of IoU_{75} and $5°2cm$ from 85.6% and 61.5% to 87.8% and 68.2% on the CAMERA25, and from 59.2% and 23.8% to 59.9% and 33.3% on the REAL275. Compared with the improvement on CAMERA25, the improvement on REAL275 is much more significant. It indicates that the proposed prior adaptation can effectively adapt the prior trained on a virtual dataset to the real environment for pose estimation. Moreover, replacing the vanilla transformer with our proposed structure regularized low-rank transformer can further improve the performance. The result demonstrates that through the guidance of object key-points, our low-rank transformer manages to leverage the most distinctive features for more effective prior adaptation, which leads to a higher pose accuracy.

Table 5. Comparison of the model reconstruction accuracy in CD metric ($\times 10^{-3}$).

Method	CAMERA25						
	bottle	bowl	camera	can	laptop	mug	mean
SPD*[26]	1.72	1.55	4.28	0.96	1.99	1.36	1.78
Ours	**1.35**	**1.30**	**3.33**	**0.87**	**1.20**	**1.17**	**1.42**
	REAL275						
	bottle	bowl	camera	can	laptop	mug	mean
SPD*[26]	4.34	1.21	8.30	1.80	2.10	**1.06**	2.99
Ours	**2.93**	**0.89**	**5.51**	**1.75**	**1.62**	1.12	**2.44**

Effect of adopted priors. Similar to SPD, we further investigate the performance of our SGPA when using different methods to generate the prior point cloud. Specifically, apart from training an encoder-decoder network on the ShapeNet to generate prior (denoted as 'Embedding' in Table 3), we use additional two different methods to generate the prior point cloud. For one, for each category, we refer to the instance model whose embedding is the closest to the average embedding of the encoder-decoder network as the prior (denoted as 'Nearest Neighbor'). For another, we randomly select one instance from each category and take its point cloud as the prior for the category (it is denoted as 'Random'). Table 3 presents the comparative results. Generally, our method is stable under different priors. This is because our prior adaptation method is based on the high-level structure similarity, which is robust to the specific prior generation method, as long as providing structure meaningful prior to the network.

Dimension of low-rank. In SGPA, we use a structure regularized low-rank transformer (see Sec. 3.3) to perform prior adaptation. In this experiment, we investigate the effect of different choices of n on the pose accuracy. We gradually reduce the value of n from 1024 to 16. Table 4 concludes the comparative results. On the CAMERA25 dataset, the pose result is relatively stable to the choice of n. When $n = 16$ or $n = 1024$, they produce nearly the same pose accuracy on CAMERA25. On the REAL275 dataset, the

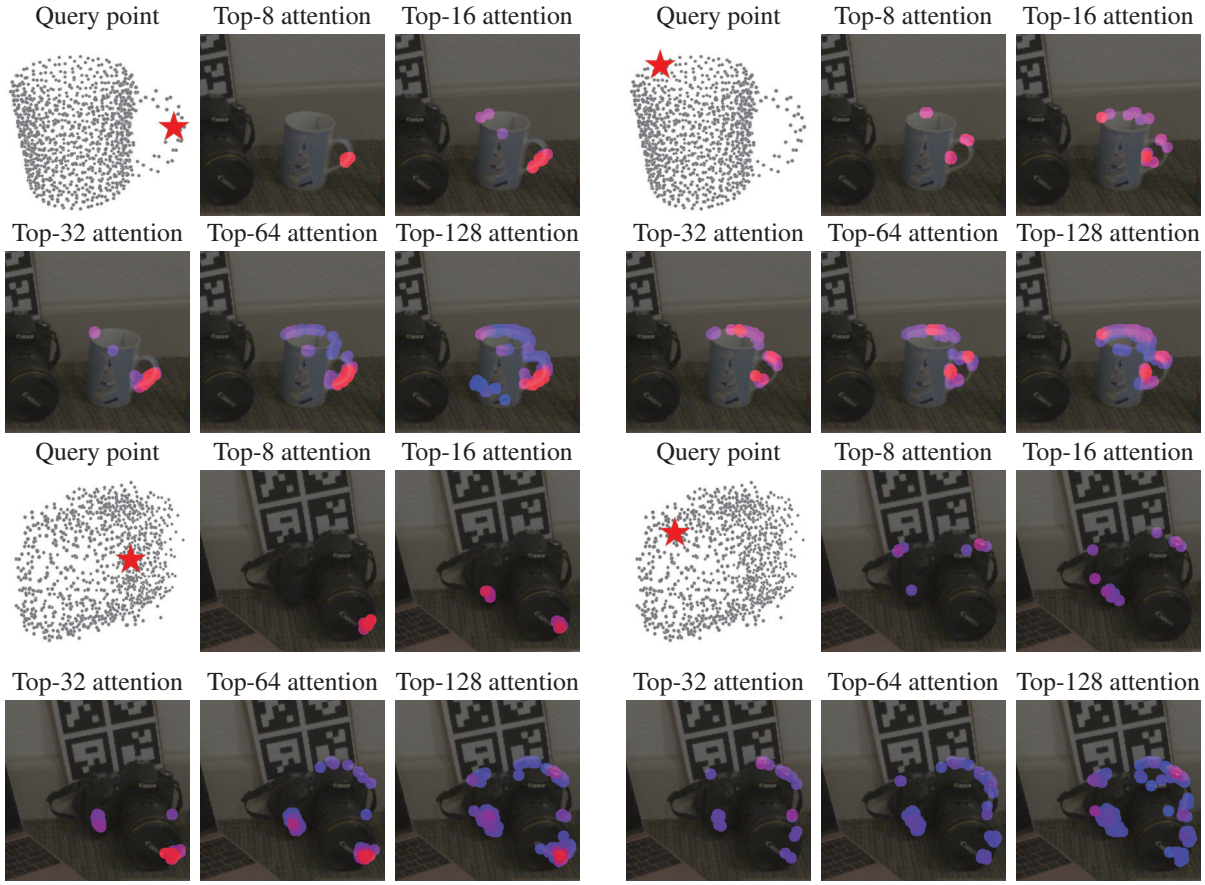

Figure 5. Visualization results of the learned attention maps from the adopted transformer network. We visualize the attention value on the position of the object point cloud. We project them on the image patch for more clear visualization. The color varies from blue to red corresponds to the attention value varies from small to large. The top attentions tend to be located at the matched region of the query point on the prior point cloud. From top-left to bottom-right, four query points are on mug handle, mug rim, camera lens, and camera body.

choice of n has a relatively large effect on the pose accuracy. When n is extremely small, we observed an obvious accuracy drop. Overall, when $N_o = N_r = 1024$, $n = 256$ receives the best performance on both datasets.

Figure 5 further visualizes the attention map learned by the transformer network, which indicates the learned relationship between the prior point cloud and object point cloud. For each point on the prior point cloud, we collect its point-wise attentions with the object point cloud and project them onto the image patch for a more clear display. As shown in Figure 5, for a query point on the prior point cloud, the learned attention tends to first focus on the corresponding part of the object (e.g., see the point on the handle of the mug and its top-8 attention map), and then spread to the whole object region to learn a global relationship (see the top-128 attention map). This result demonstrates that our network learns meaningful structure similarity between prior and object. Meanwhile, by adapting the prior feature to the observed target object through the learned structure similarity, our SGPA can also reconstruct the 3D model for the instance more accurately (see Table 5).

5. Conclusion

In conclusion, we present a novel structure-guided prior adaptation network for category-level 6D object pose estimation. It uses a transformer network to model the global structure similarity between prior and target object, based on which the object semantic information is injected into the prior feature to dynamically adapt the category-level prior to each particular object. We further propose a structure regularized low-rank transformer, in which we regularize the low-rank projection with the projection of point cloud key-points. The derived low-rank transformer therefore can leverage the feature on distinctive key-point positions for a more effective prior adaptation. Extensive experiments on two well-acknowledged benchmarks demonstrate that our method achieves dramatic performance improvements over other existing methods. This work is potentially useful for object perception and manipulation for robots, such as industrial robotics scenarios.

Acknowledgement. The work was supported by the Hong Kong Centre for Logistics Robotics.

References

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4

[2] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3364–3372, 2016. 2

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[4] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11973–11982, 2020. 1, 2, 5, 6

[5] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9121–9130, 2020. 4

[6] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[7] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8648–8657, 2019. 1

[8] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 4

[9] Xinke Deng, Yu Xiang, Arsalan Mousavian, Clemens Eppner, Timothy Bretl, and Dieter Fox. Self-supervised 6d object pose estimation for robot manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3665–3671. IEEE, 2020. 1

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[12] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11632–11641, 2020. 1, 2

[13] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1521–1529, 2017. 2

[14] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3706–3715, 2020. 2

[15] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 2

[16] Zhigang Li, Yinlin Hu, Mathieu Salzmann, and Xiangyang Ji. Robust rgb-based 6-dof pose estimation without real pose annotations. *arXiv preprint arXiv:2008.08391*, 2020. 2

[17] Fabian Manhardt, Wadim Kehl, Nassir Navab, and Federico Tombari. Deep model-based 6d pose refinement in rgb. In *European Conference on Computer Vision (ECCV)*, pages 800–815, 2018. 2

[18] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 55–64, 2020. 1

[19] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10710–10719, 2020. 2

[20] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4561–4570, 2019. 1, 2

[21] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. 5

[22] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. Instance-and category-level 6d object pose estimation. In *RGB-D Image Analysis and Processing*, pages 243–265. Springer, 2019. 2

[23] Caner Sahin and Tae-Kyun Kim. Category-level 6d object pose recovery in depth images. In *European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[24] Yongzhi Su, Jason Rambach, Nareg Minaskan, Paul Lesur, Alain Pagani, and Didier Stricker. Deep multi-state object pose estimation for augmented reality assembly. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 222–227. IEEE, 2019. 1

[25] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 292–301, 2018. 2

[26] Meng Tian, Marcelo H. Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 6, 7

[27] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):376–380, 1991. 2, 5

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[29] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066. IEEE, 2020. 2

[30] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3343–3352, 2019. 1, 2

[31] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 108–125. Springer, 2020. 2

[32] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2019. 1, 2, 5, 6

[33] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 1, 2, 3, 5, 6

[34] Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 4

[35] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1, 2

[36] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. *arXiv preprint arXiv:2012.05877*, 2020. 2

[37] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 5