SOCS: Semantically-aware Object Coordinate Space for Category-Level 6D Object Pose Estimation under Large Shape Variations

Boyan Wan* Yifei Shi* Kai Xu†
National University of Defense Technology
wanboyan@163.com {yifei.j.shi, kevin.kai.xu}@gmail.com

Abstract

Most learning-based approaches to category-level 6D pose estimation are design around normalized object coordinate space (NOCS). While being successful, NOCS-based methods become inaccurate and less robust when handling objects of a category containing significant intra-category shape variations. This is because the object coordinates induced by global and rigid alignment of objects are semantically incoherent, making the coordinate regression hard to learn and generalize. We propose Semantically-aware Object Coordinate Space (SOCS) built by warping-and-aligning the objects guided by a sparse set of keypoints with semantically meaningful correspondence. SOCS is semantically coherent: Any point on the surface of a object can be mapped to a semantically meaningful location in SOCS, allowing for accurate pose and size estimation under large shape variations. To learn effective coordinate regression to SOCS, we propose a novel multi-scale coordinate-based attention network. Evaluations demonstrate that our method is easy to train, well-generalizing for large intra-category shape variations and robust to inter-object occlusions. Code is provided at: https://github.com/wanboyan/SOCS.

1. Introduction

6D object pose estimation, i.e., determining the 3D rotation and translation of a object in the camera coordinate system, is an important computer vision task with a large body of literature [2, 27, 13, 17]. Category-level object pose estimation attempts to solve the problem without relying on the exact CAD model of the target object [28], which is hence more challenging than instance-level one. Since the seminal work of Wang et al. [28], most existing category-level works are based on a canonical representation of Normalized Object Coordinate Space (NOCS). Given an unseen object instance, they learn a neural network to map the perspective projection of the object to the NOCS of the corresponding category from which object pose can be estimated.

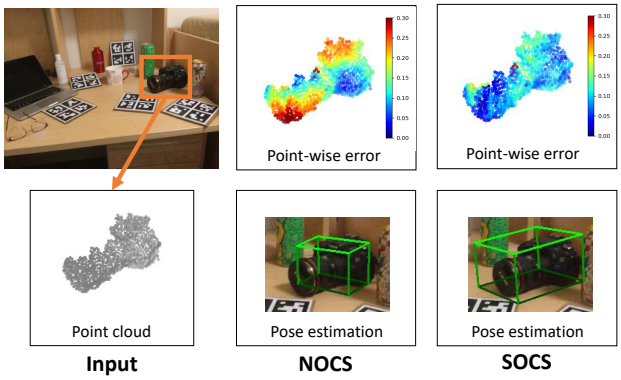


Figure 1. NOCS [28], constructed with globally aligned objects, finds difficulty in handling large intra-category shape variations. In this example, the coordinates regressed against NOCS for the long-lens camera contain much error (w.r.t. the CAD model under ground-truth pose and size) and the resulting pose is incorrect (middle). In contrast, the coordinates regressed against SOCS, built by semantically-guided non-rigid object alignment, are semantically coherent, leading to better pose estimation (right).

Given an object category, NOCS is defined by globally aligning a set of 3D object instances with normalized size and poses. It works well for objects with moderate intra-category shape variations. When handling object categories containing significant shape variations, however, NOCS-based methods become inaccurate and less robust. This is because the object coordinates induced by global and rigid alignment are not semantically coherent. For instance, a point on the lens of a long-lens camera would be mapped to a semantically incorrect point in NOCS if the NOCS was constructed with camera models of significantly varying part proportions. Such misalignment makes the mapping network hard to learn and generalize, thus causing inferior pose accuracy under large shape variations (Figure 1).

To tackle this issue, we propose Semantically-aware Object Coordinate Space (SOCS) to achieve accurate and robust category-level 6D object pose and size estimation un-

*Joint first authors
†Corresponding author

der large shape variations. Unlike NOCS which is constructed by directly aligning pose and size normalized objects of a specific category, SOCS is built by *warping-and-aligning the objects guided by a sparse set of keypoints with semantically meaningful correspondence*, leveraging the state-of-the-art category-specific keypoint selection and matching for a shape set [25]. In particular, we align all objects of a specific category in the training set to the average shape [26] of the set. We utilize 3D thin-plate spline warping [9] to ensure a smooth non-rigid deformation and hence coordinate interpolation. SOCS is therefore semantically coherent: Any point on the surface of a object can be mapped to a semantically meaningful location in SOCS, allowing for accurate pose and size estimation.

To learn the mapping from image space to SOCS effectively, we propose a novel multi-scale coordinate-based attention network. To capture the shape variation of the target object in image space, we devise a multi-scale feature extraction network with cross-attention feature aggregation. In the cross-attention module, we encode global point positions to help better extract coordinate-sensitive features. Thanks to such global positional encoding, our network is able to model 3D points in the full space, which further enables a dense point sampling in SOCS training. The latter facilitates dense coordinate estimation even for unobserved locations, which is critical to handling inter-object occlusions. To attain pose invariance, the network is trained in a contrastive fashion with a pose consistency loss.

We conducted extensive evaluations demonstrating that our method is 1) easy to train, 2) well-generalizing for large intra-category shape variations, and 3) robust to inter-object occlusions. Even with the vanilla mapping network of [28], our method is still comparable to state of the arts, clearly showing the effectiveness of SOCS. Our full method achieves state-of-the-art on the NOCS-REAL275 and ModelNet40-partial datasets, improving the $5°5$cm score by 5.6 pts on NOCS-REAL275 and $5°0.05$ score by 16 pts on ModelNet40-partial. In particular, ModelNet40-partial contains categories containing objects with large shape variations.

In summary, our work makes two contributions. *First*, we propose semantically-aligned object coordinate space (SOCS) to accommodate large intra-category shape variations for semantically coherent coordinate regression. *Second*, we propose a multi-scale attention network for learning the mapping from image space to SOCS effectively allowing for dense coordinate regression.

2. Related Work

2.1. Category-level pose estimation

Category-level pose estimation aims to predict the pose of unseen instances from a single-view image without knowing their 3D model. Existing work can be roughly classified into direct regression and correspondence-based methods. Direct regression methods estimate object pose by extracting pose-sensitive features from the input [32, 19]. The recent research focuses on exploring advanced network architectures [4], proper learning schemes [8], and different output representations [7]. Crucially, DualPoseNet [20] adopts two parallel pose decoders on top of a shared pose encoder, learning the consistency between the two brunches to impose complementary supervision. FS-Net [7] proposes a decoupled rotation output mechanism to complementarily estimate the rotation components. Correspondence-based methods first estimate the correspondence between the observed points and its coordinate in the canonical space and then optimize pose and size by postprocessing. This requires methods to extract pose-invariant point features. Wang et al. [28] present the representation of NOCS to enable the learning of pose for unseen objects. Wen et al. [29] introduce NUNOCS, which allows non-uniform scaling across three dimensions, facilitating fine-grained dense correspondences across object instances with large shape variations. Crucially, a bunch of recent works have adopted the categorical shape prior to facilitating the computation of correspondences between the observed points and their canonical coordinate [26, 5, 18]. Our method falls into the category of correspondence-based methods. However, it is different from previous work as it learns semantically-aware dense correspondences, resulting in more accurate results.

2.2. Implicit field for pose estimation

Many recent works have investigated implicitly representing 3D shapes with a continuous and differentiable implicit field implemented by neural networks. While most of the research in this field focuses on shape reconstruction, a handful of methods adopt implicit fields to estimate object pose [24, 1]. A straightforward way is to jointly reconstruct the object surface and estimate its pose [3, 23, 15, 30] with a unified framework. For example, ShAPO [12] jointly predicts object shape, pose, and size in a single-shot manner. Neural Radiance Fields (NeRF) [22] provides a mechanism for capturing complex 3D structures from only one or a few RGB images, which is also applicable to object pose estimation. iNeRF [31] estimates pose for objects with complex geometry with a pre-trained NeRF model. NeRF-Pose [14] first reconstructs the object with NeRF and then estimates the object pose. Unlike the traditional correspondence-based methods which predict 3D object coordinates at pixels of the input image, Huang et al. [11] predict canonical coordinates at any sampled 3D in the camera frustum, generating continuous neural implicit fields of canonical coordinates for instance-level pose estimation. Despite the similarity in the general concept, our method tackles the problem of category-level pose estimation where semantically-

ware cross-instance correspondences need to be estimated.

3. Method

Overview. In this section, we first describe how to generate SOCS. Then, we present the multi-scale coordinate-based attention network for SOCS estimation. In particular, a surface-independent point sampling strategy and a pose-invariant feature extraction training scheme are introduced. Last, we elaborate on the details of network inference and pose estimation.

3.1. SOCS

Existing canonical coordinate spaces for category-level 6D pose estimation, such as NOCS [28], are induced by global and rigid alignment, leading to semantic incoherency on the object coordinates. When handling object categories containing significant shape variations, NOCS-based methods become inaccurate and less robust. We introduce Semantically-aware Object Coordinate Space (SOCS) to alleviate the problem of NOCS. The coordinates in SOCS are generated by the category-specific keypoints, allowing fine-grained non-rigid coordinate alignment.

Specifically, given the shapes $\{S_i\}$ of a category in the training set, we first generate the categorical average shape S^a by using the pre-learned autoencoder [26]. The coordinates of S^a are regarded as the coordinates of SOCS. To build correspondence between the object coordinate of any object instance $\{S_i\}$ and the SOCS, we detect the semantically consistent keypoints $\{K_i\}$ and K^a for $\{S_i\}$ and S^a, respectively, by using the Skeleton Merger [25]. We denote the detected keypoints $K_i = \{k_j\}$, $j \in [1, m]$ and $K^a = \{k_j^a\}$, $j \in [1, m]$. m is the number of keypoints in a single shape, which is unified for all shapes.

Next, we compute the dense correspondence between S_i and S^a by considering the alignment of the semantically consistent keypoints. This is achieved by a 3D thin plate spline warping function [9]:

$$\begin{gathered} \Phi(x) = c + b^T x + w^T \mathbf{s}(x), \\ \mathbf{s}(x) = [\sigma(x - k_1^a), \sigma(x - k_2^a), \cdots, \sigma(x - k_m^a)]^T, \\ \sigma(x) = \|x\|_2^2 \cdot \log \|x\|_2, \end{gathered} \tag{1}$$

where $c \in \mathbb{R}^3$, $b \in \mathbb{R}^{3\times3}$, $w \in \mathbb{R}^{m\times3}$ are the parameters which are determined by optimizing the following function:

$$\min \sum_{j=1}^{m} \left\| k_j^a - \Phi(k_j) \right\|^2. \tag{2}$$

Once the parameters c, b, and w are determined, for any coordinate x in the object coordinate space, its SOCS is computed as $x^{\text{SOCS}} = \Phi(x)$.

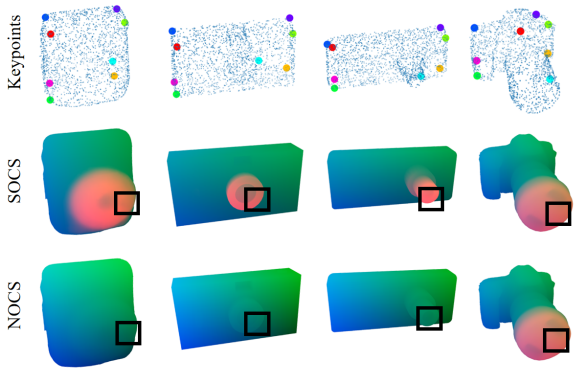


Figure 2. The proposed SOCS is more semantically meaningful, facilitating the learning of correspondence for objects with large shape variations. The coordinates in the canonical space (represented by the color) of the same semantic part in different objects are similar in SOCS (Middle row) and dissimilar in NOCS (Bottom row). Please pay special attention to the highlighted regions.

Compared to the NOCS representation and its variants [29] developed for category-level 6D pose and size estimation, the SOCS is more semantically meaningful, thus facilitating the learning of correspondence even for objects with large shape variations. See Figure 2 for an illustration.

3.2. Training of SOCS Estimation Network

In this section, we describe how to estimate the point-wise SOCS from an image. Estimating SOCS from a single-view image is non-trivial due to the potential large shape variations and the inter-object occlusions. To learn the mapping from input points to SOCS effectively, we propose a novel multi-scale coordinate-based attention network. An overview of the network architecture is shown in Figure 3.

Multi-scale coordinate-based attention network. The network contains two main components: *aggregation layers* and *propagation layers*. The *aggregation layers* extract per-point features from the point cloud. The point cloud is cropped from the depth image of the detected object. Since the task of category-level pose estimation can be challenging due to the large shape variations and severe occlusion of the input point cloud, we take 3D-GCN [21], which is able to aggregate contextual information of 3D point clouds with good performance, as the backbone. To be specific, the 3D-GCN takes the point cloud $\mathcal{P} \in \mathbb{R}^{n\times3}$ as input, generates downsampled points $\mathcal{P}^\alpha$ and extracts features $\mathcal{F}^\alpha$ at the α-th block, where $\alpha \in \{1, ..., 5\}$. Note that the aggregation layers could be any other 3D point-based network backbone according to the practical requirements. We found that 3D-GCN works best in our problem setting.

The *propagation layers* are then developed to propagate the feature from the downsampled points $\{\mathcal{P}^\alpha\}, \alpha \in \{1, ..., 5\}$ to any query point x and estimate its SOCS x^{SOCS}.

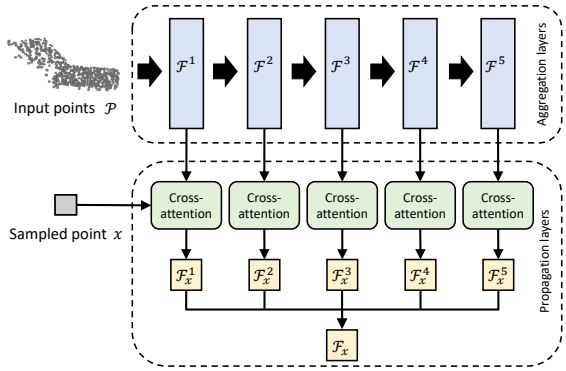


Figure 3. Given an input point cloud $\mathcal{P}$, the aggregation layers generate the features $\{\mathcal{F}^\alpha\}$, $\alpha \in [1, 5]$ at multiple network blocks. The propagation layers then propagate features from the input points to the sampled point x.

Extracting the feature of unseen points is a non-trivial task, due to the possibly infinite query locations in the 3D space. The ideal extracted feature should be both context-sensitive and coordinate-sensitive. To achieve this, we propose an implicit neural network with coordinate-based multi-scale contextual feature propagations.

We first initialize the feature vector at query point x as a zero vector, i.e., $\mathcal{F}_x^0 = \mathbf{0}$, $\mathcal{F}_x^0 \in \mathbb{R}^h$, where h is the feature length. For each block, we update the feature with a cross-attention module to aggregate feature from the nearest points (see Figure 4). Specifically, at the α-th block, we compute the k-nearest neighbors $\mathcal{N}^\alpha \in \mathbb{R}^{k\times 3}$ of x from $\mathcal{P}^\alpha$, where k is 16. We denote the features of $\mathcal{N}^\alpha$ as $\mathcal{F}_\mathcal{N}^\alpha \in \mathbb{R}^{k\times h}$. Moreover, we introduce a global point $g^\alpha = \text{Mean}(\mathcal{P}^\alpha)$ with the feature being $\mathcal{F}_g^\alpha = \text{Mean}(\mathcal{F}^\alpha)$, where $\text{Mean}(\cdot)$ is the element-wise averaging operation. The global positional encoding provides contextual information, facilitating dense coordinate estimation even on unobserved locations, which is critical to handling occlusions.

The update term on the feature $\mathcal{F}_x^{\alpha-1}$ is estimated by considering the relations to both $\mathcal{N}^\alpha$ and g^α:

$$\Delta^\alpha = \text{Softmax}\left(\frac{\left(\mathcal{F}_\mathcal{N}^\alpha W_k\right)\left(\mathcal{F}_x^{\alpha-1}W_q\right)^{\mathrm{T}} + R}{\sqrt{\mathrm{h}}}\right)\mathcal{F}_\mathcal{N}^\alpha W_v + \text{Softmax}\left(\frac{\left(\mathcal{F}_g^\alpha W_k\right)\left(\mathcal{F}_x^{\alpha-1}W_q\right)^{\mathrm{T}} + r_g}{\sqrt{\mathrm{h}}}\right)\mathcal{F}_g^\alpha W_v \tag{3}$$

where W_q, W_k, $W_v \in \mathbb{R}^{h\times h}$ are the learnable weights. $r \in \mathbb{R}^k$ denotes the influence factor of x to the points in $\mathcal{N}^\alpha$. Each element r_i in R is computed as follows by considering the relative position between the query point x and the nearest neighbor i:

$$r_i = \text{EmbLayer}(x - \mathcal{N}_i^\alpha), \tag{4}$$

where $\text{EmbLayer}(\cdot)$ is a two-layer MLP. Similarly, r_g is computed to capture the position w.r.t. the center of input points:

$$r_g = \text{EmbLayer}(x - g^\alpha). \tag{5}$$

Then, the updated feature at point x is:

$$\mathcal{F}_x^\alpha = \text{LayerNorm}(\mathcal{F}_x^{\alpha-1} + \Delta^\alpha) \tag{6}$$

where $\text{LayerNorm}(\cdot)$ is the layer normalization operation.

The extracted features at all the blocks are then concatenated: $\mathcal{F}_x = \text{Concat}(\mathcal{F}_x^1, \mathcal{F}_x^2, \mathcal{F}_x^3, \mathcal{F}_x^4, \mathcal{F}_x^5)$. The concatenated feature is utilized to estimate the SOCS:

$$\begin{aligned} x_\mathrm{X}^{\text{SOCS}} &= \text{Softmax}(\text{MLP}_\mathrm{X}(\mathcal{F}_x)), \\ x_\mathrm{Y}^{\text{SOCS}} &= \text{Softmax}(\text{MLP}_\mathrm{Y}(\mathcal{F}_x)), \\ x_\mathrm{Z}^{\text{SOCS}} &= \text{Softmax}(\text{MLP}_\mathrm{Z}(\mathcal{F}_x)), \end{aligned} \tag{7}$$

where $x_\mathrm{X}^{\text{SOCS}}$, $x_\mathrm{Y}^{\text{SOCS}}$, $x_\mathrm{Z}^{\text{SOCS}}$ are the predicted class denoting the coordinate in the axes of X, Y, Z, respectively. $\text{MLP}_\mathrm{X}(\cdot)$, $\text{MLP}_\mathrm{Y}(\cdot)$, $\text{MLP}_\mathrm{Z}(\cdot)$ represent multi-layer perceptrons.

Note that, our method is different from most existing methods where regression or classification with a small number of bins ($\mathrm{B} < 50$) for coordinate estimation are adopted. We found that using a larger number of bins (e.g. $\mathrm{B} = 256$) in our method will not lead to the training being inefficient or failing to converge. The advantage comes from the representation of SOCS, which greatly reduces the learning complexity.

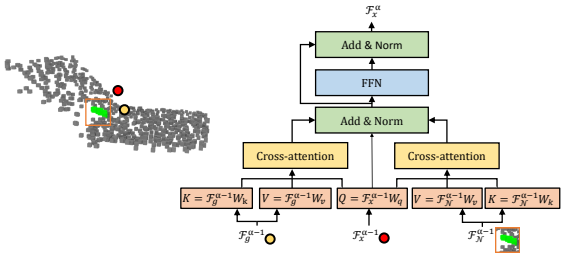
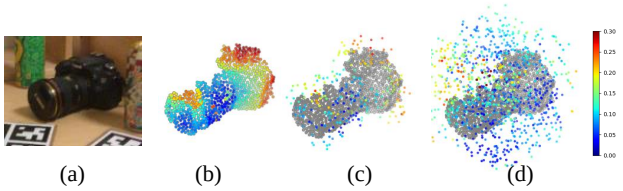


Figure 4. The network architecture of the propagation layers. For a sampled point (gray), the query point (red) updates its feature by performing cross-attention operations with the k-nearest neighbors (green) and the global point (yellow), respectively.

Surface-independent point sampling. We then describe how to sample points to feed into the multi-scale coordinate-based attention network. Several sampling strategies could be considered: 1) Sampling from the input points; 2) Surface-dependent sampling: random sampling near the input points; 3) Surface-independent sampling: random sampling in the whole 3D space. We empirically found the surface-independent sampling strategy outperforms the others, thanks to the mechanism of global positional encoding. There are two reasons for this phenomenon. First, sampling in the whole space facilitates feature aggregation in the invisible region, bringing more



(a) (b) (c) (d)

Figure 5. Illustration of the point sampling strategies. (a) The input object. (b) Sampling from the input points. (c) Surface-dependent sampling. (d) Surface-independent sampling. The color denotes the point-wise SOCS estimation error.

context information. Second, sampling in the whole space would decrease the overall pose estimation uncertainty, especially in scenarios where severe occlusion exists. Illustration of the sampling strategies is visualized in Figure 5.

Network training. Next, we describe how to train the above network. Suppose $\mathcal{X}$ is the set of sampled points, a naive loss function of the shape correspondence field estimation could be:

$$\mathcal{L}_{\text{SOCS}} = \sum_{x \in \mathcal{X}} [\mathcal{L}_{\text{CE}}(x_{\text{X}}^{\text{SOCS}}, \hat{x}_{\text{X}}^{\text{SOCS}}) + \mathcal{L}_{\text{CE}}(x_{\text{Y}}^{\text{SOCS}}, \hat{x}_{\text{Y}}^{\text{SOCS}}) + \mathcal{L}_{\text{CE}}(x_{\text{Z}}^{\text{SOCS}}, \hat{x}_{\text{Z}}^{\text{SOCS}})], \tag{8}$$

where $\mathcal{L}_{\text{CE}}(\cdot)$ is the cross entropy loss, $\hat{x}_{\text{X}}^{\text{SOCS}}$, $\hat{x}_{\text{Y}}^{\text{SOCS}}$, $\hat{x}_{\text{Z}}^{\text{SOCS}}$ denote the ground-truth. However, we found that training the network is unstable and hard to converge, especially on categories with large shape variations.

To alleviate this issue, we adopt a contrastive training fashion with a pose consistency loss to further enhance the training. The key insight is to learn the *pose-invariant feature* by transforming the input point cloud, extracting its per-point features, and making the features of the initial point cloud and the transformed point cloud consistent.

Specifically, during training, we transform the input points $\mathcal{P}$ with a random rigid transformation $\mathbf{T}_{\text{r}} = \{\mathbf{R}|\mathbf{t}\}$. We denote the transformed point cloud as $\mathcal{P}^{'} = \mathbf{T}_{\text{r}} \cdot \mathcal{P}$. $\mathcal{P}$ and $\mathcal{P}^{'}$ are then fed into the multi-scale coordinate-based attention network, respectively, to generate the per-point features. For any point $x \in \mathcal{X}$ and the transformed point $x^{'} = \mathbf{T}_{\text{r}} \cdot x$, the generated features should be consistent:

$$\mathcal{L}_{\text{consistency}} = \sum_{x \in \mathcal{X}} ||\mathcal{F}_x - \mathcal{F}^{'}_{x^{'}}||_2, \tag{9}$$

where $\mathcal{F}_x$ and $\mathcal{F}^{'}_{x^{'}}$ denote the extracted feature by the two network towers, respectively. Overall, the training loss function is: $\mathcal{L} = w_{\text{SOCS}}\mathcal{L}_{\text{SOCS}} + w_{\text{consistency}}\mathcal{L}_{\text{consistency}}$, where w_{SOCS} and $w_{\text{consistency}}$ are the pre-defined weights.

Training data preparation. To generate the training data of SOCS estimation, for each dataset, we first generate the dense SOCS for the complete 3D objects using the method described in Sec. 3.1. The dense SOCS are then transformed into the camera coordinates with the 6D object pose.

3.3. Network Inference, Pose and Size Estimation

Network inference. During the inference, given an RGB-D image with untrained object instances in it, we first perform an object detection with Mask R-CNN [10]. For each detected object, we crop the image, generate the point cloud, and feed it into the aggregation layers to generate features. Then, we densely sample points in 3D space around the input points, and extract their features with the propagation layers, predicting the SOCS for every sampled point.

Pose and size estimation. The predicted per-point coordinate in SOCS is then transformed into the camera coordinate space with the transformation of the 6D object pose and a scaling operation. The ideal transformation matrix $\mathbf{T} \in \mathbb{R}^{4\times4}$ of the pose and the scaling matrix $\mathbf{S} = \text{diag}(s_{\text{X}}, s_{\text{Y}}, s_{\text{Z}}, 1)$ should make the following function optimized:

$$\min \sum_{x \in \mathcal{X}} \|\mathbf{T} \cdot \mathbf{S} \cdot \Phi(x) - x\|^2, \tag{10}$$

where $\mathcal{X}$ represents the sampled points. Note that the scaling matrix $\mathbf{S}$ is anisotropic, so it allows more flexible and accurate size estimation compared to NOCS whose scaling matrix is isotropic. To achieve convergence toward a global optimum, we sample multiple initial T and S, optimize them respectively, and select the best one.

3.4. Implementation details

We detected 32 keypoints on each object. The multi-scale coordinate-based attention network takes $1,024$ points as input. The number of classification bins is 128. The network is optimized by a ranger optimizer, with batch size 16 and learning rate 0.001. The learning rate is annealed at 50% of the training phase using a cosine schedule. We train individual models for each category respectively. In the surface-independent sampling, we randomly sample points in a sphere with the center being the center of input points and its diameter being the diagonal length of the largest shape in the category. We set w_{SOCS} as 1 and $w_{\text{consistency}}$ as 0.1. Our method is trained and tested on an NVIDIA Tesla V100. The training takes 10 hours to converge. The inference time of one image is about 0.6 second.

4. Results and Evaluation

4.1. Experimental Datasets

We train and test our method on the NOCS-REAL275 [28] and ModelNet40-partial [16] datasets. The NOCS-REAL275 contains 4.3k training RGB-D images

Figure 6. Visual comparison of the estimated pose by our method, RBP-Pose [32], and DPDN [19].

and 2.75k testing RGB-D images captured from 6 real-world scenes. The objects belong to six object categories: bottle, bowl, can, camera, laptop, and mug. The ModelNet40-partial dataset is a synthetic dataset that contains 60k training depth images and 6k testing depth images. It contains object categories with large shape variations, such as *airplane*, *chair*, and *sofa*.

4.2. Evaluation Metrics

We use standard metrics to evaluate the performance on the two datasets, respectively. For NOCS-REAL275, we adopt the intersection over union (IoU) with a threshold of e, and the average precision of instances for which the error is less than n° for rotation and m for translation. For ModelNet40-partial, we report the rotational error, and the translational error in the form of mean, and median values. We also report the average precision of instances for which the error is less than 5° for rotation and 5cm for translation.

4.3. Performance on NOCS-REAL275

We first compare our method with the state-of-the-art on the NOCS-REAL275 dataset. The quantitative results are shown in Table 1. There are several phenomena we can observe. First, DPDN [19] slightly outperforms our method on metrics of IoU50 and IoU75, showing that their method is better than ours in terms of object detection. Second, our method outperforms all the baselines on metrics of 5°2cm, 5°5cm, 10°2cm, 10°5cm, demonstrating the effectiveness of our method on pose estimation despite the inferiority on object detection. In particular, to further study the effectiveness of the proposed SOCS and the proposed network, we replace each of them with NOCS and the network in [27] respectively (i.e., the baseline of Network in [28] + SOCS est. and Our network + NOCS est.), and conduct experiments. The results show that our full method is better than the two baselines, revealing the necessity of both the SOCS and the proposed network. We also found our method requires less training time compared to the baseline of Our network + NOCS est., demonstrating SOCS is easy to train compared to NOCS. The qualitative comparisons to the state-of-the-art are visualized in Figure 6.

4.4. Performance on ModelNet40-partial

To demonstrate the performance of our method under *large shape variations*, we conduct experiments on the ModelNet40-partial dataset. The results are reported in Table 2. We see that our method outperforms all baselines by a large margin over all the metrics, which suggests that our method is much more effective in handling categories with large shape variations. Moreover, to quantitatively analyze how our method performs under different shape variations, we conduct an additional experiment. Specifically, we generate several subsets of the *lamp* category in the ModelNet40 dataset with different degrees of shape variations. The degree of shape variations is computed as the average

Table 1. Quantitative results on the NOCS-REAL275 dataset.

Methods	Data type	Data source	IoU50 ↑	IoU75↑	5°2cm↑	5°5cm↑	10°2cm↑	10°5cm↑
NOCS [28]	RGB	Syn.+Real	0.78	0.30	0.07	0.10	0.14	0.25
SGPA [5]	RGB-D	Syn.+Real	0.80	0.62	0.36	0.40	0.61	0.71
DPDN [19]	RGB-D	Syn.+Real	**0.83**	**0.76**	0.46	0.51	0.70	0.78
GPV-Pose [8]	D	Real	0.83	0.64	0.32	0.43	-	0.73
RBP-Pose [32]	D	Real	0.83	0.68	0.38	0.48	0.63	0.79
Network in [28] + SOCS est.	RGB	Real	0.79	0.41	0.11	0.12	0.15	0.30
Our network + NOCS est.	D	Real	0.82	0.73	0.40	0.49	0.64	0.81
Ours	D	Real	0.82	0.75	**0.49**	**0.56**	**0.72**	**0.82**

Table 2. Quantitative results on the ModelNet40-partial dataset.

Methods	Data type	Data source	Rotation			Translation		
			Mean(°) ↓	Median(°)↓	5° ↑	Mean()↓	Median()↓	5°0.05 ↑
EPN [16]	D	Syn.	32.86	23.84	0.49	0.14	0.13	0.08
KPConv [16]	D	Syn.	37.48	30.86	0.24	0.11	0.08	0.06
GPV-Pose [8]	D	Syn.	30.75	30.41	0.28	0.17	0.11	0.06
RBP-Pose [32]	D	Syn.	33.09	35.25	0.26	0.08	0.13	0.10
Ours	D	Syn.	**22.53**	22.81	**0.59**	**0.03**	**0.07**	**0.26**

Table 3. Ablation studies of the key components.

MP	GP	CL	Sampling	IoU75↑	10°2cm↑
✓	✓	SI	0.59	0.56	
✓	-	✓	SI	0.63	0.58
✓	✓	-	SI	0.65	0.60
✓	✓	✓	P	0.67	0.62
✓	✓	✓	SD	0.66	0.63
✓	✓	✓	SI	**0.75**	**0.72**

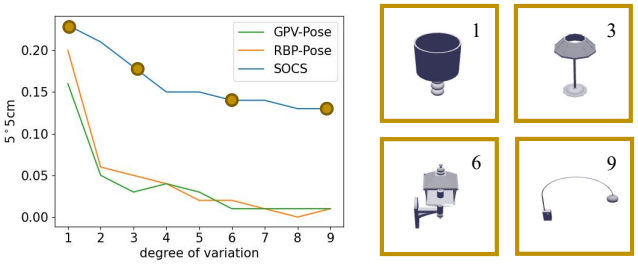


Figure 7. Left: Comparisons on subsets with different degrees of shape variations. We see our method outperforms the baselines on all subsets. Right: Examples of objects instances of different degrees of variations.

chamfer distance between every shape instance and the categorical shape prior. The results are visualized in Figure 7. It shows that our method is able to handle object instances with large shape variations, while the baselines cannot.

4.5. Ablation Studies and Parameter Setting

In Table 3, we conduct ablation and parameter setting studies to quantify the efficacy of the key components in our method. In Table 4 and 5, we study the key parameter settings. All the experiments are conducted on the NOCS-REAL275 dataset.

Network architecture. We then study the necessity of crucial network modules, i.e., the multi-block feature propagation (MP), the global position encoding in cross-attention (GP), and the consistency loss function (CL). In the experiment, we remove these crucial modules respectively, retrain the networks, and evaluate the performances. Note that, in the ablation baseline of MP, we only perform feature propagations at the last block, so it has no multi-block contextual features. The experiments show that adding any of the modules would lead to a performance improvement, confirming the effectiveness of these modules.

Sampling strategy. We have discussed the advantages of the surface-independent sampling (SI) strategy in Section 2.2. Here, we quantitatively compare it with the alternatives of sampling from the input points (P) and surface-dependent sampling (SD). We see that the network trained by the surface-independent sampling strategy outperforms the rest. Moreover, we visualize the per-point SOCS estimation error in a cross-section in Figure 8. It is clear that the estimation in most of the unseen regions is as accurate

Table 4. Effect of different numbers of keypoints.

#keypoints	IoU75↑	5°2cm↑	5°5cm↑	10°2cm↑
8	0.72	0.42	0.50	0.64
16	0.72	0.46	0.54	0.68
32	0.75	**0.49**	**0.56**	**0.72**
64	**0.75**	0.47	**0.56**	0.68
32 (ISRP [6])	0.71	0.42	0.48	0.65

Table 5. Effects of different numbers of classification bins.

#bins	IoU75 ↑	5°2cm↑	5°5cm↑	10°2cm ↑
32	0.70	0.44	0.52	0.61
64	0.71	0.47	0.55	0.64
128	**0.75**	**0.49**	**0.56**	**0.72**
256	0.73	**0.49**	0.55	**0.72**
Regression	0.69	0.43	0.50	0.66

as that near the observed surface, showing the necessity of surface-independent sampling and the efficacy of our feature propagation mechanism.

Number of keypoints. The number of keypoints is a crucial parameter that has the potential to influence the effects of SOCS. We conduct several experiments using different numbers of keypoints to generate the SOCS and retrain our network. As reported in Table 4, we see that using a relatively small number of keypoints would lead to a significant performance decrease. The reason might be that an insufficient number of keypoints would lead to inaccurate dense correspondence between object instances. We also tried to adopt an alternative keypoints extraction method, i.e., ISRP [6], instead of Skeleton Merger [25]. Results show the alternative key-point extraction method is also applicable to our method but will lead to inferior performances, implying that the quality of keypoints is crucial to our method.

Number of classification bins. We conduct a comparison with baselines that use different numbers of bins in the coordinate classification, as well as a baseline that replaces the classification with regression. As reported in Table 5, we found that using classification is a better choice compared to using regression. Besides, the performances reach their peak when the number of classification bins is 128 or 256, showing that our method is able to be compatible with a relatively large number of classification bins. This reveals that SOCS indeed simplifies and facilitates the network training.

4.6. Performance under Occlusion

Our method is designed to handle moderate inter-object occlusions. In order to verify this, we evaluate our method and compare it to the state-of-the-arts on a subset containing objects with heavy occlusions. To be specific, we select 500 RGB-D images with instances that have been heavily occluded ($\leq 30\%$ object surface can be observed) from the NOCS-REAL275 dataset. The visualization of the examples in this subset is provided in the supplemental material. In Table 6, we see that our method outperforms the state-of-the-arts by a large margin, verifying the ability of our method in terms of handling occlusions.

Table 6. Comparisons under heavy occlusion.

Method	IoU75 ↑	5°2cm↑	5°5cm↑	10°2cm ↑
RBP-Pose [32]	0.60	0.29	0.39	0.42
DPDN [19]	0.61	0.29	0.40	0.47
Ours	**0.66**	**0.38**	**0.51**	**0.55**

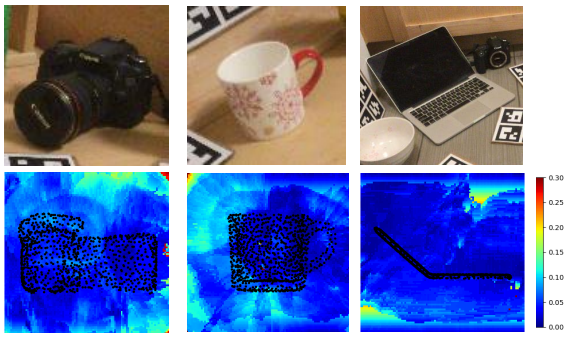


Figure 8. The per-point SOCS estimation error in a cross-section (Bottom row) of the input scene (Top row). The estimation in most of the unseen regions is as accurate as that near the observed surface, showing the necessity of surface-independent sampling and the efficacy of our feature propagation mechanism.

5. Conclusion

We have presented a method for accurate and robust category-level 6D pose and size estimation based on the novel Semantically-aware Object Coordinate Space (SOCS). Since SOCS is built by non-rigidly aligning objects based on semantically meaningful correspondences, it is semantically coherent and leads to accurate pose and size estimation under large shape variations. In future, we would like to investigate the weakly-supervised learning and sim2real transfer techniques to boost the performance of our method on more complicated categories. We would also like to apply SOCS to the category-level pose estimation of articulated objects.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported in part by the National Key Research and Development Program of China (2018AAA0102200), NSFC (62325211, 62132021, 62002379) and the Natural Science Foundation of Hunan Province of China (2023RC3011, 2023JJ20051).

References

[1] Rohith Agaram, Shaurya Dewan, Rahul Sajnani, Adrien Poulenard, Madhava Krishna, and Srinath Sridhar. Canonical fields: Self-supervised learning of pose-canonicalized neural fields. *arXiv preprint arXiv:2212.02493*, 2022. 2

[2] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014. 1

[3] Leonard Bruns and Patric Jensfelt. Sdfest: Categorical pose and shape estimation of objects from rgb-d using signed distance fields. *IEEE Robotics and Automation Letters*, 7(4):9597–9604, 2022. 2

[4] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020. 2

[5] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021. 2, 7

[6] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9121–9130, 2020. 8

[7] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021. 2

[8] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 2, 7

[9] Jean Duchon. Splines minimizing rotation-invariant seminorms in sobolev spaces. In *Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976*, pages 85–100. Springer, 1977. 2, 3

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5

[11] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 585–603. Springer, 2022. 2

[12] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 275–292. Springer, 2022. 2

[13] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020. 1

[14] Fu Li, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. *arXiv preprint arXiv:2203.04802*, 2022. 2

[15] Guanglin Li, Yifeng Li, Zhichao Ye, Qihang Zhang, Tao Kong, Zhaopeng Cui, and Guofeng Zhang. Generative category-level shape and pose estimation with semantic primitives. *arXiv preprint arXiv:2210.01112*, 2022. 2

[16] Xiaolong Li, Yijia Weng, Li Yi, Leonidas J Guibas, A Abbott, Shuran Song, and He Wang. Leveraging se (3) equivariance for self-supervised category-level object pose estimation from point clouds. *Advances in Neural Information Processing Systems*, 34:15370–15381, 2021. 5, 7

[17] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 1

[18] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2022. 2

[19] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 19–34. Springer, 2022. 2, 6, 7, 8

[20] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 2

[21] Zhi-Hao Lin, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1800–1809, 2020. 3

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[23] Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a

single image via bootstrapped radiance field inversion. *arXiv preprint arXiv:2211.11674*, 2022. 2

[24] Wanli Peng, Jianhang Yan, Hongtao Wen, and Yi Sun. Self-supervised category-level 6d object pose estimation with deep implicit shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2082–2090, 2022. 2

[25] Ruoxi Shi, Zhengrong Xue, Yang You, and Cewu Lu. Skeleton merger: an unsupervised aligned keypoint detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 43–52, 2021. 2, 3, 8

[26] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 2, 3

[27] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 1, 6

[28] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 1, 2, 3, 5, 6, 7

[29] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6401–6408. IEEE, 2022. 2, 3

[30] Yilin Wen, Xiangyu Li, Hao Pan, Lei Yang, Zheng Wang, Taku Komura, and Wenping Wang. Disp6d: Disentangled implicit shape and pose learning for scalable 6d pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 404–421. Springer, 2022. 2

[31] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2

[32] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 655–672. Springer, 2022. 2, 6, 7, 8