

# Category-Level 6D Object Pose Estimation via Cascaded Relation and Recurrent Reconstruction Networks

Jiaze Wang<sup>1\*</sup>, Kai Chen<sup>1\*</sup> and Qi Dou<sup>1,2✉</sup>

**Abstract**—Category-level 6D pose estimation, aiming to predict the location and orientation of unseen object instances, is fundamental to many scenarios such as robotic manipulation and augmented reality, yet still remains unsolved. Precisely recovering instance 3D model in the canonical space and accurately matching it with the observation is an essential point when estimating 6D pose for unseen objects. In this paper, we achieve accurate category-level 6D pose estimation via cascaded relation and recurrent reconstruction networks. Specifically, a novel cascaded relation network is dedicated for advanced representation learning to explore the complex and informative relations among instance RGB image, instance point cloud and category shape prior. Furthermore, we design a recurrent reconstruction network for iterative residual refinement to progressively improve the reconstruction and correspondence estimations from coarse to fine. Finally, the instance 6D pose is obtained leveraging the estimated dense correspondences between the instance point cloud and the reconstructed 3D model in the canonical space. We have conducted extensive experiments on two well-acknowledged benchmarks of category-level 6D pose estimation, with significant performance improvement over existing approaches. On the representatively strict evaluation metrics of  $3D_{75}$  and  $5^\circ 2cm$ , our method exceeds the latest state-of-the-art SPD [1] by 4.9% and 17.7% on the CAMERA25 dataset, and by 2.7% and 8.5% on the REAL275 dataset. Codes are available at <https://wangjiaze.cn/projects/6DPoseEstimation.html>.

## I. INTRODUCTION

Accurate 6D pose estimation has increasingly been an important yet challenging research topic in computer vision, which aims to predict the location and orientation of 3D objects [2], [3], [4]. It has extensive prospects in real-world applications such as robotic manipulation, augmented reality, navigation and 3D scene understanding. In recent years, although pioneering work in *instance-level 6D pose estimation* has made remarkable progress [5], [6], [7], [8], almost all these methods require exact 3D CAD object models for the instances. However, such an assumption is difficult, if not impossible, to be satisfied in real practice, considering the diversity of object instances as well as the cost for building a CAD model for each instance. In addition, these methods can not handle new instances with unknown CAD models, which impedes the generalizability in environments with previously object instances without CAD models.

In contrast, the aim of *category-level 6D object pose estimation* is to generate 6D poses for novel object instances

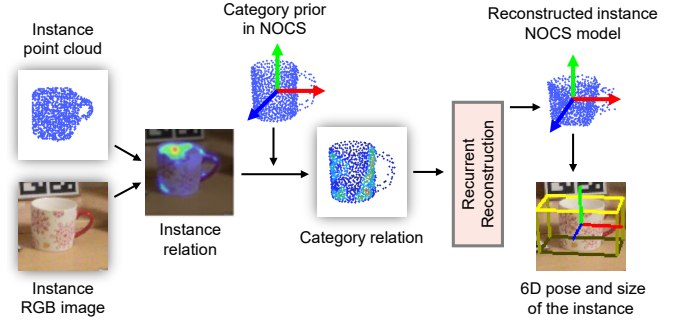


Fig. 1. The proposed category-level 6D pose estimation via cascaded relation and recurrent reconstruction networks.

of the same category, which is much more challenging. In conventional instance-level methods, object pose is estimated via correspondences between the observed instance RGB or RGB-D image and its exact CAD model. Yet, in category-level setting, such a correspondence can not be directly constructed without a specific CAD model. In order to explore such correspondence information in category-level 6D pose estimation, Wang et al. [9] innovate the Normalized Object Coordinate Space (NOCS) which is a unified coordinate system. In NOCS, the size of objects are normalized within a defined coordinate space, and the instances belonging to the same category share an identical orientation. Since the object CAD model is unknown, they reconstruct a canonical model representation in NOCS from instance observations, and build dense correspondences between the reconstructed NOCS model with instance image or point cloud for 6D pose estimation. Recently, Tian et al [1] further improve the accuracy of the reconstructed 3D model in NOCS by introducing a shape prior to the reconstruction process. They extract category features from the shape prior and concatenate instance features with the category ones, to address the shape variations across instances within the same category. The quality of reconstructed NOCS model is significantly enhanced by harnessing the shape-based category features.

Though promising progress has been witnessed, it is still extremely challenging to accurately reconstruct the 3D object models in NOCS, which plays a crucial role to boost 6D pose estimation performance. First, learning representative features from RGB-D image is essential, given that the color image and point cloud data provide complementary texture and geometry information for the objects. Capturing the inherent relations of them improves the representation capability of the instance embeddings. The relevant category feature is also valuable, which helps to model the instance

<sup>1</sup> Jiaze Wang, Kai Chen and Qi Dou are with the Department of Computer Science and Engineering at The Chinese University of Hong Kong, Hong Kong SAR, China. [qidou@cuhk.edu.hk](mailto:qidou@cuhk.edu.hk)

<sup>2</sup> Qi Dou is also with T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong SAR, China

\* Authors contributed equally

shape variations during reconstruction. Exploiting the relation between such category features with the instance ones helps to overcome intra-class shape variations as well as provide global contextual clues for NOCS model reconstruction. However, such important yet complex relations have not been well investigated in existing methods so far. Second, reconstructing the object model in the canonical NOCS space forms a dense regression task. In line with other dense regression problems [10], [11], it is difficult for a network to produce highly accurate regression results with a single step, because the network may partially focus on some object parts while overlooking the others. This hampers the 6D pose estimation accuracy, as the optimization process may be trapped into a local minimum when dealing with spatially biased correspondences.

To address above challenges, we propose a novel method of *Cascaded Relation and Recurrent Reconstruction Networks* for category-level 6D object pose estimation. As illustrated in Figure 1, our framework presents cascaded relation networks to capture the informative relations of instance RGB images, instance point cloud, and category priors, which is important for accurate canonical model reconstruction. Specifically, an instance relation network captures the relation of the input RGB image and point cloud to extract representative feature embeddings for each instance. A category relation network exploits the relation of instance features and category features to address the large shape variations. Moreover, with the relation-enhanced features, a recurrent reconstruction network is then developed to accurately reconstruct the instance model in NOCS space, which progressively refines the reconstructed models from coarse to fine. Finally, we leverage the reconstructed model and the observed point clouds to estimate object 6D pose by point matching. Our main contributions are summarized as:

- We propose a novel cascaded relation network to capture the underlying relations of multi-source inputs. Our network leverages the complementary advantages of these features for categorical object pose estimation.
- We design a recurrent reconstruction network to accurately reconstruct the instance 3D model in NOCS space. By iteratively estimating reconstruction residuals, our network progressively refines the model and the correspondence matrix.
- We conduct extensive experiments on two well-acknowledged benchmarks with dramatic performance improvement over existing methods. On the representative strict evaluation metrics of  $3D_{75}$  and  $5^\circ 2cm$ , our method exceeds the latest state-of-the-art SPD [1] by 4.9% and 17.7% on the CAMERA25 dataset, and by 2.7% and 8.5% on the REAL275 dataset.

## II. RELATED WORK

**Instance-Level 6D Pose Estimation.** Instance-level 6D object pose estimation methods can be broadly categorized into two categories of RGB-based and RGBD-based methods, according to the format of input data. Classical RGB-based methods [12], [13], [14] focus on detecting and match-

ing keypoints with known models. Current deep learning methods [15], [16], [6], [17], [8], [18], [19] improve the performance by replacing the hand-crafted keypoint detection and matching process with a data-driven learning scheme that predicts 2D keypoints on RGB images and solves object poses by PnP [20]. Instead of explicitly detecting and matching object keypoints, some methods [8] take corner points of 3D object bounding boxes as keypoints, or implicitly represent keypoints by a dense voting field [2]. These methods thus can effectively cope with low-texture environments. Other methods [21], [22], [23] propose to directly predict pose parameters from RGB observations. They extract and group pose-relevant features by CNN and regress translation and rotation with two separate multi-layer perceptrons. Although impressive, these RGB-based methods face problems in complex environments such as cluttering or occlusion [24], [2]. Lacking depth information increases the ambiguity of estimation and also hurts the pose accuracy. For RGBD-based methods, once obtained the intrinsic matrix of a RGB-D camera, we can recover the object point cloud via inverse projection [5], [3]. How to make full use of the appearance feature from the RGB image and the complementary geometry feature from point cloud is a major challenge in RGBD-based methods [25]. Algorithms such as PoseCNN [4] uses them in totally separate steps, in which RGB images are used to predict initial object poses while point clouds are utilized for ICP-based pose refinement [26]. [27] fuses the depth image as a new input channel to conventional CNNs which lacks discriminative treatments for point clouds. Recent methods [5], [7], [3] densely fuse the color embedding with the geometry embedding in a pixel-wise manner. Unfortunately, they still did not consider any global feature relations [28], [29], [30], [31], [32] during fusion.

**Category-Level 6D Pose Estimation.** Existing works on category-level 6D pose estimation are still scarce to date. Compared with instance-level object pose estimation, the category-level task is more challenging due to the large intra-class variations in aspects of texture and shape among instances. Establishing an intermediate representation that reduces such difference is a widely applied idea in existing algorithms. Sahin et al. [33] divide an object into multiple 3D skeleton structures, from which they derive a shape-invariant representation and develop a part-based random forest architecture for categorical 6D pose estimation. By integrating 3D shape estimates from a generative object model, [34] produces a distribution over predicted poses with only rotation information. Alternative methods [9] express an object in canonical coordinate space. By inferring / regressing the object canonical representation and associate it with the specific instance observation, 6D object pose can be determined without 3D CAD models. Wang et al. [9] introduce the Normalized Object Coordinate Space to represent different object instances within a category in a unified manner. Then a network is trained to predict correspondences from object pixels to points in NOCS. Subsequently, these correspondences are used with the depth map to estimate

6D pose and size by point matching. Inspired by NOCS, the first category-level pose tracker is proposed by [35]. Recent methods [36], [1] propose to leverage the category-related features to explicitly model the shape variation when reconstructing the canonical representation. In this paper, we explore the complex relations among the texture feature, geometry feature and category feature, in addition to progressively recover the object canonical model with the enhanced feature representations.

### III. METHOD

In this section, to make the contents self-contained and easy to follow, we will first introduce the useful preliminary, then briefly describe an overview of our proposed category-level 6D pose estimation framework. Next, we will describe the proposed novel cascaded relation network, recurrent reconstruction network and pose generation method in detail.

#### A. Preliminary

Given a calibrated RGB-D image, we aim to estimate a 6D pose for an object of interest, which is a rigid transformation  $[R|t]$  composed of a rotation  $R \in SO(3)$  and a translation  $t \in \mathcal{R}^3$  components. To process scenes containing multiple instances with different categories, we first employ an off-the-shelf instance segmentation network (i.e., MaskRCNN [37]) to detect and segment each individual object instance. The yielded detection bounding box is used to crop the RGB image into object patches, and the segmentation mask is leveraged to convert the depth image into object point cloud (through camera intrinsic matrix). Inspired by [1], [9], once we can reconstruct the exact instance 3D model in the NOCS canonical space, the problem of pose estimation then is reduced to determining the similarity transformation from instance point cloud to the reconstructed 3D model. In order to integrate the predicted object category information into this scheme, we further build an initial 3D canonical model for each category of objects and take it as a category-level prior feature.

#### B. Framework Overview

As shown in Figure 2, our novel pose estimation framework has three inputs: instance RGB image, instance point cloud and the corresponding category prior in NOCS. Estimating object 6D poses with the above multi-source inputs, Wang et al. [9] separately process them with independent network modules, while Tian et al. [1] integrating them by concatenating their features in latent space. In contrast, we aim to devise an effective strategy to fully leverage complementary knowledge among the provided inputs for pose estimation. Particularly, we propose a cascaded relation network that constructs the relationship context for input features in two cascaded stages. The first stage models relations between instance image feature and point cloud feature to obtain representative instance features. The second stage further correlates the instance features with category features. On top of these, a dense deformation field is regressed for the use of adjusting the category prior for

NOCS model reconstruction. Meanwhile, a correspondence matrix is also estimated to match the instance point cloud with the reconstructed model. Furthermore, we develop a recurrent reconstruction network for accurate reconstruction and matching. The network iteratively updates the reconstructed model and the category-level feature in each recurrent step. By exploiting multi-stage supervisions, it learns residuals to progressively refine the deformation field and the correspondence matrix. Finally, given the instance NOCS model and the correspondence matrix, object 6D pose and size can be generated by correspondence-based optimization.

#### C. Cascaded Relation Network

In the following, we describe the proposed cascaded relation network. Formally, we denote an instance observation by  $(I, V)$ , with  $I \in \mathcal{R}^{H \times W \times 3}$  being the image patch, and  $V \in \mathcal{R}^{N_p \times 3}$  being the point cloud recovered from the depth map.  $N_p$  is the number of points in  $V$ . Let  $V_c \in \mathcal{R}^{N_c \times 3}$  be the corresponding category prior where  $N_c$  denotes the number of points in  $V_c$ . We first resort to CNN and MLPs to extract texture and geometry features from  $I$ ,  $V$  and  $V_c$ , respectively. After that, similar to [3], we align the texture feature map to point clouds and associate each instance point with a texture feature vector. Consequently, we get the instance texture feature  $F_t \in \mathcal{R}^{C_t \times N_p}$ , the instance geometry feature  $F_g \in \mathcal{R}^{C_g \times N_p}$ , and the category feature  $F_c \in \mathcal{R}^{C_c \times N_c}$ , where  $C_*$  is the feature channels.

The relations among  $F_t$ ,  $F_g$  and  $F_c$  are important for reconstructing the NOCS model towards accurate pose estimation for an instance. On the one hand, it is supposed to capture the characteristics of the instance so that the reconstructed model can well match the observed point clouds. On the other hand, the reconstruction network should also account for certain general attributes of the category, so that the reconstruction process can overcome intra-class variations thus enhancing generalizability to unseen instances of the same category. In these regards, we propose a cascaded relation network to harness these two kinds of relations for accurate NOCS reconstruction and pose estimation.

**Instance Relation.** The instance relation network (IRN) is designed to learn the complementary knowledge between  $F_t$  and  $F_g$ , with  $F_t$  encoding instance texture and semantic features and  $F_g$  encoding its geometry information. These two features complement each other in key aspects. For example, due to absence of depth information,  $F_t$  extracted by CNN is susceptible to image background and cluttering environments, which can be alleviated in  $F_g$  from point cloud. The MLP produces  $F_g$  from disordered point clouds using a narrow receptive field, leading to less efficacy on spatial-aware representations, which in turn can be mitigated with  $F_t$ . Given these properties, we propose to capture their relations and deeply integrate them using a relation function  $\mathcal{G}^i$ . In this way, we reproduce the relation-injected texture feature and geometry feature of the instance as follows:

$$\hat{F}_t = F_t + \mathcal{G}^i(F_t, F_g), \quad \hat{F}_g = F_g + \mathcal{G}^i(F_g, F_t). \quad (1)$$

We add the original feature to the relation feature, in order

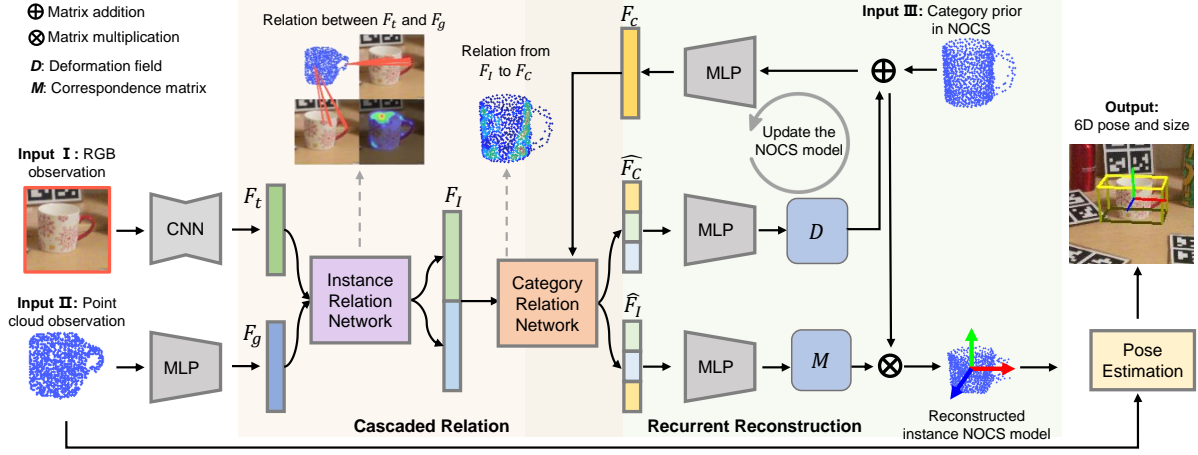


Fig. 2. Overview of our Cascaded Relation and Recurrent Reconstruction Networks. The networks are mainly composed of two networks: (1) A cascaded relation network to exploit the relation of between RGB images and point clouds, and the relation between instance features and category features. (2) An recurrent reconstruction network for canonical shape reconstruction from coarse to fine.

to drive the network to unearth as much complementary information as possible by relational learning. We concatenate  $\hat{F}_t$  and  $\hat{F}_g$  as our relation-enhanced instance embedding  $F_I$ , which is expected to represent the instance’s characteristics in the camera frame.

**Category Relation.** Next, the category relation network (CRN) aims to capture relations between the  $F_I$  and  $F_c$  that presents the category-level feature of an instance in NOCS space. The interaction between these features helps the network to accurately model the shape variation of instances from the same category. We cascade CRN behind IRN, with the careful consideration that the strong relation-enhanced instance embedding  $F_I$  should be yielded beforehand. Directly associating those original  $F_t$  and  $F_g$  with  $F_c$  may cannot fully tap the potential of relational learning in category-level. Similar to our IRN, we initiate a relation function  $\mathcal{G}^c$  to exploit relations between  $F_I$  and  $F_c$ :

$$\hat{F}_I = F_I + \mathcal{G}^c(F_I, F_c), \quad \hat{F}_c = F_c + \mathcal{G}^c(F_c, F_I). \quad (2)$$

The relation-injected features  $\hat{F}_I$  and  $\hat{F}_c$  then will be used in the subsequent model reconstruction and pose estimation.

**Choice of Relation Function  $\mathcal{G}$ .** The formulation of our cascaded relation network is flexible, which is not restricted by the specific architecture design for  $\mathcal{G}$ . Any structure capable of capturing feature relations can be easily adopted as  $\mathcal{G}$  in our framework. In this paper, we employ three representative and popular structures: MLP [38], Non-local [39], and Transformer [40]. The experimental results indicate that the cascaded relation network improves the pose estimation performance significantly, no matter which specific  $\mathcal{G}$  is utilized. Please refer to Table III for comparison results.

#### D. Recurrent Reconstruction Network

In this section, we describe the recurrent reconstruction network. With the  $\hat{F}_I$  and  $\hat{F}_c$  produced by CRN, as in [1], we regress a deformation  $D \in \mathcal{R}^{N_c \times 3}$  from the initial category model to the instance canonical model that we

aim to reconstruct. Meanwhile, we regress a correspondence matrix  $M \in \mathcal{R}^{N_p \times N_c}$  for associating the instance point cloud with the reconstructed 3D model in NOCS. Each row of  $M$  indicates a weighted correspondence between a point in  $V$  and all points in  $V_c$ . Even with our relation-enhanced features, accurately regressing all these values within one single step is still infeasible. The points at different locations may have diverse deformations. The network may have bias towards some locations but neglect the remaining points, which results in an inaccurate reconstruction model. Accordingly, the correspondence matrix will also be adversely affected.

To address this problem, we propose a recurrent reconstruction network to regress  $D$  and  $M$  from coarse to fine. Our network first predicts an initial deformation field  $D^0$  and an initial correspondence matrix  $M^0$  directly from the relation-enhanced features  $\hat{F}_c$  and  $\hat{F}_I$ . Then, taking  $D^0$  and  $M^0$  as initial values, we estimate the residuals to refine the deformation field and correspondence matrix. Specifically, we add  $D^0$  to  $V_c$  to update the NOCS model, and denote its updated category features as  $F_c^1$ . After that, we integrate  $F_c^1$  with  $F_I$  in the CRN to model the shape variation for the recurrent step and make the current reconstruction focus more on regions that are neglected in previous steps. For the first recurrent step, our network would target on estimating the deformation residual  $\bar{D}^1 = D_{gt} - D^0$  and correspondence residual  $\bar{M}^1 = (M^0)^{-1} \times M_{gt}$ . To further improve the accuracy, we can repeat the recurrent optimization multiple times. For the  $i$ -th recurrent step, its deformation and correspondence matrix can be computed as:

$$D^i = D^{i-1} + \bar{D}^i, \quad M^i = M^{i-1} \times \bar{M}^i. \quad (3)$$

The outputs of the last recurrent step are the final estimation results for the deformation field and correspondence matrix.

Notably, the loss functions to supervise the recurrent reconstruction process are crucial. We impose supervisions on  $D$  and  $M$  to drive the network learning to reconstruct



the instance NOCS model and associate it with the observed point clouds accurately.

**Reconstruction Loss.** Applying the deformation field  $D$  on  $V_c$  yields the reconstructed NOCS model  $R$ . During training, with the ground-truth NOCS model  $R_{gt}$  for each instance, we employ the reconstruction loss to penalize  $D$ . Specifically, we exploit the Chamfer distance (CD) to measure the similarity between  $R$  and  $R_{gt}$  as:

$$L_r = \sum_{i \in R^i} \min_{j \in R_{gt}^j} \|i - j\|_2^2 + \sum_{j \in R_{gt}^j} \min_{i \in R^i} \|i - j\|_2^2. \quad (4)$$

In addition, a regularization loss is further added to penalize large deformations:  $L_{\text{def}} = \frac{1}{N_c} \sum_{i \in D} \|i\|_2$ .

**Correspondence Loss.** We supervise  $M$  with a correspondence loss function inspired by [1]. After applying  $M$  on the reconstructed NOCS model, we can get the NOCS coordinate prediction for each point in  $V$ . Again, since the ground-truth NOCS coordinate for the point cloud observation is known during training, we supervise  $M$  by constraining the distance between the predicted NOCS coordinate value  $x$  and the ground-truth one  $x_{gt}$ . The correspondence loss  $L_o$  can be defined as:

$$L_o(x, x_{gt}) = \frac{1}{N_c} \begin{cases} 5(x - x_{gt})^2 & |x - x_{gt}| \leq 0.1, \\ |x - x_{gt}| - 0.05 & \text{otherwise,} \end{cases} \quad (5)$$

in which a soft  $L_1$  loss is used for robust optimization. In addition, the same regularization loss  $L_{\text{reg}}$  as the one in [1] is further adopted to constrain the sparsity of  $M$ .

**Recurrent Loss.** We combine the above two loss terms in one recurrent step. Moreover, we exploit deep supervision mechanism [41] that imposes supervision on every recurrent step. The supervision on intermediate results impels the network to learn the residuals from the ground truth, and the accumulated loss function is as:

$$L_{\text{overall}} = \sum_{k=0}^N \lambda_k \times (L_r^k + L_{\text{def}}^k + L_o^k + L_{\text{reg}}^k), \quad (6)$$

where  $L_r^k, L_{\text{def}}^k, L_o^k, L_{\text{reg}}^k$  denotes the loss for the  $k$ -th recurrent step and  $\lambda_k$  is the associated weighted factor. This overall loss is used to train the recurrent reconstruction network.

#### E. Correspondence based Pose Estimation

Feeding the instance RGB image  $I$ , instance point cloud  $V$ , and the category prior  $V_c$  into the framework, our network would output a deformation field  $\hat{D}$  and a correspondence Matrix  $\hat{M}$ . As in [1], [9], we can use  $\hat{D}$  and  $\hat{M}$  to estimate the specific 6D object pose. First of all, we apply  $\hat{D}$  on  $V_c$  to get the reconstructed NOCS model  $V_{\text{nocs}}$  for the instance. Note that  $V_{\text{nocs}}$  is in the canonical NOCS space, and the point cloud observation  $V$  is in the camera coordinate space. Once we recover  $V_{\text{nocs}}$ , estimating the 6D pose for the instance in the camera frame would be equal to finding a similarity transformation from  $V_{\text{nocs}}$  to  $V$  up to a scale factor. This similarity transformation can be estimated based on dense correspondences between  $V_{\text{nocs}}$  and  $V$ . To obtain these correspondences, we apply  $\hat{M}$  on  $V_{\text{nocs}}$  and compute a NOCS coordinate for each point in  $V$ . Finally, we

use the Umeyama algorithm [42] to estimate the similarity transformation, in which the 3D rotation and translation corresponds to the 6D object pose, and the scale factor corresponds to the object size. The RANSAC [20] is also adopted to remove outliers and achieve a robust estimation.

## IV. EXPERIMENTS

We conduct extensive experiments on two state-of-the-art benchmark datasets of category-level 6D pose estimation, i.e., CAMERA25 and REAL275. We have compared with the existing methods on this task, and outperform them by a significant margin on both datasets. We have also presented the ablation studies to analyze the individual behavior of our proposed cascaded relation network and recurrent reconstruction network.

#### A. Experimental Setup

**Datasets.** We conducted our experiments following the settings in NOCS [9]. It consists of two datasets: the CAMERA25 dataset and the REAL275 dataset, which corresponds to virtual and real environments respectively. Specifically, CAMERA25 contains 300K RGB-D images (with 25K for evaluation) which are generated by rendering and compositing synthetic objects into real scenes. REAL275 contains 4300 real-world RGB-D images from 7 scenes for training, and 2750 real-world RGB-D images from 6 scenes for evaluation (with 3 instances per category). Both datasets contain the same six categories, i.e., *bottle*, *bowl*, *camera*, *can*, *laptop* and *mug*.

**Evaluation Metrics.** Similar to [?], [1], [9], we quantitatively evaluate the estimated 6D object pose on the following metrics: **3D IoU**: It measures the accuracy of the predicted 3D object bounding box. The predicted pose and the ground truth pose can determine two 3D bounding boxes. Only when the overlapping of these two boxes is larger than a predefined threshold, the predicted pose is judged to be correct. In our experiment, we use  $3D_{50}$  and  $3D_{75}$ , which take 50% and 75% as the Intersection over Union (IoU) threshold respectively. **a°b cm**: It measures the error of predicted poses. Only when the rotation error is less than a° and the translation error is less than b cm, the pose is judged to be correct. Similar to [1], we test four different settings: 5°2 cm, 5°5 cm, 10°2 cm, and 10°5 cm. The rotation error of vertical axis is ignored for symmetrical object categories (*bottle*, *bowl* and *can*). Similar to [1], for *mug* category, we treat it as symmetrical object when the handle is not visible, otherwise as asymmetric object. In addition, the Chamfer Distance(CD) is used to evaluate the instance NOCS model reconstruction accuracy.

**Implementation Details.** The texture embedding network is PSPNet [44] with backbone of ResNet-18 [45], and the model is initialized with pre-trained models from ImageNet [46]. The image crop is resized to  $192 \times 192$ . The number of points in the input point cloud and category prior is downsampled to 1024. We develop our baseline based on the structure of SPD [1]. We train our networks for 50 epochs in total, and there are 4K iterations in each epoch.

TABLE I. Comparison of our method with current state-of-the-art methods on both benchmarks.

Method	CAMERA25						REAL275					
	$3D_{50}$	$3D_{75}$	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$	$3D_{50}$	$3D_{75}$	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$
NOCS [9]	83.9	69.5	32.3	40.9	48.2	64.6	78	30.1	7.2	10.0	13.8	25.2
SPD [1]	93.2	83.1	54.3	59.0	73.3	81.5	77.3	53.2	19.3	21.4	43.2	54.1
CASS[43]	-	-	-	-	-	-	77.7	-	-	23.5	-	58.0
SPD*	93.5	87.0	59.7	64.4	77.4	84.8	77.5	53.8	20.2	22.9	42.5	52.1
Ours	<b>93.8</b>	<b>88.0</b>	<b>72.0</b>	<b>76.4</b>	<b>81.0</b>	<b>87.7</b>	<b>79.3</b>	<b>55.9</b>	<b>27.8</b>	<b>34.3</b>	<b>47.2</b>	<b>60.8</b>

\* SPD implemented by ourselves with the open source codes.

TABLE II. Evaluation of NOCS model reconstruction results regarding the shape quality using the Chamfer Distance (CD) metric ( $\times 10^{-3}$ ).

Method	CAMERA25							REAL275						
	Bottle	Bowl	Camera	Can	Laptop	Mug	Average	Bottle	Bowl	Camera	Can	Laptop	Mug	Average
SPD [1]	1.81	1.63	4.02	0.97	1.98	1.42	1.97	3.44	1.21	8.89	1.56	2.91	<b>1.02</b>	3.17
Ours	<b>1.30</b>	<b>0.95</b>	<b>2.58</b>	<b>0.82</b>	<b>0.87</b>	<b>1.05</b>	<b>1.18</b>	<b>2.99</b>	<b>0.96</b>	<b>7.57</b>	<b>1.31</b>	<b>1.25</b>	1.19	<b>2.53</b>

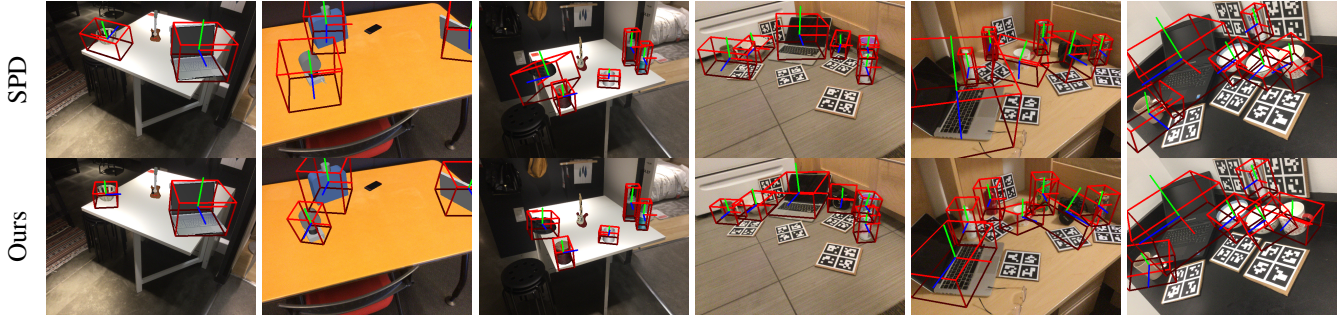


Fig. 3. Visual comparison of 6D object pose and size estimation results by our method and SPD [1]. In each row, the first three images are from CAMERA25, the remaining three images are from REAL275.

We use the ADAM[47] to train the network, where the initial learning rate is set as  $1e-4$  and with weight decay of  $1e-6$ . The learning rate is decreased by a factor of 10 for every 10 epochs. Our framework is implemented with PyTorch using 4 TITAN Xp GPUs. Section IV-B reports our results with a Transformer-based relation network and a recurrent reconstruction network with fixed 2 recurrent stages. The relevant ablation study is presented in Section IV-C. Table III focuses on cascaded relation network configurations without adding the recurrent reconstruction module. Similarly, Table IV studies recurrent steps without involving cascaded relations.

### B. Comparison with State-of-the-Art Methods

We compare with three state-of-the-art methods: NOCS [9], Shape Prior Deformation (SPD) [1] and Canonical Shape Space (CASS) [36]. To our knowledge, these are all the popular methods that address the challenging category-level 6D pose estimation task in current literature. Table I and Figure 3 present the quantitative and qualitative results.

**CAMERA25.** On this benchmark dataset, our method significantly outperforms all previous state-of-the-arts approaches across all metrics. Notably, on the two most strict metrics, we achieve 88.0% on  $3D_{75}$  and 72.0% on  $5^\circ 2cm$ , which are 18.5% and 39.7% higher than NOCS [9], and 4.9% and 17.7% higher than SPD [1]. Such large margins exceeding the current methods are attributed to our relation-enhanced representation learning and recurrent refinement of NOCS reconstruction model. On those relatively flexible metrics of  $3D_{50}$  and  $10^\circ 2cm$ , our mean average precision

(mAP) reaches a high value of 93.8% and 87.7%, indicating the promising potential for practical use.

**REAL275.** The dataset of REAL275 is much more challenging than CAMERA25 given the real-world complications and the limited amount of training data. Only 3 object instances of each category are given for training, and 3 new instances are given for testing. Thus, in consistency with [1], [9], we randomly select data from CAMERA25 and REAL275 at a ratio of 3:1 and train the network using this hybrid dataset. Under this setting, our method achieves a mAP of 55.9% for 3D IoU at 75%, and a mAP of 27.8% for pose error within  $5^\circ 2cm$ . These performances are 25.9% and 20.6% higher than NOCS [9], and 2.7% and 8.5% higher than SPD [1]. We also compare our results with the accuracy value reported in CASS [36], and our method outperforms it by 10.8% on  $5^\circ 5cm$ , as well as by 1.6% and by 2.8 on the less strict metrics of  $3D_{50}$  and  $10^\circ 5cm$ .

**NOCS Model Reconstruction.** To evaluate the quality of our reconstructed NOCS model regarding the shape of instance point cloud, we report the Chamfer Distance (CD) between our reconstructed result and the ground truth model. Table II compares our performance on the CD metric with SPD [1]. It is observed that our method can consistently improve the accuracy of reconstructions over the state-of-the-art SPD [1] across all categories on CAMERA25 and on five out of the six categories on REAL275. These analysis validate our superior quality of reconstructed NOCS model particularly with a precise shape which is important for accurate category-level pose estimation.

TABLE III. Evaluation of Cascaded Relation Network. The “-”, “M”, “N”, and “T” refer to “Without relation network”, “MLP based relation network”, “Non-Local based relation network”, and “Transformer based relation network”, respectively. For example, “T / -” denotes the instance relation network is Transformer and the category relation network is none. We report the mAP for the six different metrics.

$\mathcal{G}$	CAMERA25						REAL275					
	$3D_{50}$	$3D_{75}$	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$	$3D_{50}$	$3D_{75}$	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$
- / -	93.5	87.0	59.7	64.4	77.4	84.8	77.5	53.8	20.2	22.9	42.5	52.1
M / M	93.3	87.1	60.4	68.2	79.4	86.0	77.1	51.1	23.9	29.6	46.2	59.3
N / N	93.2	87.0	65.6	70.2	80.4	86.7	78.0	54.5	26.3	<b>30.4</b>	45.7	58.7
- / T	93.3	87	62.4	68.5	78.5	84.6	78.5	54.4	23.8	28.5	45.5	57.7
T / -	<b>94.5</b>	<b>87.3</b>	62.3	66.8	78.3	85.5	77.3	<b>57.4</b>	25.3	28.6	46.4	58.7
T / T	94.3	87.1	<b>70.9</b>	<b>75.7</b>	<b>80.5</b>	<b>87.5</b>	<b>78.7</b>	55.8	<b>26.5</b>	<b>30.4</b>	<b>46.5</b>	<b>59.6</b>

TABLE IV. Evaluation of Recurrent Reconstruction Network. We report the mAP for 6 different metrics on object pose estimation.

Step	CAMERA25						REAL275					
	$3D_{50}$	$3D_{75}$	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$	$3D_{50}$	$3D_{75}$	$5^\circ 2cm$	$5^\circ 5cm$	$10^\circ 2cm$	$10^\circ 5cm$
0	93.5	87.0	59.7	64.4	77.4	84.8	77.5	53.8	20.2	22.9	42.5	52.1
1	<b>93.7</b>	87.8	62.2	66.8	78.4	<b>85.7</b>	77.5	55.6	24.2	25.6	46.3	60.1
2	93.6	88.6	<b>64.4</b>	<b>68.5</b>	<b>78.9</b>	85.6	<b>81.9</b>	<b>57.0</b>	24.2	25.7	<b>46.4</b>	<b>60.2</b>
3	93.6	<b>88.7</b>	64.3	68.3	78.8	85.6	81.5	56.8	<b>24.5</b>	<b>26.2</b>	46.0	59.5

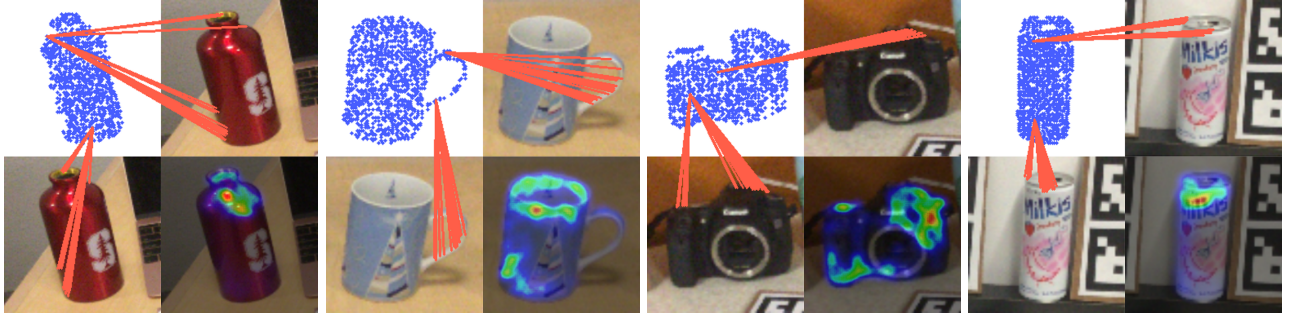


Fig. 4. Visualization of learned feature relations from instance point clouds to RGB images. The top-ranking relations (i.e., red lines) are computed from the learned network parameters (i.e., transformers in IRN). In each  $2 \times 2$  cell, red lines indicate the top-20 relations for a 3D point, and the heatmap at the lower right corner illustrates the overall relation distribution of instance point clouds.

### C. Ablation Studies

To investigate the properties of our method, we evaluate its key model components on CAMEAR25 and REAL275.

**Cascaded Relation.** In Table III, we quantitatively evaluated our cascaded relation network with different settings. Specifically, we test three well-known relational structures: MLP [38], Non-Local [39], and Transformer [40]. According to the results in Table III, we find that no matter which structure is adopted, the pose estimation results get consistent improvement over the baseline. Moreover, we observe that the Transformer-based structure achieves the best pose estimation results on both datasets, which indicates the superiority of recent Transformer modules in modelling long-range and/or cross-modal feature relations. In addition, we further remove the IRN and CRN individually and observe the change of pose estimation results (see the row of “- / T” and “T / -” respectively). The obtained results demonstrate that both IRN and CRN are important in our network with balanced contribution to the performance. Cascading them together can take full advantages of the inherent feature relations for 6D object pose estimation.

**Intuitive Visualizations of Learned Relations.** We visualize the relations from instance point clouds to RGB image that are learned by our relation network for interpretable understandings of what have been captured in the networks. As presented in Figure 4, the top-20 relations and heatmaps

would attend more to the boundary regions of the objects, which are very informative for estimating object 6D pose.

**Recurrent Reconstruction.** In Table IV, we investigate our recurrent reconstruction network on different numbers of recurrent steps. “0” denotes the baseline network without any recurrent refinement. From the mAP measurements, we find that gradually increasing the number of recurrent step progressively improves the pose estimation outcomes. The pose accuracy consistently increases in the first two iterations and gets relatively stable after that. With a 2-step recurrent, for the most strict metric  $5^\circ 2cm$ , our recurrent network improves the mAP from 59.7% to 64.4% on CAMERA25 data and from 20.2% to 24.2% on REAL275 data. These results demonstrate that harnessing our proposed recurrent reconstruction network is beneficial for boosting pose estimation accuracy.

### V. CONCLUSION

We have presented a novel cascaded relation and recurrent reconstruction framework for category-level 6D object pose and size estimation. Our approach models the relations of RGB image, point cloud, and shape prior through a devised cascaded relation network, which enables our network to learn more representative instance features and overcome shape variations of different instances. The NOCS canonical object shape is reconstructed from a coarse-to-fine manner

via a designed recurrent reconstruction network. Extensive experimental results present dramatic performance improvement over existing methods, setting new state-of-the-art results on both benchmarks. Our method has promising potential to be applied for downstream applications such as robotic manipulation of objects in real environments.

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

## REFERENCES

- [1] M. Tian, M. H. Ang Jr, and G. H. Lee, “Shape prior deformation for categorical 6d object pose and size estimation,” in *ECCV*, 2020.
- [2] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, “Pvnet: Pixel-wise voting network for 6dof pose estimation,” in *CVPR*, 2019, pp. 4561–4570.
- [3] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, “Densefusion: 6d object pose estimation by iterative dense fusion,” in *CVPR*, 2019, pp. 3343–3352.
- [4] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, “Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes,” *arXiv preprint arXiv:1711.00199*, 2017.
- [5] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, “Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation,” in *CVPR*, 2020, pp. 11 632–11 641.
- [6] Z. Li, Y. Hu, M. Salzmann, and X. Ji, “Robust rgb-based 6-dof pose estimation without real pose annotations,” *arXiv preprint arXiv:2008.08391*, 2020.
- [7] C. Song, J. Song, and Q. Huang, “Hybridpose: 6d object pose estimation under hybrid representations,” in *CVPR*, 2020, pp. 431–440.
- [8] B. Tekin, S. N. Sinha, and P. Fua, “Real-time seamless single shot 6d object pose prediction,” in *CVPR*, 2018, pp. 292–301.
- [9] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, “Normalized object coordinate space for category-level 6d object pose and size estimation,” in *CVPR*, 2019, pp. 2642–2651.
- [10] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, “Cascade cost volume for high-resolution multi-view stereo and stereo matching,” in *CVPR*, 2020, pp. 2495–2504.
- [11] T.-W. Hui, X. Tang, and C. Change Loy, “Liteflownet: A lightweight convolutional neural network for optical flow estimation,” in *CVPR*, 2018, pp. 8981–8989.
- [12] A. Collet, M. Martinez, and S. S. Srinivasa, “The moped framework: Object recognition and pose estimation for manipulation,” *The international journal of robotics research*, vol. 30, no. 10, pp. 1284–1306, 2011.
- [13] V. Ferrari, T. Tuytelaars, and L. Van Gool, “Simultaneous object recognition and segmentation from single or multiple model views,” *International journal of computer vision*, vol. 67, no. 2, pp. 159–188, 2006.
- [14] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, “Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes,” in *ACCV*. Springer, 2012, pp. 548–562.
- [15] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, “Learning 6d object pose estimation using 3d object coordinates,” in *ECCV*. Springer, 2014, pp. 536–551.
- [16] S. Hinterstoisser, S. Holzer, C. Cagniat, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, “Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes,” in *ICCV*. IEEE, 2011, pp. 858–865.
- [17] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim, “Latent-class hough forests for 3d object detection and pose estimation,” in *ECCV*. Springer, 2014, pp. 462–477.
- [18] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, “Deep object pose estimation for semantic robotic grasping of household objects,” *arXiv preprint arXiv:1809.10790*, 2018.
- [19] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin, “Movienet: A holistic dataset for movie understanding,” in *ECCV*. Springer, 2020, pp. 709–727.
- [20] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [21] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, “Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again,” in *ICCV*, 2017, pp. 1521–1529.
- [22] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, “Deepim: Deep iterative matching for 6d pose estimation,” in *ECCV*, 2018, pp. 683–698.
- [23] F. Manhardt, W. Kehl, N. Navab, and F. Tombari, “Deep model-based 6d pose refinement in rgb,” in *ECCV*, 2018, pp. 800–815.
- [24] M. Rad and V. Lepetit, “Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth,” in *ICCV*, 2017, pp. 3828–3836.
- [25] S. Song and J. Xiao, “Deep sliding shapes for amodal 3d object detection in rgb-d images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 808–816.
- [26] P. J. Besl and N. D. McKay, “Method for registration of 3-d shapes,” in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–606.
- [27] C. Li, J. Bai, and G. D. Hager, “A unified framework for multi-view multi-class object pose estimation,” in *ECCV*, 2018, pp. 254–269.
- [28] J. Wang, X. Peng, and Y. Qiao, “Cascade multi-head attention networks for action recognition,” *Computer Vision and Image Understanding*, vol. 192, p. 102898, 2020.
- [29] A. Rao, J. Wang, L. Xu, X. Jiang, Q. Huang, B. Zhou, and D. Lin, “A unified framework for shot type classification based on subject centric lens,” in *ECCV*. Springer, 2020, pp. 17–34.
- [30] A. Rao, L. Xu, Y. Xiong, G. Xu, Q. Huang, B. Zhou, and D. Lin, “A local-to-global approach to multi-modal movie scene segmentation,” in *CVPR*, 2020, pp. 10 146–10 155.
- [31] J. Wang, “Part matching with multi-level attention for person re-identification,” in *ICCVW*. IEEE, 2019, pp. 1805–1814.
- [32] J. Xia, A. Rao, Q. Huang, L. Xu, J. Wen, and D. Lin, “Online multi-modal person search in videos,” in *ECCV*. Springer, 2020, pp. 174–190.
- [33] C. Sahin and T.-K. Kim, “Category-level 6d object pose recovery in depth images,” in *ECCV*, 2018, pp. 0–0.
- [34] B. Burchfiel and G. Konidaris, “Probabilistic category-level pose estimation via segmentation and predicted-shape priors,” *arXiv preprint arXiv:1905.12079*, 2019.
- [35] C. Wang, R. Martín-Martín, D. Xu, J. Lv, C. Lu, L. Fei-Fei, S. Savarese, and Y. Zhu, “6-pack: Category-level 6d pose tracker with anchor-based keypoints,” in *ICRA*. IEEE, 2020, pp. 10 059–10 066.
- [36] D. Chen, J. Li, Z. Wang, and K. Xu, “Learning canonical shape space for category-level 6d object pose and size estimation,” in *CVPR*, 2020, pp. 11 973–11 982.
- [37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *ICCV*, 2017, pp. 2961–2969.
- [38] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, “A simple neural network module for relational reasoning,” in *NIPS*, 2017, pp. 4967–4976.
- [39] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *CVPR*, 2018, pp. 7794–7803.
- [40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [41] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, “R3net: Recurrent residual refinement network for saliency detection,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 684–690.
- [42] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 4, pp. 376–380, 1991.
- [43] X. Chen, Z. Dong, J. Song, A. Geiger, and O. Hilliges, “Category level object pose estimation via neural analysis-by-synthesis,” in *ECCV*, 2020.
- [44] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *CVPR*, 2017, pp. 2881–2890.
- [45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*. Ieee, 2009, pp. 248–255.
- [47] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.