ACR-Pose: Adversarial Canonical Representation Reconstruction Network for Category Level 6D Object Pose Estimation

Zhaoxin Fan
fanzhaoxin@ruc.edu.cn
Renmin University of China; Psyche AI Inc
Beijing, China

Zhenbo Song
songzb@njust.edu.cn
Nanjing University of Science and Technology
Nanjing, China

Zhicheng Wang
wangzhichg@163.com
Xreal
Beijing, China

Jian Xu
jian.xu@ia.ac.cn
Institute of Automation, Chinese Academy of Sciences
Beijing, China

Kejian Wu
kejian@xreal.com
Xreal
Beijing, China

Hongyan Liu
hylu@tsinghua.edu.cn
Tsinghua University
Beijing, China

Jun He*
hejun@ruc.edu.cn
Renmin University of China
Beijing, China

ABSTRACT
In the realm of category-level 6D object pose estimation, canonical 3D representation reconstruction is pivotal, yet current methods show limitations in reconstruction quality, a key step in current pose estimation pipeline. To address this, we introduce an innovative Adversarial Canonical Representation Reconstruction Network (ACR-Pose) in this paper. In particular, ACR-Pose comprises a Reconstructor, with novel sub-modules: a Pose-Irrelevant Module (PIM) for robustness to rotation and translation, and a Relational Reconstruction Module (RRM) for extracting relational information between input modalities. A Discriminator is incorporated to guide the generation of realistic canonical representations through adversarial optimization. Evaluated on the prevalent NOCS-CAMERA and NOCS-REAL datasets, our method significantly improves the performance of baseline models and achieves comparable performance with existing state-of-the-art methods, representing a promising advancement in the field of category-level 6D object pose estimation.

CCS CONCEPTS
• Computing methodologies → Vision for robotics.

KEYWORDS
6d pose, category-level, canonical representation, adversarial

*Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR ’24, June 10–14, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0619-6/24/06...$15.00
https://doi.org/10.1145/3652583.3658050

1 INTRODUCTION

6D object pose estimation is an integral technique employed in a multitude of applications, encompassing robotic grasping [3, 30, 36], augmented reality [2, 4, 38], and autonomous driving [16, 32, 33]. Essentially, it endeavors to discern the position and orientation of an object within a three-dimensional space, utilizing camera data that could be RGB (standard image data), RGBD (inclusive of depth information), or LiDAR (employing lasers to gauge distances). Despite noteworthy advancements in this domain, with a plethora of methods [13, 21, 28, 31, 34, 42, 44] being developed and successfully integrated into commercial products, most of these methodologies exhibit a significant limitation: they’re tailored to work with specific, predefined objects. This becomes a predicament in complex environments comprising a diverse range of objects, as it necessitates a separate model for each object, thereby consuming substantial memory and requiring intensive training. Such challenges underscore the imperative need for a more versatile and efficient approach.

Over the recent years, category-level 6D object pose estimation [39] has become a prominent area in both academia and industry, offering solutions to limitations associated with instance-level approaches. Typically, an RGBD observation is processed via a deep learning network to produce a canonical representation [39] of the object. This representation, often exemplified by the NOCS [39] representation, encloses a normalized 3D model that captures key characteristics of an object category, independent of their pose. The precision in reconstructing this canonical representation is
vital for accurate object pose estimation. While existing methods [5, 7, 8, 23, 25, 35, 39] have made commendable progress in category-level 6D object pose estimation using canonical representations, the quality of their reconstructions leaves room for significant enhancement. Specifically, these approaches face a triad of limitations that hinder their performance: 1) They often exhibit sensitivity to pose-related attributes, complicating the learning process for canonical-related features. 2) They tend to neglect the pivotal relational information present among different input modalities. Some of their reconstructions produce results that lack realism, posing potential hurdles to their practical application. These limitations pose serious implications for the final outcomes of 6D pose estimation, underscoring the urgency for a comprehensive solution.

In response to these challenges, we introduce a novel Adversarial Canonical Representation Reconstruction pipeline, termed ACR-Pose, designed to amplify the precision of category-level 6D object pose estimation by reconstructing high-fidelity canonical representations. As depicted in Fig. 1, the pipeline involves object detection and segmentation from RGBD images. Subsequently, a Reconstructor and a Discriminator are trained in an adversarial fashion to generate realistic canonical representations, utilizing the image patch and depth region of an object as input.

The Reconstructor plays a pivotal role in the ACR-Pose pipeline for the generation of superior quality canonical representations for accurate 6D object estimation. However, conventional canonical representation reconstruction models frequently exhibit sensitivity to pose-related characteristics, resulting in performance constraints. To mitigate this, we introduce two innovative sub-modules within the Reconstructor to learn more discerning and resilient features. The first sub-module, referred to as the Pose-Irrelevant Module (PIM), has been designed to filter out rotational and translational cues in the object observation, retaining only the shape information. This strategy is instrumental in achieving a canonical representation. By concentrating on learning canonical-related features exclusively, the PIM sub-module allows the Reconstructor to focus on the object’s shape, thereby enhancing the overall pipeline performance. The second sub-module, the Relational Reconstruction Module (RRM), draws inspiration from recent work [35] and leverages a learned shape prior as an input to understand the relational information between RGB, depth, and shape information. In doing so, the RRM sub-module elevates the pipeline’s performance by learning sturdier features and producing more precise canonical representations. Upon the Relational Reconstruction Module (RRM) learning the relational information between RGB, depth, and shape information, the canonical representation can be reconstructed using a Shape Prior Deformation step [35]. This step employs the features gleaned from the RRM to further boost the accuracy and robustness of the canonical representation.

Subsequently, 6D object pose estimation can be readily resolved by aligning the reconstructed canonical representation with the back-projected object depth using the Umeyama algorithm [37]. Moreover, though the Reconstructor is powerful, we observe that in some cases, the reconstruction results can be unrealistic, leading to inaccurate alignment results. To alleviate this limitation, we incorporate the concept of Generative Adversarial Networks (GANs) [48, 49] into our pipeline and propose the use of a Discriminator to guide the Reconstructor in an adversarial manner. This approach allows the Reconstructor to generate more realistic canonical representations, thereby further enhancing the overall accuracy and robustness of the pipeline. Experimental results on both synthetic and real scenarios demonstrate that our proposed ACR-Pose pipeline can significantly improve the performance of the baseline models and the final model achieve comparable performance with state-of-the-art methods.

In summary, we make three key contributions. 1) We introduce ACR-Pose, an innovative network leveraging adversarial training for the first time in category-level 6D object pose estimation, marking a significant development in computer vision and robotics. 2) We also develop two new sub-modules in ACR-Pose: the Pose-Irrelevant and the Relational Reconstruction Modules, enabling effective learning of high-quality, canonical-related and relational features for credible reconstruction. 3) Finally, through extensive experimentation, we demonstrate that our model outperforms existing methods on synthetic and real-world datasets, underscoring its effectiveness and robustness.

2 RELATED WORK

2.1 6D Object Pose Estimation

While instance-level 6D object pose estimation methods [28, 31, 46] offer precise pose estimation for specific objects, their generalization capabilities remain confined to those objects, necessitating a unique model for each object [9]. Conversely, category-level pose estimation provides broader object generalization, prompting our investigation into this approach. Initial deep learning-based category-level methodologies, such as the Normalized Object Coordinate Space (NOCS) representation [39], addressed the substantial shape variance amongst objects. Subsequent works, including the Canonical Shape Space (CASS) [5], attracted significant attention. Later developments, such as SPD [35], Lee et al. [18], and Dual-PoseNet [23], sought to refine these methods, utilizing strategies like shape prior deformation, simultaneous prediction of NOCS representation, and metric-scale object shape from RGB observations, and dual-pose networks for improved pose estimation accuracy. Recent approaches like CR-Net [40] and SGPA [6] employed attention mechanisms and Transformers for enhanced feature learning, while GPV-Pose [8], SAR-Net [22], and CARTRE [25] improved existing methods from various perspectives. Nonetheless, the majority of...
these methodologies employ the reconstruction of canonical representation for accurate 6D object pose estimation. In this paper, we also adopt the reconstruction of NOCS representation but seek to uncaps its full potential through our proposed ACR-Pose, focusing on enhancing the quality of the reconstructed canonical representation via adversarial training.

2.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [10, 26] have found extensive application in diverse generative tasks such as style translation [15], image generation [24], and human reconstruction [17]. GANs typically comprise a Generator and a Discriminator, wherein the Generator predicts the task objective while the Discriminator discerns real samples from the Generator’s output. A wealth of GAN-based methodologies have been proposed, including conditional GANs [10], deep convolutional GANs (DCGANs) [43], Wasserstein GANs (WGANs) [1], and CycleGANs [49]. Recently, GANs have been incorporated into 3D vision tasks, such as point cloud up-sampling [20] and unsupervised point cloud completion [45]. Notably, some methods like [27] have utilized GANs to predict dense coordinates for instance-level 6D object pose estimation. However, to the best of our knowledge, the application of GANs in category-level 6D pose estimation remains unexplored. In light of the demonstrated versatility and potency of Generative Adversarial Networks (GANs) in various tasks, we leverage these principles in ACR-Pose, a new approach for category-level 6D object pose estimation. Unlike previous methods, ACR-Pose employs GANs in canonical representation reconstruction, utilizing adversarial training between a Reconstructor and an MLP Discriminator to enhance the generation of essential relational information and improve 3D feature learning.

3 METHOD

3.1 Overview

Our work is dedicated to reconstructing the 3D canonical representation, such as the Normalized Object Coordinate Space (NOCS) [39], from an RGBD input. Our pipeline, summarized in Figure 2, is divided into three components.

In the initialization phase, a detector extracts the image patch \( I \in \mathbb{R}^{U \times V \times 3} \) and depth region \( D \in \mathbb{R}^{U \times V} \) of the target object from the RGBD input. In parallel, a shape prior \( S \in \mathbb{R}^{N_c \times 3} \) is learned for each category using an encoder-decoder network, similar to the approach of Tian et al. [35].

The second stage is the Reconstructor, the core of our proposed approach. The image patch is input into an Image Encoder backbone \( E_I \) to extract instance RGB image features \( f_I = E_I(I) \in \mathbb{R}^{U \times V \times C} \), where \( U \) and \( V \) are the image dimensions. The depth region is back-projected into a point cloud and fed into the Pose-Irrelevant Module (PIM) \( E_D \) to extract canonical-related instance point features \( f_D = E_D(D) \in \mathbb{R}^{N_p \times C} \), where \( N_p \) is the number of observed points. The shape prior is processed through a Shape Prior Encoder \( E_s \) to yield high-dimensional category prior features \( f_c = E_s(S) \in \mathbb{R}^{N_c \times C} \). The features from the three modalities are integrated and enhanced by the Relational Reconstruction Module (RRM), \( f_{combined} = RRM(f_I, f_D, f_c) \).

Finally, the features learned by both modules are used in the application of the Shape Prior Deformation function \( \phi \) to reconstruct the NOCS representation, \( N_{rec} = \phi(f_{combined}) \). The Discriminator, the third component of our pipeline, employs an adversarial scheme during training. Specifically, the reconstructed NOCS representation \( N_{rec} \) and the ground-truth NOCS \( N_{gt} \) are evaluated by the Discriminator \( D \) to assess their quality levels, \( Q_{rec}, Q_{gt} = D(N_{rec}), D(N_{gt}) \). This guides our Reconstructor to generate NOCS representations that are as realistic as possible, a process we refer to as Adversarial Reconstruction.

At the inference stage, the Umeyama algorithm \( U \) is applied to the reconstructed NOCS representation \( N_{rec} \) and the back-projected observed object depth \( D \) to recover the 6D object pose, \( P = U(N_{rec}, D) \). Further details of the Pose-Irrelevant Module (PIM) and the Relational Reconstruction Module (RRM) in the Reconstructor will be provided in the subsequent sections.

3.2 Pose-Irrelevant Module in Reconstructor

Our venture in reconstructing the Normalized Object Coordinate Space (NOCS) representation is propelled by the need to extract a canonical representation, one that is devoid of both rotation and translation. Traditional methodologies, such as those proposed by Wang et al. [39] and Tian et al. [35], typically use Multilayer Perceptrons (MLPs) to extract features from the depth input. This results in an encoded output that encompasses shape, translation, and rotation-related information. In the context of our objective, however, the information related to translation and rotation introduces unnecessary noise, thus requiring removal.

To address this, we introduce the Pose-Irrelevant Module (PIM). The PIM is a dual-component mechanism that consists of a normalization filter, employed to negate the effects of translation, and a Rotation-Invariant Network, fashioned to suppress the influence of rotation. The details of PIM is as follows:

1. Normalization Filter: The first step in the PIM pipeline involves the normalization filter. This filter mitigates the effects of translation by calculating:

\[
\hat{P} = P - \bar{P},
\]

where \( P \) represents the back-projected point cloud and \( \bar{P} \) its geometric center.

2. Rotation-Invariant Convolution: The normalized point cloud \( \hat{P} \) is then processed through several layers of Rotation-Invariant Convolution to derive rotation-invariant features. The process includes 4 stages:

**Mean Geometric Center Calculation:** The mean geometric center of the group is calculated as:

\[
G = \frac{1}{K+1} \sum_{j=0}^{K} x_{ij}
\]

where \( x_{ij} \) represents each point in the group, including \( x_{i} \) itself.

**Nearest and Farthest Points Determination** The nearest and farthest points relative to \( x_{i} \) are identified as:

\[
x_{in} = \arg\min_{x_{ij} \in \text{group}} ||x_{i} - x_{ij}||
\]

\[
x_{if} = \arg\max_{x_{ij} \in \text{group}} ||x_{i} - x_{ij}||
\]
1. Relational Instance Feature, \( f_{\text{ins}} \in \mathbb{R}^{N_p \times 2C} \): This feature encapsulates the point-wise features of the observed object, including texture and geometric properties. It is constructed by concatenating each point’s point feature and RGB image feature from \( f_p \) and \( f_i \), respectively, thereby forming a \( N_p \times 2C \) feature matrix. Following a dimensionality reduction via a Multilayer Perceptron (MLP), a graph \( G(V, E) \) is constructed within the feature space, where \( V \) denotes nodes (each node representing an observed point) and \( E \) signifies edges. An EdgeConv operation \([41]\) is then used to learn features from this graph. This feature not only enriches the semantics in the feature space but also suppresses noise in RGBD images and enhances the learning of local details. It is computed as:

\[
x_{ij} = \arg \max_{x_{ij} \in \text{group}} \| x_i - x_j \|_2
\]

Rotation-Invariant Feature Calculation: For each point \( x_{ij} \) in this group, nine rotation-invariant geometric features and two spherical signals are calculated. These features are processed using a Multi-Layer Perceptron (MLP) and a Squeeze-and-Excitation (SE) block \([14]\) to learn the rotation-invariant convolutional kernels, denoted by \( \mathcal{K} \).

Feature Updating: The learned kernels \( \mathcal{K} \) are applied to the input point-wise features \( \mathcal{F} \) through a PointConv operation to update these features:

\[
\mathcal{F}_{\text{updated}} = \text{PointConv}(\mathcal{F}, \mathcal{K})
\]

The resulting output from the PIM, denoted as \( f_p \in \mathbb{R}^{N_p \times C} \), represents the canonical-related features, where \( N_p \) is the number of observed points in the depth region or the point cloud. It is then integrated with features derived from the remaining two modalities to accomplish the reconstruction of the NOCS representation. The implementation of the PIM effectively filters out extraneous information, thereby facilitating a more accurate and robust representation.

3.3 Relational Reconstruction Module in the Reconstructor

The Relational Reconstruction Module (RRM) is a critical component of the Reconstructor, engineered to extract superior features from three distinct data modalities. These features are pivotal in the precise reconstruction of the Normalized Object Coordinate Space (NOCS) representation. The RRM encompasses three discrete graph convolutional networks that learn and merge three types of relational features in the feature space, each fulfilling a unique function:

1. Relational Instance Feature. \( f_{\text{ins}} \in \mathbb{R}^{N_p \times 2C} \): This feature encapsulates the point-wise features of the observed object, including texture and geometric properties. It is constructed by concatenating...
The design of RRM facilitates the effective integration of relational graphs’ connectivity and message passing capabilities are determined by the embedded high-level relational information, thereby supplying a richer context for shaping the prior. It is computed as:

\[ v_{\text{ins}} = \text{MLP}_{\text{global}}(f_{\text{ins}}) \]  
\[ v_c = \text{MLP}_{\text{global}}(f_c) \]

2. Relational Deformation Feature. \( f_d \in \mathbb{R}^{N_c \times 3} \): This feature is derived from the global feature vectors \( v_{\text{ins}} \) and \( v_c \), as well as prior features \( f_c \). The feature functions as a mechanism to deform the shape prior into a canonical model of the observed object. The graph network, constructed on nodes representing object pixels and their multi-modal features, augments the fusion of these features. The graph’s connectivity and message passing capabilities are determined by the embedded high-level relational information, thereby supplying a richer context for shaping the prior. It is computed as:

\[ f_d = \text{MLP}_{\text{up}}(\text{EdgeConv}(\text{MLP}_{\text{down}}([v_{\text{ins}}, v_c, f_c]))) \]

3. Relational Assignment Feature. \( f_a \in \mathbb{R}^{N_a \times 3} \): This feature is employed in the Shape Prior Deformation step to assign the deformed shape prior to the NOCS representation. It establishes a correspondence between the shape prior and the NOCS representation, leveraging observed instance points and shape variation. The feature is calculated from the global feature vectors and the relational instance feature, as follows:

\[ f_a = \text{MLP}_{\text{up}}(\text{EdgeConv}(\text{MLP}_{\text{down}}([v_{\text{ins}}, v_c, f_{\text{ins}}]))) \]

To summarize, the RRM utilizes three types of relational features, each contributing distinctively to the final object reconstruction. The design of RRM facilitates the effective integration of relational information between features of different modalities. The construction of graphs within the feature space is advantageous in learning more intricate semantics about object shape and category specialties. These learned features play a critical role in the Shape Prior Deformation step for the final reconstruction, significantly contributing to top-tier NOCS representation reconstruction, as substantiated by experimental analysis.

3.4 Shape Prior Deformation in Reconstructor

Building upon the work of Tian et al. [35], we leverage the Shape Prior Deformation approach to reconstruct the NOCS representation. Our implementation, however, introduces significant advancements, most notably the adversarial reconstruction framework and the novel PIM and RRM.

In the Shape Prior Deformation operation, two transformation matrices—deformation field \( M_d \in \mathbb{R}^{N_c \times 3} \) and assignment matrix \( M_a \in \mathbb{R}^{N_a \times N_c} \)—are learned. The relational deformation feature is processed through an MLP to obtain \( M_d \), while \( M_a \) is derived from the relational assignment features via another MLP. The NOCS representation is subsequently reconstructed as:

\[ P_{\text{NOCS}} = M_a (P_c + M_d) \]  

where \( P_c \) is the shape prior. This two-stage operation first deforms \( P_c \) into the canonical instance model of the observed points, before mapping it to the NOCS representation. This fosters robustness in the reconstruction process, marking the culmination of our Reconstructor operation.

3.5 Discriminator and Adversarial Training

Our Reconstructor, while adept at predicting NOCS representations, may still produce unrealistic reconstructions. To rectify this and further improve reconstruction quality, we introduce an adversarial reconstruction strategy, deploying a Discriminator to guide the Reconstructor.

The Discriminator, equipped with a MLP, a global max-pooling operation, and two fully connected layers, is trained to distinguish between real and reconstructed NOCS representations. The Discriminator’s loss function is defined as:

\[ L_d = (\mathbb{E}(P_{\text{NOCS}}) - 1)^2 + (\mathbb{E}(\hat{P}_{\text{NOCS}}))^2 \]  
\[ L_g = (\mathbb{E}(\hat{P}_{\text{NOCS}}) - 1)^2 \]

During training, a competitive dynamic unfolds between the Discriminator and the Reconstructor, enhancing the realism of the reconstructed NOCS representation. Hence, our adversarial strategy contributes substantially to the accuracy of 6D object pose estimation.

3.6 Loss Function

Beyond adversarial losses \( L_d \) and \( L_g \), we employ multiple loss functions to enhance the quality of the reconstructed NOCS representations. The smooth L1 loss \( L_{\text{corr}} \) is used to ensure accurate correspondence between the reconstructed and the ground-truth representations. The chamfer distance loss \( L_{\text{cd}} \) retains appearance information, measuring distance between the deformed shape prior and the object’s canonical model. We also introduce a cross-entropy loss \( L_{\text{entro}} \) for a peak distribution of the assignment matrix \( M_a \), along with an L2 regularization loss \( L_{\text{reg}} \) on \( M_a \) to prevent deformation collapse. The final loss function is:

\[ L = \gamma_1 L_d + \gamma_2 L_g + \gamma_3 L_{\text{corr}} + \gamma_4 L_{\text{cd}} + \gamma_5 L_{\text{entro}} + \gamma_6 L_{\text{reg}} \]

Here, \( \gamma_1 \) to \( \gamma_6 \) are balance terms.

4 EXPERIMENTS

4.1 Datasets

We conduct rigorous experiments using the NOCS-CAMERA and NOCS-REAL datasets, as proposed by Wang et al. [39]. Currently, these datasets are recognized as the most authoritative and broadly utilized benchmarks for evaluating category-level 6D object pose estimation methodologies. The NOCS-CAMERA dataset is a synthetic dataset comprising 300K RGBD images (with 25K designated for evaluation purposes), generated by integrating rendered synthetic objects into actual scenes. The NOCS-REAL dataset, conversely, is derived from real-world scenarios and encompasses 4.3K real-world RGBD images from 7 different scenes for training and 2.75K real-world RGBD images from 6 scenes for evaluation. Both datasets consist of six distinct categories namely, bottle, bowl, camera, can,
We implement the advanced ACR-Pose methodology utilizing the PyTorch framework. To optimize the model, we employ the Adam optimizer with a batch size set at 96. For object detection, we adopt the Mask-RCNN [11]. The Image Encoder is constructed as a PSPNet [47], underpinned by a ResNet-18 [12] backbone. Concurrently, the Shape Prior Encoder is built as a Multilayer Perceptron (MLP). The category instances are sourced from the ShapeNet dataset, and the vertices of the ground truth and estimated 3D models post their alignment based on the estimated poses.

### 4.2 Implementation details

We implement the advanced ACR-Pose methodology utilizing the PyTorch framework. To optimize the model, we employ the Adam optimizer with a batch size set at 96. For object detection, we adopt the Mask-RCNN [11]. The Image Encoder is constructed as a PSPNet [47], underpinned by a ResNet-18 [12] backbone. Concurrently, the Shape Prior Encoder is built as a Multilayer Perceptron (MLP). The category instances are sourced from the ShapeNet dataset, and the feature dimension, denoted as \( C \), is set as 64. Throughout the training process, we undertake an alternate update of the parameters for both the Reconstructor and the Discriminator. The initial learning rates for the Reconstructor and the Discriminator are set at 0.0001 and 0.00001, respectively. During the exclusive training on the NOCS-CAMERA train set, the learning rate undergoes decay at the 10th, 30th, and 40th epochs, with respective decay rates of 0.5, 0.1, and 0.01 with respect to the initial learning rate. During the fine-tuning process, the initial learning rates are similarly set to 0.0001 and 0.00001, and they undergo halving at the 5th epoch. The image patch is resized to a dimension of \( 192 \times 192 \), and the number of object points (back-projected object depth), denoted as \( P \), is 1024. We also set the number of shape prior points to 1024, while the number of adjacent neighbors, represented as \( K \) in the feature graphs, is 36. The balance terms, designated as \( \gamma_1 \) through \( \gamma_5 \) in the loss function, are respectively set as 0.1, 0.1, 1.0, 5.0, 0.0001, and 0.01. In the fine-tuning phase, we reinitialize the Discriminator and randomly select data from NOCS-CAMERA and NOCS-REAL at a 3:1 ratio for network training, in accordance with [35, 39]. All experiments are conducted on a single A6000 GPU. We follow [35] to use IoU50, IoU75, \( 5° \)cm, \( 5° \)scm, \( 10° \)cm2 and \( 10° \)scm as our metrics and report the mean Average Precision (mAP).

### 4.3 Comparison Results

**Improvement over baseline models:** We adopt two baseline methods, SPD [35] and SGPA [6], for a comprehensive comparison with our proposed method. SPD, which uses MLPs as the backbone for learning features from the point cloud, is chosen as the baseline for our naive version, referred to as ACR-Pose-MLP. We also propose an enhanced version of our method, ACR-Pose-PN2, which employs PointNet++ [29] as the backbone to extract more expressive canonical-related features. For ACR-Pose-PN2, SGPA, which also uses PointNet++, is selected as the baseline. The detailed comparison is provided in Table 1. As shown in Table 1, ACR-Pose-MLP significantly outperforms SPD across all metrics on both NOCS-REAL and NOCS-CAMERA datasets. For the NOCS-REAL dataset, ACR-Pose-MLP improves the IoU50, IoU75, \( 5° \)cm, and \( 10° \)cm2 metrics by 5.5%, 12.8%, 12.3%, and 11.6%, respectively, compared to SPD. For the NOCS-CAMERA dataset, the improvements are even more significant: 0.6%, 6.8%, 16.1%, 15.1%, and 9.3% for IoU50, IoU75, \( 5° \)cm, \( 5° \)scm, and \( 10° \)cm2 metrics, respectively. These results demonstrate the effectiveness of our adversarial reconstruction scheme and the improved performance of our ACR-Pose-MLP method.
Figure 4: Visualization result of ACR-Pose. Green boxes are the ground-truth and red boxes are our predictions. The top two rows are results on the NOCS-CAMERA dataset and the bottom two rows are results on the NOCS-REAL dataset.

model over SPD. The comparison between SGPA and ACR-Pose-PN2 also reveals superior performance of our method on several metrics. For the NOCS-REAL dataset, our ACR-Pose-PN2 model beats SGPA by 2.2%, 4.7%, and 1.4% on IoU50, IoU75, and 5°2cm metrics, respectively. However, it slightly underperforms in the 10°2cm and 10°5cm metrics, which suggests areas for further improvement. On the NOCS-CAMERA dataset, ACR-Pose-PN2 outperforms SGPA across all metrics with improvements of 0.5%, 1.5%, 0.9%, 0.6%, 1.8%, and 1.1% for IoU50, IoU75, 5°2cm, 5°5cm, 10°2cm, and 10°5cm metrics, respectively. In summary, our proposed methods, ACR-Pose-MLP and ACR-Pose-PN2, demonstrate superior performance over the corresponding baselines, SPD and SGPA, in most metrics across both NOCS-REAL and NOCS-CAMERA datasets. These results confirm the effectiveness and robustness of our adversarial reconstruction scheme in improving 6D object pose estimation.

Comparison with state-of-the-art methods: We also compare our method with recently sota methods. As shown in Table 2, our methods - ACR-Pose-MLP and ACR-Pose-PN2 - demonstrate comparable, and in some instances superior, performance to the current state-of-the-art methods. In the NOCS-CAMERA dataset, ACR-Pose-MLP and ACR-Pose-PN2 achieved the highest score in IoU75 and 5°2cm metrics, with scores of 89.9 and 71.6, respectively. These results indicate that our methods excel in handling strict metrics, emphasizing their precision in determining pose estimation. Moreover, ACR-Pose-PN2 outperformed all other methods in the 5°5cm evaluation with a score of 75.1, demonstrating its superior capability in capturing object pose at a 5° angle within a 5 cm range. Simultaneously, ACR-Pose-MLP achieved the highest score of 82.6 in the 10°2cm metric, underlining its accuracy at a broader angle and stricter distance. When tested on the NOCS-REAL dataset, which is more challenging due to the presence of real-world noise, our methods still exhibited robust performance. Specifically, the ACR-Pose-PN2 method achieved the highest scores in IoU75, 5°2cm, and 5°5cm metrics, with scores of 66.6, 36.7, and 41.3, respectively. These results underscore our methods’ adaptability to real-world scenarios and their superior performance even under strict evaluation metrics. In conclusion, our methods not only match but also outperform existing state-of-the-art methods in various critical evaluation metrics, highlighting their effectiveness and robustness in both synthetic and real-world datasets. These results position our methods as promising solutions for category-level object pose estimation tasks.

Qualitative results: Expanding on our quantitative evaluation, we offer a qualitative visual assessment of our proposed method, as depicted in Fig. 4. Our model demonstrates its proficiency in accurately predicting 6D object poses. The visualized bounding boxes exhibit a substantial overlap with the ground-truth boxes, indicating the precision of our object pose estimation even in diverse scenarios. This high-level of consistency between the predicted and actual object poses underscores the robustness and reliability of our model, especially in its ability to handle both synthetic and real-world situations. Despite the promising results, we recognize limitations, including instances of object truncation, occlusion, and occasional detection failures, as shown in Fig. 5. These areas for refinement will be the focus of our future work.

4.4 Ablation Study

This subsection presents the ablation experiments on the NOCS-CAMERA dataset. These experiments were designed around the ACR-Pose-MLP framework, in which we methodically removed
the innovative components from ACR-Pose and then sequentially reintroduced them to evaluate their individual impacts on the final 6D object pose estimation. The results are shown in Table 3.

**Significance of the relational reconstruction module:** The RRM is a fundamental component of our Reconstructor, specifically designed to exploit relational information among three separate modalities. It consists of three graphs, each tailored to learn relational instance features, relational deformation features, and relational assignment features, correspondingly. A review of rows 4, 6, and 7 in Table 3 indicates that each type of feature incrementally contributes to the model’s accuracy. This improvement is credited to the message propagation across graphs in the feature space, enabling the model to learn robust semantic features. These features effectively capture both specific object shape attributes and common category-level characteristics, leading to superior NOCS representations. The results validate our hypothesis that relational features are crucial for achieving high-accuracy NOCS representation reconstruction.

**Role of the pose-irrelevant module:** The PIM, a unique feature of our model, is crucial for learning canonical-related features that efficiently eliminate rotation and translation information, which are irrelevant to the reconstruction of high-quality NOCS representations. When integrated into our baseline model, all evaluation metrics recorded an increase ranging from 1% to 3%. This improvement is attributed to the fact that the model, devoid of rotation and translation distractions, is able to concentrate on learning information about the shape of the object’s visible part, a key aspect for NOCS representation reconstruction.

**Impact of adversarial reconstruction:** The adversarial reconstruction strategy boosts the quality and realism of the NOCS representations. The Discriminator’s involvement is limited to the training phase, making this an efficient approach for performance enhancement. Evidence from the 5th row of Table 3 supports that the adversarial training scheme significantly enhances the model’s capacity to learn high-quality and more realistic canonical representations. The addition of the Discriminator makes the reconstructed canonical representations more realistic, thereby improving the robustness and performance of the subsequent Umeyama algorithm [37].

## 5 CONCLUSION

In this paper, we present an Adversarial Canonical Representation Reconstruction Network, the ACR-Pose, for accurate category-level 6D object pose estimation. Our work focuses on reconstructing high-quality canonical representations for the observed objects, given the RGB patch, the depth region, and the learned shape prior. To achieve so, we propose an adversarial reconstruction scheme, which competitively trains a Reconstructor and a Discriminator to improve reconstruction performance. Besides, within our reconstructor, a Pose-Irrelevant Module is introduced to better extract the shape information of the objects, as well as a Relational Reconstruction Module that leverages the relational information among the three input sources. Experiments are conducted on synthetic and real datasets, where the effectiveness of each novel module is demonstrated.

## 6 ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (NSFC) under Grant Nos. 62072459 and 62172421.
REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Leon Bottou. 2017. Wasserstein genera-
tive adversarial networks. In Int’l Conf. on machine learning. PMLR, 214–223.

[2] Ronald T Azuma. 1997. A survey of augmented reality. Presence: teleoperators &
virtual environments 6, 4 (1997), 355–385.

[3] Antonio Bicchi and Vijay Kumar. 2000. Robotic grasping and contact: A review. In Proceedinggs 2000 ICRA. Millennium Conf. IEEE Int’l Conf. on Robotics and Automa-
tion. Symposium Proceedings. (Cat. No.00CH37065), Vol. 1. IEEE, 348–353.

[4] Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Cervello, Ernesto Damiani,
and Misa Ivkovic. 2011. Augmented reality technologies, systems and applications.

[5] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks.

[6] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xi-

[7] Lei Ke, Shichao Li, Yanan Sun, Yu-Wing Tai, and Chi-Keung Tang. 2020. Gsnet:

[8] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In Proceed-
inggs of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 6781–6791.

[9] Zhaoxun Fan, Yanzhi Zhu, Yulin He, Qi Sun, Hongyan Liu, and Jun He. 2021. Deep
Learning on Monocular Object Pose Detection and Tracking: A Comprehensive Overview. In arXiv preprint arXiv:2105.14292 (2021).

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial
nets. In Advances in neural information processing systems 27 (2014).

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual
learning for image recognition. In Proceedings of the IEEE Conf. on computer vision
and pattern recognition. 770–778.

[13] Yisheng He, Shijing Lu, Yimin Li, and Yongdong Yu. 2019. Dual head for cat-
cegory-level 6d object pose estimation. In Proceedings of the IEEE/CVF Conf. on Computer Vision
and Pattern Recognition. 18221–28229.

[14] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator archi-
tecture. In ECCV '18: Proceedings of the 12th European Conference on Computer Vision–ECCV 2018. 17th European Conf., Tel Aviv, Israel, October 23–27, 2018.

[15] Weichi Chen, Xiangyang Hu, and Misa Ivkovic. 2011. Augmented reality technolo-
gies, systems and applications. Springer, 348–353.

[16] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet:
Deep learning on point sets in a metric space. In Proceedings of the IEEE Conf. on computer vision
and pattern recognition. 4561–4570.

[17] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++:
Deep hierarchical feature learning on point sets in a metric space. Advances in
neural information processing systems 30 (2017).

[18] Ashutosh Saxena, Justin Dienerwey, and Andrew Y Ng. 2008. Robotic grasping
of novel objects using vision. The Int’l Journal of Robotics Research 27, 3 (2008), 157–173.

[19] Chen Song, Jiaru Song, and Qixing Huang. 2020. Hybridpose: 6d object pose estimation
under hybrid representations. In Proceedings of the IEEE/CVF Conf. on computer vision
and pattern recognition. 431–440.

[20] Xihan Song, Peng Peng, Dingzhu Zhou, Rui Zhu, Chengyu Guan, Yuchao Dai, Hao
Su, Hongdong Li, and Ruigang Yang. 2019. ApolloCar: A large 3d car instance
understanding benchmark for autonomous driving. In Proceedings of the IEEE/CVF
Conf. on Computer Vision and Pattern Recognition. 5452–5461.

[21] Rui Sun, Henrik Kretzschmar, Xavers Dotwaltla, Aurelien Chouraou, Vijayasai
Patnaka, Paul Tsui, James Guo, Yin Zhou, Yuning Chi, Benjamin Caue, et al. 2020.
Scalability in perception for autonomous driving: Waymo open dataset. In Proceed-
inggs of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2446–2454.

[22] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. 2018. Real-time seamless single
shot 6d object pose prediction. In Proceedings of the IEEE Conf. on computer vision
and pattern recognition. 292–301.

[23] Meng Tian, Marcello H Ang, and Gim Hee Lee. 2020. Shape prior deformation for
categorical 6d object pose and size estimation. In European Conf. on Computer Vision.
Springer, 530–546.

[24] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox,
and Stan Birchfield. 2018. Deep object pose estimation for semantic robotic
grasping of household objects. arXiv preprint arXiv:1809.10790 (2018).

[25] Shiniyi Ueyama. 1991. Least-squares estimation of transformation parameters
between two point patterns. IEEE Transactions on Pattern Analysis & Machine
Intelligence 13, 8 (1991), 736–380.

[26] Petr Vávra, Jan Roman, Pavel Zonca, Peter Ihaň, Martin Němec, Jayant Kumar,
Nagy Habib, and Ahmed El-Gendi. 2017. Recent development of augmented
reality in surgery: a review. Journal of Healthcare engineering 2017 (2017).

[27] He Wang, Sinath Sridhar, Jingwei Huang, Julian Valentin, Shurang Jahan,
and Leonidas J Guibas. 2019. Normalized object coordinate space for category-level 6d
object pose and size estimation. In Proceedings of the IEEE/CVF Conf. on Computer Vision
and Pattern Recognition. 2642–2651.

[28] Wenyuan Zhang, Jiemin Fang, Xinggang Wang, and Wenyu Liu. 2021. Ef-
tive adversarial network. In Proceedings of the IEEE/CVF Conf. on Computer Vision
and Pattern Recognition. 3227–3236.

[29] Meng Tian, Marcello H Ang, and Gim Hee Lee. 2020. Shape prior deformation for
categorical 6d object pose and size estimation. In Proceedings of the IEEE/CVF Conf. on Computer Vision
and Pattern Recognition. 2642–2651.

[30] Junhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi,
Chai Kiat Yeoh, Bo Dai, and Chen Change Loy. 2021. Unsupervised 3D Shape
Completion through GAN Inversion. In Proceedings of the IEEE/CVF Conf. on Computer Vision
and Pattern Recognition. 1768–1777.

[31] Wei Wang, Mingling Niu, Jinhua Zhu, and Yanqi Chen. 2020. Point-grasping
for 6d pose estimation. In IEEE/RSJ Int’l Conf. on Intelligent Robots and Systems (IROS).
IEEE, 4807–4814.

[32] Yang Yu, Zhiqiang Gong, Ping Zhong, and Jiaxin Shan. 2017. Unsupervised
image-to-image translation using cycle-consistent adversarial networks. In Proceed-
inggs of the IEEE Conf. on computer vision and pattern recognition. 1768–1777.

[33] Meng Tian, Marcello H Ang, and Gim Hee Lee. 2020. Shape prior deformation for
categorical 6d object pose and size estimation. In Proceedings of the IEEE/CVF Conf. on Computer Vision
and Pattern Recognition. 2642–2651.

[34] Yew Wang, Yongbin Sun, Zewei Liu, Sanjay E. Sarma, Michael M Bronstein,
and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. Acm
Transactions On Graphics (tog) 38, 5 (2019), 1–12.

[35] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2017.
Posecnn: A convolutional neural network for 6d object pose estimation in cluttered
scenarios. arXiv preprint arXiv:1711.01999 (2017).

[36] Yang Yu, Zhiqiang Gong, Ping Zhong, and Jiaxin Shan. 2017. Unsupervised
representation learning with deep convolutional neural network for remote
sensing images. In Int’l Conf. onImage and Graphics. Springer, 97–108.

[37] Sergey Zakharyev. Ivan Shugurov, and Shohanad Ilic. 2019. Dpdp: 6d object
pose detector and refiner. In Proceedings of the IEEE/CVF Int’l Conf. on Computer Vision.
1941–1950.

[38] Tianzi Jiang, Xiaolong Wang, and Chen Change Loy. 2021. Unsupervised
3D Shape Completion through GAN Inversion. In Proceedings of the IEEE/CVF Conf. on Computer Vision
and Pattern Recognition. 1768–1777.

[39] Wenqiang Zhang, Jiemin Fang, Xianggang Wang, and Wenyu Liu. 2021. Ef-
ticientpose: Efficient human pose estimation with neural architecture search.
Computational Visual Media 7, 3 (2021), 335–347.

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017.
Pyramid scene parsing network. In Proceedings of the IEEE Conf. on computer vision
and pattern recognition. 2881–2890.

[41] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2016. Energy-based generative
adversarial network. arXiv preprint arXiv:1609.02162 (2016).

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired
image-to-image translation using cycle-consistent adversarial networks. In Proceed-
inggs of the IEEE Int’l Conf. on computer vision. 2223–2232.