Robust Category-Level 6D Pose Estimation with Coarse-to-Fine Rendering of Neural Features

Wufei Ma[1], Angtian Wang[1], Alan Yuille[1], and Adam Kortylewski[1,2,3]

[1] Johns Hopkins University, Baltimore MD 21218, USA
{wma27,angtianwang,ayuille1,akortyl1}@jhu.edu
[2] Max Planck Institute for Informatics, Saarbrücken, Germany
[3] University of Freiburg, Germany

Abstract. We consider the problem of category-level 6D pose estimation from a single RGB image. Our approach represents an object category as a cuboid mesh and learns a generative model of the neural feature activations at each mesh vertex to perform pose estimation through differentiable rendering. A common problem of rendering-based approaches is that they rely on bounding box proposals, which do not convey information about the 3D rotation of the object and are not reliable when objects are partially occluded. Instead, we introduce a coarse-to-fine optimization strategy that utilizes the rendering process to estimate a sparse set of 6D object proposals, which are subsequently refined with gradient-based optimization. The key to enabling the convergence of our approach is a neural feature representation that is trained to be scale- and rotation-invariant using contrastive learning. Our experiments demonstrate an enhanced category-level 6D pose estimation performance compared to prior work, particularly under strong partial occlusion.

Keywords: Category-level 6D pose estimation, Render-and-Compare

1 Introduction

Estimating the 3D position and 3D orientation of objects is an important requirement for a comprehensive scene understanding in computer vision. Real-world applications, such as augmented reality (AR) or robotics, require vision systems to generalize in new environments that may contain previously unseen and partially occluded object instances. However, most prior work on 6D pose estimation focused on the "instance-level" task, where exact CAD models of the object instances are available [35, 23, 19, 10, 12]. Moreover, the few prior methods on "category-level" 6D pose estimation often either rely on a ground truth depth map [31, 20], which are practically hard to obtain in many application areas, or rely on 2D bounding box proposals [38, 28], which are not reliable in challenging occlusion scenarios [30] (see also our experimental results).

Recent work introduced generative models of neural network features for image classification [14] and 3D pose estimation [28], which have the ability to learn category-level object models that are highly robust to partial occlusion.

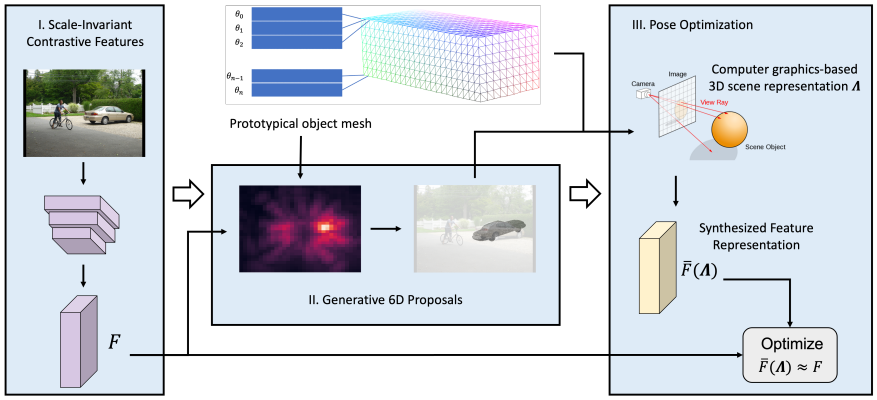

Fig. 1: Overview of our coarse-to-fine 6D pose estimation. We propose to train a neural representation that is invariant to instance-specific details, 3D rotation, and changes in the object scale. With the help of the scale-invariant contrastive features, we can efficiently search generative 6D proposals in the coarse stage and then refine the initial 6D poses with pose optimization in the fine stage.

Intuitively, these models are composed of a convolutional neural network [17] and a Bayesian generative model of the neural feature activations. The invariance properties of the neural features enable these models to generalize despite variations in instance-specific details such as changes in the object shape and texture. Moreover, the generative model can be augmented with an outlier model [11] to avoid being distorted by local occlusion patterns.

In this work, we build on and significantly extend generative models of neural network features to perform category-level 6D pose estimation from a single RGB image. In particular, we follow neural mesh models [28] and represent an object category as a cuboid mesh and learn a generative model of the neural feature activations at each mesh vertex to perform pose estimation through a render-and-compare process. The core problem of such a rendering-based approach to pose estimation is to search efficiently through the combinatorially large space of the 6D latent parameters, because the iterative rendering process is rather costly compared to simple feed-forward regression approaches. Related work addresses this problem by first estimating 2D object bounding boxes with a proposal network [18, 31, 12], but these are not reliable under partial occlusion and truncation [30]. Instead, we address this problem by extending neural mesh models with scale-invariant features and a coarse-to-fine render-and-compare optimization strategy, which retains the robustness to partial occlusion (Figure 1). In particular, we first use a coarse search strategy, in which we render the model in a set of pre-defined initial poses that are evenly distributed over the entire search space and select a sparse set of candidate initializations with low reconstruction loss. This process generates 6D object proposals robustly under

occlusion, as it relies on the generative object model. Subsequently, the 6D proposals are refined with gradient-based render-and-compare optimization.

The key to making the coarse-to-fine optimization efficient and accurate (i.e., not missing out on small or partially occluded objects) is to learn a feature representation that induces large convergence basins in the optimization process. To this end, we introduce a contrastive learning framework [9, 32, 1] to learn features that are invariant to instance-specific details (such as changes in the shape and texture), as well as to changes in the 3D pose and scale of the object.

We evaluate our model on the PASCAL3D+ dataset [34] and the ObjectNet3D dataset [33], which are challenging real-world datasets of outdoor and indoor scenes, respectively. Our experiments demonstrate that our model outperforms strong object detection and pose estimation baseline models. Our model further demonstrates exceptional robustness to partial occlusion compared to all baseline methods on the Occluded PASCAL3D+ dataset [30].

The main contributions of our work are:

1. We introduce a render-and-compare approach for category-level 6D pose estimation and adopt a coarse-to-fine pose estimation strategy that is accurate and highly robust to partial occlusion.
2. We use a contrastive learning framework to train a feature representation invariant to instance-specific details, 3D rotation, and changes in object scale.
3. The invariant features enable a coarse-to-fine render-and-compare optimization, which involves novel generative 6D object proposals and a subsequent gradient-based pose refinement.
4. Our method outperforms previous methods on the PASCAL3D+, ObjectNet3D dataset, and we demonstrate the robustness of our model to partial occlusion on Occluded PASCAL3D+ dataset. We further show the efficacy of our proposed modules in the ablation study.

2 Related Work

Category-level 3D pose estimation. Category-level 3D pose estimation assumes the bounding box of the object is given and predicts the 3D object pose. Previous methods can be categorized into two groups, keypoint-based methods and render-and-compare methods. Keypoint-based methods [22, 38] first detect semantic keypoints and then predict 3D object pose by solving a Perspective-n-Point problem. Render-and-compare methods [28] predict the 3D pose by fitting a rigid transformation of the mesh model that minimizes the reconstruction error between a predicted feature map and a rendered feature map. 3D pose estimation methods often exploit the inductive bias that the principal points of the objects are close to the image center and the objects have a similar scale.

Category-level 6D pose estimation. Category-level 6D pose estimation is a more challenging problem and involves object detection and pose estimation without knowing the accurate 3D model or the textures of the testing objects. Previous methods [31, 3, 20] often investigate this problem in the RGBD setting.

Depth maps help these models to infer the 3D location of the objects and at the same time resolve the scale ambiguities [3]. However, depth annotations are often hard to obtain, which limits the practicality of these methods. In this work, we investigate category-level 6D pose estimation from monocular RGB images and show that our method can robustly estimate 6D object poses under partial occlusion and truncation.

Feature-level render-and-compare. Render-and-compare methods minimize the reconstruction error between a predicted feature representation and a representation rendered from a 3D scene (e.g., a 3D mesh model $\mathcal{M}$ and the corresponding 3D pose m). Previous methods follow similar formulations but differ in the feature representation and the optimization algorithms. Wang et al. [31] proposed to hard-code the features as the normalized 3D coordinates and predict the object pose by solving a rigid transformation between the 3D model $\mathcal{M}$ and the predicted coordinate map with the Umeyama algorithm [27]. NeMo [28] learns contrastive features for the 3D model $\mathcal{M}$ and solves 3D object pose with the objects centered and rescaled. Iwase et al. [12] found that features with only 3 channels are sufficient for instance-level 6D pose estimation and proposed to learn the features with a differentiable Levenberg-Marquardt (LM) optimization.

In a broader context, the feature-level render-and-compare process can be interpreted as an *approximate* analysis-by-synthesis [7, 8] approach to computer vision. Analysis-by-synthesis has several advantages over purely discriminative methods as it enables efficient learning [29] and largely enhances robustness in out-of-distribution situations, particularly when objects are partially occluded in image classification [15, 16, 37, 36], object detection [30], scene understanding [25, 21], face reconstruction [5] and human detection [6], as well as when objects are viewed from unseen 3D poses [28]. Our work extends the approximate analysis-by-synthesis approach to category-level 6D pose estimation.

3 Method

This section presents our main contributions. First, we review the render-and-compare approach for pose estimation in Section 3.2. Then we introduce the learning of scale-invariant contrastive features in Section 3.3. In Section 3.4, we introduce a coarse-to-fine optimization strategy that uses a generative model to generate 6D object proposals in the coarse stage and then refines the initial 6D poses with a render-and-compare pose optimization. We discuss a multi-object reasoning module in Section 3.5 that enables our model to accurately detect occluded and truncated objects and as well as complicated multi-object scenes.

3.1 Notation

We denote a feature representation of an input image I as $\zeta(I) = F^l \in \mathbb{R}^{H \times W \times c}$. Where l is the output of layer l of a deep convolutional neural network ζ, with c being the number of channels in layer l. $f_i^l \in \mathbb{R}^c$ is a feature vector in F^l

at position i on the 2D lattice $\mathcal{P}$ of the feature map. In the remainder of this section, we omit the superscript l for notational simplicity because this is fixed a-priori in our model.

3.2 Prior Work: Render-And-Compare for Pose Estimation

Our work builds on and significantly extends neural mesh models (NMMs) [12, 28], which are themselves 3D extensions of Compositional Generative Networks [13]. Neural mesh models define a probabilistic generative models $p(F \mid \mathfrak{N})$ of the real-valued feature activations F using a 3D neural mesh representation $\mathfrak{N}$. The neural mesh $\mathfrak{N} = \{\mathcal{V}, \mathcal{E}, \mathcal{C}\}$ is represented by a set of vertices $\mathcal{V} = \{V_i \in \mathbb{R}^3\}_{i=1}^N$ and learnable features for each vertex $\mathcal{C} = \{C_i \in \mathbb{R}^c\}_{i=1}^N$, where c is the number of channels in layer l. Given the object pose (or camera viewpoint) m, we can render the neural mesh model $\mathfrak{N}$ into feature maps using rasteriation, i.e., $\bar{F}(m) = \mathfrak{R}(\mathfrak{N}, m) \in \mathbb{R}^{H \times W \times D}$. The neural mesh model defines the likelihood of a target feature map $F \in \mathbb{R}^{H \times W \times D}$ as

$$p(F \mid \mathfrak{N}, m, \mathcal{B}) = \prod_{i \in \mathcal{FG}} p(f_i \mid \mathfrak{N}, m) \prod_{i' \in \mathcal{BG}} p(f_{i'} \mid \mathcal{B}) \tag{1}$$

where the foreground $\mathcal{FG}$ is the set of all positions on the 2D lattice $\mathcal{P}$ of the feature map F that are covered by the rendered neural mesh model and the background $\mathcal{BG}$ contains those pixels that are not covered by the mesh. The foreground likelihood is defined as a Gaussian distribution $p(f_i \mid \mathfrak{N}, m) = \mathcal{N}(f_i \mid C_r, \sigma_r^2 I)$. The correspondence between the image feature f_i and the vertex feature C_r is determined through the rendering process. Background features are modeled using a simple background model that is defined by a Gaussian distribution $p(f_{i'} \mid \mathcal{B}) = \mathcal{N}(f_{i'} \mid b, \sigma^2 I)$ with $\mathcal{B} = \{b, \sigma\}$, which can be estimated with maximum likelihood from the background features. The training of the generative model parameters $\{\mathfrak{N}, B\}$ and the feature extractor is done by maximum likelihood estimation (MLE) from the training data. At test time, we can infer the object pose m by minimizing the negative log-likelihood of the model w.r.t. the pose m with gradient descent

$$\begin{aligned}
\mathcal{L}_{\text{NLL}}(F, \mathfrak{N}, m, \mathcal{B}) &= -\ln p(F \mid \mathfrak{N}, m, \mathcal{B}) \\
&= -\sum_{i \in \mathcal{FG}} \left(\ln \left(\frac{1}{\sigma_r \sqrt{2\pi}} \right) - \frac{1}{2\sigma_r^2} \| f_i - C_r \|^2 \right) \\
&\quad - \sum_{i' \in \mathcal{BG}} \left(\ln \left(\frac{1}{\sigma \sqrt{2\pi}} \right) - \frac{1}{2\sigma^2} \| f_{i'} - b \|^2 \right)
\end{aligned} \tag{2}$$

Assuming unit variance [28], i.e., $\sigma_r = \sigma = 1$, the loss function reduce to the mean squared error (MSE) between vertex features and the target feature map

$$\mathcal{L}_{\text{NLL}}(F, \mathfrak{N}, m, \mathcal{B}) = \frac{1}{2} \sum_{i \in \mathcal{FG}} \| f_i - C_r \|^2 + \frac{1}{2} \sum_{i' \in \mathcal{BG}} \| f_i - b \|^2 + \textit{const.} \tag{3}$$

Fig. 2: Illustration of our object-centric data augmentation strategy, to generate feature activations across several scales, which are essential for our contrastive learning framework.

Previous works adopted this general framework for category-level 3D pose estimation [28, 31] and instance-level 6D pose estimation [12], thereby using different types of learnable features $\mathcal{C}$ and optimization algorithms. In this work, we extend this framework to category-level 6D pose estimation from a single RGB image, which requires us to overcome additional challenges. Specifically, we need to address the challenge that the learnable feature representation $\mathcal{C}$ needs to account for the large variations in object scale, as well as the intra-category variation in terms of the object shape and texture properties.

3.3 Learning Scale-Invariant Contrastive Features

In this work, we propose to account for the variations in the object scale, shape, and appearance by learning contrastive features that are invariant to these variations. This will enable us to estimate the 6D object pose by optimizing the maximum likelihood formulation in Equation 3 directly with gradient-based optimization. We demonstrate the efficacy of our scale-invariant contrastive features in Figure 3 and quantitatively in Section 4.2.

Contrastive Learning of Scale-Invariant Features. One of the major challenges in 6D pose estimation is the variation in object scales. Due to the nature of convolution layers in the feature extractor ζ, nearby and distant objects could yield very different feature activations in F. Unfortunately, annotations of 6D poses for small objects are limited. Therefore, we use data augmentation to learn scale-invariant features from object-centric samples.

Specifically, given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we prepare the training sample as follows. First, we resize the image with scale s and obtain a new image with size $\frac{H}{s} \times \frac{W}{s}$. Then texture images from the Describable Textures Dataset (DTD) [4] are used to pad the image back to $H \times W$. We update the distance annotation d of the object assuming a pinhole camera model, such that the distance annotation can be computed as $d' = d \cdot s$. The augmented data is depicted in Figure 2.

In order for the CNN backbone ζ to extract feature invariant to instance-specific details and to avoid local optima in the loss landscapes of the reconstruction loss, we train the feature extractor ζ using contrastive learning to learn features that are distributed akin to the probabilistic generative model as

defined in Equations 1-3. We achieve this by adopting a contrastive loss:

$$\mathcal{L}_{\text{contrastive}} = -\sum_{i \in \mathcal{FG}} \sum_{j \in \mathcal{FG} \setminus \{i\}} \|f_i - f_j\|^2 - \sum_{i \in \mathcal{FG}} \sum_{j \in \mathcal{BG}} \|f_i - f_j\|^2 \tag{4}$$

which encourages the features of different vertices to be discriminative from each other and features of the object vertices distinct from the features in the background. Our full model is trained by optimizing $\mathcal{L}_{\text{contrastive}}$ in a contrastive learning framework, where we update the parameters of the feature extractor ζ and the vertices features $\mathcal{C}$ in the neural mesh model jointly.

MLE Learning of the Neural Mesh Model (NMM). We train the parameters $\mathcal{C}$ of the probabilistic generative model through maximum likelihood estimation (MLE) by minimizing the negative log-likelihood of the feature representations over the whole training set (Equation 3). The correspondence between the feature vectors f_i and vertices r is computed using the annotated 6D pose. To reduce the computational cost of optimizing Equation 3, we follow [1] and update $\mathcal{C}$ in a moving average manner.

Convergence properties. The benefits of the scale-invariant contrastive features are two-fold. First, the ground truth 6D pose is very close to the global minimum of the reconstruction loss in all six dimensions. We illustrate the 6D loss landscapes in Figure 3. Each curve corresponds to one of the six dimensions of the 6D pose and is centered at the ground truth pose. The large convergence basins that can be observed allow us to search for object proposals from simply sparse sampling and to evaluate a pre-defined set of 6D poses, without the need of a first-stage model widely used by related works [18, 31, 12]. Second, the loss landscapes are generally smooth around the global minimum. This contrasts with the keypoint-based methods that fit a rigid transformation between two groups of keypoints [18, 31] and the render-and-compare methods over RGB space [2, 26] with many local minima on the optimization surface.

3.4 Coarse-to-Fine 6D Pose Estimation

Previous methods for 3D object detection or 6D pose optimization are built on top of a 2D region proposal network or refine predictions from a separate pose estimation network. Although this approach was empirically effective, the performance of the hybrid model is largely limited by the 2D region proposal network or the initial pose estimation network. The first-stage networks are unreliable for objects with out-of-distribution textures or shapes, or even miss the object if the object is partially occluded or truncated.

Therefore, we propose a coarse-to-fine 6D pose estimation strategy that searches generative 6D proposals in the coarse stage and then refines the initial 6D poses with pose optimization in the fine stage. The overview of our coarse-to-fine strategy is depicted in Figure 4. Since the generative 6D proposals are built on the generative neural mesh models and scale-invariant contrastive

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

Fig. 3: We visualize the loss landscapes of the pose optimization with scale-invariant contrastive features. (a) shows the input image. (b) shows the reconstruction loss $\max_r \|C_r - f_i\|^2$ for each pixel. (c) visualizes the predicted 6D pose. Each curve in (d) corresponds to one of the six dimensions of the 6D pose and is centered at the ground truth pose. We can see with the help the scale-invariant contrastive features, the pose optimization has a clear global minimum near the ground truth pose and is easy to optimize. This further allows us to search for generative 6D proposals, as described in Section 3.4.

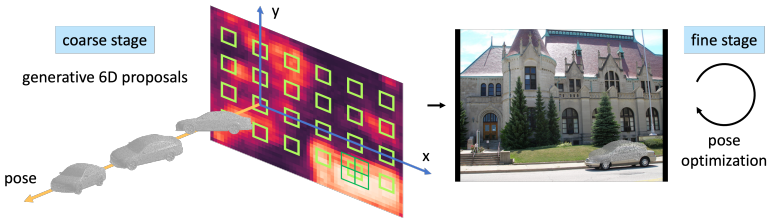

Fig. 4: Overview of our proposed coarse-to-fine 6D pose estimation. We search for generative 6D proposals in the coarse stage and then refine the initial 6D poses with pose optimization in the fine stage. Since the generative 6D proposals are built on the generative neural mesh models and the scale-invariant contrastive features, they are robust to partial occlusion and truncation and are easy to optimize. Note that the CAD models are for visualization only and are not used in any part of our model.

features, they are robust to partial occlusion and truncation. Moreover, this coarse-to-fine strategy can largely benefit subsequent pose optimization. The generative 6D proposals are often located at regions near global optimum that makes effective gradients toward the ground truth pose. We compare Faster R-CNN 6D proposals and our generative 6D proposals quantitatively in Section 4.4. We further visualize loss landscapes of different 6D object proposals in the supplementary material.

Coarse stage: generative 6D proposals. With the scale-invariant contrastive features, the pose optimization has a clear global minimum near the ground truth location, and the loss landscapes are smooth with decent gradients around the global minimum (see Figure 3). This nice property allows us to search for generative 6D proposals from a sparse sampling over six dimensions. Given a 6D pose sample, we estimate the reconstruction loss in Equation 3 and predict

generative 6D proposals with non-maximum suppression. Since the structure of the 3D model $\mathcal{M}$ and the sampled 6D poses are consistent across all testing samples, the 2D coordinates and visibility of the vertices can be pre-computed and cached. We further adopt a strategy to speed up by searching over the 2D locations first and then the other four dimensions, which is detailed in the supplementary materials. This allows us to predict generative 6D proposals that are robust to partial occlusion and truncation and are easy to optimize with a negligible complexity overhead.

Fine stage: pose optimization. The goal of pose optimization is to refine the initial 6D pose of an object, which can be either predicted by a stand-alone pose estimation network or from our generative 6D proposals. We maximize the feature correlation between the predicted features F from the feature extract ζ and the rendered features $\bar{F}$ with respect to the object pose. Since the ground truth $\mathcal{FG}$ and $\mathcal{BG}$ is unknown, we approximate the maximum likelihood problem in Equation 3 with a one-hot map $\mathbf{Z}$ to denote the foreground and background regions in the feature map:

$$\mathbf{Z}_i = \begin{cases} 1 & \text{if } \|f_i - C_r\|^2 \geq \|f_i - b\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Finally, we minimize $\mathcal{L}_{\text{NLL}}$ with respect to object pose m with gradient descent

$$\mathcal{L}_{\text{NLL}}(F, \mathfrak{N}, m, \mathcal{B}) = \frac{1}{2} \sum_i \left(\mathbb{1}_{Z_i=1} \|f_i - C_r\|^2 + \mathbb{1}_{Z_i=0} \|f_i - b\|^2 \right) + const. \tag{6}$$

3.5 Multi-Object Reasoning

One challenge when extending 3D pose estimation to 6D pose estimation is the existence of multiple objects in the image. Therefore, we propose a multi-object reasoning module that can resolve mutual occlusion and can be applied on top of any render-and-compare methods. The motivation is that we need to assign the pixels in the feature maps to different instances, and our multi-object reasoning module resembles related methods in instance segmentation [37].

Given multiple generative 6D proposals, we run the pose optimization gradient descent for a small number of epochs. If the rendered feature maps of two objects overlap, we recover the occlusion order by running pixel-level competition, and for each overlapping region, only one object is considered as the foreground object and the other objects are considered as background. We use a one-hot map to record the multi-object reasoning results, where

$$\mathbf{Z} \in \mathbb{Z}^{H \times W \times k}, \quad Z_{i,j,k} = \begin{cases} 1 & \text{if the } k\text{-th object is the foreground object} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Then we run the pose optimization again given the occlusion ordering $\mathbf{Z}$. We visualize the results of the multi-object occlusion reasoning in Figure 5.

4 Experiments

In this section, we investigate the performance of our approach in challenging 6D pose estimation datasets and compare its performance to related methods. We first describe the experimental setup in Section 4.1. Then we study the performance of our model in Section 4.2. We visualize some qualitative examples in Section 4.3. Finally, we run ablation study experiments on the generative 6D proposals and the multi-object reasoning module in Section 4.4.

4.1 Experimental Setup

Datasets. We evaluate our model on PASCAL3D+ dataset [34], Occluded PASCAL3D+ dataset [30], and ObjectNet3D dataset [33]. PASCAL3D+ dataset contains objects from 12 man-made categories, and each object is annotated with 3D pose, 2D centroid, and object distance. The ImageNet subset of the PASCAL3D+ dataset contains 11045 images for training and 10812 images for evaluation, and the PASCAL VOC subset contains 4293 images for training and 4212 images for validation. Occluded PASCAL3D+ is based on the ImageNet subset of the PASCAL3D+ dataset, and partial occlusion is simulated by superimposing occluders on top of the objects and the background. We also experimented on ObjectNet3D dataset, which consists of 100 categories with 17101 training images and 19604 testing images. Following [38, 28], we compare the 6D pose estimation performance on 18 categories.

Evaluation metrics. Category-level 6D pose estimation estimates both the 3D pose (azimuth, elevation, and in-plane rotation) and the 3D location of the visible objects. In our experiments, we adopt the pose estimation error and the average distance metric (ADD) for evaluation. Following [38], the pose estimation error measures the angle between the predicted rotation matrix and the ground truth rotation matrix $\Delta(R_{\text{pred}}, R_{\text{gt}}) = \frac{\|\text{logm}(R_{\text{pred}}^{\top} R_{\text{gt}})\|_{\mathcal{F}}}{\sqrt{2}}$. Average distance (ADD) is a widely used metric to measure the translation of the keypoints between the ground truth pose and the predicted pose. For the PASCAL VOC images, we also evaluate the mean average precision (mAP) at $(\pi/3, 5.0)$.

Implementation details. Our model includes a contrastive feature backbone and a corresponding neural mesh model. The feature extractor is a ResNet50 model with two upsampling layers, so the output feature map is $\frac{1}{8}$ of the input resolution. The neural mesh model is a category-wise cuboid model with around 1100 vertices. The scale of the cuboid mesh model is the average of the scales of the sub-category mesh models, and the vertices are sampled uniformly across six faces. Our model is trained for 1200 epochs with random horizontal flip and 2D translation and takes around 20 hours on one NVIDIA RTX Titan GPU. During inference, the pose optimization with multi-object reasoning takes 4.1 seconds on average per object.

Table 1: Quantitative results of 6D pose estimation on PASCAL3D+ dataset.

Subset	Model	Pose Acc ($\frac{\pi}{6}$) ↑	Pose Acc ($\frac{\pi}{18}$) ↑	Median Pose Error ↓	Median ADD ↓	mAP ↑
ImageNet	FRCNN+Cls	78.90	37.35	0.22	0.74	-
ImageNet	FRCNN+NeMo	66.06	28.44	0.33	1.84	-
ImageNet	RTM3DExt	74.94	39.56	0.23	0.92	-
ImageNet	Ours	**81.45**	**47.68**	**0.19**	**0.53**	-
PASCAL VOC	FRCNN+Cls	38.98	15.05	1.38	2.04	0.11
PASCAL VOC	FRCNN+NeMo	40.13	**19.17**	1.40	2.14	0.32
PASCAL VOC	RTM3DExt	18.04	8.12	6.28	20.0	0.11
PASCAL VOC	Ours	**45.32**	18.09	**0.65**	**1.87**	**0.43**

Table 2: Quantitative results of 6D pose estimation on ObjectNet3D dataset.

Model	Pose Acc ($\frac{\pi}{6}$) ↑	Pose Acc ($\frac{\pi}{18}$) ↑	Median Pose Error ↓	Median ADD ↓
RTM3DExt	38.44	16.61	2.50	4.87
Ours	**52.47**	**16.65**	**0.49**	**1.95**

Baseline models. Since we know of no other 6D pose estimation methods for category-level 6D pose estimation from a single RGB image, we compare our model with related works in 3D object detection and 3D pose estimation and extend them to the 6D pose estimation setting.

RTM3D is one of the state-of-the-art models for monocular 3D object detection. It predicts a 3D bounding box (i.e., location, rotation, and scale) by minimizing the reprojection error between the regressed 2D keypoints and the corners of the 3D cuboid. To extend RTM3D to 6D pose estimation, we fix the cuboid dimensions and fit a rigid 6D transformation.

We further compare our approach with two-stage models that predict object proposals in the first stage and then estimate object poses from the proposed RoIs. We adopt Faster R-CNN [24] for object detection. Two methods are considered for pose estimation. Following previous works [38, 28], we formulate the pose estimation as a classification problem and predict the object pose from the RoI features from the Faster R-CNN backbone. Based on the reported results [38, 28], we also consider the state-of-the-art 3D pose estimation model, NeMo [28], where we optimize the 3D object pose from the predicted 2D bounding box. The two models are denoted as "FRCNN+Cls" and "FRCNN+NeMo" respectively.

4.2 Quantitative Results

6D pose estimation on PASCAL3D+ and ObjectNet3D dataset. Table 1 shows the 6D pose estimation results on the ImageNet and PASCAL VOC subsets of the PASCAL3D+ dataset. Compared to the ImageNet images, the PASCAL VOC subset is more challenging as there are multiple objects with occlusion, truncation, as well as a larger variance in the object scale and location.

Our model outperforms all baseline models in both the pose error and the average distance metric. To show our model can be applied to different man-made in-door and out-door categories, we also experiment on the ObjectNet3D dataset, and the results are shown in Table 2. Despite the considerable number of occluded and truncated images in ObjectNet3D dataset, our model achieves reasonable accuracy and outperforms the competitive baseline by a wide margin.

Robust 6D pose estimation on the Occluded PASCAL3D+ dataset. In order to investigate the robustness under occlusion, we further evaluate each model on Occluded PASCAL3D+ dataset under different occlusion levels. The quantitative results are reported in Table 3. As we can see, our model achieves superior performance across all occlusion levels and shows a wider performance gap compared to the performance on the un-occluded images.

Table 3: Quantitative results of 6D pose estimation on the Occluded PASCAL3D+ dataset.

Subset	Level	Method	Pose Acc ($\frac{\pi}{6}$) ↑	Pose Acc ($\frac{\pi}{18}$) ↑	Median Pose Error ↓	Median ADD ↓
ImageNet	1	FRCNN+Cls	61.48	26.11	0.33	1.07
ImageNet	1	FRCNN+NeMo	48.34	17.46	0.55	1.90
ImageNet	1	RTM3DExt	43.55	17.68	0.82	3.29
ImageNet	1	Ours	**66.63**	**30.84**	**0.31**	**0.77**
ImageNet	2	FRCNN+Cls	41.95	14.75	0.75	1.47
ImageNet	2	FRCNN+NeMo	34.33	9.65	1.05	2.03
ImageNet	2	RTM3DExt	21.27	7.24	3.14	5.00
ImageNet	2	Ours	**47.95**	**16.25**	**0.56**	**1.22**
ImageNet	3	FRCNN+Cls	22.42	**5.58**	2.01	1.95
ImageNet	3	FRCNN+NeMo	18.19	3.32	2.40	2.35
ImageNet	3	RTM3DExt	10.17	3.11	3.14	19.92
ImageNet	3	Ours	**27.43**	5.30	**1.07**	**1.94**

4.3 Qualitative Examples

Figure 5 shows some qualitative examples of our proposed model on PASCAL3D+ dataset. As we can see, our method can robustly estimate 6D poses for objects varying in scales and textures and is robust to partial occlusion.

4.4 Ablation Study

Generative 6D proposals. Unlike previous works that are based on 2D region proposal networks, we introduce generative 6D proposals that are robust to partial occlusion and truncation and are easy to optimize. We run ablation study experiments on the object proposal methods and compare the performance of our model using (1) generative 6D proposals ("Ours w/ GP"), or (ii) Faster R-CNN object proposals ("Ours w/ FRCNN"). The quantitative results on the

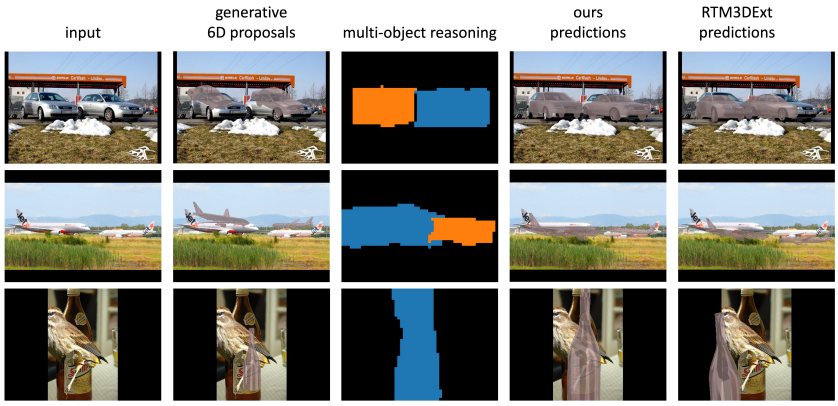

Fig. 5: Qualitative examples of our proposed model.

Occluded PASCAL3D+ dataset are reported in Table 4. As we can see, we can significantly improve the performance of our model in almost all tests with the generative 6D proposals. Note that "Ours w/ FRCNN" naturally benefits from model ensembling, and "Ours w/ FRCNN" can achieve a better pose estimation only in occlusion level 1 when the Faster R-CNN model can predict highly accurate poses without any refinement. In occlusion levels 2 and 3, "Ours w/ GP" significantly outperforms "Ours w/ FRCNN" in terms of both object location and pose estimation.

Table 4: Ablation study on the 6D proposal method. We compare the performance of our proposed model using Faster R-CNN 6D proposals and generative 6D proposals (GP) on the PASCAL3D+ dataset.

Level	Method	Pose Acc ($\frac{\pi}{6}$) ↑	Pose Acc ($\frac{\pi}{18}$) ↑	Median Pose Error ↓	Median ADD ↓
1	Ours w/ FRCNN	65.79	**34.56**	**0.28**	0.95
1	Ours w/ GP	**66.63**	30.84	0.31	**0.77**
2	Ours w/ FRCNN	45.43	**17.46**	0.64	1.30
2	Ours w/ GP	**47.95**	16.25	**0.56**	**1.22**
3	Ours w/ FRCNN	23.53	5.27	1.67	2.05
3	Ours w/ GP	**27.43**	**5.30**	**1.07**	**1.94**

Number of pre-defined initial poses. Far all categories, we uniformly sample a sparse set of initial poses over the space of 6D poses. In Table 5, we ablate on the number of initial poses used to search the generative 6D proposals.

Multi-object reasoning. In order to estimate 6D poses of multiple objects with render-and-compare, we propose a multi-object reasoning module to correctly assign pixels in the feature map to each object proposal. We quantitatively

Table 5: Ablation on the number of initial poses used to search the generative 6D proposals. The default setting in the paper is bold.

Init. 3D Pose	Init. 3D Loc	Acc ($\frac{\pi}{6}$)	Acc ($\frac{\pi}{18}$)	Median Err	Median ADD
12x3x3	3x3x3	69.77	35.82	0.28	0.65
6x2x2	9x9x9	71.64	38.75	0.26	0.63
12x3x3	**9x9x9**	81.45	47.68	0.19	0.53
18x6x6	9x9x9	83.46	49.28	0.18	0.51
12x3x3	12x12x12	84.54	51.39	0.17	0.49

Table 6: Ablation study on the multi-object reasoning. We compare the performance of our proposed model with and without the multi-object reasoning on the PASCAL VOC subset of PASCAL3D+ dataset.

Method	Pose Acc ($\frac{\pi}{6}$) ↑	Pose Acc ($\frac{\pi}{18}$) ↑	Median Pose Error ↓	Median ADD ↓	mAP ↑
Ours w/o reasoning	44.89	17.63	0.66	**1.87**	0.41
Ours	**45.32**	**18.09**	**0.65**	**1.87**	**0.43**

compare the performance of our model with and without an object reasoning module on the PASCAL subset of the PASCAL3D+ dataset. As shown in Table 6, we can effectively improve the performance of our model with the multi-object reasoning module.

5 Conclusions

In this work, we consider the problem of category-level 6D pose estimation from a single RGB image. We find that previous methods built on 2D region proposal networks are less robust to partial occlusion and truncation, and the predicted initial poses are harder to optimize. Therefore, we propose a coarse-to-fine 6D pose optimization strategy where we search generative 6D proposals in the coarse stage and then refine them with pose optimization in the second stage. Both stages of our coarse-to-fine 6D pose estimation are built on our scale-invariant contrastive features and are hence robust to partial occlusion and truncation. We demonstrate the superiority of our approach compared to related works on several challenging datasets.

Acknowledgements. AK acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under Grant No. 468670075. AY acknowledges the Institute for Assured Autonomy at JHU with Grant IAA 80052272, ONR N00014-21-1-2812, NSF grant BCS-1827427.

References

1. Bai, Y., Wang, A., Kortylewski, A., Yuille, A.: Coke: Localized contrastive learning for robust keypoint detection. arXiv preprint arXiv:2009.14115 (2020)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. p. 187–194. SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., USA (1999). https://doi.org/10.1145/311535.311556, https://doi.org/10.1145/311535.311556
3. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category level object pose estimation via neural analysis-by-synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 139–156. Springer International Publishing, Cham (2020)
4. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)
5. Egger, B., Schönborn, S., Schneider, A., Kortylewski, A., Morel-Forster, A., Blumer, C., Vetter, T.: Occlusion-aware 3d morphable models and an illumination prior for face image analysis. International Journal of Computer Vision **126**(12), 1269–1287 (2018)
6. Girshick, R., Felzenszwalb, P., McAllester, D.: Object detection with grammar models. Advances in neural information processing systems **24** (2011)
7. Grenander, U.: A unified approach to pattern analysis. In: Advances in computers, vol. 10, pp. 175–216. Elsevier (1970)
8. Grenander, U.: Elements of pattern theory. JHU Press (1996)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
10. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
11. Huber, P.J.: Robust statistics, vol. 523. John Wiley & Sons (2004)
12. Iwase, S., Liu, X., Khirodkar, R., Yokota, R., Kitani, K.M.: Repose: Fast 6d object pose refinement via deep texture rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3303–3312 (October 2021)
13. Kortylewski, A., He, J., Liu, Q., Cosgrove, C., Yang, C., Yuille, A.L.: Compositional generative networks and robustness to perceptible image changes. In: 2021 55th Annual Conference on Information Sciences and Systems (CISS). pp. 1–8. IEEE (2021)
14. Kortylewski, A., He, J., Liu, Q., Yuille, A.L.: Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
15. Kortylewski, A., Liu, Q., Wang, A., Sun, Y., Yuille, A.: Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. International Journal of Computer Vision pp. 1–25 (2020)
16. Kortylewski, A., Liu, Q., Wang, H., Zhang, Z., Yuille, A.: Combining compositional models and deep networks for robust object classification under occlusion.

In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1333–1341 (2020)
17. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10), 1995 (1995)
18. Li, P., Zhao, H., Liu, P., Cao, F.: Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 644–660. Springer (2020)
19. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: Deepim: Deep iterative matching for 6d pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
20. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3560–3569 (October 2021)
21. Moreno, P., Williams, C.K., Nash, C., Kohli, P.: Overcoming occlusion with inverse graphics. In: European Conference on Computer Vision. pp. 170–185. Springer (2016)
22. Pavlakos, G., Zhou, X., Chan, A., Derpanis, K.G., Daniilidis, K.: 6-dof object pose from semantic keypoints. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 2011–2018 (2017). https://doi.org/10.1109/ICRA.2017.7989233
23. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016)
25. Romaszko, L., Williams, C.K., Moreno, P., Kohli, P.: Vision-as-inverse-graphics: Obtaining a rich 3d explanation of a scene from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 851–859 (2017)
26. Schönborn, S., Egger, B., Morel-Forster, A., Vetter, T.: Markov chain monte carlo for automated face image analysis. International Journal of Computer Vision **123**(2), 160–183 (2017)
27. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence **13**(04), 376–380 (1991)
28. Wang, A., Kortylewski, A., Yuille, A.: Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=pmj131uIL9H
29. Wang, A., Mei, S., Yuille, A.L., Kortylewski, A.: Neural view synthesis and matching for semi-supervised few-shot learning of 3d pose. Advances in Neural Information Processing Systems **34**, 7207–7219 (2021)
30. Wang, A., Sun, Y., Kortylewski, A., Yuille, A.L.: Robust object detection under occlusion with context-aware compositionalnets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12645–12654 (2020)
31. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

32. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
33. Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S.: Objectnet3d: A large scale database for 3d object recognition. In: European conference on computer vision. pp. 160–176. Springer (2016)
34. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE winter conference on applications of computer vision. pp. 75–82. IEEE (2014)
35. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017)
36. Xiao, M., Kortylewski, A., Wu, R., Qiao, S., Shen, W., Yuille, A.: Tdmpnet: Prototype network with recurrent top-down modulation for robust object classification under partial occlusion. In: European Conference on Computer Vision. pp. 447–463. Springer (2020)
37. Yuan, X., Kortylewski, A., Sun, Y., Yuille, A.: Robust instance segmentation through reasoning about multi-object occlusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11141–11150 (June 2021)
38. Zhou, X., Karpur, A., Luo, L., Huang, Q.: Starmap for category-agnostic keypoint and viewpoint estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 318–334 (2018)