Universal Features Guided Zero-Shot Category-Level Object Pose Estimation

Wentian Qu[1,2,3], Chenyu Meng[1,2], Heng Li[3], Jian Cheng[1,2], Cuixia Ma[1,2], Hongan Wang[1,2], Xiao Zhou[4], Xiaoming Deng[1,2*], Ping Tan[3*]

[1]Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Hong Kong University of Science and Technology
[4]Aerospace Information Research Institute, Chinese Academy of Sciences
{wentian2019, chengjian, cuixia, hongan, xiaoming}@iscas.ac.cn, mengchenyu21@mails.ucas.ac.cn,
lh.heng.li@connect.ust.hk, zhouxiao@aircas.ac.cn, pingtan@ust.hk

Abstract

Object pose estimation, crucial in computer vision and robotics applications, faces challenges with the diversity of unseen categories. We propose a zero-shot method to achieve category-level 6-DOF object pose estimation, which exploits both 2D and 3D universal features of input RGB-D image to establish semantic similarity-based correspondences and can be extended to unseen categories without additional model fine-tuning. Our method begins with combining efficient 2D universal features to find sparse correspondences between intra-category objects and gets initial coarse pose. To handle the correspondence degradation of 2D universal features if the pose deviates much from the target pose, we use an iterative strategy to optimize the pose. Subsequently, to resolve pose ambiguities due to shape differences between intra-category objects, the coarse pose is refined by optimizing with dense alignment constraint of 3D universal features. Our method outperforms previous methods on the REAL275 and Wild6D benchmarks for unseen categories.

Project Page — https://iscas3dv.github.io/universal6dpose/

Introduction

Object pose estimation, which aims to estimate the orientation and location of an object in 3D space, is a long-standing challenge and plays a key role in AR/VR and robotics applications. Instance-level methods (Labbé et al. 2020; Xiang et al. 2017; Li et al. 2018) achieve pose estimation for seen instances. To deal with unseen instances of the same category, category-level methods (Zhang et al. 2022; Wang et al. 2019a; He et al. 2022) introduce a mean shape for a category object with improved model attention to texture and geometry details, while these methods require retraining or model fine-tuning when applied to unseen categories. To address this limitation, we aim to address zero-shot category-level object pose estimation from input RGB-D, which utilizes both 2D and 3D universal features from pre-trained models to achieve pose estimation for unseen categories.

Witnessed by a strong ability to establish correlations (Amir et al. 2021; Zhang et al. 2022) with 2D universal features extracted by the foundation models (Caron et al. 2021; Oquab et al. 2023; Rombach et al. 2022) pre-trained on large-scale datasets, recent work has explored 2D universal features to estimate object pose, such as FoundPose (Örnek et al. 2023), FoundationPose (Wen et al. 2024) and Zero-Pose (Goodwin et al. 2022). FoundPose needs to know the model of the instance object in advance, and FoundationPose needs to use multi-view images to model the object before pose estimation. These settings require strong object shape priors and are difficult to extend to unseen categories. Zero-Pose establishes 2D correspondences for object pose estimation using DINOv1 (Caron et al. 2021). We find that DINOv1 is less effective in building correspondences and is severely affected by object pose (Fig. 1 (b)). The correspondence will degrade if the pose difference between the reference image and the target image is large (Fig. 3). They also do not consider shape differences between intra-category objects, resulting in biased correspondences and inaccurate pose estimation (Fig. 1 (c)). Similarly to 2D universal features, 3D universal features, extracted by pre-trained models such as DGCNN (Wang et al. 2019c) can provide effective geometric clues for correspondence, while they are barely used in object pose estimation. Inspired by this, we utilize both 2D and 3D universal features to solve the 6-DOF pose estimation for unseen categories without model training or fine-tuning, achieving a zero-shot category-level pose estimation. In particular, we do not need to know the 3D model of the instance objects. Our method offers superior generalizability over traditional instance-level and category-level methods (Fig. 1 (a)).

In this paper, we design a coarse-to-fine framework for accurate 6-DOF pose estimation. At the coarse stage, it identifies sparse correspondences to solve an initial coarse object pose. Given an input RGB-D image, we use a reference model of the interested category to render reference images and extract 2D universal features from both the target and rendered reference images. We then calculate the cosine similarity map between the 2D features and use cyclical distance to select Top-k correspondences. Combined with the depth map and camera intrinsics, we choose the Top-k keypoints in the camera coordinate and calculate the transformation from the reference to the target space to get the initial coarse 6-DOF object pose by a least-squares solution. To deal with the problem of feature correspondence degradation of 2D universal features if the initial pose deviates

*indicates corresponding author.

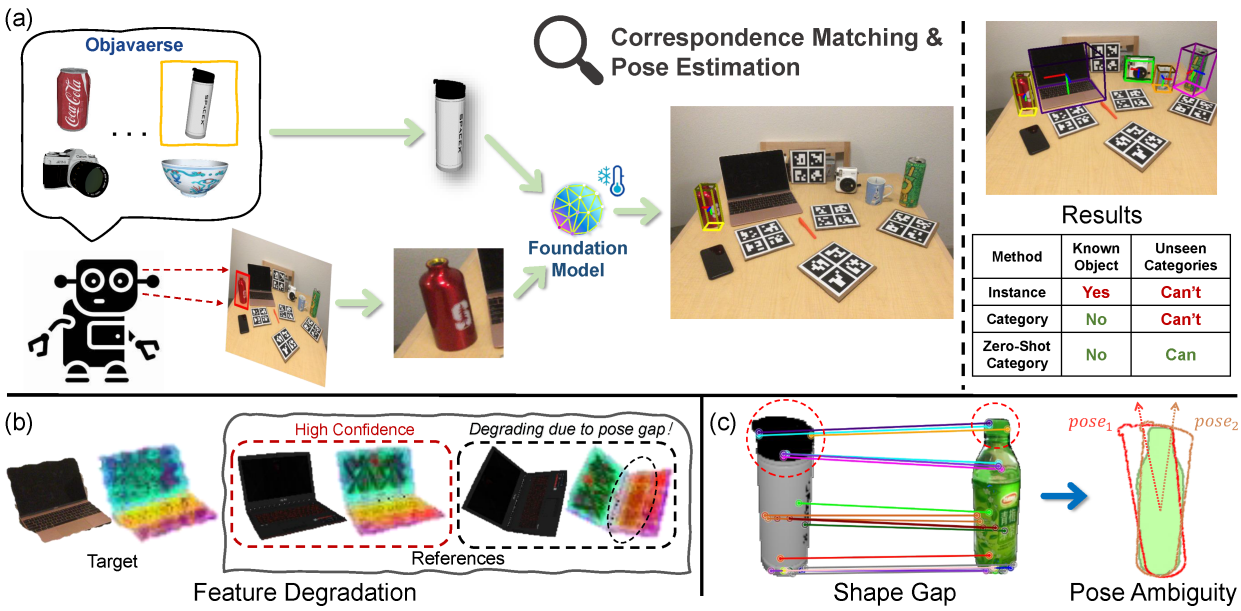

Figure 1: (a) We propose a zero-shot pose estimation method for unseen categories using universal features and obtain accurate results for multi-category scenes. Our method offers cost-efficient and superior generalization ability over traditional instance-level and category-level methods. (b) The correspondence with universal features degrades when pose has large gaps. (c) The shape gap between objects will cause pose ambiguity in optimization. These challenges affect the accuracy of pose estimation.

much from the target pose, we use an iterative strategy to optimize the correspondence and coarse pose. After the coarse pose estimation, we map the reference model to the target image space to perform pose refinement with pixel-wise optimization. In order to resolve pose ambiguities due to shape differences between intra-category objects during the optimization, we employ 3D universal features extracted from the point cloud to refine the 6-DOF object pose and the reference model iteratively by dense pixel-level registration.

The main contribution of our method can be summarized as follows: 1) We propose a 2D/3D universal features guided zero-shot category-level object pose estimation with coarse-to-fine optimization. To deal with the correspondence degrade issue of 2D universal features, we use an iterative strategy to optimize the correspondence and coarse pose; 2) During the pose refinement, to handle pose ambiguity due to intra-category shape difference, we employ 3D universal features to refine the 6-DOF object and the shape of reference model by dense pixel-level registration; 3) Experiments on the REAL275 (Wang et al. 2019b) and Wild6D (Ze and Wang 2022) benchmarks demonstrate that our method establishes robust correspondences based on pretrained 2D/3D universal features, resulting in accurate pose estimation based on coarse-to-fine optimization.

Related Work

Instance-Level Object Pose Estimation. Instance-level object pose estimation methods regard each object as an independent entity with known object shapes. They directly regress the object pose within each RoI through characteristics (Labbé et al. 2020; Xiang et al. 2017; Li et al. 2018), or estimate the object pose through 2D-3D correspondences based on conventional PnP (Hodan, Barath, and Matas 2020; Peng et al. 2019; Tekin, Sinha, and Fua 2018) or PnP with network learning (Hu et al. 2020; Wang et al. 2021). However, a major challenge for instance-level methods is that they struggle to estimate poses on unseen objects.

Category-Level Object Pose Estimation. Category-level object pose estimation methods divide objects into different categories (Wang et al. 2022; Jung et al. 2024), emphasizing the commonality between objects. They perform well in 6-DoF pose estimation for unknown object shape (Chen et al. 2024; Di et al. 2022; Lin et al. 2024), and can directly learn the pose distribution of objects using the shape prior (Burchfiel and Konidaris 2019; Sahin and Kim 2018). Previous methods either use a Normalized Object Coordinate Space (NOCS) representation (Wang et al. 2019a) to estimate object pose (Chen et al. 2020; Ze and Wang 2022; Chen et al. 2021), or learn the pose distribution using geometric priors such as shape (Tian, Ang, and Lee 2020), symmetry (Lin et al. 2021) and keypoints (Lin et al. 2022b). The existing category-level methods suffer from model generalization to unseen categories.

Zero-Shot Object Pose Estimation. In order to remove the time-consuming dataset collection requirement, several works leverage the foundation models such DINO (Caron et al. 2021; Oquab et al. 2023) and Stable Diffusion (SD) (Rombach et al. 2022) for pose estimation (Goodwin et al. 2022; Chen et al. 2023). Zero-shot instance-level methods (Örnek et al. 2023; Fan et al. 2023; Labbé et al. 2023) estimate the object pose of the seen instance through feature matching based on 2D universal features such as DINO.

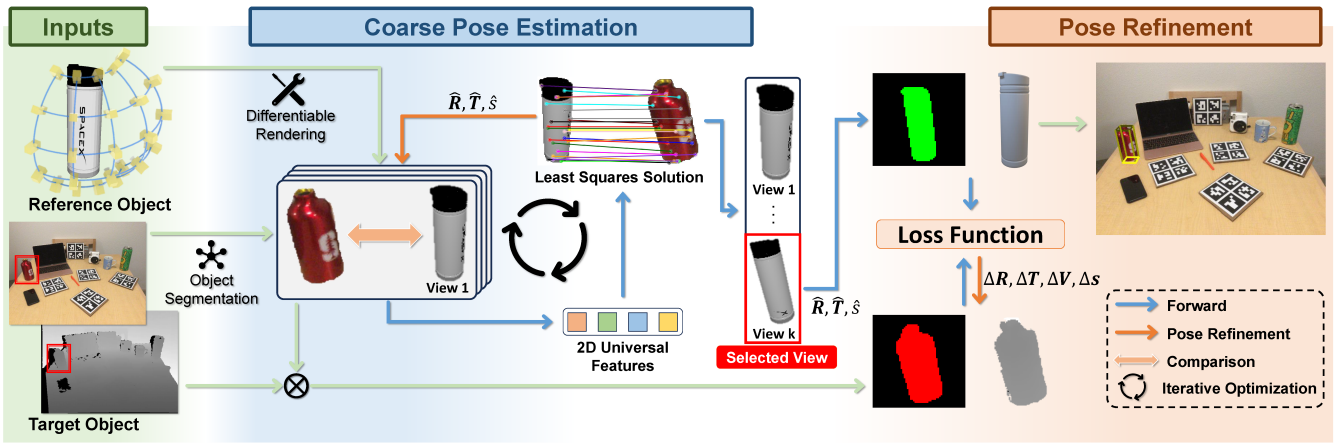

Figure 2: Overview. Our framework includes a keypoint-level coarse pose estimation module and a pixel-level pose refinement module. In the first module, we establish the correspondences between image pairs based on the 2D universal features and calculate the coarse pose using least squares in an iterative manner. In the second module, we use pixel-level optimization combined with 3D universal features to refine the pose and shape of reference model to obtain the fine pose.

However, these methods cannot be applied to unknown instances. In order to overcome the generalization limitation of these instance-level methods on known object instances, the zero-shot category-level method (Goodwin et al. 2022) is designed to be generalized to unseen objects based on universal features DINOv1. However, this method is affected by less effective features, especially in varied pose and shape, resulting in inaccurate pose results, and it is verified that DINOv1 is inferior to more advanced universal features such as DINOv2 and the combination with Stable Diffusion features (Zhang et al. 2023; Luo et al. 2023; Tang et al. 2023). Our method has two differences from (Goodwin et al. 2022). First, to deal with the inefficient of DINOv1 in which the feature similarity performance drops under a large pose gap, we conduct a strong universal feature combined with DINOv2 and SD, and design an iterative module to conduct feature matching under step-by-step optimized pose. Second, to address the shape gaps, we design a pixel-level pose refinement framework, which utilizes 3D universal features to jointly optimize the object's pose and shape with a new universal alignment constraint.

Method

This paper exploits multi-modal (both 2D and 3D) universal features to estimate object pose on unseen categories. Our method strives to address the following three challenges, 1) to overcome the correspondence degradation of 2D universal features caused by large pose differences; 2) to exploit the effect of 3D universal features for category-level object pose estimation; 3) to handle the pose ambiguity caused by shape differences between the reference and target object.

Our zero-shot category-level pose estimation works with multi-modal universal features in a coarse-to-fine manner (See Fig. 2), which first utilizes 2D universal features for keypoint-level coarse pose estimation and then uses 3D universal features for pixel-level pose refinement. To solve the first challenge, we propose an iterative optimization to establish more accurate correspondences based on the rendered reference image under updated pose. To solve the last two issues, we perform a pixel-level pose refinement based on 3D universal features with a new universal alignment constraint to resolve the ambiguity between shape and pose optimization.

Multi-Modal Universal Features

2D Universal Features. The 2D universal features are extracted by pre-trained image models such as DINOv1 (Caron et al. 2021), DINOv2 (Oquab et al. 2023), and Stable Diffusion (SD) (Rombach et al. 2022). They provide effective texture prior information to establish semantic correspondence between images. Given a target image $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, we use Mask R-CNN (He et al. 2017) to predict the shape token of the interested object and retrieve a reference mesh model from Objaverse (Deitke et al. 2023) to render the reference image. Given the image, we extract the universal 2D features $\mathbf{F}$ by sending them to DINOv1, DINOv2 and SD:

$$\mathbf{F} = (\alpha_{D1}\|\mathbf{F}_{D1}\|_2, \alpha_{D2}\|\mathbf{F}_{D2}\|_2, \alpha_{SD}\|\mathbf{F}_{SD}\|_2) \quad (1)$$

where $\mathbf{F}_{D1}$, $\mathbf{F}_{D2}$, $\mathbf{F}_{SD}$ denotes the features from DINOv1, DINOv2 and SD, and α denotes the hyperparameter that balances the proportions of different 2D universal features. We conduct a systematic study on pose estimation over different combinations of 2D universal features and observe that the combination of DINOv2 and SD can leverage the strengths of both local and global semantic similarity to generate robust correspondences, which result in an accurate pose.

3D Universal Features. DGCNN (Wang et al. 2019c) pre-trained on 3D point cloud datasets could extract 3D universal features $\mathbf{F}_{3d}$ containing geometric information. As shown in the red dashed box of Fig. 4(a), the 3D universal features aligned based on the initial pose contain more semantic similarity in geometric details, which can further refine object pose and resolve the shape ambiguities.

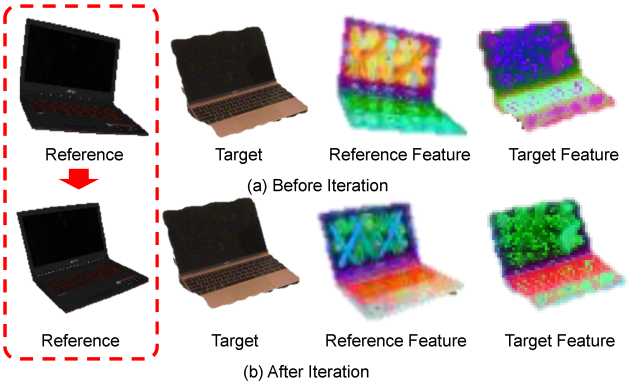

Figure 3: Feature Performance Drop and Effect of Iterative Estimation. When there are large pose differences between objects, the 2D universal features similarity degrade. After iterative optimization, as the objects are gradually aligned, the correspondence between the objects become smoother, which support to calculate an accurate pose.

Keypoint-Level Coarse Pose Estimation

We extract the 2D features for reference and target images and leverage them to estimate the coarse pose transformation by computing the sparse keypoint-level correspondence.

Establish Correspondence. Specifically, we render four images $\{\mathbf{I}_r\}$ of the reference object from the front, back, and two sides. Then, we extract 2D feature for all the reference images $\{\mathbf{I}_r\}$ and the target image $\mathbf{I}_t$, denoted by $\{\mathbf{F}(\mathbf{I}_r)\}$ and $\mathbf{F}(\mathbf{I}_t)$, respectively. A score matrix $\mathbf{S}$ can be defined based on the cosine similarity between the features:

$$\mathbf{S}(p,q) = d_{cos}(\mathbf{F}(\mathbf{I}_t)_p, \mathbf{F}(\mathbf{I}_r)_q), p \in [1, \mathbf{N}_t], q \in [1, \mathbf{N}_r], \quad (2)$$

where $\mathbf{N}_t$ and $\mathbf{N}_r$ are the number of patches in the target and the reference feature maps, $d_{cos}(\cdot,\cdot)$ is the cosine similarity between two vectors. The cyclical distance matrix $\mathbf{D} \in \mathbb{R}^{N_t}$ based on $\mathbf{S}$ can be used to compute the correspondence for each patch pair between the reference and target features:

$$\mathbf{D}_{p,p\in[1:\mathbf{N}_t]} = d(p, \arg\max_{p'\in[1:N_t]} \mathbf{S}(p', \arg\max_{q\in[1:N_r]} \mathbf{S}(p,q))), \quad (3)$$

where $d(\cdot,\cdot)$ is the L2 distance. We select M correspondence based on ascending order of $\mathbf{D}$ for each image pair.

Iterative Coarse Pose Estimation. For each reference and target image pair, we can lift the 2D keypoints into 3D camera space to get the reference point cloud $\mathbf{P}_r \in \mathbb{R}^{N\times 3}$ and target 3D point cloud $\mathbf{P}_t \in \mathbb{R}^{N\times 3}$ using the depth map and camera intrinsics. Then we use Umeyama (Umeyama 1991) with RANSAC (Fischler and Bolles 1981) to compute the updated object pose for each reference image:

$$(\hat{\mathbf{R}}, \hat{\mathbf{T}}, \hat{s}) = \arg\min_{\mathbf{R},\mathbf{T},s} \frac{1}{2} \sum_{m=1}^{M} \|\mathbf{P}_t^m - (s\mathbf{R}\mathbf{P}_r^m + \mathbf{T})\|_2^2, \quad (4)$$

where $\hat{\mathbf{R}} \in \mathbb{R}^{3\times3}$,$\hat{\mathbf{T}} \in \mathbb{R}^3$, $\hat{s} \in \mathbb{R}$ represent the object rotation, translation, and scale, respectively. We find that the semantic similarity based on 2D universal features will degrade when the pose between reference and target objects has large differences (Fig. 3 (a)). To solve this problem, we render the reference model with the updated coarse pose and establish more accurate correspondences to iteratively optimize the coarse pose. During the iteration, as the objects are gradually aligned, the correspondence between the objects become more consistent, which supports the calculation of an accurate pose (Fig. 3 (b)). We define the confidence for each reference image by averaging the cosine similarity of the M correspondences, and we choose the result with the highest confidence as the final coarse pose output.

Pixel-Level Pose Refinement

Although the keypoint-level method can achieve good results, estimating accurate object pose with a standard reference shape model for intra-category objects is still challenging, especially when the intra-category shape gap is large. Moreover, the keypoint-level method only uses sparse keypoints, thus it does not utilize dense geometric information to reduce the pose searching space. To address these two issues, we propose a pose refinement module by jointly optimizing object shape and pose (Fig. 4 (a)), a dense pixel-level optimization built on 3D universal features. After optimizing shape and pose, the objects are more accurately aligned (Fig. 4 (b)).

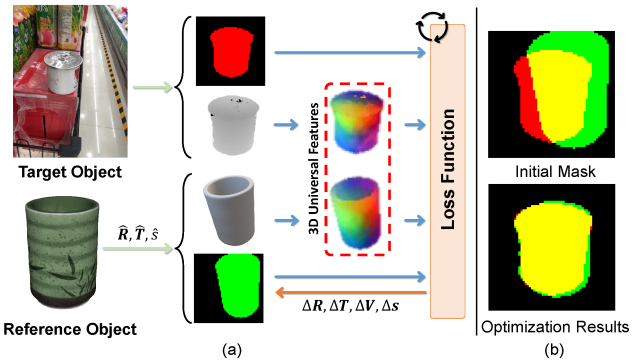

Figure 4: (a) Pose Refinement. Based on the coarse pose as initialization, the reference model can be warped to the target space to obtain the initial mask and extract 3D universal features. Then we optimize the coarse pose and shape by minimizing the loss function. (b) After pose refinement stage, the pose and shape of the reference model are more accurately aligned with the target object.

3D Universal Features Extraction. We use the coarse pose to align the reference object with the target object, and then feed them to the pre-trained model DGCNN to extract 3D universal features, respectively. Compared with 2D features, the 3D universal features can measure the 3D geometric semantic similarity for intra-category objects, which tackle the searching ambiguity in pose and shape.

Optimization Parameters. During pose refinement for each instance, we aim to optimize the rotation $\Delta\mathbf{R} \in \mathbb{R}^{3\times3}$, translation $\Delta\mathbf{T} \in \mathbb{R}^3$, independent deformation $\Delta\mathbf{V} \in \mathbb{R}^{N\times3}$ for each vertex and scale $\Delta\mathbf{s} \in \mathbb{R}^3$. The optimized object rotation and translation can be expressed as $\bar{\mathbf{R}} =$

$\Delta\mathbf{R} \times \hat{\mathbf{R}}, \bar{\mathbf{T}} = \Delta\mathbf{T} + \hat{\mathbf{T}}$, where we abbreviate $\hat{s} \cdot \hat{\mathbf{R}}$ as $\hat{\mathbf{R}}$ in Eq. 4. We define the reference mesh vertices in model coordinate as $\mathbf{V}$ and the optimized vertices can be formulated as $\bar{\mathbf{V}} = e^{\Delta\mathbf{s}}(\mathbf{V} + \Delta\mathbf{V})$.

Loss Function. We define the loss to constrain the output pose and object shape to guarantee feasible results, including pose optimization loss L_p and regularization loss L_r. The purpose of pose optimization loss is to optimize the pose and shape of the reference object to align accurately with the target object, and regularization loss is used to constrain the optimized reference object to be close to the initial state. The total loss function can be defined as: $L = L_p + L_r$.

Pose Optimization Loss L_p. We solve the relative motion between the target and reference objects by minimizing a registration objective function defined to match projected masks, 3D universal features and shapes. Our pose optimization loss including mask loss L_m, Chamfer loss L_c and universal alignment loss L_g as: $L_p = \alpha_m L_m + \alpha_c L_c + \alpha_g L_g$. The mask loss measures the difference between the reference mask $\mathbf{M}_r$ and the target mask $\mathbf{M}_t$, which can be calculated as: $L_m = 1 - \frac{\mathbf{M}_r \cap \mathbf{M}_t}{\mathbf{M}_r \cup \mathbf{M}_t}$. We use Chamfer loss L_c to enforce that the position and shape difference from the reference point cloud to the target point cloud is small. The 3D universal features reflect the relative positional relationship between the sampling points and the global geometry, which contain more shape details. Therefore, we propose a new universal alignment loss L_g to ensure that the 3D universal features between the reference object and the target object are consistent, which can solve the ambiguity when optimizing the pose and shape of the reference model. We calculate the cosine similarity of 3D universal features of the reference model and the target point cloud, and then make the 3D positions between high-confidence keypoint pairs as close as possible. The universal alignment loss can be defined as: $L_g = \frac{1}{N_g} \sum_{(p^r, p^t) \in N_g} \|p^r - p^t\|_2^2 \cdot d_{cos}^{3D}(p^r, p^t)$, where p^r and p^t represent the reference and target object point cloud, respectively, $d_{cos}^{3D}(p^r, p^t)$ represents the cosine similarity of 3D universal features between p^r and p^t, N_g represents the number of point pairs that satisfy $d_{cos}^{3D}(p^r, p^t) > 0.8$.

Regularization Loss L_r. During the optimization, we also want to constrain the optimized reference model not to deviate greatly from the initial state. We use pose regularization loss to enforce the refined pose to be close to the initial pose and use center point regularization loss to reduce the object displacement before and after optimization. The deformation regularization loss is used to constrain the deformation of the vertices to be small. In addition, we follow Pytorch3D (Ravi et al. 2020) to minimize geometric distortion with normal, edge, and Laplacian constraints.

Experiment

Experimental Setup

Datasets. We select Wild6D (Ze and Wang 2022) and REAL275 (Wang et al. 2019a) for category-level object pose estimation. Wild6D provides 5,166 videos across 1,722 different objects with five categories (including bottles, bowls, cameras, laptops and mugs). REAL275 contains six testing scenes with six categories(including bottles, bowls, cameras, cans, laptops and mugs). Both benchmarks provide RGB-D images, and the foreground segmentation and shape token are obtained by Mask R-CNN (He et al. 2017).

Evaluation Metrics. We follow REAL275 (Wang et al. 2019a) to use intersection over union (IoU) with a threshold of 25% ($IOU_{0.25}$) and 50% ($IOU_{0.5}$) to evaluate 3D object localization. For object pose estimation, we report the average pose accuracy of unseen categories where a pose is considered accurate when the translation errors are smaller than m cm and rotation errors are smaller than n°. We also compare the inference speed (s) of one frame.

Baselines. We compare our method with three types of baselines: 1) *Supervised Methods* train the models with ground truth pose annotations, including DPDN (Lin et al. 2022a), VI-Net (Lin et al. 2023)and SPD (Tian, Ang, and Lee 2020). 2) *Self-Supervised Methods* train the models without ground truth pose annotations, including Self-Pose (Zhang et al. 2022), Wild6D (Ze and Wang 2022) and SSC-6D (Peng et al. 2022). 3) *Zero-Shot Methods* without additional training including Zero-Pose (Goodwin et al. 2022) and MegaPose (Labbé et al. 2023).

Experimental Settings. During comparison with baselines, the pose estimation models are not trained on the tested categories. We follow REAL275 (Wang et al. 2019a) and Wild6D (Ze and Wang 2022) to divide the training set and the test set. We adapt the leave-1 strategy for supervised methods, which selects one category as the test set and use the remaining categories to train the model. We conduct leave-1 experiments for each category and finally take the average of them. We directly use the official pre-trained models to conduct the leave-p experiments for self-supervised methods, as they have the property of per-subject-per-train. We select the model for each category, test it on other unseen categories, and finally take the average of evaluation metrics. For the zero-shot method, we test directly on all categories without model training or fine-tuning. During testing, we do not provide supervised and self-supervised methods with a reference model used in the zero-shot method as the shape prior. Because these methods have not seen such category shapes during training, it is impossible to establish an effective association with the target object and will result in poor prediction results.

Implementation Details

We set $\alpha_{D1}, \alpha_{D2}, \alpha_{SD}$ to 0, 0.7, 0.3 in Eq.1, set the α_m, α_g, α_c to 1, 1 and 0.1 respectively. We iteratively update the object pose 2 times to obtain the coarse object pose and run RANSAC up to 1,000 times for each iteration to handle outliers. In the pose refinement stage, we use Adam (Kingma and Ba 2014) as the optimizer to minimize the loss function. We test our method on a single GeForce RTX 4090, costing 11.7 GB memory in coarse pose estimation stage and costing 5.5 GB memory in pose refinement stage for each instance. Please refer to the supplementary file for more details.

Comparison Results

We show comparison results on REAL275 and Wild6D in Tab. 1 and Fig. 5. Supervised and self-supervised methods

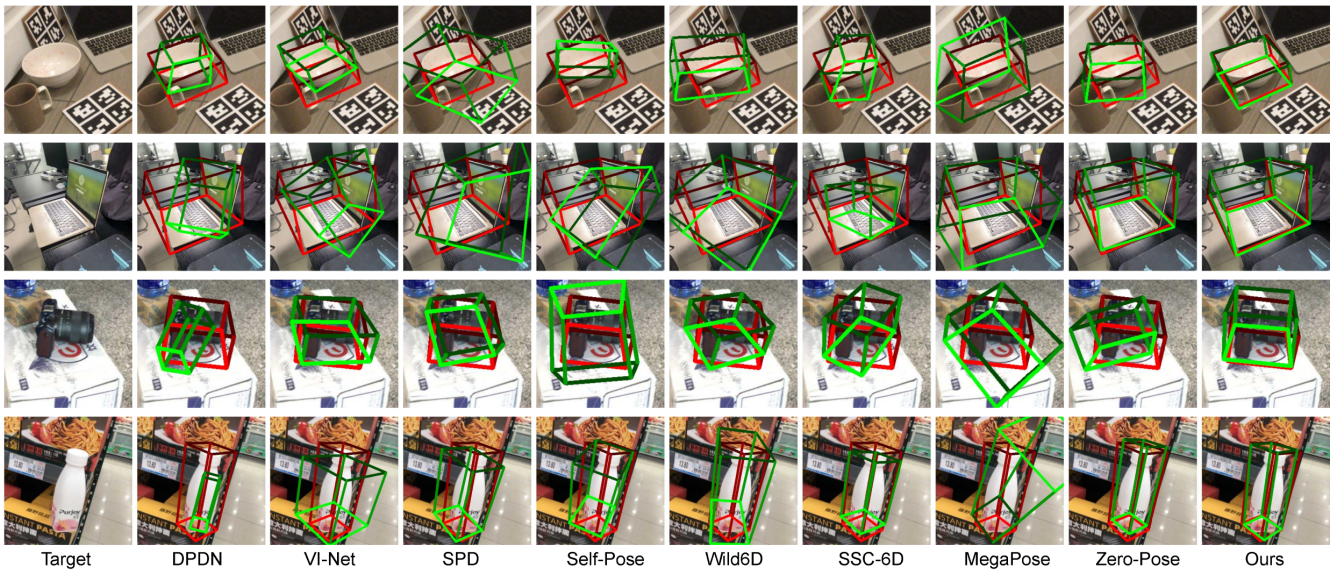

Figure 5: Qualitative results on REAL275 and Wild6D. The red box represents the ground truth, and the green box represents the estimation. Previous methods exhibit large errors when applied to unseen categories due to the significant texture and shape differences. Our method demonstrates strong generalization on unseen categories with accurate pose estimation.

Method	REAL275			WILD6D			
	$IOU_{0.25/0.5}$	$5°2cm$ / $5°5cm$	$10°2cm$ / $10°5cm$	$IOU_{0.25/0.5}$	$5°2cm$ / $5°5cm$	$10°2cm$ / $10°5cm$	Speed
DPDN	70.35/12.36	5.11/8.27	12.86/20.32	74.73/15.77	6.82/11.86	15.58/27.44	0.01
VI-Net	51.79/13.81	19.47/**37.52**	22.59/50.79	60.59/25.40	42.14/**66.15**	44.44/72.84	0.01
SPD	45.91/8.81	4.18/6.49	10.33/16.88	37.95/8.55	1.99/6.10	5.76/16.35	0.01
Self-Pose	60.54/ 6.96	0.40/0.48	1.44/1.98	63.34/ 3.61	0.10/0.98	0.38/3.44	0.01
Wild6D	48.84/12.89	5.26/7.71	12.28/18.64	55.50/15.36	10.10/16.42	20.23/36.32	0.01
SSC-6D	59.84/11.21	4.73/6.54	12.41/17.36	69.81/15.82	6.30/10.29	16.33/27.70	0.08
MegaPose	5.93/0.27	0.28/0.93	0.45/2.16	0/0	0/0	0/0.01	0.92
Zero-Pose	**82.51**/58.34	21.36/22.94	43.84/49.82	86.30/55.36	24.89/43.56	41.38/71.09	0.97
Ours	80.06/**63.49**	**30.61**/33.23	**50.15/57.74**	**88.46/67.16**	**47.69**/60.58	**59.46/80.47**	3.83

Table 1: Quantitative results on REAL275 and Wild6D. Zero-Shot methods show strong generalization ability on unseen categories and our cascaded coarse-to-fine optimization strategy is more effective in accurate pose prediction

over-fit the seen category shape priors, which make them difficult to accurately predict the exact geometry of unseen categories, leading to significant errors and pose drifting. These non-foundation models cannot establish effective semantic similarities on unseen categories with complex texture and shape variations. VI-Net overfits symmetric objects and obtains a better $5°5cm$ result. However, when the tolerance for rotation error increases ($10°5cm$), its accuracy becomes worse than ours. Note that MegaPose fails in this experimental setup. We analyze that this instance-level zero-shot method cannot establish effective associations between different instances and the difference in scale between the reference object model and the target object leads to a large translation error. Our method consistently outperforms Zero-Pose in almost all evaluation metrics, because our universal features can find more consistent correspondences (last row of Tab. 3), and the cascaded coarse-to-fine optimization strategy is more effective in accurate pose prediction (second and fifth row of Tab. 2).

Ablation Results

Effect of Iterative Optimization. To investigate the effect of iterative optimization in coarse pose estimation stage, we evaluate the pose accuracy at different iterative steps. As shown in Tab. 2 (Iter.), by comparing the results in the first and second rows, we conclude that iterative optimization can significantly improve the performance of the object pose (i.e. $5°2cm$ and $5°5cm$). The iterative optimization can effectively address the problem of correspondence degradation if the pose deviates much from the target pose: As the estimated object pose improves, better correspondences will be established between the target object and the reference object, making the pose estimation more accurate (Fig. 3).

Effect of Pose Refinement Module. We compare the results with and without the pose refinement module to investigate the effect of the pose refinement module. As shown in the

Ite.	Ref.	L_g	Def.	REAL275			WILD6D		
				$IOU_{0.25/0.5}$	5° $2/5cm$	10° $2/5cm$	$IOU_{0.25/0.5}$	5° $2/5cm$	10° $2/5cm$
1	×	×	×	77.49/59.22	24.16/27.10	47.43/54.39	87.17/60.36	32.79/52.63	49.08/79.55
2	×	×	×	80.00/58.39	28.56/30.94	48.33/55.27	88.17/62.77	42.71/57.58	55.99/**80.69**
2	✓	×	✓	80.04/63.44	30.35/33.13	49.91/57.66	88.45/67.11	47.48/60.37	59.35/80.42
2	✓	✓	×	80.04/58.94	26.74/29.18	44.05/54.18	88.32/64.79	45.83/59.27	58.00/80.19
2	✓	✓	✓	**80.06/63.49**	**30.61/33.23**	**50.15/57.74**	**88.46/67.16**	**47.69/60.58**	**59.46**/80.47

Table 2: Ablation studies on the influence of the iterative optimization and pose refinement module. After iteration, objects can be further aligned to establish smoother correspondences to get performance improvements. After pose refinement, the shapes between the objects are more similar, thereby ensuring more accurate object alignment to get precise pose estimation.

	Method	$IOU_{0.25/0.5}$	5° $2/5cm$	10° $2/5cm$
REAL275	v1	82.41/59.06	21.34/23.03	44.49/50.68
	v1+SD	81.19/58.20	21.35/23.06	44.46/50.34
	v2	76.47/56.96	21.71/24.95	45.04/53.00
	All	**84.62/62.77**	23.36/25.02	47.02/53.77
	v2+SD	77.49/59.22	**24.16/27.10**	**47.43/54.39**
WILD6D	v1	85.36/56.20	23.81/42.18	41.68/70.85
	v1+SD	85.53/56.30	23.68/42.77	41.61/70.77
	v2	83.87/56.90	30.94/46.86	46.36/74.05
	ALL	86.46/57.73	24.51/43.72	42.03/72.45
	v2+SD	**87.17/60.36**	**32.79/52.63**	**49.08/79.55**

Table 3: Ablation studies on different combinations of 2D universal features. We find that the feature combination of DINOv2 and SD will comprehensively utilize both global and local context to obtain more accurate estimation results.

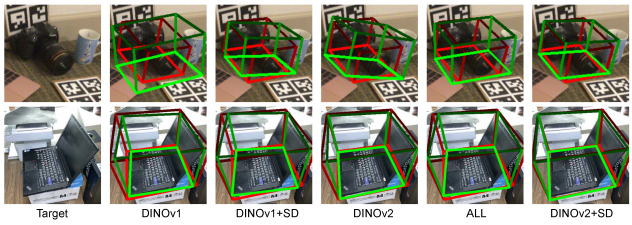

Figure 6: Qualitative results on different combinations of 2D universal features.

last row of Tab. 2, the pose refinement can further improve the pose accuracy. A critical step in pose refinement is shape deformation. Fig. 7 (a) shows the reference shape before and after deformation. The camera has been optimized to low the top bulge, and shrunk the overall body. After shape deformation, the bias in correspondences caused by shape differences is reduced, resulting in a better alignment and more accurate pose estimation. By comparing the third row and the fifth row in Tab. 2, we observe that our universal alignment loss L_g allows for more accurate pose. Conceptually, 3D universal features can effectively build correspondences with geometric semantic similarity, thereby constraining the ambiguity caused by optimizing the shape and pose.

Effect of 2D Universal Features. We evaluate different combinations of image features for object pose estimation without iteration and pose refinement and the results are

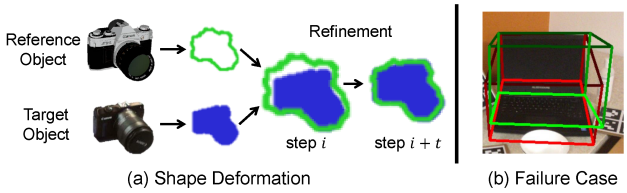

Figure 7: (a) After the shape optimization, the reference object shape will become closer to the target object shape, resulting in more accurate pose. (b) The occlusion makes the shape of the target object incomplete, causing translation and rotation errors.

shown in Tab. 3 and Fig. 6. This experiment shows that DINOv2 is more effective than DINOv1 in capturing semantic similarity, making the predicted scale and rotation more accurate. Furthermore, combining DINOv2 with Stable Diffusion can make semantic correspondences smoother because SD features can take care of global relevance. Note 'ALL' will establish more correspondences on the boundary of object which results in a oversize scale estimation, which results in higher IOU but inaccurate pose. Therefore, we combine DINOv2 and Stable Diffusion features to establish the most accurate correspondences for coarse pose estimation.

Limitations and Failure Cases. Computation time is the bottleneck of our method. In real applications, we can use the full pipeline to locate objects in the first frame and perform efficient pose tracking with few fitting steps in 'Pose Refinement' initialized with pose of previous frame. The occlusion will also affect the accuracy of the pose estimation results (e.g., as shown in Fig. 7 (b)). We can use the inpainting method (Suvorov et al. 2022) to complete the image to establish correspondences, and then use the visible point cloud for registration.

Conclusion

We propose a new universal features guided zero-shot category-level object pose estimation method in coarse-to-fine fashion. It can estimate the 6D pose of objects from unseen categories without additional model fine-tuning. Our method efficiently utilizes 2D and 3D pre-trained universal features to achieve strong generalization capabilities. It can potentially help many applications deal with unseen categories and avoid additional model training or fine-tuning.

Acknowledgments

This work was supported in part by National Science and Technology Major Project (2022ZD0117904), National Natural Science Foundation of China (62473356,62373061), Beijing Natural Science Foundation (L232028), and CAS Major Project (RCJJ-145-24-14). Heng Li and Ping Tan are supported by the project HKPC22EG01-E from the Hong Kong Industrial Artificial Intelligence & Robotics Centre (FLAIR).

References

Amir, S.; Gandelsman, Y.; Bagon, S.; and Dekel, T. 2021. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3): 4.

Burchfiel, B.; and Konidaris, G. 2019. Probabilistic category-level pose estimation via segmentation and predicted-shape priors. *arXiv preprint arXiv:1905.12079*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, D.; Li, J.; Wang, Z.; and Xu, K. 2020. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11973–11982.

Chen, J.; Sun, M.; Bao, T.; Zhao, R.; Wu, L.; and He, Z. 2023. ZeroPose: CAD-model-based zero-shot pose estimation. *arXiv preprint arXiv:2305.17934*, 2.

Chen, W.; Jia, X.; Chang, H. J.; Duan, J.; Shen, L.; and Leonardis, A. 2021. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1581–1590.

Chen, Y.; Di, Y.; Zhai, G.; Manhardt, F.; Zhang, C.; Zhang, R.; Tombari, F.; Navab, N.; and Busam, B. 2024. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9959–9969.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Di, Y.; Zhang, R.; Lou, Z.; Manhardt, F.; Ji, X.; Navab, N.; and Tombari, F. 2022. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6781–6791.

Fan, Z.; Pan, P.; Wang, P.; Jiang, Y.; Xu, D.; Jiang, H.; and Wang, Z. 2023. POPE: 6-DoF Promptable Pose Estimation of Any Object, in Any Scene, with One Reference. *arXiv preprint arXiv:2305.15727*.

Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.

Goodwin, W.; Vaze, S.; Havoutis, I.; and Posner, I. 2022. Zero-shot category-level object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 516–532. Springer.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, Y.; Fan, H.; Huang, H.; Chen, Q.; and Sun, J. 2022. Towards self-supervised category-level object pose and size estimation. *arXiv preprint arXiv:2203.02884*.

Hodan, T.; Barath, D.; and Matas, J. 2020. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11703–11712.

Hu, Y.; Fua, P.; Wang, W.; and Salzmann, M. 2020. Single-stage 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2930–2939.

Jung, H.; Wu, S.-C.; Ruhkamp, P.; Zhai, G.; Schieber, H.; Rizzoli, G.; Wang, P.; Zhao, H.; Garattoni, L.; Meier, S.; et al. 2024. HouseCat6D-A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22498–22508.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Labbé, Y.; Carpentier, J.; Aubry, M.; and Sivic, J. 2020. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, 574–591. Springer.

Labbé, Y.; Manuelli, L.; Mousavian, A.; Tyree, S.; Birchfield, S.; Tremblay, J.; Carpentier, J.; Aubry, M.; Fox, D.; and Sivic, J. 2023. MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare. In *Conference on Robot Learning*, 715–725. PMLR.

Li, Y.; Wang, G.; Ji, X.; Xiang, Y.; and Fox, D. 2018. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 683–698.

Lin, H.; Liu, Z.; Cheang, C.; Zhang, L.; Fu, Y.; and Xue, X. 2021. Donet: Learning category-level 6d object pose and size estimation from depth observation. *arXiv preprint arXiv:2106.14193*, 4: 11–12.

Lin, J.; Wei, Z.; Ding, C.; and Jia, K. 2022a. Category-level 6D object pose and size estimation using self-supervised deep prior deformation networks. In *European Conference on Computer Vision*, 19–34. Springer.

Lin, J.; Wei, Z.; Zhang, Y.; and Jia, K. 2023. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, 14001–14011.

Lin, X.; Yang, W.; Gao, Y.; and Zhang, T. 2024. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21040–21049.

Lin, Y.; Tremblay, J.; Tyree, S.; Vela, P. A.; and Birchfield, S. 2022b. Single-stage keypoint-based category-level object pose estimation from an RGB image. In *2022 International Conference on Robotics and Automation (ICRA)*, 1547–1553. IEEE.

Luo, G.; Dunlap, L.; Park, D. H.; Holynski, A.; and Darrell, T. 2023. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. *arXiv preprint arXiv:2305.14334*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Örnek, E. P.; Labbé, Y.; Tekin, B.; Ma, L.; Keskin, C.; Forster, C.; and Hodan, T. 2023. FoundPose: Unseen Object Pose Estimation with Foundation Features. *arXiv preprint arXiv:2311.18809*.

Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; and Bao, H. 2019. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4561–4570.

Peng, W.; Yan, J.; Wen, H.; and Sun, Y. 2022. Self-supervised category-level 6D object pose estimation with deep implicit shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2082–2090.

Ravi, N.; Reizenstein, J.; Novotny, D.; Gordon, T.; Lo, W.-Y.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sahin, C.; and Kim, T.-K. 2018. Category-level 6d object pose recovery in depth images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent Correspondence from Image Diffusion. *arXiv preprint arXiv:2306.03881*.

Tekin, B.; Sinha, S. N.; and Fua, P. 2018. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 292–301.

Tian, M.; Ang, M. H.; and Lee, G. H. 2020. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 530–546. Springer.

Umeyama, S. 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380.

Wang, G.; Manhardt, F.; Tombari, F.; and Ji, X. 2021. Gdrnet: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16611–16621.

Wang, H.; Sridhar, S.; Huang, J.; Valentin, J.; Song, S.; and Guibas, L. J. 2019a. 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*, 2642–2651.

Wang, H.; Sridhar, S.; Huang, J.; Valentin, J.; Song, S.; and Guibas, L. J. 2019b. 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*, 2642–2651.

Wang, P.; Jung, H.; Li, Y.; Shen, S.; Srikanth, R. P.; Garattoni, L.; Meier, S.; Navab, N.; and Busam, B. 2022. Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21222–21231.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019c. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Wen, B.; Yang, W.; Kautz, J.; and Birchfield, S. 2024. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17868–17879.

Xiang, Y.; Schmidt, T.; Narayanan, V.; and Fox, D. 2017. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*.

Ze, Y.; and Wang, X. 2022. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *Advances in Neural Information Processing Systems*, 35: 27469–27483.

Zhang, J.; Herrmann, C.; Hur, J.; Cabrera, L. P.; Jampani, V.; Sun, D.; and Yang, M.-H. 2023. A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. *arXiv preprint arXiv:2305.15347*.

Zhang, K.; Fu, Y.; Borse, S.; Cai, H.; Porikli, F.; and Wang, X. 2022. Self-supervised geometric correspondence for category-level 6d object pose estimation in the wild. *arXiv preprint arXiv:2210.07199*.