

Gen6D: Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images

Yuan Liu¹, Yilin Wen¹, Sida Peng², Cheng Lin³,
 Xiaoxiao Long¹, Taku Komura¹, and Wenping Wang⁴

¹ The University of Hong Kong

² Zhejiang University

³ Tencent

⁴ Texas A&M University

Abstract. In this paper, we present a generalizable model-free 6-DoF object pose estimator called Gen6D. Existing generalizable pose estimators either need the high-quality object models or require additional depth maps or object masks in test time, which significantly limits their application scope. In contrast, our pose estimator only requires some posed images of the unseen object and is able to accurately predict poses of the object in arbitrary environments. Gen6D consists of an object detector, a viewpoint selector and a pose refiner, all of which do not require the 3D object model and can generalize to unseen objects. Experiments show that Gen6D achieves state-of-the-art results on two model-free datasets: the MOPED dataset and a new GenMOP dataset. In addition, on the LINEMOD dataset, Gen6D achieves competitive results compared with instance-specific pose estimators. Project page: <https://liuyuan-pal.github.io/Gen6D/>.

Keywords: 6-DoF Object Pose Estimation; Camera Pose Estimation

1 Introduction

Estimating the orientation and location of an object in 3D space is a preliminary and necessary step for many tasks involving interaction with the object. In the last decade, 3D vision has witnessed tremendous development ranging from robotics, games, to VR/AR. These applications raise new demands for the 6-DoF object pose estimation, requiring a pose estimator to be generalizable, flexible, and easy-to-use. However, existing methods suffer from several restrictive conditions. Most methods [28,70,64] can only be used for a specific object or category same as the training data. Some methods [57,29,74,72,43,41,73] can generalize to unseen objects, but they rely on high-quality target 3D models [57,29,74,72,43], or additional depth maps [41] and masks [41,73] at test time. These requirements severely limit the practical applications of the existing pose estimators.

To meet demands in practical applications, we argue that such a pose estimator should have the following properties. 1) **Generalizable.** The pose estimator can be applied on an arbitrary object without training on the object or its category. 2) **Model-free.** When generalizing to an unseen object, the estimator

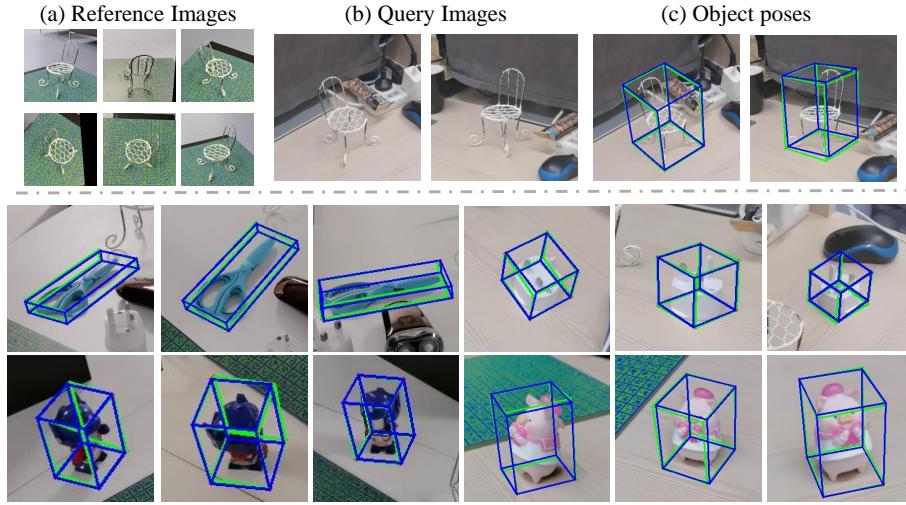


Fig. 1. Given (a) reference images of an object with known poses and (b) query images containing the same object with unknown poses, our pose estimator is able to accurately estimate (c) their object poses in the query images, where green color means ground-truth and blue color means estimation. Note that all objects are unseen in the training set and the same estimator is applied for all objects.

only needs some reference images of this object with known poses to define the object reference coordinate system, as shown in Fig. 1 (a), but does not rely on a 3D model of the object. 3) **Simple inputs.** When estimating object poses, the estimator only takes RGB images as inputs without requiring additional object masks or depth maps.

To the best of our knowledge, there is no existing pose estimator satisfying all the above three properties simultaneously. Thus, in this paper, we propose a simple but effective pose estimator, called **Gen6D**, which possesses the three properties above. Given input reference images of an arbitrary object with known poses, Gen6D is able to directly predict its object pose in any query images, as shown in Fig. 1. In general, an object pose can be estimated by directly predicting rotation/translation by regression [70,27,59], solving a Perspective-n-Points (PnP) problem [42,47] or matching images with known poses [58,69,57]. Direct prediction of rotation and translation by regression is mostly limited to a specific instance or category, which has difficulty in generalizing to unseen objects. Meanwhile, due to the lack of 3D models, PnP-based methods do not have 3D keypoints to build 2D-3D correspondences so that they are incompatible with model-free setting. Hence, we apply image-matching in our framework for pose estimation, which can generalize to unseen objects by learning a general image similarity metric.

In Gen6D, we propose a novel image-matching based framework to estimate the object pose in a coarse-to-fine manner. The framework consists of an object

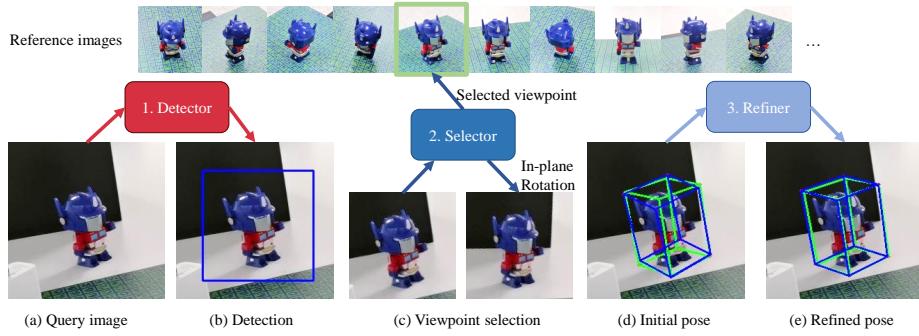


Fig. 2. Overview. The proposed pose estimator consists of a detector which detects the object in the query image, a viewpoint selector which selects the most similar viewpoint from reference images, and a pose refiner which refines the initial pose into an accurate object pose.

detector, a viewpoint selector and a pose refiner, as shown in Fig. 2. Given reference images and a query image, an object detector first detects the object regions by correlating the reference images with the query image, which is similar to [1]. Then, a viewpoint selector matches the query image against the reference images to produce a coarse initial pose. Finally, the initial pose is further refined by a pose refiner to search for an accurate object pose.

A challenge is how to design a viewpoint selector when the reference images are sparse and contain cluttered background. Existing image-matching methods [58, 69, 57, 22, 2] have difficulty in handling this problem due to two problems. First, these image-matching methods embed images into feature vectors and compute similarities using distances of feature vectors, in which cluttered background interferes the embedded feature vectors and thus severely degenerates the accuracy. Second, given a query image, there may not be a reference image with exactly the same viewpoint as the query image. In this case, there will be multiple plausible reference images and the selector has to select the one with the nearest viewpoint as the query image, which usually are very ambiguous as shown in Fig. 3.

To address these problems in viewpoint selection, we propose to use neural networks to pixel-wisely compare the query image with every reference image to produce similarity scores and select the reference image with highest similarity score. This pixel-wise comparison enables our selector to concentrate on object regions and reduces the influence of cluttered background. Furthermore, we add global normalization layers and self-attention layers to share similarity information across different reference images. These two kinds of layers enable every reference images to commute with each other, which provides context information for the selector to select the most similar reference image.

The main challenge of developing our pose refiner is the unavailability of the object model. Existing pose refiners [29, 74] are based on rendering-and-comparison, which render an image on the input pose and then match the ren-

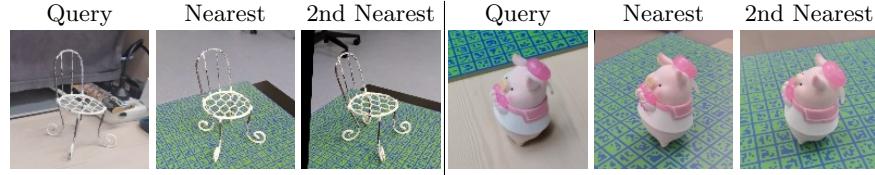


Fig. 3. Query images and reference images have cluttered background. The reference image with the nearest viewpoint looks very similar as the one with second nearest viewpoint, which brings challenges for the selector to correctly select the nearest one.

dered image with the query image to refine the input pose. However, without the object model, rendering high-quality images on arbitrary poses is difficult, which makes these refinement methods infeasible in the model-free setting.

To address this problem, we propose a novel 3D volume-based pose refinement method. Given a query image and an input pose, we find several reference images that are near to the input pose. These reference images are projected back into 3D space to construct a feature volume. Then, the constructed feature volume is matched against the features projected from the query image by a 3D CNN to refine the input pose. In comparison with previous pose refiners [29,74], our pose refiner avoids rendering any new images. Meanwhile, the constructed 3D feature volume enables our method to infer the 3D pose refinement in the 3D space. In contrast, previous pose refiners [29,74] only rely on 2D image features to regress a 3D relative pose, which are less accurate especially for unseen objects.

To validate the effectiveness of our generalizable model-free pose estimator, we introduce a new dataset, called General Model-free Object Pose Dataset (GenMOP), which contains video sequences of objects in different environments and lighting conditions. We choose one sequence as reference images and the rest sequences of the same object as test query images. Experiments show that without training on these objects, our method still outperforms instance-specific estimator PVNet [42] on the GenMOP dataset and another model-free MOPED [41] dataset. We also evaluate our method on the LINEMOD dataset [23], on which our generalizable pose estimator achieves comparable results as instance-specific estimators which needs to be trained with excessive rendered images.

2 Related works

2.1 Specific object pose estimator

Most object pose estimators [70,58,42,28,24,68,14,63,34,54,27,45,26,25,46,55] are instance-specific, which cannot generalize to unseen objects and usually require a 3D model of the object to render extensive images for training. Recent instance-specific pose estimators [40,6,33] reconstruct the object model implicitly in the pipeline so that they are model-free. Category-specific pose estimators [64,11,67,13,30,10,60,8,31,9,16,15] can generalize to objects in the same category and also do not require the object model. However, they are still unable to

predict poses for objects in unseen categories. In comparison, Gen6D is generalizable, which makes no assumption of the category or the instance of the object and also does not need the 3D model of the object.

2.2 Generalizable object pose estimator

Generalizable pose estimators mostly require an object model either for shape embedding [72,43,12,44] or template matching [22,2,69,57,21,37,75] or rendering-and-comparison [29,74,38,4,57,18]. To avoid using 3D models, recent works [73,41] utilize the advanced neural rendering techniques [36] to directly render from posed images for pose estimation. However, current rendering methods are only able to render images under exactly the same appearance like lighting conditions, which degenerates the accuracy under varying appearance. To remedy this, these methods have to resort to additional depth maps [41] or object masks [41,73] to achieve robustness. There are also some works focusing on estimating poses of unseen objects using RGBD sequences [66,38,52,20,5,17]. In contrast to these methods, Gen6D is model-free and does not require depth maps or masks. There are also concurrent works [56,51] of generalizable pose estimation.

2.3 Instance detection

Instance detection aims to detect a given object with some images of the object [1,35,19,39]. There are some instance detection methods which also estimate viewpoints [71,3] for novel category in one- or few-shot setting. The detector of Gen6D is inspired by [1], which uses correlation to find the object region. The target of Gen6D is to estimate the 6-DoF object pose, which is different from these methods for detection or category-level viewpoint estimation.

3 Method

Given N_r images of an object with known camera poses, called **reference images**, our target is to predict the pose of the object in a **query image**. The object pose here means a translation \mathbf{t} and a rotation \mathbf{R} that transform the object coordinate \mathbf{x}_{obj} to the camera coordinate $\mathbf{x}_{cam} = \mathbf{R}\mathbf{x}_{obj} + \mathbf{t}$. All the intrinsics parameters of images are already known.

Data normalization. For every object, we can estimate a rough size of the object by triangulating points from reference images or simply unprojecting reference images to find an intersection. The center of triangulated points or the center of the 3D intersection region is regarded as the object center. Then, the object coordinate system is normalized so that the object center locates at the origin and the object size is 1, which means the whole object resides inside a unit sphere at the origin. This data normalization ensures the feature volume constructed by our pose refiner in Sec. 3.3 will contain the target object. More details about the normalization can be found in the supplementary materials.

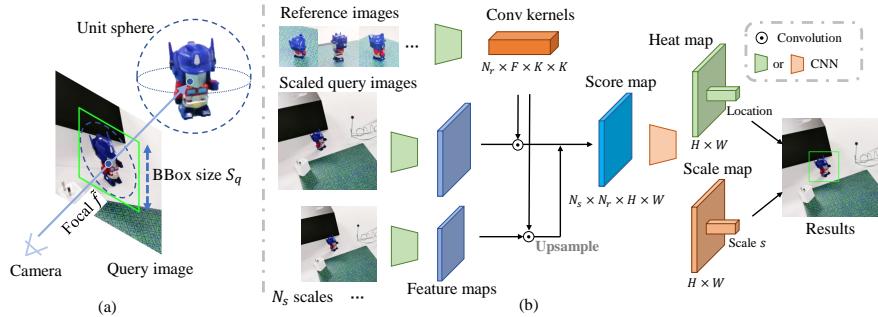


Fig. 4. (a) Detection outputs. Depth can be computed from the bounding box size S_q , which along with the 2D projection of the object center determine the location of the object center. (b) Architecture of the detector. We use features of reference images as kernels to convolve features of multi-scale query images to get score maps. Score maps are further processed by a CNN to produce a heat map about the object center and a scale map to determine the bounding box size.

Overview. As shown in Fig. 2, the proposed Gen6D pose estimator consists of an object detector, a view selector and a pose refiner. The object detector crops the object region and estimates an initial translation (Sec. 3.1). The view selector finds an initial rotation by selecting the most similar reference image and estimating an in-plane rotation (Sec. 3.2). The initial translation and rotation are used in the pose refiner to iteratively estimate an accurate pose (Sec. 3.3).

3.1 Detection

The query image is usually very large and the object only occupies a small region on the query image. To focus on the object, we apply a correlation-based instance detector similar to [1]. We decompose the detection problem into two parts, i.e. finding the 2D projection q of the object center and estimating a compact square bounding box size S_q that encloses the unit sphere. As shown in Fig. 4 (a), such a compact bounding box size is used in computing the depth of the object center by $d = 2\tilde{f}/S_q$, where 2 is the diameter of the unit sphere and \tilde{f} is a virtual focal length by changing the principle point to the estimated projection q . The projection q and the depth d will determine the location of the object center, which provides an initial translation for the object pose.

The design of our detector is shown in Fig. 4 (b). We extract feature maps on both the reference images and the query image by a VGG [53]-11 network. Then, the feature maps of all reference images are regarded as convolution kernels to convolve with the feature map of the query image to get score maps. To account for scale differences, we conduct such convolution at N_s predefined scales by resizing the query images to different scales. Based on the multi-scale score maps, we regress a heat map and a scale map. We select the location with the max value on the heat map as the 2D projection of the object center and use the

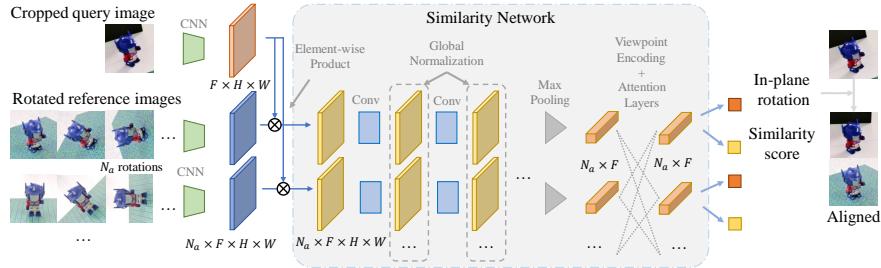


Fig. 5. Architecture of the viewpoint selector. We compute the element-wise product of every reference image with the query image to get a score map, on which a similarity network is applied to compute an in-plane rotation and a similarity score for this reference image. Note that in the similarity network, we use global normalization layers and a transformer to share information across reference images.

scale value s at the same location on the scale map to compute the bounding box size $S_q = S_r * s$, where S_r is the size of reference images.

With the detected 2D projections and scales, we compute the initial 3D translations and crop the object regions for subsequent processing. More details about architecture and training of detector networks can be found in the supplementary materials.

3.2 Viewpoint Selection

Viewpoint selection aims to select a reference image whose viewpoint is the nearest to the query image. Meanwhile, we will estimate an in-plane rotation between the query image and the selected reference image. We approximately regard the viewpoint of the selected reference image as the viewpoint of the query image, which along with the estimated in-plane rotation forms an initial rotation for the object pose.

As shown in Fig. 5, we design a viewpoint selector to compare the query image with every reference image to compute similarity scores. Specifically, we first extract feature maps by applying a VGG [53]-11 on reference images and the query image. Then, for every feature map of reference images, we compute its element-wise product with the feature map of the query image to produce a correlation score map. Finally, the correlation score map is processed by a similarity network to produce a similarity score and a relative in-plane rotation to align the query image with the reference image. In our viewpoint selector, we have three special designs.

In-plane rotation. To account for in-plane rotations, every reference image is rotated by N_a predefined angles and all rotated versions are used in the element-wise product with the query image.

Global normalization. For every feature map produced by the similarity network, we normalize it with the mean and variance computed from all feature maps of reference images. Such a global normalization helps our selector select

the relatively most similar reference image because it allows the distribution of feature maps to encode the context similarity and amplifies the similarity differences among different reference images. For every reference image, max-pooling is applied on its feature map to produce a similarity feature vector.

Reference view transformer. We apply a transformer on the similarity feature vectors of all reference images, which includes the positional encoding of their viewpoints and attention layers over all similarity feature vectors. Such a transformer lets feature vectors communicate with each other to encode contextual information [61,49,65], which is helpful to determine the most similar reference image. The outputs of reference view transformer will be used to regress a similarity score and an in-plane rotation angle for each reference image. The viewpoint of the reference image with highest score will be selected.

With the selected viewpoint and the estimated in-plane rotation, we estimated an initial rotation for the object pose, which will be refined by the pose refiner. More details about the network and training can be found in the supplementary materials.

3.3 Pose refinement

Combining the translation estimated by the object detector and the rotation estimated by the viewpoint selector, we get an initial coarse object pose. This initial pose is further refined by a 3D volume-based pose refiner.

Specifically, since the objects are already normalized inside an unit sphere at the origin, we build a volume within the unit cube at the origin with $S_v^3 = 32^3$ vertices. As shown in Fig. 6 (a), to construct the features on these vertices, we first select $N_n = 6$ reference images that are near to the input pose. We extract feature maps on these selected reference images by a 2D CNN. Then, these feature maps are unprojected into the 3D volume and we compute the mean and variance of features among all reference images as features for volume vertices. For the query image, we also extract its feature map by the same 2D CNN, unproject feature map into the 3D volume using the input pose and concatenate the unprojected query features with the mean and variance of reference image features. Finally, we apply a 3D CNN on the concatenated features of the volume to predict a pose residual to update the input pose.

Similarity approximation. Instead of regressing the rigid pose residual directly, we approximate it with a similarity transformation, as shown in Fig. 6 (b). The approximate similarity transformation consists of a 2D in-plane offset, a scale factor and a residual 3D rotation. The reason of using this approximation is that it avoids direct regression of the 3D translation from the red circle to the solid green circle in Fig. 6, which is out of the scope of the feature volume. Instead, we regress a similarity transformation from red circle to dotted green circle, which can be easily inferred from the features defined in the volume. More details can be found in the supplementary materials. In our implementation, we apply the refiner iteratively 3 times by default.

Discussion. The key difference between our volume-based refiner and other pose refiners [29,74,57] is that our pose refiner does not require rendering an

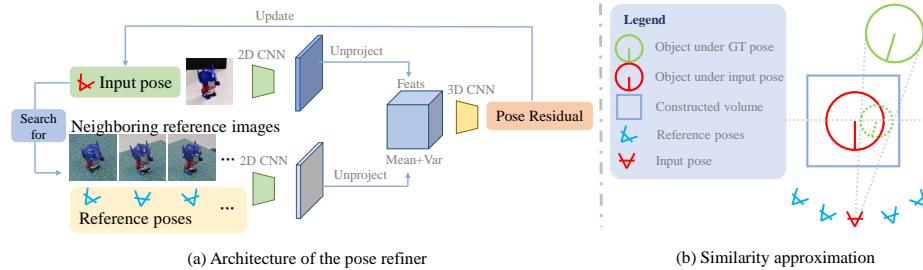


Fig. 6. (a) Architecture of our pose refiner. (b) A 2D diagram to illustrate the similarity transformation approximation. Though the ground-truth pose residual from the input object pose (solid red circle) to the ground-truth object pose (solid green circle) is a rigid transformation, we can approximate this rigid transformation by a similarity transformation inside the feature volume. Our pose refiner predicts the similarity transformation, which transforms the input red circle to the dotted green circle. Then, the similarity transformation can be converted to a rigid transformation.

image on the input pose, which thus is more suitable for the model-free pose estimation. Meanwhile, since the 3D volume is constructed by multiple reference images with different poses, our volume-based refiner is able to know the image features under different poses and infer how pose changes affect the image features for unseen objects. In comparison, previous pose refiners [29,74,57] only compare a rendered image with the input query image to compute a pose residual. Such a 2D image does not provide enough 3D structure information to infer how pose changes affect image patterns, especially for unseen objects. Thus, it is hard for these methods to predict correct pose residuals for unseen objects.

4 Experiments

4.1 GenMOP Dataset

To validate the effectiveness of the proposed method, we collect a dataset called General Model-free Object Pose Dataset (GenMOP). GenMOP dataset consists of 10 objects ranging from flatten objects like “scissors” to thin structure objects like “chair” as shown in Fig. 7. For each object, two video sequences of the same object are collected in different environments like backgrounds and lighting conditions. Every video sequence is split into ~ 200 images. For each sequence, we apply COLMAP [50] to reconstruct the camera poses in each sequences separately and manually label keypoints on the object for cross-sequence alignment. More details about the GenMOP dataset can be found in the supplementary.

4.2 Protocol

We evaluate Gen6D pose estimator on the GenMOP dataset, the LINEMOD [23] dataset and the MOPED [41] dataset.

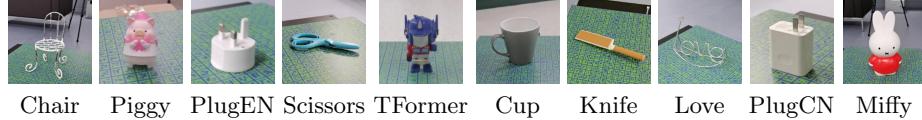


Fig. 7. Objects in the GenMOP dataset. The first 5 objects are used in test and the last 5 objects are used in training.

GenMOP. On the GenMOP dataset, we select one video sequence as reference images and the other video sequence in a different environment as test query images, both of which contain ~ 200 images.

LINEMOD. The LINEMOD [23] dataset is a widely-used dataset for object pose estimation. On the LINEMOD dataset, we follow the commonly-used train-test split as [59]. We select the training images (~ 180) as reference images and all the rest ~ 1000 test images as query images for evaluation.

MOPED. The MOPED [41] dataset is intended for model-free object pose estimation. Since the MOPED dataset is generated automatically by depth fusion and point cloud registration, object poses in some sequences are not very accurate. Thus, we manually select reliable subsets from 5 objects for evaluation. For each object, there are 200-600 reference images and 100-300 query images.

Training datasets. The training dataset of Gen6D estimator consists of: 1) Rendered images from ~ 2000 ShapeNet [7] models, 2) Google Scanned Object dataset rendered by [65] with 1023 objects, 3) 5 objects from the GenMOP dataset and 4) 5 objects (ape/can/holepuncher/iron/phone) from the LINEMOD dataset. Note we only train a single model and test its performance on the unseen objects on GenMOP, LINEMOD and MOPED dataset.

Metrics. We adopt the widely-used Average Distance (ADD) [23] and the projection error as metrics. On the ADD, we compute the recall rate with 10% of the object diameter (ADD-0.1d) and the AUC in 0-10cm (ADD-AUC). On the projection error, we compute the recall rate at 5 pixels (Prj-5).

4.3 Results on GenMOP

For comparison, we choose the generalizable image-matching based ObjDesc [69] and two instance-specific estimators PVNet [42] and RLLG [6] as baseline methods. Quantitative results are shown in Table 1 and some qualitative results are shown in Fig. 1. More qualitative results are in the supplementary.

Baseline implementation. For the generalizable template-matching ObjDesc [69], we use the same training dataset as Gen6D. In testing, we crop the object region by our object detector and then use ObjDesc to select the most similar reference image to the query image. The pose of the selected reference image is regarded as the pose of the query image. All objects used in evaluation are unseen for Gen6D and ObjDesc in training. For instance-specific estimators PVNet [42] and RLLG [6], we have to train different models for different objects separately. On every test object, the reference images for Gen6D are used as

Table 1. Performance on the GenMOP dataset. “General” means generalizable or not. “Ours w/o Ref.” means not using the pose refiner in the Gen6D estimator.

Metrics	Method	General	Object Name				avg.	
			Chair	PlugEN	Piggy	Scissors		
ADD-0.1d	PVNet [42]	✗	49.50	2.33	77.89	44.40	19.84	38.79
	RLLG [6]	✗	0.70	1.28	1.01	3.45	0.79	2.71
	ObjDesc [69]	✓	3.50	5.14	14.07	1.25	7.54	8.55
	Ours w/o Ref.	✓	14.00	7.48	39.70	16.81	11.51	17.90
	Ours	✓	61.50	19.63	75.38	32.76	62.70	50.39
Prj-5	PVNet [42]	✗	15.00	30.37	83.42	96.55	59.52	56.97
	RLLG [6]	✗	2.00	4.67	17.59	35.78	7.94	13.59
	ObjDesc [69]	✓	4.00	10.75	4.52	18.53	8.33	9.23
	Ours w/o Ref.	✓	11.50	40.65	33.17	34.05	64.29	36.73
	Ours	✓	55.00	72.90	92.96	93.53	98.81	82.64

training set for PVNet and RLLG. However, only ~ 200 reference images are not enough to produce reasonable results so we additionally label the object masks on these reference images and cut the object to randomly paste on backgrounds from COCO [32] to enlarge their training set. For PVNet, we use its 8 corners of the 3D bounding box as keypoints for voting because no model is available.

Comparison with baselines. 1) Both ObjDesc [69] and “Ours w/o Ref” select the most similar reference image to estimate the object pose. The results show that our viewpoint selector is able to select more accurate viewpoint than ObjDesc. However, only selecting the best reference viewpoint is not enough for predicting accurate poses because the reference images do not cover all possible viewpoints. 2) With further pose refinement, our Gen6D estimator is able to produce better results than the instance-specific methods PVNet and RLLG on average. The main reason is that for PVNet and RLLG, these reference images are not enough for training a very accurate pose estimator. In contrast, Gen6D well adapts into this setting with limited reference images of a novel object. Our pose refiner is able to learn generalizable features for accurate pose refinement.

4.4 Results on LINEMOD [23]

We further report results in ADD-0.1d on the LINEMOD [23] dataset in Table 2 and show qualitative results in Fig. 8. For baselines, we include the instance-specific pose estimators [58,62,74,6,42,70,29] and a generalizable estimator Pose-From-Shape (PFS) [72]. The instance-specific estimators are either trained on the synthetic data of the object (“synthetic training”) [58,62,74] or trained on both the synthetic and real data of the object (“real training”) [42,70,74]. PFS [72] is trained on ShapeNet [7], which embeds an object shape into a feature vector and applies the embedded feature vector on a query image to predict an object pose. For [72,74,70], we also include their reported performance using pose refiners DeepIM [29] or DPOD [74], both of which are trained on the synthetic data or

Table 2. ADD-0.1d on LINEMOD [23] dataset. “Training” means what kind of training set is used. “Synthetic” means the model only uses synthetic data of the given object for training; “Real” means the model is trained on both the synthetic images and real images of the given model; “No” means the model is not trained on any data of the test object. “GT-BBox” means a model uses the ground-truth bounding box or not to produce its performance. “Refine” means the pose refiner.

Training	Name	GT-BBox	Refine	Object Name					Avg.	
				cat	duck	bvise	cam	driller		
Synthetic	AAE [58]	✗	No	17.90	4.86	20.92	30.47	23.99	60.47	26.44
	Self6D [62]	✗	No	57.90	19.60	75.20	36.90	67.00	68.20	54.13
	DPOD [74]	✗	No	32.36	26.12	66.76	24.22	66.60	57.26	45.55
	DPOD [74]	✗	DPOD [74]	65.10	50.04	72.69	34.76	73.32	74.27	61.70
Real	PFS [72]	✓	DeepIM [29]	54.10	48.60	63.80	40.00	75.30	55.30	56.18
	PVNet [42]	✗	No	79.34	52.58	99.90	86.86	96.43	99.33	85.74
	PoseCNN [70]	✗	DeepIM [29]	82.10	77.70	97.40	93.50	95.00	96.84	95.19
	DPOD [74]	✗	DPOD [74]	94.71	86.29	98.45	96.07	98.80	97.50	90.53
Gen	PFS [72]	✓	No	15.40	8.20	25.10	12.10	18.60	6.50	14.32
	Ours	✓	No	94.11	81.31	99.52	94.31	96.33	93.38	93.16
	Ours	✗	No	15.97	7.89	25.48	22.06	17.24	35.80	20.74
	Ours	✗	Volume	60.68	40.47	77.03	66.67	67.39	89.83	67.01

real data of the test object. Ground-truth bounding box is used in PFS to crop the object region for the pose estimation. For baselines, we use the performance reported in their paper for comparison.

The results in Table 2 show that: 1) In comparison with the generalizable pose estimator PFS [72], Gen6D outperforms PFS [72] with or without subsequent pose refinement. Note the PFS [72] uses the DeepIM [29] refiner which actually is trained on the synthetic data of the test object while our volume-based refiner is not trained on the test object at all. 2) In comparison with instance-specific estimators [74,62,58] with synthetic training on the test object, Gen6D clearly outperforms all these methods. 3) However, Gen6D performs worse than instance-specific estimators [42,70,74] with real training. The main reason is the inaccurate estimation of the depth. Since the object is usually very far away from the camera and small scale difference (1-2 pixels) will result in a huge offset in the depth direction. Without training on the object, Gen6D cannot perceive such subtle scale changes, which results in worse performance. 4) With ground-truth bounding box, Gen6D achieves comparable results as the instance-specific estimators [42,70,74] with real training because such ground-truth bounding boxes provide correct depths.

4.5 Results on MOPED [41]

On the MOPED dataset, we compare Gen6D with Latent-Fusion [41] and PVNet [42]. Latent-Fusion [41] is also a generalizable pose estimator which does not require training on the test object but needs depth and object masks on query images. We use the official codes and the pretrained weights of Latent-Fusion [41]

Table 3. ADD-AUC on the MOPED dataset with threshold 0-10cm. “LF” means Latent-Fusion [41]. “General” means the pose estimator is trained on the specific object or not. “Input” means the required type of query images at test time.

Method	General	Input	Object Name					avg.
			B.Drill	D.Dude	V.Mug	T.Plane	R.Aid	
LF [41]	✓	RGBD	74.11	75.40	38.27	54.95	62.97	61.14
PVNet [42]	✗	RGB	49.49	43.30	67.78	48.61	72.92	56.42
Ours	✓	RGB	64.87	59.23	50.95	69.83	72.03	63.38



Fig. 8. Qualitative results on the LINEMOD [23] dataset. Ground-truth poses are drawn in green while predicted poses of Gen6D are drawn in blue.

for evaluation. For training PVNet [42], we apply the same strategy as used on the GenMOP dataset. Table 3 reports ADD-AUC on the MOPED dataset, which shows that Gen6D outperforms both baselines on average while Gen6D only uses simple RGB inputs and does not require training on the object.

4.6 Analysis

Ablation study on the viewpoint selector. To demonstrate the designs in the viewpoint selector, we conduct ablation studies on the GenMOP dataset and results are shown in Table 4. Without global normalization and reference view transformer, our viewpoint selector already outperforms the baseline image embedding method ObjDesc [69] by a large margin because our selector pixel-wisely compare the query image with reference images to compute similarity scores, which is more robust to clutter backgrounds. Meanwhile, adding the global normalization or the reference view transformer further improves the results because they exchange information between reference images to help the selector choose the relatively most similar reference image.

Analysis on the pose refiner. To demonstrate the advantage of our volume-based refiner on unseen objects over other rendering-and-comparison based refiners [29,74,57], we report results on the GenMOP in Table 4. For the baseline refiner DeepIM [29], we regard the reference image selected by our selector as the rendered image and use DeepIM to match it with the query image to update the pose, i.e. one step refinement. Note further refinement with more steps using DeepIM are infeasible because there is no object model to render a new image on the updated pose. The DeepIM refiner is trained on the same training data

Table 4. Ablations on the GenMOP dataset. “GN” means the global normalization used in view selector. “RVT” means the reference view transformer. “+ DeepIM Ref.” means using the refiner DeepIM [29] to refine the pose for one step. “+Volume Ref.” means refinement with our volume-base refiner for one step.

Metrics	Method	Object Name					avg.
		Chair	PlugEN	Piggy	Scissors	TFormer	
ADD-0.1d	ObjDesc [69]	3.50	5.14	14.07	1.25	7.54	8.55
	w/o GN and RVT	8.50	13.08	36.18	14.66	1.98	14.88
	w/o RVT	14.50	10.75	36.18	14.22	11.51	17.43
	Full selector	14.00	7.48	39.70	16.81	11.51	17.90
	+ DeepIM Ref.	12.50	6.54	29.15	18.10	31.35	19.53
	+ Volume Ref.	50.50	9.81	55.28	24.57	52.78	38.59
Prj-5	ObjDesc [69]	4.00	10.75	4.52	18.53	8.33	9.23
	w/o GN and RVT	7.00	40.19	20.60	28.88	54.76	30.28
	w/o RVT	16.00	46.73	31.66	24.57	55.16	34.82
	Full selector	11.50	40.65	33.17	34.05	64.29	36.73
	+ DeepIM Ref.	4.50	5.23	18.50	61.64	73.81	42.16
	+ Volume Ref.	44.00	71.03	92.96	84.48	95.24	77.54

as our volume-based refiner. The results show that our volume-based refiner has better generalization ability on unseen objects than DeepIM.

More analysis. We provide more analysis about reference image number, comparison with finetuned DeepIM [29], refinement iterations, ablation on training data and symmetric objects in the supplementary material.

Limitations. The generalization ability of Gen6D mainly comes from matching image patterns in the viewpoint selection and the pose refinement. Thus, Gen6D requires enough diverse training data to learn general image matching for accurate pose estimation and it performs worse with limited training data. Meanwhile, Gen6D is not specially designed to handle occlusions and the performance may degenerate when severe occlusions exist.

Running time. To process an image of size 540×960 , Gen6D estimator costs ~ 0.64 second in total on a 2080Ti GPU, in which the object detector costs ~ 0.1 second, the viewpoint selector costs ~ 0.04 second and the refiner with 3 times refinement costs ~ 0.5 second.

5 Conclusion

In this paper, we propose an easy-to-use 6-DoF pose estimator Gen6D for unseen objects. To predict poses for unseen objects, Gen6D does not require the object model but only needs some posed images of the object to predict its pose in arbitrary environments. In Gen6D, we design a novel viewpoint selector and a novel volume-based pose refiner. Experiments demonstrate the superior performance of Gen6D estimator in predicting poses for unseen objects in the model-free setting.

References

1. Ammirato, P., Fu, C.Y., Shvets, M., Kosecka, J., Berg, A.C.: Target driven instance detection. arXiv preprint arXiv:1803.04610 (2018)
2. Balntas, V., Doumanoglou, A., Sahin, C., Sock, J., Kouskouridas, R., Kim, T.K.: Pose guided rgbd feature learning for 3d object pose estimation. In: ICCV (2017)
3. Banani, M.E., Corso, J.J., Fouhey, D.F.: Novel object viewpoint estimation through reconstruction alignment. In: CVPR (2020)
4. Busam, B., Jung, H.J., Navab, N.: I like to move it: 6d pose estimation as an action decision process. arXiv preprint arXiv:2009.12678 (2020)
5. Cai, D., Heikkilä, J., Rahtu, E.: Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In: CVPR (2022)
6. Cai, M., Reid, I.: Reconstruct locally, localize globally: A model free method for object pose estimation. In: CVPR (2020)
7. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
8. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6d object pose and size estimation. In: CVPR (2020)
9. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: ICCV (2021)
10. Chen, W., Jia, X., Chang, H.J., Duan, J., Shen, L., Leonardis, A.: Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In: CVPR (2021)
11. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category level object pose estimation via neural analysis-by-synthesis. In: ECCV (2020)
12. Dani, M., Narain, K., Hebbalaguppe, R.: 3dposelite: A compact 3d pose estimation using node embeddings. In: WACV (2021)
13. Deng, X., Geng, J., Bretl, T., Xiang, Y., Fox, D.: icaps: Iterative category-level object pose and shape estimation. IEEE Robotics and Automation Letters (2022)
14. Di, Y., Manhardt, F., Wang, G., Ji, X., Navab, N., Tombari, F.: So-pose: Exploiting self-occlusion for direct 6d pose estimation. In: CVPR (2021)
15. Di, Y., Zhang, R., Lou, Z., Manhardt, F., Ji, X., Navab, N., Tombari, F.: Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. arXiv preprint arXiv:2203.07918 (2022)
16. Goodwin, W., Vaze, S., Havoutis, I., Posner, I.: Zero-shot category-level object pose estimation. arXiv preprint arXiv:2204.03635 (2022)
17. Gou, M., Pan, H., Fang, H.S., Liu, Z., Lu, C., Tan, P.: Unseen object 6d pose estimation: A benchmark and baselines. arXiv preprint arXiv:2206.11808 (2022)
18. Grabner, A., Wang, Y., Zhang, P., Guo, P., Xiao, T., Vajda, P., Roth, P.M., Lepetit, V.: Geometric correspondence fields: Learned differentiable rendering for 3d pose refinement in the wild. In: ECCV (2020)
19. Gu, Q., Okorn, B., Held, D.: Ossid: Online self-supervised instance detection by (and for) pose estimation. IEEE Robotics and Automation Letters (2022)
20. He, Y., Wang, Y., Fan, H., Sun, J., Chen, Q.: Fs6d: Few-shot 6d pose estimation of novel objects. In: CVPR (2022)
21. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient response maps for real-time detection of textureless objects. T-PAMI **34**(5), 876–888 (2011)

22. Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: ICCV (2011)
23. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: ACCV (2012)
24. Hodan, T., Barath, D., Matas, J.: Epos: Estimating 6d pose of objects with symmetries. In: CVPR (2020)
25. Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., et al.: Bop: Benchmark for 6d object pose estimation. In: ECCV (2018)
26. Hodaň, T., Sundermeyer, M., Drost, B., Labb  , Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: Bop challenge 2020 on 6d object localization. In: ECCV (2020)
27. Hu, Y., Fua, P., Wang, W., Salzmann, M.: Single-stage 6d object pose estimation. In: CVPR (2020)
28. Labb  , Y., Carpentier, J., Aubry, M., Sivic, J.: Cosopose: Consistent multi-view multi-object 6d pose estimation. In: ECCV (2020)
29. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: Deepim: Deep iterative matching for 6d pose estimation. In: ECCV (2018)
30. Lin, J., Li, H., Chen, K., Lu, J., Jia, K.: Sparse steerable convolutions: An efficient learning of se (3)-equivariant features for estimation and tracking of object poses in 3d space. NeurIPS (2021)
31. 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: ICCV (2021)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll  r, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
33. Liu, X., Iwase, S., Kitani, K.M.: Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In: CVPR (2021)
34. Liu, X., Jonschkowski, R., Angelova, A., Konolige, K.: Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In: CVPR (2020)
35. Mercier, J.P., Garon, M., Giguere, P., Lalonde, J.F.: Deep template-based object instance detection. In: WACV (2021)
36. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
37. Nguyen, V.N., Hu, Y., Xiao, Y., Salzmann, M., Lepetit, V.: Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions. In: CVPR (2022)
38. Okorn, B., Gu, Q., Hebert, M., Held, D.: Zephyr: Zero-shot pose hypothesis rating. In: ICRA (2021)
39. Osokin, A., Sumin, D., Lomakin, V.: Os2d: One-stage one-shot object detection by matching anchor features. In: ECCV (2020)
40. Park, J., Cho, N.I.: Dprost: 6-dof object pose estimation using space carving and dynamic projective spatial transformer. arXiv preprint arXiv:2112.08775 (2021)
41. Park, K., Mousavian, A., Xiang, Y., Fox, D.: Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In: CVPR (2020)
42. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6-dof pose estimation. In: CVPR (2019)

43. Pitteri, G., Bugeau, A., Ilic, S., Lepetit, V.: 3d object detection and pose estimation of unseen objects in color images with local surface embeddings. In: ACCV (2020)
44. Pitteri, G., Ilic, S., Lepetit, V.: Cornet: generic 3d corners for 6d pose estimation of new objects without retraining. In: ICCVW (2019)
45. Pitteri, G., Ramamonjisoa, M., Ilic, S., Lepetit, V.: On object symmetries and 6d pose estimation from images. In: 3DV (2019)
46. Poniatkin, G., Labb  , Y., Russell, B., Aubry, M., Sivic, J.: Focal length and object pose estimation via render and compare. In: CVPR (2022)
47. Rad, M., Lepetit, V.: Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: CVPR (2017)
48. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: CVPR (2021)
49. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR (2020)
50. Sch  nberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
51. Shugurov, I., Li, F., Busam, B., Ilic, S.: Osop: A multi-stage one shot object pose estimation framework. In: CVPR (2022)
52. Simeonov, A., Du, Y., Tagliasacchi, A., Tenenbaum, J.B., Rodriguez, A., Agrawal, P., Sitzmann, V.: Neural descriptor fields: Se (3)-equivariant object representations for manipulation. arXiv preprint arXiv:2112.05124 (2021)
53. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
54. Song, C., Song, J., Huang, Q.: Hybridpose: 6d object pose estimation under hybrid representations. In: CVPR (2020)
55. Su, Y., Saleh, M., Fetzer, T., Rambach, J., Navab, N., Busam, B., Stricker, D., Tombari, F.: Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In: CVPR (2022)
56. Sun, J., Wang, Z., Zhang, S., He, X., Zhao, H., Zhang, G., Zhou, X.: OnePose: One-shot object pose estimation without CAD models. CVPR (2022)
57. Sundermeyer, M., Durner, M., Puang, E.Y., Marton, Z.C., Vaskevicius, N., Arras, K.O., Triebel, R.: Multi-path learning for object pose estimation across domains. In: CVPR (2020)
58. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. In: ECCV (2018)
59. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6d object pose prediction. In: CVPR (2018)
60. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: ECCV (2020)
61. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
62. Wang, G., Manhardt, F., Shao, J., Ji, X., Navab, N., Tombari, F.: Self6d: Self-supervised monocular 6d object pose estimation. In: ECCV (2020)
63. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: CVPR (2021)
64. 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: CVPR (2019)

65. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: CVPR (2021)
66. Wen, B., Bekris, K.: Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In: IROS (2021)
67. Wen, Y., Li, X., Pan, H., Yang, L., Wang, Z., Komura, T., Wang, W.: Disentangled implicit shape and pose learning for scalable 6d pose estimation. arXiv preprint arXiv:2107.12549 (2021)
68. Wen, Y., Pan, H., Yang, L., Wang, W.: Edge enhanced implicit orientation learning with geometric prior for 6d pose estimation. In: IROS (2020)
69. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: CVPR (2015)
70. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Robotics: Science and Systems (2018)
71. Xiao, Y., Marlet, R.: Few-shot object detection and viewpoint estimation for objects in the wild. In: ECCV (2020)
72. Xiao, Y., Qiu, X., Langlois, P.A., Aubry, M., Marlet, R.: Pose from shape: Deep pose estimation for arbitrary 3d objects. In: BMVC (2019)
73. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: inferf: Inverting neural radiance fields for pose estimation. In: IROS (2021)
74. Zakharov, S., Shugurov, I., Ilic, S.: Dpod: 6d pose object detector and refiner. In: ICCV (2019)
75. Zhao, C., Hu, Y., Salzmann, M.: Fusing local similarities for retrieval-based 3d orientation estimation of unseen objects. arXiv preprint arXiv:2203.08472 (2022)

A Overview

This supplementary material has the following contents.

1. Sec. B contains implementation details of the data normalization (Sec. B.1), the object detector (Sec. B.2), the viewpoint selector (Sec. B.3), the pose refiner (Sec. B.4) and the inference process.
2. Sec. C includes experiment details about the training set, the GenMOP dataset and the reference/query splits.
3. Sec. D provides more qualitative results on the LINEMOD [23] dataset, the MOPED [41] dataset and the GenMOP dataset.
4. Sec. E contains additional analysis on comparison with finetuned DeepIM [29], reference image number, unevenly distributed reference images, refinement iterations, ablations on training data, symmetric objects and imperfect view selection.

B Implementation detail

B.1 Data normalization

Object size and center. On the GenMOP dataset, the object size and object center are roughly estimated from reconstructed points of COLMAP [50]. On the LINEMOD [23] dataset and the MOPED [41] dataset, we directly use the provided 3D model to compute the object size and the object center. Note we do not need an exact object size and an object center.

Normalization of object coordinate system. Given the object size d and the object center c , we normalize the object coordinate system by

$$x_{norm} = (x - c)/d \times 2, \quad (1)$$

where x_{norm} is the normalized coordinate and x is the original coordinate.

Normalization of reference images. Since the raw reference images may not look at the object, we normalize these reference images by warping the image with a homography and changing the intrinsic correspondingly, as shown in Fig. 9. The normalized reference images all look at the object and they enclose the projection of the unit sphere at the origin. Note we can do this normalization because we have already normalized the object coordinate and we know the poses of reference images.

B.2 Detector

Networks architecture. The detailed architecture is shown in Fig. 10. The input reference images are resized to 120×120 . We actually use three levels of feature maps from the query image and the corresponding three levels of Conv kernels from reference images. We conduct convolutions on the feature map at each level with the corresponding conv kernel at the same level. $N_s = 5$ scales are used and we downsample the input query image with $\sqrt{2}$ as the factor.

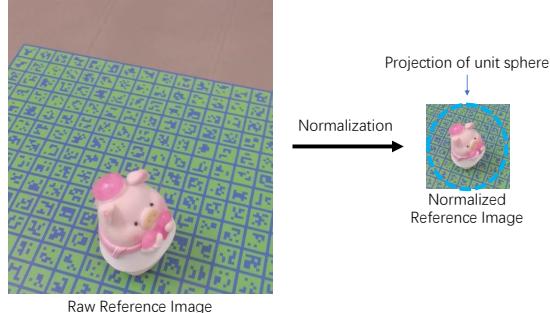


Fig. 9. Raw reference images are normalized to look at the object. The projections of the unit sphere are enclosed by the reference images.

Loss. Given the heat map H , we regard all values of the heat map as logits and use a binary classification loss to supervise the heat map prediction. Here, we first project the object center to the image using the ground-truth object pose. Then, a pixel on the heat map is assumed to be correct if it is within 1.5-pixel distance from the object center projection. Otherwise, it is incorrect.

$$\ell_{heat} = \sum_p -\mathbb{1}(\|p - c_{prj}\|_2 < 1.5) \log \sigma(H(p)) - (1 - \mathbb{1}(\|p - c_{prj}\|_2 < 1.5)) \log(1 - \sigma(H(p))), \quad (2)$$

where $\mathbb{1}$ is an indicator function, $c_{prj} \in \mathbb{R}^2$ is the 2D projection of the object center, p is a pixel on the heat map, σ is the Sigmoid function, $H(p)$ means the heat value on the pixel p . To supervise the scale map prediction, we apply a scale loss to minimize the L2 distance between the ground-truth scale and the predicted scale in the log space. Note we only apply such scale loss on the pixels within 1.5-pixel distance away from the projected object center.

$$\ell_{scale} = \sum_p \mathbb{1}(\|p - c_{2d}\|_2 < 1.5) \|\log s_{gt} - S(p)\|_2^2, \quad (3)$$

where s_{gt} is the ground-truth scale, S is the scale map, $S(p)$ means the scale value at pixel p . s_{gt} can be computed from the distance l between the camera center and the object center by

$$s_{gt} = \frac{2f}{lS_r}, \quad (4)$$

where f is a virtual focal length by changing the principle point to the object center projection c_{prj} , 2 is the diameter of the unit sphere and $S_r = 120$ is the size of reference image.

B.3 Selector

Network architecture. The detailed architecture of the selector is shown in Fig. 11. The input query image is cropped according to the detection results and

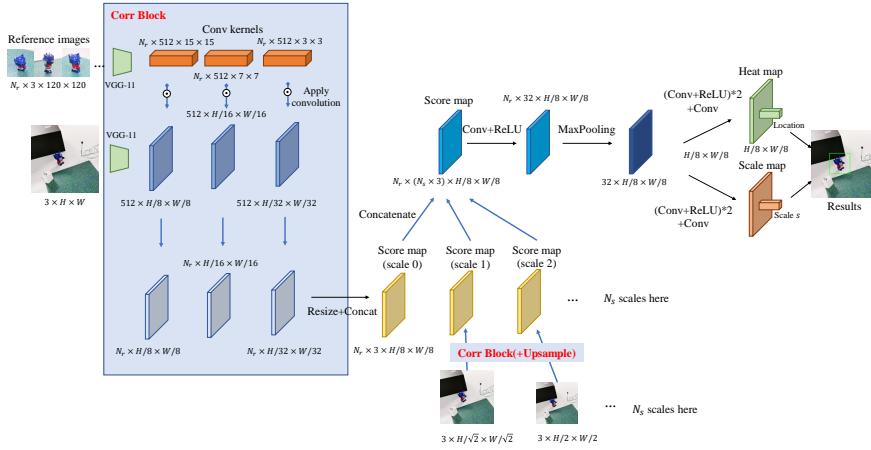


Fig. 10. Detailed pipeline of the detector. All “Conv” layers use 3×3 kernel size.

resized to 128×128 . All reference images are also cropped to the size 128×128 . We also use three levels of feature maps to conduct the pixel-wise product. The viewpoint is represented by a 3-dimensional vector. The in-plane rotation is a scalar of the clock-wise rotation angle. The final similarity score for a reference image is also a scalar. We use $N_a = 5$ rotation angles in $[-\pi/2, \pi/2]$ to rotate every reference image.

Loss. To compute the loss, we first introduce how we compute the ground-truth similarity between two viewpoints. In the object coordinate system, we denote the camera locations of all reference images as $\{u_i \in \mathbb{R}^3 | i = 1, \dots, N_r\}$ where N_r is the number of reference images. Note that the object center is the origin so that $-u_i$ is the vector from the camera to the object center. In training, we know the query camera pose and compute the camera location v of the query image in the object coordinate system. Both u_i and v are normalized by $\tilde{u}_i = u_i / \|u_i\|_2$ and $\tilde{v} = v / \|v\|_2$. Then, the ground-truth viewpoint similarity between i -th reference image and the query image is their dot product $\tilde{u}_i \cdot \tilde{v}$.

We normalize the ground-truth viewpoint similarity to $[0,1]$, which is denoted as \tilde{s}_j . Then, the similarity loss is

$$\ell_{sim} = \sum_j BCE(s_j, \tilde{s}_j) \quad (5)$$

where BCE means binary cross entropy loss and we force the predicted score to be consistent with the ground-truth viewpoint similarity.

To train the in-plane rotation angle prediction, we adopt a L2 loss,

$$\ell_{angle} = \|\alpha_j - \alpha_{gt}\|_2^2, \quad (6)$$

where α_j is the predicted in-plane angle on the ground-truth nearest reference image and α_{gt} is the ground-truth in-plane rotation between the query image and the reference image.

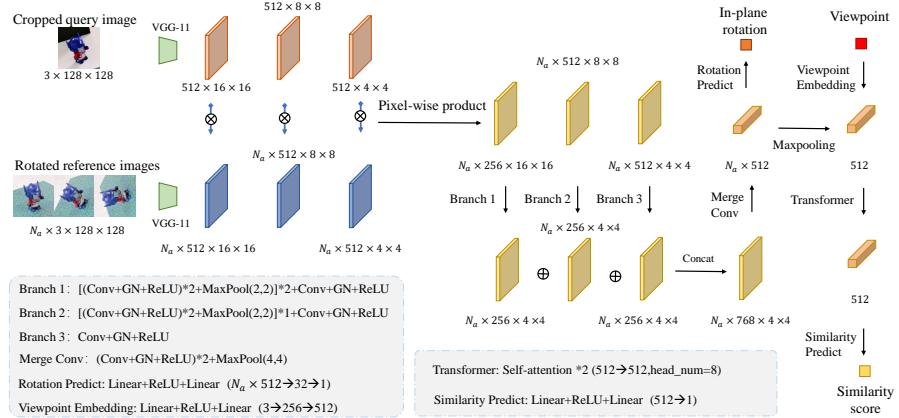


Fig. 11. Detailed pipeline of the selector. We only draw one reference image in the figure for clear visualization. “Conv” means a convolution layer with 3×3 kernels. “GN” means the global normalization that normalize the feature maps using the mean and variance computed from feature maps of all reference images. “Transformer” is applied among all feature vectors of reference images.

B.4 Refiner

Network architecture. The detailed architecture for the refiner is shown in Fig. 12. In the figure, we show the 2D CNN and the 3D CNN separately. In the 2D CNN, both reference images and query images are resized to 128×128 and the final feature map for unprojection is extracted from three levels of feature maps. In the 3D CNN, we first embed reference mean, reference variance and query features separately. Then, they are concatenated and processed by several 3D convolution layers. Final 3D pose residuals are regressed by linear layers from the flatten feature vector of the 3D volume. We use $N_n = 6$ neighboring reference images in refinement.

Loss. To train the refiner, we first sample 32^3 voxel points in the unit cube in the object coordinate system. Then, these points are transformed to the input camera coordinate system by the input pose. The loss is the distance between the sample points transformed by the ground-truth similarity transformation and the sample points transformed by the predicted similarity transformation.

$$\ell_{ref} = \sum_k^{32^3} \|s_{pr} R_{pr}(p_k + t'_{pr}) - s_{gt} R_{gt}(p_k + t'_{gt})\|_2, \quad (7)$$

where $p_k \in \mathbb{R}^3$ is the coordinate of a sample point in the input camera coordinate, s_{pr} and s_{gt} are predicted scale and ground-truth scale respectively, R_{pr} and R_{gt} are the predicted rotation and the ground-truth rotation respectively, t'_{pr} and t'_{gt} are the predicted 2D translation and the ground-with 2D translation respectively. We set the third element of t'_{pr} and t'_{gt} to 0 to form a 3D-translation. Note the

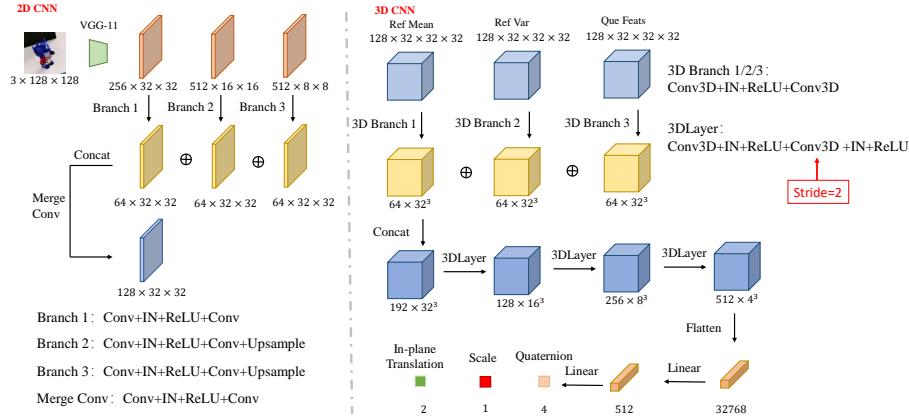


Fig. 12. Detailed architecture of the refiner. “IN” means instance normalization.

predicted and ground-truth similarity transformations are used in transforming the object (red circle in Fig. 13) with the input pose to the object with the ground-truth pose (green dotted circle in Fig. 13).

Similarity approximation. We discuss how to convert the similarity transformation (red circle to dotted green circle in Fig. 13) to the rigid transformation (red circle to solid green circle in Fig. 13). Denoting the similarity transformation as (R, s, t) , our target is to compute the rigid transformation (R, t') . Actually, we only need to convert (s, t) to t' because rotation R is the same in two transformations. Assuming the red circle center is (c_x, c_y, c_z) in the input camera coordinate, then the center of the green dotted circle is $(c_x + t_x, c_y + t_y, z)$ where t_x and t_y are the first and the second element of t respectively. The depth affects the scale, so the center of the green solid circle is $((c_x + t_x)/s, (c_y + t_y)/s, z/s)$. The final rigid translation is $t' = ((c_x + t_x)/s - c_x, (c_y + t_y)/s - c_y, z/s - z)$.

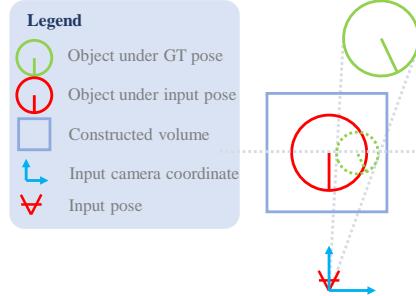


Fig. 13. Diagram to illustrate the transformation. Note all transformations are applied on coordinates in the input camera coordinate system.

B.5 Inference details

In inference, we apply pose refiner iteratively for 3 times. For the input data, not all reference images are used in the object detector and the viewpoint selector.

Instead, we use farthest point sampling to sample 32 and 64 images for the detector and the selector respectively.

C Experimental setting

Training set for Gen6D. The training objects on the LINEMOD [23] dataset are “ape”, “can”, “holepuncher”, “iron” and “phone”. There are ~ 1200 images per object and we randomly use them as reference images and query images. The training objects on the GenMOP dataset are “cup”, “knife”, “love”, “plugCN” and “Miffey”, each of which contains ~ 200 reference images and ~ 200 query images. On every object from ShapeNet [7], we render 1024 images. On every object from Google Scanned Objects, we use 512 rendered images from IBR-Net [65] for training. To train the 2D object detector, we additionally use the CO3D [48] for training.

Reference/query split on training sets. On the LINEMOD [23] dataset, the ShapeNet dataset and the Google Scanned objects dataset, we randomly select 128 images by farthest point sampling on camera locations as reference images while the other images are selected as query images. On the GenMOP dataset, we use images from one video as reference images while images from the other video as query images.

Reference/query split on test sets. On the LINEMOD [23] dataset, we use the training set of previous instance-specific estimators [59,42] as reference images and the other images are selected as query images. On the GenMOP dataset, we also use images from one video sequence as reference images and images from the other video sequence as query images. On the MOPED dataset [41], we use the provided reference video sequences as reference images and the other video sequences as query images. We list the number of reference images and query images in Table 5. Note that reference images are used in inference of Gen6D but not in training the Gen6D estimator while instance-specific estimators like PVNet [42] actually use these reference images to train their models.

GenMOP. On every object of the GenMOP dataset, we collect two 1-minute videos in different environments by a cell-phone. On each video, we sample 1 image per 10 frames, which results in ~ 200 images for every video. On each video sequence, we apply COLMAP [50] to recover the extrinsics and intrinsics of all cameras. Then, in every sequence, we manually select two images, label 4 keypoints on the selected images, and use triangulation to compute the 3D points. For two difference sequences of the same object, we compute the transformation from the triangulated 3D points to align them.

D Qualitative Results

More qualitative results on the LINEMOD dataset, the MOPED [41] dataset and the GenMOP dataset are shown in Fig. 18, Fig. 19 and Fig. 20 respectively.

Table 5. Numbers of reference images and test query images on different datasets.

	GenMOP				
	Chair	PlugEN	Piggy	Scissors	TFormer
Reference	212	199	227	202	206
Query	200	214	199	232	252
LINEMOD [23]					
	cat	duck	bvise	cam	driller
Reference	177	189	183	182	179
Query	1002	1065	1032	1020	1009
MOPED [41]					
	B.Drill	D.Dude	V.Mug	T.Plane	R.Aid
Reference	355	451	606	662	394
Query	215	297	91	250	58

E More analysis

E.1 Comparison with finetuned DeepIM

Based on the pretrained generalizable DeepIM [29] model in the experiments of the main paper, we further finetune it separately on every test object using the reference images of the object as the training set. The performance of finetuned DeepIM model is shown in Table 6, which shows that finetuning DeepIM brings significant improvements but still underperforms the Gen6D model which does not train on the object.

Table 6. Performance on the GenMOP dataset. “General” means generalizable or not. “DeepIM [29]-Ft” means we finetune DeepIM models on every object’s reference images separately.

Metrics	Method	General	Object Name					avg.
			Chair	PlugEN	Piggy	Scissors	TFormer	
ADD-0.1d	DeepIM [29]	✓	12.50	6.54	29.15	18.10	31.35	19.53
	DeepIM [29]-Ft	✗	65.50	36.92	62.31	19.40	38.49	44.52
	Ours	✓	61.50	19.63	75.38	32.76	62.70	50.39
Prj-5	DeepIM	✓	4.50	52.34	18.50	61.64	73.81	42.16
	DeepIM [29]-Ft	✗	40.00	70.56	82.37	73.28	98.41	72.92
	Ours	✓	55.00	72.90	92.96	93.53	98.81	82.64

E.2 Fewer reference images

To show the performance of Gen6D estimator with less reference images, we reduce the reference images on the GenMOP dataset from 128 to 8 by farthest point sampling (FPS) as shown in Table 7. Sampling 128 images by FPS even slightly improves the performance because FPS makes the view distribute evenly. With 64 reference images, the Gen6D estimator still produces similar results as with all images. When only 16 or 8 reference images retain, the performance reduces reasonably. We further provide detailed Prj-5 of different objects on the GenMOP dataset in Table 8. This show that the suitable view numbers for different objects are different. 16 reference views on “piggy” still produce 91% Prj-5 while 16 views on “chair” will reduce Prj-5 to 34%.

Table 7. Performance of the Gen6D estimator on the GenMOP dataset with different numbers of reference images. Two metrics are averaged among all objects. “All” means using all images (~ 200) from the sequence as reference images.

Metrics	Reference image number					
	All	128	64	32	16	8
ADD-0.1d	50.39	51.30	49.68	39.45	33.56	27.55
Prj-5	82.64	83.28	82.30	81.03	73.51	37.33

Table 8. Prj-5 on GenMOP dataset of Gen6D with different reference image numbers.

# Ref. Img.	Object Name					avg.
	Chair	PlugEN	Piggy	Scissors	TFormer	
All	55.00	72.90	92.96	93.53	98.81	82.64
128	58.00	69.63	96.48	92.67	99.60	83.28
16	34.00	69.63	90.95	79.74	93.25	73.51
8	11.00	61.68	28.14	26.29	59.52	37.33

E.3 Uneven reference image distribution

By default, reference images are preferred to be distributed evenly around the object as shown in Fig. 14 (a), which benefits the viewpoint selection and the refinement using neighboring views. When only a part of viewpoints are available like Fig. 14 (b), Gen6D is not able to accurately predict object poses on images captured from uncovered region. We show the performance of Gen6D using reference images of Fig. 14 (b) in Table 9.

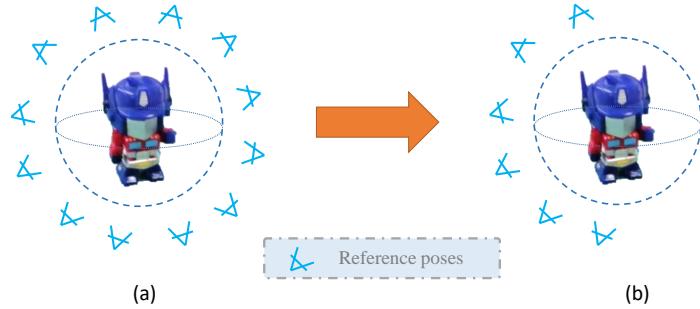


Fig. 14. (a) Reference images are homogeneously-distributed around the object. (b) Reference images only distribute in the Y- half space.

Table 9. Performance of Gen6D on the GenMOP dataset using homogeneously-distributed reference images (“Even”) or only images whose camera centers are in the Y- space (Fig. 14(b)) (“Partial”).

Metrics	Ref. Img.	Object Name					avg.
		Chair	Plug	EN	Piggy	Scissors	
ADD-0.1d	Even	61.50	19.63	75.38	32.76	62.70	50.39
	Partial	43.00	8.53	56.28	5.17	30.95	28.79
Prj-5	Even	55.00	72.90	92.96	93.53	98.81	82.64
	Partial	32.50	7.48	73.37	17.24	47.22	35.56

E.4 Refinement iterations

In Table 10, we show results on the GenMOP dataset with different refinement iterations. The results show that applying 1 refinement iteration already greatly improves performance from 17.90 to 38.59 on the ADD-0.1d metric. Further applying 2 refinement iteration will continuously improve the results while using 3 iterations does not.

Table 10. Results of our Gen6D estimator on the GenMOP dataset. “#Refine” means the number of refinement iterations used to produce the results.

Metrics	#Refine	Object Name					avg.
		Chair	PlugEN	Piggy	Scissors	TFormer	
ADD-0.1d	0	14.00	7.48	39.70	16.81	11.51	17.90
	1	50.50	9.81	55.28	24.57	52.78	38.59
	2	62.00	29.91	80.90	37.07	56.75	53.32
	3	61.50	19.63	75.38	32.76	62.70	50.39
Prj-5	0	11.50	40.65	33.17	34.05	64.29	36.73
	1	44.00	71.03	92.96	84.48	95.24	77.54
	2	51.50	72.90	94.97	94.40	99.60	82.67
	3	55.00	72.90	92.96	93.53	98.81	82.64

E.5 Ablations on training data

To show the effects of different training data, we show the performance of Gen6D using different training sets in Table 11. Training only on synthetic datasets suffers from the domain gap between the real data and synthetic data. Using real data for training greatly improves the results. LINEMOD [23] brings more obvious improvements than adding GenMOP. The main reason is that LINEMOD has more images (~ 1200) on every object while GenMOP only has ~ 200 reference images and ~ 200 query images for training.

E.6 Symmetric objects

Gen6D is able to predict poses for symmetric objects. We evaluate Gen6D on two unseen symmetric objects from LINEMOD [23], i.e. “glue” and “eggbox”. Qualitative results are shown in Fig. 15 and quantitative results are shown in Table 12. The results show that our method is able to achieve reasonable performance on symmetric objects. The main reason is that Gen6D is based on matching query images with reference images. Though symmetry makes multiple feasible poses for a query image, the selector and refiner of Gen6D are able to find reference images near to one of such feasible poses.

E.7 Imperfect viewpoint selection

Table 11. Performance of Gen6D on the GenMOP and LINEMOD [23] datasets with different training sets. “Syn.” means the ShapeNet [7] and Google Scanned Objects [65] datasets. “GMP” means 5 training objects from the GenMOP dataset. “LM” means 5 training objects from the LINEMOD [23] dataset. All test objects are not in the training set.

Metrics	Trainset	Object Name					avg.
		Chair	PlugEN	Piggy	Scissors	TFormer	
ADD-0.1d	Syn.+GMP+LM	61.50	19.63	75.38	32.76	62.70	50.39
	Syn.+GMP	39.00	11.21	71.86	37.93	39.68	39.94
	Syn.	30.00	21.96	61.81	24.57	30.16	33.70
Prj-5	Syn.+GMP+LM	55.00	72.90	92.96	93.53	98.81	82.64
	Syn.+GMP	34.50	77.57	92.46	91.38	89.68	77.18
	Syn.	10.05	77.57	68.84	74.57	98.41	65.98
Metrics	Trainset	cat	duck	bvise	cam	driller	avg.
ADD-0.1d	Syn.+GMP+LM	60.68	40.47	77.03	66.67	67.39	62.45
	Syn.+GMP	40.92	16.24	62.11	45.59	48.76	42.72
	Syn.	31.04	11.64	61.63	35.39	54.61	38.86



Fig. 15. Qualitative results on symmetric objects.

Table 12. Results on symmetric objects of the LINEMOD [23] dataset. For the ADD metric, we report “ADD-S-0.1d” which computes the nearest distance between the object points transformed by ground-truth pose and the estimated pose. Note PVNet [42] is trained on the specific test object with both synthetic and real images while our Gen6D is not trained on the test object.

Name	Prj-5		ADD-S-0.1d	
	PVNet [42]	Ours	PVNet [42]	Ours
Eggbox	99.34	97.84	99.15	98.40
Glue	98.45	96.24	95.66	87.16

As discussed in the Section 1 in the main paper, it would be challenging for selector to select the most similar reference image when there is no reference image with an exactly same viewpoint as the query image. To show this, we show Fig. 16, where the x-axis shows the viewpoint difference between the ground-truth nearest reference image and the query image while the y-axis shows the viewpoint difference between the selected reference image and the query image. The viewpoint difference is computed as the angle between the query viewpoint and the reference viewpoint by $\arccos \tilde{u} \cdot \tilde{v}$. With the increase of viewpoint difference between the ground-truth reference image and the query image, the viewpoint difference between the selected reference image and the query image is also increasing. However, the proposed selector is able to select more accurate reference image than the baseline ObjDesc [69]. Figure 17 also shows some examples.

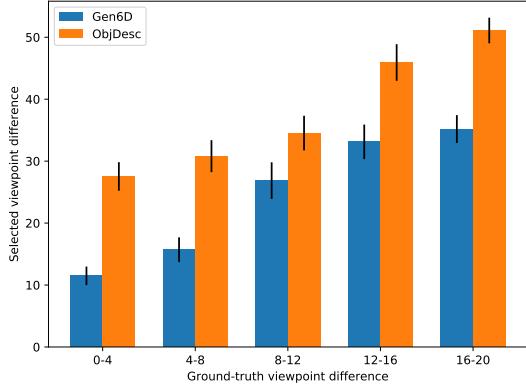


Fig. 16. Viewpoint difference between the selected reference image and the query image (y-axis); Viewpoint difference between the ground-truth reference image and the query image (x-axis).

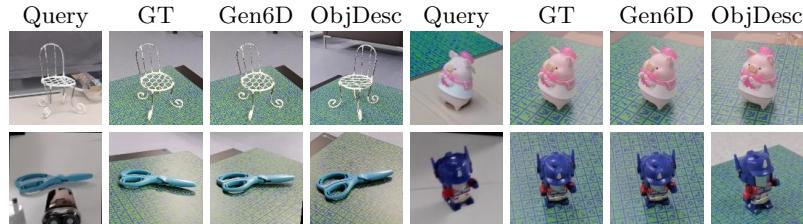


Fig. 17. The input query image, the ground-truth reference image with nearest viewpoint, the reference images selected by Gen6D and ObjDesc [69].



Fig. 18. Additional qualitative results on the LINEMOD [23] dataset. Ground-truth poses are drawn in green while predicted poses are drawn in blue.

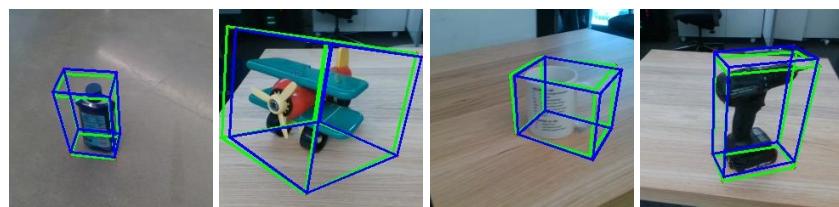


Fig. 19. Qualitative results on the MOPED [41] dataset. Ground-truth poses are drawn in green while predicted poses are drawn in blue.

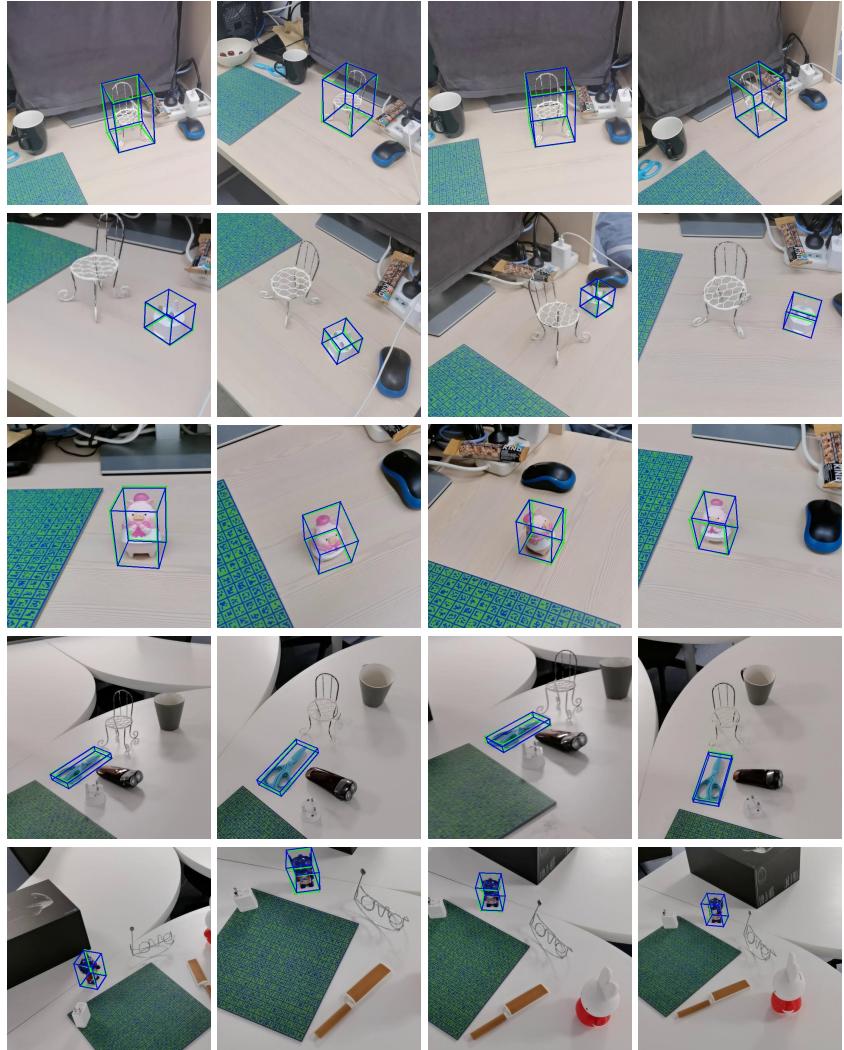


Fig. 20. More qualitative results of Gen6D on the GenMOP dataset.