# FS6D: Few-Shot 6D Pose Estimation of Novel Objects

Yisheng He[1] Yao Wang[2] Haoqiang Fan[2] Jian Sun[2] Qifeng Chen[1]
[1]Hong Kong University of Science and Technology [2]Megvii Technology

## Abstract

*6D object pose estimation networks are limited in their capability to scale to large numbers of object instances due to the close-set assumption and their reliance on high-fidelity object CAD models. In this work, we study a new open set problem; the few-shot 6D object poses estimation: estimating the 6D pose of an unknown object by a few support views without extra training. To tackle the problem, we point out the importance of fully exploring the appearance and geometric relationship between the given support views and query scene patches and propose a dense prototypes matching framework by extracting and matching dense RGBD prototypes with transformers. Moreover, we show that the priors from diverse appearances and shapes are crucial to the generalization capability under the problem setting and thus propose a large-scale RGBD photorealistic dataset (ShapeNet6D) for network pre-training. A simple and effective online texture blending approach is also introduced to eliminate the domain gap from the synthesis dataset, which enriches appearance diversity at a low cost. Finally, we discuss possible solutions to this problem and establish benchmarks on popular datasets to facilitate future research. [project page]*

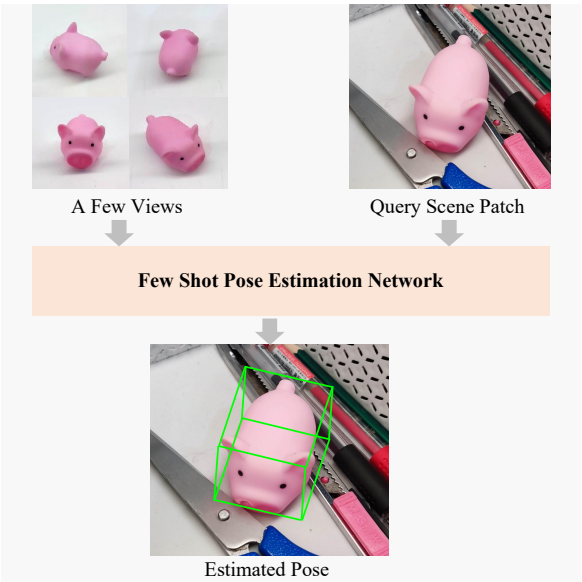


Figure 1. **The few-shot 6D pose estimation problem**. Given a few RGBD views of a *novel* objects with pose labels. The few-shot pose estimation network aims to estimate 6D pose of that object in a novel query scene without extra training. No precise CAD models are required as well.

## 1. Introduction

6D object pose estimation aims to predict a rigid transformation from the object coordinate system to the camera coordinate system, which benefits various applications, including robotic manipulation, augmented reality, autonomous driving, etc. The explosive development of deep learning has brought significant improvement to this problem. With recent works [15, 16] reaching nearly 99% recall accuracy on existing benchmarks [18, 22, 58], one may get the impression that the 6D object pose problem has been solved, which is not the case. We argue that the current problem has been simplified with strict restrictions. They are under the *close-set* assumption that the training and testing data are drawn from the same object space, which, however, does not adhere to the real dynamic worlds. Moreover, extravagant high-fidelity CAD models and large-scale datasets are required for training to obtain good performance on new objects under the current instance-level pose estimation setting.

The recently proposed category-level pose estimation task [55] loosens the restriction with generalizability to novel objects within the same categories. However, it is still limited in the *close-set* assumption of predefined categories. Instead, in this work, we study a new *open-set* problem, the few-shot 6D object pose estimation: estimating 6D pose of unknown objects by only a few views of the objects without extra training. As shown in Figure 1, in our setting, only a few labeled RGBD images of novel objects are provided, and no high-fidelity CAD models are required. The goal of the problem is to bridge the capability gap between machine learning algorithms and flexible human visual systems that can locate and estimate the pose of a novel object given only several views of it. Besides, it has a wide range of

real-world applications in robotic vision systems, i.e., fast registration of novel objects for robotic manipulation and home robots.

Under the observation that human beings utilize both appearance and geometric information to match and locate a new object, we propose a dense RGBD prototypes matching framework to tackle the problem. Specifically, transformers are utilized to fully explore the semantic and geometric relationship between the query scene patch and the support views of novel objects. Moreover, we point out that large-scale datasets' diverse shape and appearance priors are essential to empower networks to generalize on novel objects. Therefore, we introduce a large-scale photorealistic dataset (ShapeNet6D) with diverse shapes and appearances for prior learning. To our knowledge, ours (800K images of 12K objects) is the largest and most diverse dataset for 6D pose algorithms. To bridge the domain gap between rendered RGB images and real-world scenes, we introduce a simple and effective *online* texture blending augmentation, which further enriches the appearance diversity and facilitates network performance at a low cost.

To summarize, the contributions of this work are:

- We introduce a challenging *open-set* problem, the few-shot 6D object pose estimation, and establish a benchmark to study it.
- We formulate the problem by dense RGBD prototypes matching and introduce FS6D-DPM, which fully leverage appearance and geometric information to tackle the problem.
- Datasets: We introduce ShapeNet6D, a large-scale photorealistic dataset with diverse shapes and appearances for prior learning of few-shot 6D pose estimation algorithms. We also introduce an online texture blending augmentation to obtain scenes of texture-rich objects without domain gaps at a low cost.

## 2. Related Work

### 2.1. 6D Object Pose Estimation in Close-Set Setting

Instance-level pose estimation retrieves pose parameters of *known* object instances. Matching based approaches [13, 17, 24, 47, 59] requires precise CAD models to render thousands of templates and establish hand-craft or learned codebook for matching. Learning-based approaches includes direct pose regression [53,58], dense correspondence exploration [29] and recent keypoint-based approaches [15, 16, 38], which improve the performance by large margins. Despite compelling results, these approaches can only deal with scenarios of *known objects* with high-fidelity CAD models. Instead, the recent category-level pose estimation [55] improves the generalizability by estimating unseen object instances within the *known categories*. Normalized Object Coordinate Space (NOCS) [55] or shape deformation based [14, 48] approaches are proposed. However, both traditional instance- and category-level pose estimation problems are under the *close-set* setting, assuming that the training and testing data are within the same predefined instance or category spaces. While such *close-set* setting does not adhere to the real dynamic world, we instead define a new *open-set* problem, the few-shot 6D pose estimation. Algorithms developed in our *open-set* setting can be flexibly applied to unknown objects without extra training with only a few labeled RGBD images, no matter they are within the trained categories or not.

### 2.2. Possible Few-Shot Pose Estimation Solutions

**Local Image Feature Matching.** Local feature matching can establish the correspondence between two images for the few-shot pose estimation problem. Existing methods can be categorized into detector-based [33–35, 42, 43] and detector-free [28, 31, 41, 46]. While these algorithms only leverage the grey-scale images, the performance drops on texture-less objects. Instead, we fully leverage both the appearance and the geometric information and generalize well in more scenarios.

**Point Cloud Registration.** One line of point cloud registration algorithms solve the problem by detecting 3D keypoints [1, 27], extracting feature descriptors [7, 8, 12, 20, 40, 40] and estimating the relative transformation. Several end-to-end approaches [57] are also proposed. However, these algorithms heavily rely on fine point clouds and fail on objects that are not captured by depth sensors, i.e., reflective ones. Instead, we fully leverage the complementary information in RGBD images for dense prototypes extraction and matching to retrieve better object pose parameters.

### 2.3. Metric learning on few-shot learning problems

Metric learning techniques have been applied to several few-shot learning problems, including classification [11, 44, 52] and segmentation [10, 32, 49, 61]. The representative prototypical network [44] for classification map the support and query images into a global embedding space and then retrieve the class label of query image based on the support embedding, named prototype. The recent metric learning-based approaches in more challenging segmentation areas utilize similar technique but output per-pixel prediction on the query images by matching per-pixel query features with global average prototypes [10, 49, 64] or part-level prototypes [32, 61]. While sparse support prototypes are enough to solve the above problems, few-shot pose estimation requires more dense correspondence exploration on pixel-level support prototypes and query features, which is more challenging.

# 3. Proposed Method

## 3.1. Problem Formulation

We introduce the problem setting of the few-shot 6D object pose estimation and the derived domain generalization problem.

**The few-shot 6D object pose estimation.** We formulate the new *open-set* task, the few-shot 6D pose estimation as follows. Given $k$ support RGBD patches $P = \{p_1, p_2, ..., p_k\}$ of a novel object with pose parameters as support frames, the inference task is to retrieve the 6D pose parameters of that novel object in the query novel scene image $I$. Compared to current *close-set* setting, the proposed *open-set* one eliminates the reliance on precise CAD models and focuses on the generalizability of trained models on unseen objects. Specifically, once the model is trained, we expect to apply it on novel scenes of novel objects by a few views without extra training. It bridges the gap between machine learning algorithms and flexible human visual systems. Moreover, it enables real-world applications, i.e., fast registration of new objects for robotic manipulation and service home robots.

The generalization requirement of the *open-set* problem also derive another interesting research question to study:

**Domain generalization.** The domain generalization targets to reduce the domain gaps between models trained on the synthesis and real-world data. It has been introduced to the 6D pose estimation field to deal with the lack of data [26, 36, 45, 54]. However, this field is less explored as existing real-world benchmarking datasets [21, 22] for the *close-set* problem has been well established: real-world training data for the object to be estimated is available. While existing datasets are small with limited objects, in our few-shot *open-set* setting, diversity of shape and appearance are crucial to the generalizability of few-shot 6D object pose estimation algorithms. However, capturing and labeling such a large-scale real-world dataset is not practical due to the high cost (money and time). It is crucial to fully leverage the geometry and appearance diversity in our large-scale photorealistic datasets and generalize to the real world. The domain generalization problem is thus an important problem to study for the few-shot 6D pose estimation.

## 3.2. Datasets

The prior learned from large-scale datasets is crucial to the performance and generalizability of few-shot learning algorithms. ImageNet [9], for example, has been widely used for network pre-training in several few-shot learning tasks, i.e., object detection and segmentation. While 2D vision tasks rely more on the semantic prior in RGB images, for the few-shot 6D object pose estimation, both shape and semantic prior are crucial for the generalizability of the network. However, existing datasets [21, 22, 58] for 6D object pose estimation are small and lack diversity in shape and appearance to provide enough prior for the generalization capability. Therefore, we keep their role as real-world benchmark datasets and propose a new large-scale dataset, ShapeNet6D, with diverse shapes and appearances for prior learning.

| Dataset | Modality | $N_{cat}$ | $N_{obj}$ | $N_{img}$ |
|---|---|---|---|---|
| LineMOD [18] | RGBD | - | 15 | 18,273 |
| YCB-V [4] | RGBD | - | 21 | 133,936 |
| TLESS [21] | RGBD | - | 30 | 47,664 |
| NOCS-REAL [55] | RGBD | 6 | 42 | 80,000 |
| NOCS-CAMERA [55] | RGBD | 6 | 1,085 | 300,000 |
| ShapeNet6D | RGBD | **51** | **12,490** | **800,000** |

Table 1. **Statistics of Different Datasets.** ShapeNet6D is diverse in shape and appearance, which is crucial to the generalizability of few-shot 6D pose algorithms. $N_{cat}$: number of category; $N_{obj}$: number of object instance. $N_{img}$: number of images.

### 3.2.1 ShapeNet6D

The proposed ShapeNet6D is a large-scale photorealistic dataset containing RGBD scene images of more than 12K object instances from the ShapeNet [5] repository. Each scene image is labeled with ground truth information for the 6D pose estimation problem, including instance semantic segmentation and pose parameters of each object. As we demonstrate empirically, the diversity of shape and appearance is crucial for the network to generalize. While it is not practical to collect and label such a large-scale, diverse dataset in the real world due to the high cost (time and money), we instead generate photorealistic images by physically-based rendering. Our approach is inspired by the successful application of photorealistic datasets in [22, 62, 63] while improving the diversity of object shape and appearance. Specifically, we utilize the physically-based rendering engine, Blender[1] that simulates the flow of light energy by ray tracing to render realistic scene images. To arrange a scene to render, we first randomly select several objects from ShapeNet, apply random material and texture, and drop them into a box with the PyBullet physics engine integrated into Blender. To enrich the variety of the background, we randomly selected physically-based rendering material from the HDRI Haven[2] and applied them to the wall of the box. Random environment lights are also added to generate diverse lighting conditions. Finally, the RGBD scene image is rendered from a random camera pose, and the ground truth instance semantic segmentation labels and pose parameters of each object are also obtained. Statistics about ShapeNet6D compared to existing 6D pose benchmark datasets are shown in Table 1. ShapeNet6D is

[1] https://www.blender.org
[2] https://hdrihaven.com/hdris

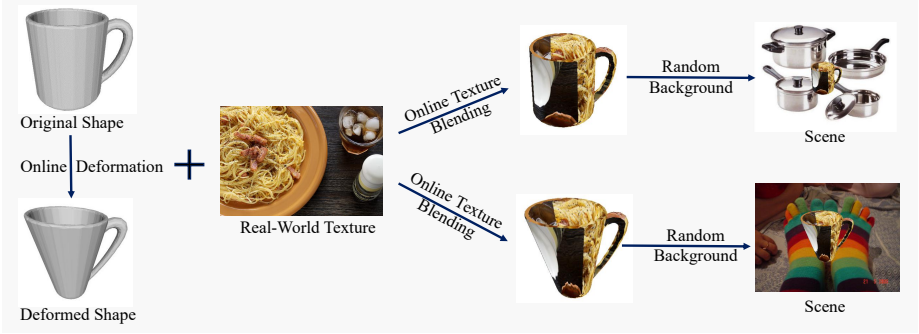


Figure 2. **Online data augmentation.** The *online* texture blending augmentation generates texture by directly blending the real-world image to the object mesh model. No extra artificial simulation is applied, i.e., simulated lighting and the domain of real-world RGB images is preserved. Along with the online deformation augmentation [6], we can obtain data with diverse appearances and shapes at a low cost.

on a larger scale and is more diverse in shape and appearance, which provides better prior to the few-shot pose estimation problem as we showed empirically.

#### 3.2.2 Online texture blending

As one of the crucial clues to solve the few-shot 6D pose estimation problem, the texture field is also essential to the performance of the few-shot 6D object pose estimation. However, it is labor-intensive and time-consuming to generate textures and materials for objects that can be rendered to be photorealistic. The rendered RGB images tend to have more significant domain gaps between the real world as well. Moreover, to produce photorealistic images, time- and computation-consuming techniques like ray tracing are required. Therefore, the images should be pre-processed offline and stored before network training, which costs a lot of storage space for a large-scale dataset. On the other hand, real-world RGB images captured from various cameras are easy to access, i.e., ImageNet [9], and MS-COCO [30]. It motivates us to leverage efficient texture wrapping techniques to generate scenes of objects with rich real-world texture to serve as *online* data argumentation. Specifically, the mesh is first unwrapped to obtain a UV map. For each triangle, we get the UV coordinate of each vertex and then utilize it to determine the UV coordinate of each pixel by linear interpolation during rasterization. The UV coordinate is then applied to lookup the color value from a texture map randomly sampled from the real-world ImageNet [9], and MS-COCO [30]. Previous works [22,37] render images with artificial simulations, i.e., Beckmann model [2] , which change the domain and cause domain gaps. Instead, we applied no simulation, so the composite images are kept in the real domain, i.e., the lighting condition, sensors noise of the real-world images are preserved. Moreover, such a simple blending strategy can be implemented fast to serve *online*. Moreover, we can combine it with *online* shape deformation [6] to produce data with rich appearance and shape diversity for training, as shown in Figure 2.

### 3.3. FS6D-DPM

#### 3.3.1 Preliminaries

**Prototypes-based few-shot learning.** We first briefly introduce the prototypes-based algorithms for few-shot learning. It has been successfully applied to various few-shot 2D vision tasks, i.e., classification and semantic segmentation. Specifically, a pre-trained Siamese backbone is utilized for feature extraction from the support and the query images. Then, global average pooling is applied on the extracted support feature maps to obtain the support prototypes. This global average prototype is then applied to calculate the similarity between the global features (in classification) or dense pixel-wise features (in semantic segmentation) extracted from the query image for prediction. However, these tasks' global-to-global or global-to-local correspondence is not enough to recover 6D object pose parameters. This work, instead, proposes a dense prototypes extraction module to establish the local-to-local correspondence between the support RGBD images and the query scene patch for pose estimation.

**Transformer [51].** Transformers networks are first introduced in Natural Language Processing and are brought into many vision tasks. The multi-head attention mechanism enables it to capture the long-term dependency even on an unordered set. Specifically, given three vectors as inputs, namely query $Q$, key $K$, and value $V$. The attention mechanism is to retrieve information $I$ from the value s.t. the similarity between $Q$ and $K$, denoted as:

$$I_{retrieved} = softmax(QK^T)V. \tag{1}$$

Gifted with the capability of capturing long-term dependency, the Transformer networks [51, 56] have been successfully applied to aggregate contextual information in the local feature matching [43, 46] and point cloud registration [23] field. In this work, we further extend it to dense

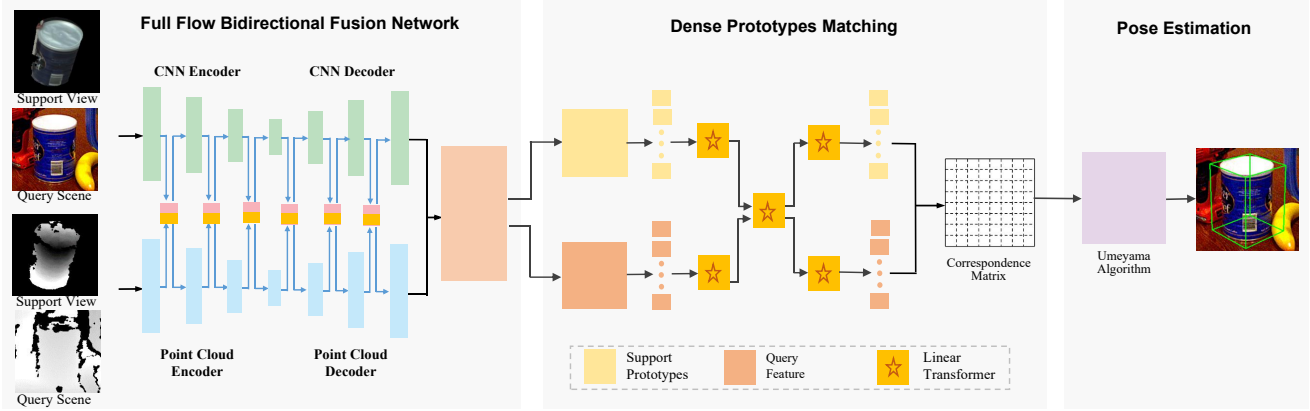


Figure 3. **Overview of our pipeline.** A Siamese full flow bidirectional fusion network [15] is utilized to extract rich appearance and geometric features from the support view and the query scene patch, respectively. The extracted features are then fed into self- and cross-attention modules to obtain dense support prototypes and query features for correspondence reasoning. Finally, the Umeyama algorithm [50] is applied to recover the pose parameters of novel objects in the query scene patch.

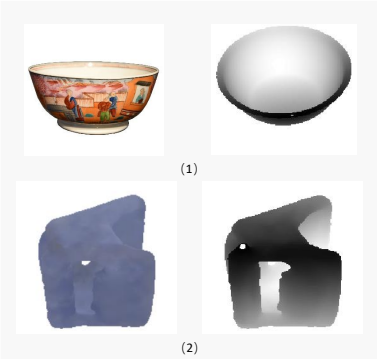


Figure 4. **Complementary information in RGBD images for few-shot 6D pose estimation.** (1) Texture information in RGB images is crucial cue for objects with smooth surface. (2) Geometric information in depth images is important cue for textureless objects.

RGBD prototypes matching for few-shot 6D pose estimation.

### 3.3.2 Overview

To build a few-shot pose estimation algorithm that can generalize well to novel objects, it is crucial to fully explore the semantic and geometric relationship between the given support views and the query scene patch, as shown in Figure 4. In this section, we introduce our dense prototypes matching framework to tackle this challenging problem. As shown in Figure 3, our framework consists of three main parts. Firstly, a Siamese RGBD feature extraction backbone is utilized to extract rich semantic and geometric features for each pixel/point. Then, a dense prototypes extraction network based on transformers is applied to extract dense RGBD prototypes from the support view and point-wise local features from the query scene patch for similarity calculation. Finally, after the correspondence between dense prototypes and scene features is established, the Umeyama algorithm [50] is leveraged to estimate the 6D pose parameters.

### 3.3.3 Feature Extraction Backbone

The first step is to extract rich semantic and geometric features from the given RGBD images. As a fundamental problem, many works [15, 53, 60] have studied this representation learning task. Recently, FFB6D [15] introduce a full flow bidirectional fusion network for 6D pose estimation and significantly improve the performance of *close-set* pose estimation. Specifically, bidirectional local feature fusion blocks are added into each encoding and decoding layer to bridge the information gap and improve the quality of extracted semantic and geometric features (see [15] for details). In this work, we leverage FFB6D to build a Siamese network for feature extraction from the support images and the query scenes.

### 3.3.4 Dense Prototypes Extraction and Matching

Now we have obtained dense features from the Siamese feature extraction backbone. We then extract dense support prototypes and query features to calculate the similarity and establish the correspondence. To extract descriptive and representative dense RGBD prototypes from the support views and dense query features from query scenes, it is crucial to fully leverage the structural geometric information residing in point clouds and semantic information abiding in RGB images. Besides, contextual information

between the support shot and the query patch is also essential to improve the precision of similarity calculation and correspondence exploration.

Considering the power of transformers on long-term dependency capturing, we utilize the optimized Linear Transformers [56] to serve the above two purposes. As shown in the middle part of Figure 3, we first establish self-attention on the extracted feature maps to strengthen the geometric and semantic information residing in the extracted dense prototypes and dense query features. We regard the extracted features as query, key, and value and fed them into the Linear Transformer networks to enhance the semantic and geometric features. Meanwhile, a cross-attention module is also applied to explore the contextual information between the support prototypes and the query scene features. Precisely, to extract contextual information from the support prototypes to the query scene features, we took each scene feature as a query and the dense prototypes as keys and values to the Linear Transformers. Contextual information from query scene features to support prototypes is enhanced similarly. With extracted contextual information, another self-attention modules are applied to enhance the geometric and semantic features further. In this way, we obtain dense support prototypes and query features with rich semantic, geometric and contextual information. Unlike prototype-based few-shot classification and segmentation algorithms that calculate the similarity by cosine distance, we follow local feature matching pipelines [43] to establish the dense correspondence by calculating $C(i,j) = \langle P(i), Q(j) \rangle$ with $P(i)$ the $i_{th}$ prototype, $Q(j)$ the $j_{th}$ query feature and $\langle \cdot, \cdot \rangle$ the inner product. The Sinkhorn Algorithm [39] is applied for differentiable optimization as well.

#### 3.3.5 Pose Parameters Estimation

After the correspondence between the dense prototypes and the query scene features is established, we utilize the Umeyama [50] algorithms to recover the pose parameters. Specifically, given a set of matched pairs $\mathcal{M} = \{(p_i, q_i), 1 \leq i \leq N\}$ with $p_i$, $q_i$ the 3D coordinate of matched prototypes and queries, the Umeyama algorithms estimate the rotation $R$ and translation $T$ by minimizing:

$$L_{lsq} = \sum_{i=1}^{N} ||q_i - (Rp_i + T)||_2^2. \tag{2}$$

To eliminate the influence of outliers. The RANSAC algorithms are also applied.

Given $K$ support views of a novel object, we can obtain $K$ predicted pose parameters along with their losses. We select the one with minimum loss as our final prediction.

# 4. Experiments

## 4.1. Benchmark Datasets

The LineMOD [18] and the YCB-Video [4] are two popular datasets for 6D object pose estimation. The LineMOD dataset contains 13 videos of 13 low-textured objects, while the YCB-Video dataset consists of 92 RGBD videos of 21 YCB objects. For the few-shot pose estimation problem, we select 16 shots for each object for pose estimation. We also follow the strategy of other well-established few-shot problems, i.e., segmentation, and split the dataset into different groups. Specifically, we split the objects into three groups for each dataset and select one for testing and the remaining two for training each time (see the supplementary material for details).

## 4.2. Evaluation Metrics

The average distance metrics ADD and ADDS are widely used for performance evaluation of 6D pose estimation. For an object $\mathcal{O}$ consists of vertexes $v$, the ADD of asymmetric objects with the predicted pose $R$, $T$ and ground truth pose $R^*$, $T^*$ is calculated by:

$$\text{ADD} = \frac{1}{m} \sum_{v \in \mathcal{O}} ||(Rv + T) - (R^* v + T^*)||. \tag{3}$$

For symmetric objects, the ADDS based on the closest point distance is defined as:

$$\text{ADDS} = \frac{1}{m} \sum_{v_1 \in \mathcal{O}} \min_{v_2 \in \mathcal{O}} ||(Rv_1 + T) - (R^* v_2 + T^*)||. \tag{4}$$

In the YCB-Video dataset, the area under the accuracy-threshold curve obtained by varying the distance threshold (ADDS and ADD AUC) is reported following [15, 16, 58]. In the LineMOD datasets, we report the distance less than 10% objects diameter recall (ADD-0.1d) as in [19, 38].

## 4.3. Baselines

Possible solutions to the few-shot 6D object pose estimation problem include local image feature matching, point cloud registration, and template matching. We select the state-of-the-art solution in each direction as our baseline.

**LoFTR** [46] is a detector-free deep learning architecture for local image feature matching. It uses the self- and cross-attention layers in Transformers to obtain high-quality matches.

**PREDATOR** [23] is a neural architecture for pairwise 3D point cloud registration with deep attention to the overlap region. It learns to detect the overlap region between two unregistered scans and focus on that region when sampling feature points.

**Template Matching.** Template matching approaches [13, 17, 24] discrete pose estimation problem into classification problem. These approaches rely on CAD models to

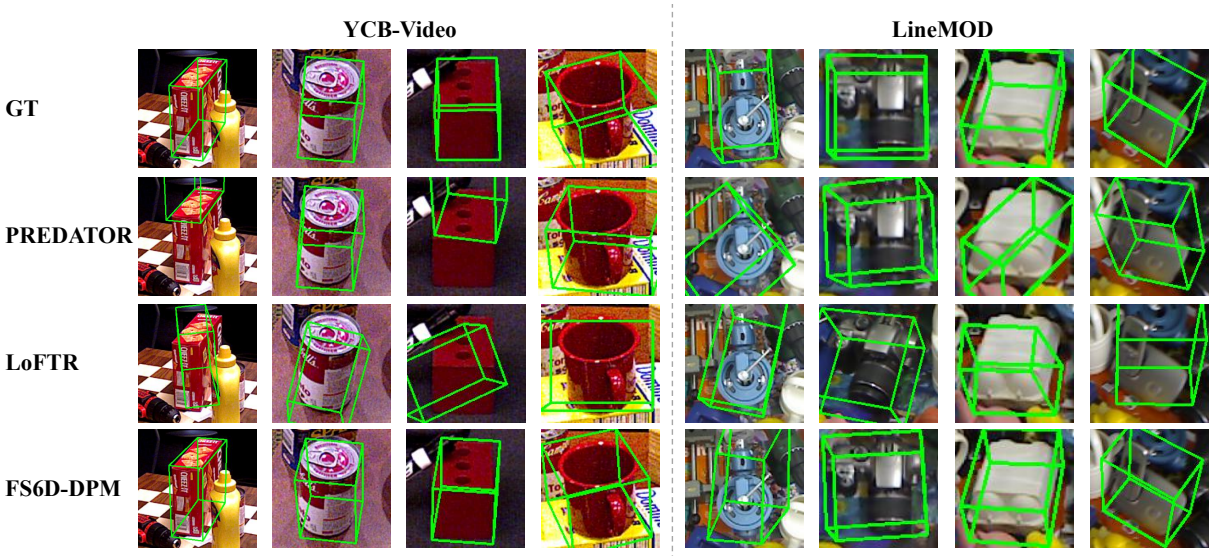


Figure 5. Qualitative results on the YCB-Video (left) and the LineMOD (right) datasets. We visualize the results of PREDATOR [23], LoFTR [46] and the proposed FS6D-DPM. The ground truths are also visualized in the first row.

| Group | Object | PREDATOR [23] | | LoFTR [46] | | TP-UB | | FS6D-DPM | |
|---|---|---|---|---|---|---|---|---|---|
| | | ADDS | ADD | ADDS | ADD | ADDS | ADD | ADDS | ADD |
| 0 | 002 master chef can | 73.0 | 17.4 | 87.2 | **50.6** | 62.2 | 21.4 | **92.6** | 36.8 |
| | 003 cracker box | 41.7 | 8.3 | 71.8 | **25.5** | 65.6 | 5.0 | **83.9** | 24.5 |
| | 004 sugar box | 53.7 | 15.3 | 63.9 | 13.4 | 66.7 | 21.5 | **95.1** | **43.9** |
| | 005 tomato soup can | 81.2 | 44.4 | 77.1 | 52.9 | 75.2 | 43.1 | **93.0** | **54.2** |
| | 006 mustard bottle | 35.5 | 5.0 | 84.5 | 59.0 | 47.1 | 4.0 | **97.0** | **71.1** |
| | 007 tuna fish can | 78.2 | 34.2 | 72.6 | **55.7** | 72.8 | 38.4 | **94.5** | 53.9 |
| | 008 pudding box | 73.5 | 24.2 | 86.5 | 68.1 | 86.3 | 18.4 | **94.9** | **79.6** |
| 1 | 009 gelatin box | 81.4 | 37.5 | 71.6 | **45.2** | 90.9 | 43.2 | **98.3** | 32.1 |
| | 010 potted meat can | 62.0 | 20.9 | 67.4 | 45.1 | 59.8 | 28.9 | **87.6** | **54.9** |
| | 011 banana | 57.7 | 9.9 | 24.2 | 1.6 | 79.2 | 54.5 | **94.0** | **69.1** |
| | 019 pitcher base | 83.7 | 18.1 | 58.7 | 22.3 | 17.5 | 0.7 | **91.1** | **40.4** |
| | 021 bleach cleanser | 88.3 | **48.1** | 36.9 | 16.7 | 20.3 | 0.6 | **89.4** | 44.1 |
| | 024 bowl | 73.2 | **17.4** | 32.7 | 1.4 | 30.7 | 0.0 | **74.7** | 0.9 |
| | 025 mug | 84.8 | 29.5 | 47.3 | 23.6 | 46.0 | 13.9 | **86.5** | **39.2** |
| 2 | 035 power drill | 60.6 | 12.3 | 18.8 | 1.3 | 42.3 | 0.7 | **73.0** | **19.8** |
| | 036 wood block | 70.5 | 10.0 | 49.9 | 1.4 | 13.5 | 1.3 | **94.7** | **27.9** |
| | 037 scissors | 75.5 | 25.0 | 32.3 | 14.6 | **89.5** | **71.8** | 74.2 | 27.7 |
| | 040 large marker | 81.8 | 38.9 | 20.7 | 8.4 | 82.5 | 51.9 | **97.4** | **74.2** |
| | 051 large clamp | **83.0** | 34.4 | 24.1 | 11.2 | 49.0 | 20.0 | 82.7 | **34.7** |
| | 052 extra large clamp | **72.9** | **24.1** | 15.0 | 1.8 | 50.2 | 9.4 | 65.7 | 10.1 |
| | 061 foam brick | 79.2 | 35.5 | 59.4 | 31.4 | 91.8 | **60.5** | **95.7** | 45.8 |
| MEAN | | 71.0 | 24.3 | 52.5 | 26.2 | 59.0 | 24.2 | **88.4** | **42.1** |

Table 2. Quantitative evaluation of different few-shot 6D pose baselines on the YCB-Video dataset. Among them, the proposed FS6D-DPM fully leverages the appearance and geometric information achieves the best performance. TP-UB: upper bound of template approaches.

| Group | PREDATOR [23] | LoFTR [46] | TP-UB | FS6D-DPM |
|---|---|---|---|---|
| | ADD-0.1d | ADD-0.1d | ADD-0.1d | ADD-0.1d |
| 0 | 55.1 | 38.0 | 8.1 | **70.0** |
| 1 | 40.4 | 30.4 | 10.0 | **86.8** |
| 2 | 46.8 | 30.3 | 13.2 | **93.4** |
| Mean | 48.0 | 33.4 | 10.1 | **83.4** |

Table 3. Quantitative evaluation of different few-shot 6D pose baselines on the LineMOD dataset. The proposed FS6D-DPM that fully leverages the appearance and geometric information achieves the best performance. TP-UB: upper bound of template-based approach.

| Object | w/o OTB | w/ OTB |
|---|---|---|
| | ADD | ADD |
| 002 master chef can | 23.4 | **50.0** |
| 003 cracker box | 15.1 | **42.0** |
| 004 sugar box | 12.3 | **52.5** |
| 005 tomato soup can | 52.8 | **74.7** |
| 006 mustard bottle | 55.4 | **75.4** |
| 007 tuna fish can | 54.5 | **56.5** |
| 008 pudding box | 34.4 | **42.2** |
| 009 gelatin box | 50.7 | **94.2** |
| 010 potted meat can | 38.7 | **54.8** |
| Mean | 37.5 | **60.3** |

Table 4. Effect of online texture blending. w/o OTB: without online texture blending; w/ OTB: with online texture blending.

| Group | from scratch | pretrained | pretrained + finetuned |
|---|---|---|---|
| 0 | 62.8 | **73.9** | 70.0 |
| 1 | 57.7 | 77.9 | **86.8** |
| 2 | 75 | 86.1 | **93.4** |
| Mean | 65.2 | 79.3 | **83.4** |

Table 5. Effect of ShapeNet6D for pre-training on the LineMOD dataset. The variety of shape and appearance priors improves generalizability by large margins.

generate thousands of templates and retrieve the closest one to the scene. However, we eliminate the dependency of precise object CAD models in our problem. Besides, capturing, labeling, and storing thousands of support shots are time- and storage-consuming. We assign the view with rotation closest to the ground truth and the center shift as translation to reveal the upper bound of these approaches.

For a fair comparison, all baselines and the proposed one are not equipped with iterative refinement, e.g., ICP [3].

### 4.4. Training and Implementation

We crop object patches with ground-truth bounding boxes for our model and resize them to $255 \times 255$ as input. The correspondence is optimized by negative log-likelihood loss [43]. For a fair comparison, we pretrained all models on ShapeNet6D with online data augmentation for two epochs and fine-tuned on benchmark datasets for five epochs. We select 16 different views for each object as support images.

### 4.5. Benchmark Results

**Results on LineMOD and YCB-Video datasets.** Quantitative results on the YCB-Video and the LineMOD dataset are shown in Table 2 and Table 3 respectively. Thanks to the joint reasoning of appearance and geometric relationship between the support and query images, our method outperforms the state-of-the-art local image feature matching method and point cloud registration algorithms by large margins. Some qualitative results are shown in Figure 5.

**Domain generalization.** As is shown in Table 5, our model trained on ShapeNet6D with online data augmentation is $4.1\%$ behind the fine-tuned one. Considering the small shape and appearance diversity in the LineMOD dataset, compared with ShapeNet6D, we think the performance drop mainly comes from the domain gap. More future works are expected to bridge this gap to fully explore the power of shape and appearance diversity in ShapeNet6D, e.g., designing domain invariant algorithms.

### 4.6. Ablation Study

**Effect of pre-training on the large-scale ShapeNet6D.** As shown in Table 5, FS6D-DPM trained on ShapeNet6D outperforms the one trained from scratch on the LineMOD dataset by a large margin ($+11\%$), proving the efficacy of the shape and appearance diversity resides in ShapeNet6D.

**Effect of online texture blending.** As shown in Table 4, the proposed online texture blending provides diverse texture prior and improves the performance on texture-rich objects in the YCB-Video dataset by large margins.

## 5. Discussion and Limitations

In this work, we study a challenging *open-set* problem, the few-shot 6D object pose estimation. We point out the essence of appearance and geometric information to tackle the problem and propose FS6D-DPM as a solid baseline to solve it. Furthermore, we show that prior from diverse shapes and appearances are crucial to the generalizability of few-shot 6D pose estimation algorithms and introduce a large-scale dataset (ShapeNet6D) for network pre-training. An online texture blending augmentation is proposed to bridge the domain gap as well.

However, there are still some limitations in this work. Firstly, we focus on the pose estimation problem and rely on object detection algorithms to crop out the region of interested objects. Though various off-the-shelf few-shot object detection algorithms [25] are available, a joint framework is more practical. Secondly, despite being diverse in shape and appearance, the proposed large-scale ShapeNet6D is synthesis, and the domain gaps problem is not tackled yet. Future directions include domain invariant pose estimation algorithms or large-scale real-world datasets. Lastly, there is still a significant performance gap between few-shot algorithms and those trained under the *close-set* setting. We expect more future research, e.g., leveraging 3D keypoint-based techniques [15, 16] to bridge this gap.

**Acknowledgements** This work is supported by Guangzhou Okay Information Technology with the project GZETDZ18EG05.

## References


[1] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6359–6367, 2020. 2

[2] Petr Beckmann and Andre Spizzichino. The scattering of electromagnetic waves from rough surfaces. *Norwood*, 1987. 4

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 8

[4] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 3, 6

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[6] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. 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*, pages 1581–1590, 2021. 4

[7] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8966, 2019. 2

[8] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 195–205, 2018. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 4

[10] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018. 2

[11] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017. 2

[12] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5545–5554, 2019. 2

[13] Chunhui Gu and Xiaofeng Ren. Discriminative mixture-of-templates for viewpoint classification. In *European Conference on Computer Vision*, pages 408–421. Springer, 2010. 2, 6

[14] Yisheng He, Haoqiang Fan, Haibin Huang, Qifeng Chen, and Jian Sun. Towards self-supervised category-level object pose and size estimation. *arXiv preprint arXiv:2203.02884*, 2022. 2

[15] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1, 2, 5, 6, 8

[16] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11641, 2020. 1, 2, 6, 8

[17] Stefan Hinterstoisser, Cedric Cagniart, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit. Gradient response maps for real-time detection of textureless objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):876–888, 2011. 2, 6

[18] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *2011 international conference on computer vision*, pages 858–865. IEEE, 2011. 1, 3, 6

[19] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 6

[20] Stefan Hinterstoisser, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige. Going further with point pair features. In *European conference on computer vision*, pages 834–848. Springer, 2016. 2

[21] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 3

[22] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *European Conference on Computer Vision*, pages 577–594. Springer, 2020. 1, 3, 4

[23] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 4, 6, 7

[24] Daniel P Huttenlocher, Gregory A Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993. 2, 6

[25] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 8

[26] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1529, 2017. 3

[27] Jiaxin Li and Gim Hee Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 361–370, 2019. 2

[28] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[29] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7678–7687, 2019. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[31] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 2

[32] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020. 2

[33] David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999. 2

[34] Zixin Luo, Tianwei Shen, Lei Zhou, Siyu Zhu, Runze Zhang, Yao Yao, Tian Fang, and Long Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–183, 2018. 2

[35] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6589–6598, 2020. 2

[36] Fabian Manhardt, Wadim Kehl, Nassir Navab, and Federico Tombari. Deep model-based 6d pose refinement in rgb. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 800–815, 2018. 3

[37] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10710–10719, 2020. 4

[38] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 2, 6

[39] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 6

[40] Fabio Poiesi and Davide Boscaini. Distinctive 3d local deep descriptors. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5720–5727. IEEE, 2021. 2

[41] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *arXiv preprint arXiv:1810.10510*, 2018. 2

[42] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[43] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2, 4, 6, 8

[44] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 2

[45] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2686–2694, 2015. 3

[46] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 2, 4, 6, 7

[47] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13916–13925, 2020. 2

[48] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pages 530–546. Springer, 2020. 2

[49] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Annals of the History of Computing*, (01):1–1, 2020. 2

[50] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Computer Architecture Letters*, 13(04):376–380, 1991. 5, 6

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4

[52] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016. 2

[53] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d

object pose estimation by iterative dense fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3343–3352, 2019. 2, 5

[54] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *European Conference on Computer Vision*, pages 108–125. Springer, 2020. 3

[55] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 1, 2, 3

[56] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 4, 6

[57] Yue Wang and Justin M Solomon. Prnet: Self-supervised learning for partial-to-partial registration. *arXiv preprint arXiv:1910.12240*, 2019. 2

[58] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1, 2, 3, 6

[59] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3d objects. *BMVC 2019*, 2019. 2

[60] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018. 5

[61] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 763–778. Springer, 2020. 2

[62] Lei Yang, Yan Zi Wei, Yisheng He, Wei Sun, Zhenhang Huang, Haibin Huang, and Haoqiang Fan. ishape: A first step towards irregular shape instance segmentation. *arXiv preprint arXiv:2109.15068*, 2021. 3

[63] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[64] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019. 2