DVMNet: Computing Relative Pose for Unseen Objects Beyond Hypotheses

Chen Zhao[1] Tong Zhang[1*] Zheng Dang[1] Mathieu Salzmann[12]
[1]EPFL [2]ClearSpace SA
{chen.zhao, tong.zhang, zheng.dang, mathieu.salzmann}@epfl.ch

Abstract

Determining the relative pose of an object between two images is pivotal to the success of generalizable object pose estimation. Existing approaches typically approximate the continuous pose representation with a large number of discrete pose hypotheses, which incurs a computationally expensive process of scoring each hypothesis at test time. By contrast, we present a Deep Voxel Matching Network (DVMNet) that eliminates the need for pose hypotheses and computes the relative object pose in a single pass. To this end, we map the two input RGB images, reference and query, to their respective voxelized 3D representations. We then pass the resulting voxels through a pose estimation module, where the voxels are aligned and the pose is computed in an end-to-end fashion by solving a least-squares problem. To enhance robustness, we introduce a weighted closest voxel algorithm capable of mitigating the impact of noisy voxels. We conduct extensive experiments on the CO3D, LINEMOD, and Objaverse datasets, demonstrating that our method delivers more accurate relative pose estimates for novel objects at a lower computational cost compared to state-of-the-art methods. Our code is released at: https://github.com/sailor-z/DVMNet/.

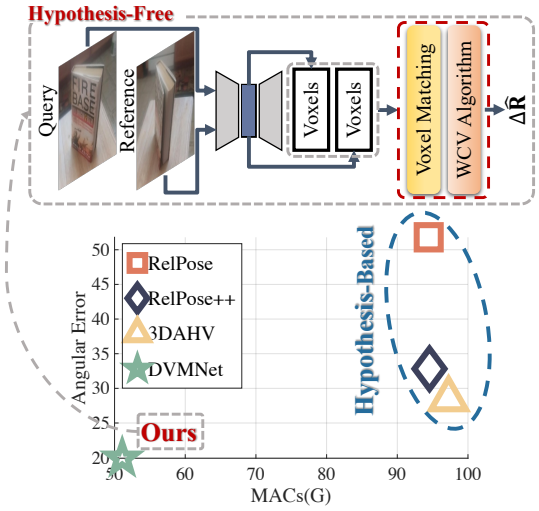

Figure 1. **Advantages of our DVMNet compared to hypothesis-based methods.** Hypothesis-based techniques approximate the relative object pose by scoring numerous pose hypotheses, leading to a high computational cost. By contrast, our DVMNet computes the pose in a hypothesis-free fashion by robustly matching voxelized 3D representations of the reference and query images via a Weighted Closest Voxel (WCV) algorithm. Our method strikes a favorable balance between computational cost and accuracy in relative object pose estimation, as measured by multiply-accumulate operations (MACs) and angular error.

1. Introduction

Object pose estimation plays a crucial role in 3D computer vision and robotics tasks [2, 11, 23, 33], aiming to estimate the 3D rotation and 3D translation of an object depicted in an RGB image. The vast majority of existing methods work under the assumption that the training and testing data include the same object instances, thereby limiting their applicability to scenarios that involve previously unseen objects. Recently, generalizable object pose estimation [21, 29, 32] has received growing attention, showcasing the potential to generalize to unseen objects from new categories without retraining the network. In pursuit of this generalization capability, existing methods leverage densely sampled images depicting unseen objects in diverse poses, serving as references. Object pose estimation is then carried out through template matching [21, 29] or by establishing 2D-3D correspondences [14, 32]. Unfortunately, the effectiveness of these methods strongly depends on the references densely covering the viewpoints of unseen objects, making them inapplicable to practical scenarios where only sparse reference views are available.

In this context, a few methods [18, 47, 52] highlight the importance of relative object pose estimation. Unlike previous approaches in generalizable object pose estimation, these methods focus on estimating the relative pose of an unseen object between two images, i.e., a single reference image of the object and the query image. In this

*Corresponding author.

paper, we also work in this setting, motivated by the practical ease of obtaining a single reference image for a new object. One plausible solution is to compute the relative pose based on 2D-2D correspondences [13]. However, the single-reference scenario tends to yield a significant viewpoint gap between the reference and the query. Existing studies [18, 52] have shown that image-matching techniques [26, 31] are sensitive to such pose differences. To handle this issue, the prior methods [18, 47, 52] follow an alternative strategy of scoring multiple pose hypotheses for the input reference-query pair, and predicting the pose as the hypothesis with the highest score. However, this strategy comes with the drawback of requiring numerous pose hypotheses to achieve reasonable accuracy, e.g., 500,000 in [18], which thus induces a computational burden. Moreover, we empirically found that these approaches occasionally produce unnaturally large errors. One plausible explanation is their failure to model the continuous nature of the object pose space, as they primarily concentrate on learning to score discrete hypotheses.

To overcome these drawbacks, we present a novel Deep Voxel Matching Network (DVMNet) that computes the relative pose of unseen objects in a *hypothesis-free* and end-to-end manner. Our approach starts by voxelizing the query and reference images in a dedicated autoencoder. The encoder network lifts 2D image features to 3D voxels, leveraging cross-view 3D information. The decoder network reconstructs a masked object image from the voxels, encouraging the learned voxels to account for the object. Subsequently, we align the query and reference voxels based on a score matrix that measures the voxel similarities. To handle unreliable voxels due to background, varying illumination, and other potential nuisances, we present a Weighted Closest Voxel (WCV) algorithm to facilitate robust pose estimation. In this algorithm, each voxel-voxel correspondence is assigned a confidence score computed by utilizing both the 3D voxel objectness map and the 2D object mask learned by the autoencoder. The relative object pose is then computed by solving a weighted least-squares problem. Such an end-to-end learning mechanism eliminates the necessity for voxel-wise annotations and allows the network to directly learn pose-aware features from RGB images. As illustrated in Fig. 1, our DVMNet requires significantly fewer multiply-accumulate operations (MACs) while achieving smaller angular errors than its hypothesis-based competitors.

We perform comprehensive experiments on the CO3D [25], LINEMOD [15], and Objaverse [9] datasets. Our method yields more accurate and robust relative pose estimates for previously unseen objects than both image-matching and hypothesis-based competitors. We also conduct ablation studies where the results demonstrate the effectiveness of the key components in our framework. In short, our contributions are twofolds: First, we tackle the problem of relative pose estimation for unseen objects in a *hypothesis-free* manner by introducing a deep voxel matching network. Second, we present a weighted closest voxel algorithm that robustly and end-to-end computes the object pose from voxel-voxel correspondences.

2. Related Work

Instance-Level Object Pose Estimation. The majority of previous deep learning approaches to object pose estimation [24, 30, 34, 35, 43] tackle the problem at an instance level, holding an assumption that the training and testing data depict the same object instances. Since the appearance of an object instance in different poses typically exhibits limited variations, these methods provide highly accurate object pose estimates. Nevertheless, they struggle to generalize to previously unseen objects during testing without retraining the network, as has been observed in the literature [21, 29, 32]. This limitation constrains their applicability in real-world scenarios that often involve diverse object instances. This has been remedied to a degree by category-level object pose estimation methods [7, 19, 36]. In this scenario, the testing images comprise new object instances from specific categories already included in the training data. Although these methods have achieved promising generalization ability within the predefined object categories, they become ineffective when facing objects from entirely new categories.

Generalizable Object Pose Estimation. To tackle the scenario of unseen objects from new categories, there has been growing interest in generalizable object pose estimation. When a textured 3D mesh is available for an unseen object, some approaches [29, 42, 51] suggest generating synthetic images as references by rendering the 3D mesh from various viewpoints. Given a query image that depicts this object, a template matching paradigm is utilized to identify the most similar reference and approximate the object pose in the query as that of the selected reference. Some methods bypass the need for 3D meshes by assuming the availability of multiple real reference images. Object pose estimation is then carried out by employing either a template matching strategy [21] or a 3D object reconstruction technique [14, 32]. Nevertheless, all of these methods rely on having access to dense-view reference images, which limits their applicability in scenarios where only sparse views of references are available.

Relative Object Pose Estimation. Recently, several studies [18, 37, 47, 52] have highlighted the importance of relative object pose estimation. These methods stand out in generalizable object pose estimation due to their key advantage of requiring only a single reference image. The objective of these methods is to estimate the relative object pose between the input query image and the reference. Since the

single-reference assumption tends to result in a large object pose difference between the query and the reference, unseen object pose estimation becomes more challenging. Intuitively, one could establish pixel-pixel correspondences between the two images and compute the relative object pose based on multi-view geometry [13]. However, as reported in the literature [18, 52] and also in our experiments, even the state-of-the-art image-matching techniques [12, 26, 31] cannot deliver accurate pose estimates when confronted with large object pose differences. To address this, existing methods [18, 47, 52] suggest approximating the relative object pose via a discrete set of pose hypotheses, and learning to maximize scores of the positive hypotheses. Since object pose lies in a continuous space [53], accurately approximating the pose necessitates a vast number of pose hypotheses, which makes the hypothesis-based approaches computationally expensive. Moreover, scoring discrete samples lacks an understanding of the continuous pose distribution, leading to failure cases with unnaturally high pose estimation errors. By contrast, we present a hypothesis-free technique that is capable of computing the relative object pose in a single pass via deep voxel matching.

3. Method

3.1. Problem Formulation

We tackle the problem of estimating the relative pose $\Delta\mathbf{P}$ of a previously-unseen object between a query image $\mathbf{I}_q$ and a reference image $\mathbf{I}_r$. In this scenario, the objects present in the testing set Ω_{test} differ from those in the training set Ω_{train}, and the goal is to handle the unseen objects without retraining the network. This requires addressing the challenges of generalization to unseen objects in Ω_{test} and robustness to the large object pose difference between $\mathbf{I}_q$ and $\mathbf{I}_r$.

Since the relative object translation estimation can be facilitated by using zero-shot object detectors [17, 39, 50], our focus is on predicting the relative object rotation, i.e., $\Delta\mathbf{R} \in SO(3)$, which is more challenging. This choice aligns with existing literature [42, 44, 51, 52]. Previous hypothesis-based approaches [18, 47, 52] approximate $\Delta\mathbf{R}$ by sampling discrete pose hypotheses and maximizing the scores of positive samples. The problem is formulated as

$$\Delta\hat{\mathbf{R}} = \underset{\Delta\mathbf{R}_i \in \mathcal{R}}{\arg\max}\, f(\mathbf{I}_q, \mathbf{I}_r, \Delta\mathbf{R}_i), \tag{1}$$

where $\mathcal{R}$ denotes the set of discrete pose hypotheses. Achieving a decent approximation accuracy requires a large number of hypotheses, e.g., 500,000 in [18]. By contrast, we present a hypothesis-free technique that computes $\Delta\mathbf{R}$ in a single pass as $\Delta\hat{\mathbf{R}} = g(\mathbf{I}_q, \mathbf{I}_r)$.

3.2. Motivation

Drawing inspiration from the success of pixel-pixel correspondences in image matching [26, 45, 49] a natural approach to avoiding the use of pose hypotheses would be to compute the relative object pose based on 2D correspondences. However, recent studies [18, 52] have observed that such an image-matching strategy is unreliable in the scenario of object pose estimation. We trace this limitation back to the fact that image-matching methods are not fully differentiable w.r.t. the pose. Specifically, some approaches [45, 46, 48] encode a notion of consistency among the pixel-pixel correspondences utilizing the essential matrix. However, computing the pose from the essential matrix leads to multiple solutions [13]. The pose estimation is thus detached from the learning process as a post-processing step. Notably, in the context of object pose estimation, the pre-generated correspondences tend to be unreliable in the presence of challenges such as large object pose differences and textureless objects. Therefore, the isolated pose estimation step in the two-stage design becomes less effective.

To address this issue, we propose to lift the input images to voxelized 3D representations [40] and perform the matching process in 3D space. Therefore, the computation of the relative object pose from the resulting voxel-voxel matches becomes a differentiable operation. This characteristic enables us to directly supervise the entire framework with the actual quantity we aim to predict, i.e., the relative object pose. Subsequently, we elaborate on the steps involved in the presented hypothesis-free mechanism.

3.3. Image Voxelization

To achieve the pose estimation from voxel-voxel correspondences with only RGB images as input, we first need to lift each RGB image to a set of 3D voxels. To enable such a voxelization, we introduce an autoencoder network depicted in Fig. 2, which includes a 2D-3D encoder and a 3D-2D decoder. Specifically, we employ a pretrained vision transformer [41] to convert the query and reference images to 2D feature embeddings denoted as $\mathbf{F}_q$ and $\mathbf{F}_r$, respectively. Considering the difficulty of lifting 2D images to 3D representations, we incorporate a cross-attention module to capture cross-view 3D information. We take the feature embedding $\mathbf{F}_q$ as an example (a symmetric process is carried out for $\mathbf{F}_r$). The cross-attention module [41] is defined as

$$\tilde{\mathbf{F}}_q^l = \text{MHSA}(\text{LN}(\mathbf{F}_q^{l-1})) + \mathbf{F}_q^{l-1}, \tag{2}$$

$$\hat{\mathbf{F}}_q^l = \text{MHCA}(\text{LN}(\tilde{\mathbf{F}}_q^l), \text{LN}(\mathbf{F}_r^{l-1})) + \tilde{\mathbf{F}}_q^l, \tag{3}$$

$$\mathbf{F}_q^l = \text{FFN}(\text{LN}(\hat{\mathbf{F}}_q^l)) + \hat{\mathbf{F}}_q^l, \tag{4}$$

where MHSA stands for a multi-head self-attention layer, MHCA represents a multi-head cross-attention layer that

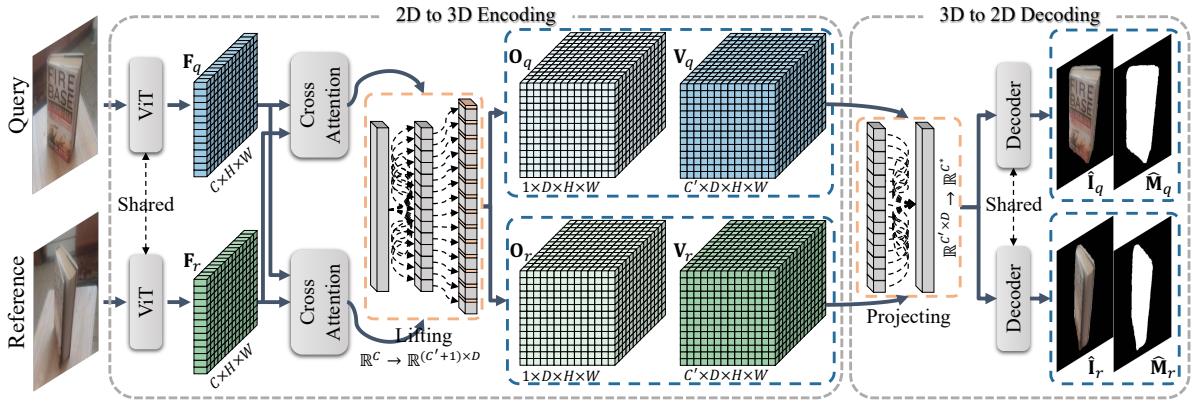

Figure 2. **Network architecture of our autoencoder.** The encoder takes two RGB images, query and reference, as input and lifts their 2D feature embeddings to 3D voxels by leveraging cross-view 3D information. $\mathbf{O}_q$ and $\mathbf{O}_r$ represent the learned 3D objectness maps account for robust object pose estimation. The decoder then reconstructs the masked object images from the voxels, allowing the voxels to encode the object patterns.

takes $\tilde{\mathbf{F}}_q^l$ as *query* and $\mathbf{F}_r^{l-1}$ as *key* and *value*, LN denotes layer normalization [3], and FFN is a feed-forward network that includes MLPs. The resulting $\hat{\mathbf{F}}_q^l$ then serves as the input to the next cross-attention module. Consequently, the output of the last cross-attention module contains object features depicted from two different viewpoints, thus incorporating 3D information.

Benefiting from such a 3D-aware encoding process, we voxelize the image feature embeddings via a simple reshaping process. Note that, to facilitate the robust pose estimation that will be introduced in Sec. 3.5, we predict an objectness score for each voxel, which reflects the significance of the voxel to the relative object pose estimation. Therefore, the actual reshaping process is conducted as $\mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{(C^{'}+1)\times D\times H\times W}$, where $C = (C^{'}+1)\times D$. As shown in Fig. 2, we denote the resulting 3D objectness maps and 3D volumes as $\mathbf{O}_q, \mathbf{O}_r \in \mathbb{R}^{1\times D\times H\times W}$ and $\mathbf{V}_q, \mathbf{V}_r \in \mathbb{R}^{C^{'}\times D\times H\times W}$, respectively.

As our approach does not rely on object segmentation, the learned voxel representations might be affected by the background of the query and reference images. To alleviate this issue, we introduce an object-aware decoding process over $\mathbf{V}_q$ and $\mathbf{V}_r$. Concretely, $\mathbf{V}_q$ and $\mathbf{V}_r$ are projected to 2D space by aggregating the voxels along the depth direction as $\mathbb{R}^{C^{'}\times D\times H\times W} \rightarrow \mathbb{R}^{C^*\times H\times W}$, where $C^* = C^{'}\times D$. The resulting 2D feature embeddings are then fed into a decoder that contains several self-attention modules [10] from which the object images $\hat{\mathbf{I}}_q$ and $\hat{\mathbf{I}}_r$ without background are produced. The object masks $\hat{\mathbf{M}}_q$ and $\hat{\mathbf{M}}_r$ are additionally predicted to provide auxiliary information that benefits the following robust object pose estimation.

We supervise the training of the autoencoder with an image-level loss function defined as

$$L_{ae} = L_{img} + L_{mask}, \tag{5}$$

$$L_{img} = L_{mse}(\hat{\mathbf{I}}_q, \hat{\mathbf{I}}_q^{gt}) + L_{mse}(\hat{\mathbf{I}}_r, \hat{\mathbf{I}}_r^{gt}), \tag{6}$$

$$L_{mask} = L_{bce}(\hat{\mathbf{M}}_q, \hat{\mathbf{M}}_q^{gt}) + L_{bce}(\hat{\mathbf{M}}_r, \hat{\mathbf{M}}_r^{gt}), \tag{7}$$

where L_{mse} is the mean squared error loss, L_{bce} indicates the binary cross entropy loss, $(\hat{\mathbf{I}}_q^{gt}, \hat{\mathbf{I}}_r^{gt})$ denote the ground-truth foreground images, and $(\hat{\mathbf{M}}_q^{gt}, \hat{\mathbf{M}}_r^{gt})$ represent the ground-truth object masks.

3.4. Object Pose from Deep Voxel Matching

According to multi-view geometry [4, 13, 38], the relative object pose can be computed by solving a least-squares problem expressed in terms of voxel-voxel correspondences. Specifically, as we focus on 3D rotation estimation, the least-squares problem is formulated as

$$E(\Delta\mathbf{R}) = \frac{1}{N}\sum_{i=1}^{N} ||\Delta\mathbf{R}\mathbf{x}_r^i - \mathbf{x}_q^i||_2, \tag{8}$$

where $\mathbf{x}_r^i \in \mathbf{X}_r$ and $\mathbf{x}_q^i \in \mathbf{X}_r$ stand for the 3D coordinates of the i-th reference and query voxels, respectively. The optimal $\Delta\hat{\mathbf{R}}$ is then determined as

$$\Delta\hat{\mathbf{R}} = \underset{\Delta\mathbf{R}_i\in SO(3)}{\arg\min} -2\sum_{i=1}^{N} {\mathbf{x}_q^i}^T \Delta\mathbf{R}_i \mathbf{x}_r^i, \tag{9}$$

As suggested in [4], this problem can be solved by performing a singular value decomposition (SVD) of a covariance matrix as

$$\mathbf{H} = \sum_{i=1}^{N} \mathbf{x}_r^i {\mathbf{x}_q^i}^T, \tag{10}$$

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{11}$$

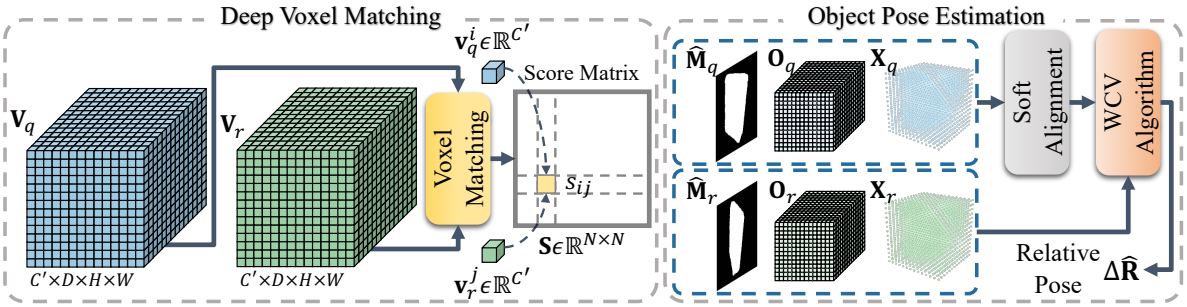

Figure 3. **Computing relative object pose from 3D voxels.** The feature similarities of $\mathbf{V}_q$ and $\mathbf{V}_r$ are computed, which results in a score matrix $\mathbf{S}$. A soft assignment is performed based on $\mathbf{S}$ over the query object mask $\hat{\mathbf{M}}_q$, the 3D objectness map $\mathbf{O}_q$, and the 3D coordinates $\mathbf{X}_q$. The aligned query and reference voxels are then fed into a Weighted Closest Voxel (WCV) algorithm that estimates the relative object pose in a robust and end-to-end manner.

where $\mathbf{H}$ indicates the covariance matrix. The closed-form solution to the least-squares problem is given by $\Delta\hat{\mathbf{R}} = \mathbf{V}\mathbf{U}^T$. Consequently, the key aspect of this problem is to align the 3D voxel coordinates $\mathbf{X}_q$ with $\mathbf{X}_r$.

Inspired by the studies [16, 19, 38] showing that object pose estimation benefits from the end-to-end training, we carry out the alignment in a differentiable fashion. As illustrated in Fig. 3, the alignment is conducted based on a deep voxel matching module. Specifically, we compute a score matrix $\mathbf{S}$ whose entry s_{ij} indicates the cosine similarities between two voxels as

$$s_{ij} = \frac{\mathbf{v}_q^i \cdot \mathbf{v}_r^j}{||\mathbf{v}_q^i||_2 ||\mathbf{v}_r^j||_2}, \tag{12}$$

where $\mathbf{v}_q^i \in \mathbb{R}^{C'}$ and $\mathbf{v}_r^j \in \mathbb{R}^{C'}$ denote the i-th voxel in $\mathbf{V}_q$ and the j-th voxel in $\mathbf{V}_r$, respectively. The alignment is then achieved as

$$\mathbf{X}_q^{'} = p(\mathbf{S}/\tau)\mathbf{X}_q, \tag{13}$$

where $p(\cdot)$ represents the softmax process and τ is a predefined temperature.

3.5. Weighted Closest Voxel Algorithm

Note that our task differs from standard point cloud registration [1, 28, 38], which typically operates on 3D point clouds sampled from 3D object meshes [6] or captured using specific sensors [8]. Here, by contrast, we work with 3D volumes lifted from 2D images, and some voxels could thus be outliers since the corresponding 2D image patches may depict nuisances such as the background. The presence of these outliers may impact the accuracy of the relative object pose estimated from the voxel matches. To address this challenge, we introduce a weighted closest voxel algorithm that enables robust relative object pose estimation.

Concretely, the objective is to mitigate the effect of unreliable voxel matches. We thus incorporate a weight vector

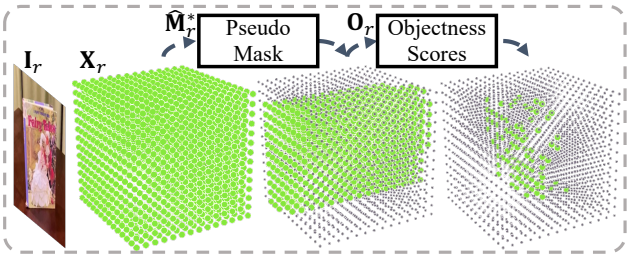

Figure 4. **Illustration of the voxel weights.** Each colored dot indicates the voxel position in 3D space. The green dots in the middle and right cubes represent the voxels with larger weights. The voxel weights in the middle cube are computed based on the replicated object mask, while the weights in the right cube are updated by integrating the 3D objectness map with the object mask.

into the pose estimation process, modifying Eq. 10 as

$$\mathbf{H} = \sum_{i=1}^{N} \omega^i \mathbf{x}_r^i \mathbf{x}_q^{i\,T}, \tag{14}$$

where $\omega^i \in (0, 1)$ denotes the weight of the i-th voxel pair. This makes the subsequent relative object pose estimation aware of the reliability of each voxel pair. We determine the weight vector by utilizing both the object mask and voxel objectness information produced by the encoder network. Specifically, we first replicate $\hat{\mathbf{M}}_q$ and $\hat{\mathbf{M}}_r$ D times along the depth dimension, which creates *pseudo* 3D masks, $\hat{\mathbf{M}}_q^*, \hat{\mathbf{M}}_r^* \in \mathbb{R}^{1\times D\times H\times W}$. These pseudo 3D masks contribute to alleviating the influence of voxels that depict the background. The weight of each voxel pair is then determined as

$$\mathbf{W}_m = h\left(\frac{p(\mathbf{S}/\tau)\hat{\mathbf{M}}_q^* + \hat{\mathbf{M}}_r^*}{2\lambda}\right), \tag{15}$$

where $h(\cdot)$ indicates the sigmoid function, and λ is a manually defined temperature. Additionally, to mitigate the redundancies naturally introduced by the replication process

over $\hat{\mathbf{M}}_q$ and $\hat{\mathbf{M}}_r$, we integrate the resulting pseudo masks with the 3D objectness maps. The final weight vector of all pairwise voxels is determined as $\mathbf{W} = \mathbf{W}_o \odot \mathbf{W}_m$, where $\odot$ indicates the Hadamard product, and $\mathbf{W}_o$ is obtained by carrying out Eq. 15 over $\mathbf{O}_q$ and $\mathbf{O}_r$.

Fig. 4 provides an example of the estimated voxel weights. The dots denote the 3D voxel positions and the voxels assigned with larger weights are colored in green within the middle and right cubes. In the right cube, the green dots roughly depict a 3D surface that corresponds to the object visible in the 2D image. Note that our DVMNet is trained without relying on ground-truth 3D object models. This observation thus demonstrates that the voxels crucial in determining the relative object pose are aware of the 3D object shape information. The complete DVMet is trained end-to-end with a loss function defined as $L = L_{ae} + L_{pose}$ with

$$L_{pose} = ||q(\Delta\hat{\mathbf{R}}) - q(\Delta\mathbf{R}^{gt})||, \tag{16}$$

where $\Delta\mathbf{R}^{gt}$ is the ground-truth relative object pose, and $q(\cdot)$ is a function that converts a rotation matrix to a 6D continuous representation [53].

4. Experiments

4.1. Implementation Details

In the presented autoencoder, we use 3 cross-attention modules in the 2D-3D encoder and 3 self-attention modules in the 3D-2D decoder. In the relative object pose estimation module, we normalize the 3D coordinates of the voxels to an interval of $[-1, 1]$ with a mean of $\mathbf{0}$. We set the temperatures τ and λ in Eq. 13 and Eq. 15 to 0.1 and 1.0, respectively. We train our network in an end-to-end fashion on an A100 GPU, employing the AdamW [22] optimizer with a batch size of 64 and a learning rate of 10^{-5}. Since we focus on predicting the relative object rotation, we crop the object from the RGB image by utilizing the ground-truth object bounding box, following the implementation in [18, 42, 47, 51, 52].

4.2. Relative Object Pose Estimation on CO3D

We first evaluate our method on the CO3D dataset [25], which has been commonly utilized in the literature [18, 47, 52]. This dataset contains 18,619 video sequences that depict 51 object categories. To evaluate the generalization ability of the network to unseen objects, we follow the setting in [47], training the network on 41 object categories and testing it on the other 10 categories. The performance is measured by the mean angular error $e \in [0°, 180°]$ of the estimated relative object rotation, which is defined as

$$e = \arccos\left(\frac{\mathrm{tr}(\Delta\hat{\mathbf{R}}^{\mathrm{T}}\Delta\mathbf{R}^{gt}) - 1}{2}\right). \tag{17}$$

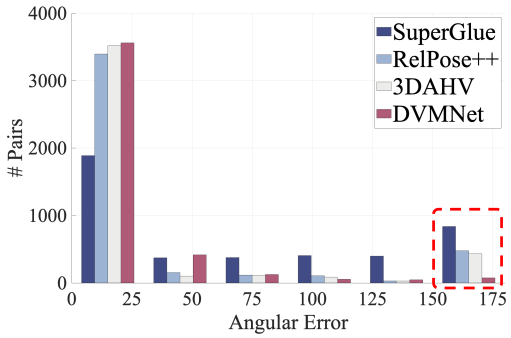

Figure 5. **Histogram depicting the distribution of angular errors.** The image pairs in the testing set are divided into distinct groups based on the angular errors obtained by a specific pose estimation approach. Each bar in the histogram represents the count of image pairs within a particular group. Our DVMNet yields much fewer unnaturally large errors than image-matching and hypothesis-based methods.

We compare our approach with state-of-the-art techniques including image-matching methods, SuperGlue [26], LoFTR [31], and ZSP [12], hypothesis-based methods, RelPose [47], RelPose++ [18], and 3DAHV [52], and a direct regression method implemented in [18].

As reported in Table 1, DVMNet delivers superior relative pose estimation performance for unseen objects, outperforming both the image-matching and hypothesis-based competitors by at least $8.49°$ in terms of mean angular error. To shed more light on the robustness of the evaluated approaches, we categorize the testing image pairs into different groups according to the corresponding angular errors observed when applying a particular relative object pose estimation method. We count the number of image pairs in each group and show the results in Fig. 5. Our method results in a higher number of image pairs with smaller angular errors. More importantly, as highlighted by the red dash box in Fig. 5, both image-matching and hypothesis-based methods exhibit large angular errors for some image pairs. By contrast, our DVMNet results in fewer failure instances, thus demonstrating better robustness.

Furthermore, as argued in Sec. 3, our hypothesis-free strategy is more efficient than the hypothesis-based techniques in relative object pose estimation. We thus assess their computational cost, utilizing the multiply-accumulate operations (MACs). For hypothesis-based methods, all sampled hypotheses are processed in parallel. The results shown in Fig. 1 indicate the benefits of our hypothesis-free DVMNet, which requires considerably fewer MACs than the hypothesis-based competitors. To further substantiate this advantage, we provide detailed results in Fig. 6, where the hypothesis-based methods are evaluated with the number of pose samples varying from 1,000 to 500,000. Note that for 3DAHV, the maximum number is 100,000 because

Method	Ball	Book	Couch	Frisbee	Hotdog	Kite	Remote	Sandwich	Skateboard	Suitcase	**Mean**
SuperGlue [26]	83.55	71.02	45.14	68.67	88.74	56.46	78.58	73.64	72.14	76.74	71.47
LoFTR [31]	82.51	77.33	60.57	78.39	85.05	70.03	89.74	77.77	74.33	90.73	78.64
ZSP [12]	88.09	90.09	64.07	79.08	99.62	72.71	98.61	89.09	89.41	95.03	86.66
Regress [18]	47.56	52.91	39.12	50.16	51.28	52.33	43.85	52.89	51.59	29.11	47.08
RelPose [47]	56.96	55.89	40.71	54.11	64.20	69.43	42.89	59.05	42.32	32.50	51.80
RelPose++ [18]	36.42	35.64	20.00	36.27	33.62	33.63	34.83	36.93	40.60	20.32	32.82
3DAHV [52]	34.83	31.21	22.12	31.30	35.39	34.96	24.73	26.97	26.81	**16.13**	28.44
DVMNet	**28.31**	**21.98**	**19.01**	**23.23**	**21.45**	**17.50**	**11.39**	**19.63**	**20.14**	16.85	**19.95**

Table 1. **Experimental results on CO3D [25].** We report the angular errors of the estimated relative object rotations. All testing object categories were unseen during training. The best results are shown in bold fonts.

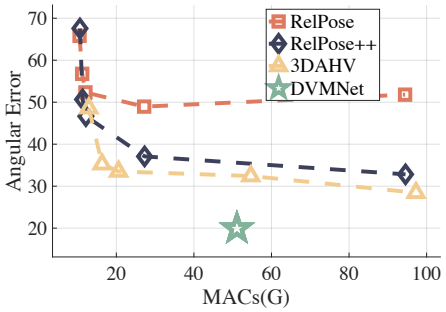

Figure 6. **Comparison with hypothesis-based methods.** We measure the computational cost as multiply-accumulate operations (MACs). The results for hypothesis-based methods are shown with varying numbers of pose samples, ranging from 1,000 to 500,000. For 3DAHV, we set the maximum number to be 100,000 due to the computational resource constraints.

of our computational resource constraints. As shown in Fig. 6, one can enhance the efficiency of the hypothesis-based methods by reducing the number of samples. However, this efficiency gain comes at the cost of sacrificing pose estimation accuracy. By contrast, our method achieves a good trade-off between efficiency and pose estimation accuracy. We also evaluate the time consumption on an A100 GPU. On average, DVMNet processes a pair of images in 23ms. Despite benefiting from parallel estimation, RelPose++ and 3DAHV still cost 29ms and 35ms, respectively, which are slower than our method.

4.3. Relative Object Pose Estimation on GROP

Recently, a new benchmark called GROP for relative pose estimation of unseen objects was introduced in [52]. This benchmark comprises two datasets, i.e., Objaverse [9] and LINEMOD [15]. Both synthetic and real images with diverse object poses are considered. We develop experiments on these two datasets, following the same setup as described in [52]. More concretely, the synthetic images are generated by rendering the object models of the Objaverse dataset from different viewpoints [20]. Several sequences of calibrated real images that depict 13 texture-less household objects are provided from the LINEMOD dataset. The testing set encompasses 128 objects from Objaverse and 5 objects from LINEMOD. The images containing these objects are excluded from the training data, ensuring that all testing objects are previously unseen. All evaluated approaches are trained and tested on the same predefined image pairs, leading to a fair comparison.

Table 2 provides the angular errors of the estimated relative object poses on the LINEMOD and Objaverse datasets. In the synthetic scenarios of Objaverse, DVMNet outperforms the previous methods by at least 7.92° in terms of mean angular error. In the real scenarios of LINEMOD, DVMNet achieves the smallest angular error for most of the testing objects and reduces the mean angular error by at least 4.73° compared to the other approaches. Moreover, we visualize the object pose depicted in the query image and show qualitative results in Fig. 7. The query object pose is determined as $\mathbf{R}_q = \Delta\mathbf{R}\mathbf{R}_r$, where $\Delta\mathbf{R}$ is the relative object pose and $\mathbf{R}_r$ denotes the object pose in the reference image. The ground-truth and predicted query object poses are represented as green and blue arrows, respectively. It is evident from Fig. 7 that the poses estimated by employing our DVMNet are consistently more similar to the ground truth than those obtained with the baselines.

4.4. Ablation Studies

As a critical component of DVMNet, the presented WCV algorithm plays a pivotal role in achieving hypothesis-free and end-to-end relative object pose estimation. To substantiate the effectiveness of the WCV algorithm, we develop comprehensive ablation studies on the CO3D dataset.

We first replace the WCV algorithm with a pose regression module. More concretely, we perform global average pooling over $\mathbf{V}_q$ and $\mathbf{V}_r$. The resulting feature embeddings are concatenated and passed through three fully connected layers to predict the 6D continuous representation of relative object rotation. We maintain all the other components in our framework unchanged to ensure a fair comparison. This alternative approach is also able to predict the rela-

LINEMOD	SuperGlue [26]	LoFTR [31]	ZSP [12]	Regress [18]	RelPose [47]	RelPose++ [18]	3DAHV [52]	**DVMNet**
Cat	67.28	88.06	79.61	54.21	53.72	47.77	50.99	**31.70**
Ben.	58.52	70.80	74.07	52.03	62.32	44.67	38.16	**34.00**
Cam.	58.11	87.13	79.65	51.04	59.91	44.31	41.92	**33.18**
Dri.	65.16	78.85	76.35	52.83	57.61	47.95	**32.65**	46.29
Duck	74.90	97.63	83.43	55.44	55.15	48.65	44.03	**38.91**
Mean	64.79	84.49	78.62	53.11	57.75	46.67	41.55	**36.82**
Objaverse	SuperGlue [26]	LoFTR [31]	ZSP [12]	Regress [18]	RelPose [47]	RelPose++ [18]	3DAHV [52]	**DVMNet**
Mean	102.40	134.05	107.20	55.90	80.39	33.49	28.11	**20.19**

Table 2. **Experimental results on the GROP benchmark [52].** The methods are evaluated in terms of angular error on the LINEMOD [15] and Objaverse [9] datasets. The testing data comprises 5 objects from LINEMOD and 128 objects from Objaverse. All images containing these objects are omitted from the training set.

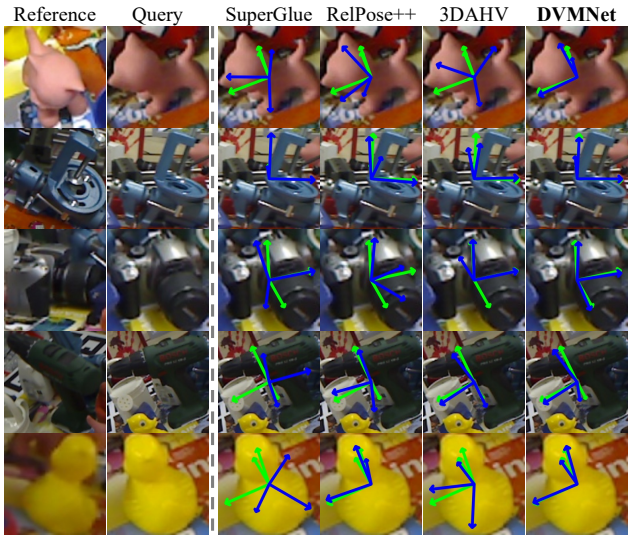

Figure 7. **Qualitative results on LINEMOD [15].** We compute the object pose in the query image utilizing the relative object pose. The green and blue arrows represent the ground-truth object poses and the estimated ones, respectively.

WCV	2D Mask	Voxel Objectness	Angular Error
✗	✗	✗	31.78
✓	✗	✗	21.64
✓	✓	✗	20.92
✓	✗	✓	20.07
✓	✓	✓	19.95

Table 3. **Effectiveness of the presented WCV algorithm.** We report the mean angular errors on the CO3D dataset. The second row indicates the scenario where the WCV algorithm is replaced with a pose regression module. The third row presents a closest voxel algorithm without weights involved.

tive object pose in a hypothesis-free and end-to-end fashion. However, as shown in Table 3, the mean angular error on the CO3D dataset increases by more than 10° when the regression module is employed, showcasing the importance of the WCV algorithm in the presented hypothesis-free mechanism. Furthermore, we evaluate three counterparts of the WCV algorithm, i.e., a closest voxel algorithm without weights, a WCV algorithm with only replicated 2D object masks, and a WCV algorithm with only 3D objectness maps. The final weights of the voxel pairs are determined as $\mathbf{W}_m$ and $\mathbf{W}_o$ in the last two counterparts, respectively. The closest voxel algorithm delivers the worst results among these three variants, revealing that the pose estimation process is affected by the potential outliers. The optimal performance is achieved by leveraging both 2D object masks and 3D voxel objectness maps, which thus demonstrates the effectiveness of these components in DVMNet.

5. Conclusion

In this paper, we have introduced DVMNet, a novel deep voxel matching network for relative pose estimation of unseen objects. Unlike hypothesis-based approaches that rely on a multitude of discrete pose samples, DVMNet computes relative object poses in a hypothesis-free and end-to-end manner. We determine the object pose by solving a weighted least-squares problem based on voxel-voxel correspondences. The voxels are obtained by lifting the query and reference images to 3D via a dedicated autoencoder. We then align query voxels with the reference ones and apply a weighted closest voxel (WCV) algorithm for robust pose estimation. Comprehensive experiments on the CO3D, Objaverse, and LINEMOD datasets have demonstrated that our method excels in efficiently delivering accurate relative poses for unseen objects. In future work, we plan to explore the integration of DVMNet with zero-shot object detectors for 6D relative object pose estimation.

Acknowledgment. This work was funded in part by the Swiss National Science Foundation via the Sinergia grant CRSII5-180359 and the Swiss Innovation Agency (Innosuisse) via the BRIDGE Discovery grant 40B2-0_194729.

References

[1] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7163–7172, 2019. 5

[2] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997. 1

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 4

[5] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Proceedings of the European Conference on Computer Vision*, pages 536–551. Springer, 2014. 2

[6] 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. 5

[7] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11973–11982, 2020. 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 5

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 7, 8

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[12] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 516–532. Springer, 2022. 3, 6, 7, 8

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 3, 4

[14] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022. 1, 2

[15] 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. 2, 7, 8, 1

[16] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2930–2939, 2020. 5

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 1

[18] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 1, 2, 3, 6, 7, 8

[19] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *Proceedings of the European Conference on Computer Vision*, pages 19–34. Springer, 2022. 2, 5

[20] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 7

[21] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[23] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2633–2651, 2015. 1

[24] 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

[25] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2, 6, 7

[26] 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, 3, 6, 7, 8

[27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1

[28] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, page 435. Seattle, WA, 2009. 5

[29] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022. 1, 2

[30] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022. 2

[31] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 2, 3, 6, 7, 8

[32] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 1, 2

[33] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning*, 2018. 1

[34] 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

[35] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 2

[36] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2

[37] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2

[38] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3523–3532, 2019. 4, 5

[39] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023. 3

[40] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *Proceedings of the European Conference on Computer Vision*, pages 230–247. Springer, 2020. 3

[41] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 3

[42] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3109–3118, 2015. 2, 3, 6

[43] 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. 2

[44] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3D objects. In *British Machine Vision Conference (BMVC)*, 2019. 3

[45] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2666–2674, 2018. 3

[46] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5845–5854, 2019. 3

[47] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 592–611. Springer, 2022. 1, 2, 3, 6, 7, 8

[48] Chen Zhao, Zhiguo Cao, Chi Li, Xin Li, and Jiaqi Yang. Nm-net: Mining reliable neighbors for robust feature correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 215–224, 2019. 3

[49] Chen Zhao, Yixiao Ge, Feng Zhu, Rui Zhao, Hongsheng Li, and Mathieu Salzmann. Progressive correspondence pruning by consensus learning. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 6464–6473, 2021. 3

[50] Chen Zhao, Yinlin Hu, and Mathieu Salzmann. Locposenet: Robust location prior for unseen object pose estimation. *arXiv preprint arXiv:2211.16290v2*, 2022. 3

[51] Chen Zhao, Yinlin Hu, and Mathieu Salzmann. Fusing local similarities for retrieval-based 3d orientation estimation of unseen objects. In *Proceedings of the European Conference on Computer Vision*, pages 106–122. Springer, 2022. 2, 3, 6

[52] Chen Zhao, Tong Zhang, and Mathieu Salzmann. 3d-aware hypothesis & verification for generalizable relative object pose estimation. *arXiv preprint arXiv:2310.03534*, 2023. 1, 2, 3, 6, 7, 8

[53] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3, 6, 1