DCL-Net: Deep Correspondence Learning Network for 6D Pose Estimation

Hongyang Li[1*], Jiehong Lin[1,2*], and Kui Jia[1,3†]

[1] South China University of Technology
[2] DexForce Co. Ltd.
[3] Peng Cheng Laboratory
{eeli.hongyang,lin.jiehong}@mail.scut.edu.cn, kuijia@scut.edu.cn

Abstract. Establishment of point correspondence between camera and object coordinate systems is a promising way to solve 6D object poses. However, surrogate objectives of correspondence learning in 3D space are a step away from the true ones of object pose estimation, making the learning suboptimal for the end task. In this paper, we address this shortcoming by introducing a new method of *Deep Correspondence Learning Network* for direct 6D object pose estimation, shortened as *DCL-Net*. Specifically, DCL-Net employs dual newly proposed *Feature Disengagement and Alignment (FDA) modules* to establish, in the feature space, partial-to-partial correspondence and complete-to-complete one for partial object observation and its complete CAD model, respectively, which result in aggregated pose and match feature pairs from two coordinate systems; these two FDA modules thus bring complementary advantages. The match feature pairs are used to learn confidence scores for measuring the qualities of deep correspondence, while the pose feature pairs are weighted by confidence scores for direct object pose regression. A confidence-based pose refinement network is also proposed to further improve pose precision in an iterative manner. Extensive experiments show that DCL-Net outperforms existing methods on three benchmarking datasets, including YCB-Video, LineMOD, and Oclussion-LineMOD; ablation studies also confirm the efficacy of our novel designs. Our code is released publicly at https://github.com/Gorilla-Lab-SCUT/DCL-Net.

Keywords: 6D Pose Estimation, Correspondence Learning

1 Introduction

6D object pose estimation is a fundamental task of 3D semantic analysis with many real-world applications, such as robotic grasping [7,44], augmented reality [27], and autonomous driving [8,9,21,42]. Non-linearity of the rotation space of $SO(3)$ makes it hard to handle this nontrivial task through direct pose regression from object observations [6,11,15,18,24–26,39,45,47]. Many of the data-driven methods [3,14,20,23,28,31,33,34,38,41] thus achieve the estimation by learning point correspondence between camera and object coordinate systems.

*Equal contribution
†Corresponding author

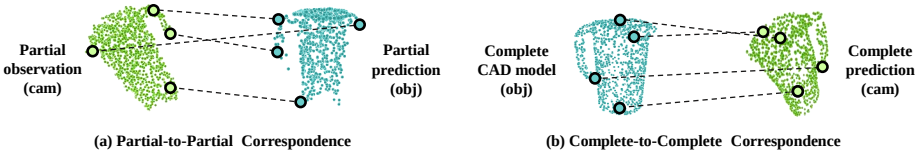

Fig. 1. Illustrations of two kinds of point correspondence between camera coordinate system (cam) and object coordinate system (obj). Best view in the electronic version.

Given a partial object observation in camera coordinate system along with its CAD model in object coordinate one, we show in Fig. 1 two possible ways to build point correspondence: i) inferring the observed points in object coordinate system for partial-to-partial correspondence; ii) inferring the sampled points of CAD model in camera coordinate system for complete-to-complete correspondence. These two kinds of correspondence show different advantages. The partial-to-partial correspondence is of higher qualities than the complete-to-complete one due to the difficulty in shape completion, while the latter is more robust to figure out poses for objects with severe occlusions, which the former can hardly handle with.

While these methods are promising by solving 6D poses from point correspondence (*e.g.*, via a PnP algorithm), their surrogate correspondence objectives are a step away from the true ones of estimating 6D object poses, thus making their learnings suboptimal for the end task [40]. To this end, we present a novel method to realize the above two ways of correspondence establishment in the feature space via dual newly proposed *Feature Disengagement and Alignment (FDA) modules*, and directly estimate object poses from feature pairs of two coordinate systems, which are weighted by confidence scores measuring the qualities of deep correspondence. We term our method as *Deep Correspondence Learning Network*, shortened as *DCL-Net*. Fig. 2 gives the illustration.

For the partial object observation and its CAD model, DCL-Net firstly extracts their point-wise feature maps in parallel; then dual Feature Disengagement and Alignment (FDA) modules are designed to establish, in feature space, the partial-to-partial correspondence and the complete-to-complete one between camera and object coordinate systems. Specifically, each FDA module takes as inputs two point-wise feature maps, and disengages each feature map into individual pose and match ones; the match feature maps of two systems are then used to learn an attention map for building deep correspondence; finally, both pose and match feature maps are aligned and paired across systems based on the attention map, resulting in pose and match feature pairs, respectively. DCL-Net aggregates two sets of correspondence together, since they bring complementary advantages, by fusing the respective pose and match feature pairs of two FDA modules. The aggregated match feature pairs are used to learn confidence scores for measuring the qualities of deep correspondence, while the pose ones are weighted by the scores to directly regress object poses. A confidence-based pose refinement network is also proposed to further improve the results of DCL-

Net in an iterative manner. Extensive experiments show that DCL-Net outperforms existing methods for 6D object pose estimation on three well-acknowledged datasets, including YCB-Video [4], LineMOD [16], and Occlusion-LineMOD [3]; remarkably, on the more challenging Occlusion-LineMOD, our DCL-Net outperforms the state-of-the-art method [13] with an improvement of 4.4% on the metric of ADD(S), revealing the strength of DCL-Net on handling with occlusion. Ablation studies also confirm the efficacy of individual components of DCL-Net. Our technical contributions are summarized as follows:

- We design a novel *Feature Disengagement and Alignment (FDA) module* to establish deep correspondence between two point-wise feature maps from different coordinate systems; more specifically, FDA module disengages each feature map into individual pose and match ones, which are then aligned across systems to generate pose and match feature pairs, respectively, such that deep correspondence is established within the aligned feature pairs.
- We propose a new method of *Deep Correspondence Learning Network* for direct regression of 6D object poses, termed as DCL-Net, which employs dual FDA modules to establish, in feature space, partial-to-partial correspondence and complete-to-complete one between camera and object coordinate systems, respectively; these two FDA modules bring complementary advantages.
- Match feature pairs of dual FDA modules are aggregated and used for learning of confidence scores to measure the qualities of correspondence, while pose feature pairs are weighted by the scores for estimation of 6D pose; a confidence-based pose refinement network is also proposed to iteratively improve pose precision.

2 Related Work

6D Pose Estimation from RGB Data This body of works can be broadly categorized into three types: i) holistic methods [11,15,18] for directly estimating object poses; ii) keypoint-based methods [28,33,34], which establish 2D-3D correspondence via 2D keypoint detection, followed by a PnP/RANSAC algorithm to solve the poses; iii) dense correspondence methods [3,20,23,31], which make dense pixel-wise predictions and vote for the final results.

Due to loss of geometry information, these methods are sensitive to lighting conditions and appearance textures, and thus inferior to the RGB-D methods.

6D Pose Estimation from RGB-D Data Depth maps provide rich geometry information complementary to appearance one from RGB images. Traditional methods [3, 16, 32, 37, 43] solve object poses by extracting features from RGB-D data and performing correspondence grouping and hypothesis verification. Earlier deep methods, such as PoseCNN [45] and SSD-6D [19], learn coarse poses firstly from RGB images, and refine the poses on point clouds by using ICP [2] or MCN [22]. Recently, learning deep features of point clouds becomes an efficient

way to improve pose precision, especially for methods [39,47] of direct regression, which make efforts to enhance pose embeddings from deep geometry features, due to the difficulty in the learning of rotations from a nonlinear space. Wang *et al.* present DenseFusion [39], which fuses local features of RGB images and point clouds in a point-wise manner, and thus explicitly reasons about appearance and geometry information to make the learning more discriminative; due to the incomplete and noisy shape information, Zhou *et al.* propose PR-GCN [47] to polish point clouds and enhance pose embeddings via Graph Convolutional Network. On the other hand, dense correspondence methods show the advantages of deep networks on building the point correspondence in Euclidean space; for example, He *et al.* propose PVN3D [14] to regress dense keypoints, and achieve remarkable results. While promising, these methods are usually trained with surrogate objectives instead of the true ones of estimating 6D poses, making the learning suboptimal for the end task.

Our proposed DCL-Net borrows the idea from dense correspondence methods by learning deep correspondence in feature space, and weights the feature correspondence based on confidence scores for direct estimation of object poses. Besides, the learned correspondence is also utilized by an iterative pose refinement network for precision improvement.

3 Deep Correspondence Learning Network

Given the partial object observation $\mathcal{X}_c$ in the camera coordinate system, along with the object CAD model $\mathcal{Y}_o$ in the object coordinate one, our goal is to estimate the 6D pose $(\boldsymbol{R}, \boldsymbol{t})$ between these two systems, where $\boldsymbol{R} \in SO(3)$ stands for a rotation, and $\boldsymbol{t} \in \mathbb{R}^3$ for a translation.

Fig. 2 gives the illustration of our proposed *Deep Correspondence Learning Network* (dubbed *DCL-Net*). DCL-Net firstly extracts point-wise features of $\mathcal{X}_c$ and $\mathcal{Y}_o$ (cf. Sec. 3.1), then establishes correspondence in feature space via *dual Feature Disengagement and Alignment modules* (cf. Sec. 3.2), and finally regresses the object pose $(\boldsymbol{R}, \boldsymbol{t})$ with confidence scores based on the learned deep correspondence (cf. Sec. 3.3). The training objectives of DCL-Net are given in Sec. 3.4. A confidence-based pose refinement network is also introduced to iteratively improve pose precision (cf. Sec. 3.5).

3.1 Point-wise Feature Extraction

We represent the inputs of the object observation $\mathcal{X}_c$ and its CAD model $\mathcal{Y}_o$ as $(\boldsymbol{I}^{\mathcal{X}_c}, \boldsymbol{P}^{\mathcal{X}_c})$ and $(\boldsymbol{I}^{\mathcal{Y}_o}, \boldsymbol{P}^{\mathcal{Y}_o})$ with $N_{\mathcal{X}}$ and $N_{\mathcal{Y}}$ sampled points, respectively, where $\boldsymbol{P}$ denotes a point set, and $\boldsymbol{I}$ denotes RGB values corresponding to points in $\boldsymbol{P}$. As shown in Fig. 2, we use two parallel backbones to extract their point-wise features $\boldsymbol{F}^{\mathcal{X}_c}$ and $\boldsymbol{F}^{\mathcal{Y}_o}$, respectively. Following [12], both backbones are built based on 3D Sparse Convolutions [10], of which the volumetric features are then converted to point-level ones; more details about the architectures are given in the supplementary material. Note that for each object instance, $\boldsymbol{F}^{\mathcal{Y}_o}$ can be pre-computed during inference for efficiency.

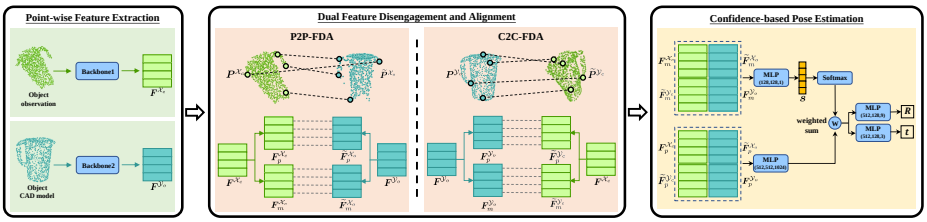

Fig. 2. An illustration of DCL-Net. Given object observation and its CAD model, DCL-Net first extracts their point-wise features $\boldsymbol{F}^{\mathcal{X}_c}$ and $\boldsymbol{F}^{\mathcal{Y}_o}$, separately; then dual Feature Disengagement and Alignment (FDA) modules are employed to establish, in feature space, partial-to-partial correspondence and complete-to-complete one between camera and object coordinate systems, respectively, which result in aggregated pose and match feature pairs; the match feature pairs are used to learn confidence scores $\boldsymbol{s}$ for measuring the qualities of deep correspondence, while the pose ones are weighted by $\boldsymbol{s}$ for estimating 6D object pose $(\boldsymbol{R}, \boldsymbol{t})$. Best view in the electronic version.

3.2 Dual Feature Disengagement and Alignment

The key to figure out the pose between the object observation and its CAD model lies in the establishment of correspondence. As pointed out in Sec. 1, there exist at least two ways to achieve this goal: i) learning the partial point set $\widetilde{\boldsymbol{P}}^{\mathcal{X}_o}$ in object system from complete $\boldsymbol{P}^{\mathcal{Y}_o}$ to pair with $\boldsymbol{P}^{\mathcal{X}_c}$, *e.g.*, $(\boldsymbol{P}^{\mathcal{X}_c}, \widetilde{\boldsymbol{P}}^{\mathcal{X}_o})$, for partial-to-partial correspondence; ii) inferring the complete point set $\widetilde{\boldsymbol{P}}^{\mathcal{Y}_c}$ in camera coordinate system from partial $\boldsymbol{P}^{\mathcal{X}_c}$ to pair with $\boldsymbol{P}^{\mathcal{Y}_o}$, *e.g.*, $(\widetilde{\boldsymbol{P}}^{\mathcal{Y}_c}, \boldsymbol{P}^{\mathcal{Y}_o})$, for complete-to-complete correspondence.

In this paper, we propose to establish the correspondence in the deep feature space, from which *pose feature pairs* along with *match feature pairs* can be generated for the learning of object pose and confidence scores, respectively. Fig. 2 gives illustrations of the correspondence in both 3D space and feature space. Specifically, we design a novel *Feature Disengagement and Alignment (FDA) module* to learn the pose feature pairs, *e.g.*, $(\boldsymbol{F}_p^{\mathcal{X}_c}, \widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o})$ and $(\widetilde{\boldsymbol{F}}_p^{\mathcal{Y}_c}, \boldsymbol{F}_p^{\mathcal{Y}_o})$ *w.r.t* the above $(\boldsymbol{P}^{\mathcal{X}_c}, \widetilde{\boldsymbol{P}}^{\mathcal{X}_o})$ and $(\widetilde{\boldsymbol{P}}^{\mathcal{Y}_c}, \boldsymbol{P}^{\mathcal{Y}_o})$, respectively, and the match feature pairs, *e.g.*, $(\boldsymbol{F}_m^{\mathcal{X}_c}, \widetilde{\boldsymbol{F}}_m^{\mathcal{X}_o})$ and $(\widetilde{\boldsymbol{F}}_m^{\mathcal{Y}_c}, \boldsymbol{F}_m^{\mathcal{Y}_o})$, which can be formulated as follows:

$$\boldsymbol{F}_p^{\mathcal{X}_c}, \boldsymbol{F}_m^{\mathcal{X}_c}, \widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o}, \widetilde{\boldsymbol{F}}_m^{\mathcal{X}_o}, \widetilde{\boldsymbol{P}}^{\mathcal{X}_o} = \texttt{FDA}(\boldsymbol{F}^{\mathcal{X}_c}, \boldsymbol{F}^{\mathcal{Y}_o}), \tag{1}$$

$$\boldsymbol{F}_p^{\mathcal{Y}_o}, \boldsymbol{F}_m^{\mathcal{Y}_o}, \widetilde{\boldsymbol{F}}_p^{\mathcal{Y}_c}, \widetilde{\boldsymbol{F}}_m^{\mathcal{Y}_c}, \widetilde{\boldsymbol{P}}^{\mathcal{Y}_c} = \texttt{FDA}(\boldsymbol{F}^{\mathcal{Y}_o}, \boldsymbol{F}^{\mathcal{X}_c}). \tag{2}$$

We term the partial-to-partial (1) and complete-to-complete (2) FDA modules as P2P-FDA and C2C-FDA modules, respectively.

Feature Disengagement and Alignment Module Feature Disengagement and Alignment (FDA) module takes point-wise feature maps of different coordinate systems as inputs, disengages each feature map into pose and match ones,

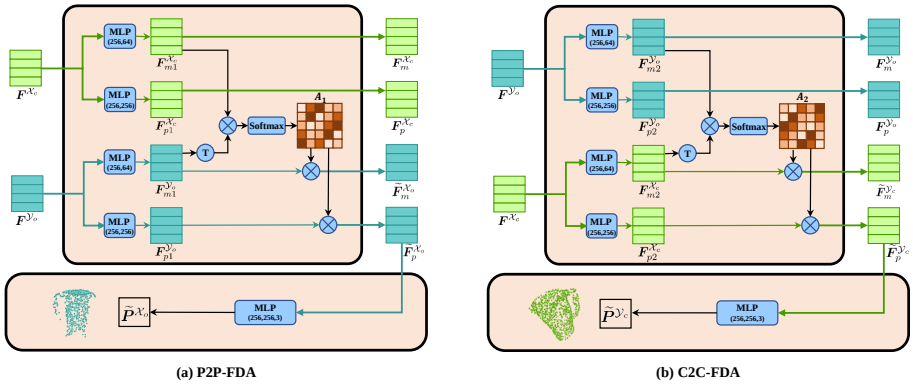

Fig. 3. Illustrations of dual Feature Disengagement and Alignment modules. "T" denotes matrix transposition, and "×" denotes matrix multiplication.

which are then aligned across systems to establish deep correspondence. Fig. 3 gives illustrations of both P2P-FDA and C2C-FDA modules, where network specifics are also given.

We take P2P-FDA module (1) as an example to illustrate the implementation of FDA. Specifically, as shown in Fig. 3, we firstly disengage $\boldsymbol{F}^{\mathcal{X}_c}$ into a pose feature $\boldsymbol{F}_{p1}^{\mathcal{X}_c}$ and a match one $\boldsymbol{F}_{m1}^{\mathcal{X}_c}$:

$$\boldsymbol{F}_{p1}^{\mathcal{X}_c} = \texttt{MLP}(\boldsymbol{F}^{\mathcal{X}_c}), \boldsymbol{F}_{m1}^{\mathcal{X}_c} = \texttt{MLP}(\boldsymbol{F}^{\mathcal{X}_c}), \tag{3}$$

where $\texttt{MLP}(\cdot)$ denotes a subnetwork of Multi-layer Perceptron (MLP). The same applies to $\boldsymbol{F}^{\mathcal{Y}_o}$, and we have $\boldsymbol{F}_{p1}^{\mathcal{Y}_o}$ and $\boldsymbol{F}_{m1}^{\mathcal{Y}_o}$. The match features $\boldsymbol{F}_{m1}^{\mathcal{X}_c}$ and $\boldsymbol{F}_{m1}^{\mathcal{Y}_o}$ is then used for the learning of an attention map $\boldsymbol{A}_1 \in \mathbb{R}^{N_{\mathcal{X}} \times N_{\mathcal{Y}}}$ as follows:

$$\boldsymbol{A}_1 = \texttt{Softmax}(\boldsymbol{F}_{m1}^{\mathcal{X}_c} \times \texttt{Transpose}(\boldsymbol{F}_{m1}^{\mathcal{Y}_o})), \tag{4}$$

where $\texttt{Transpose}(\cdot)$ denotes tensor transposition, and $\texttt{Softmax}(\cdot)$ denotes softmax operation along columns. Each element $a_{1,ij}$ in $\boldsymbol{A}_1$ indicates the match degree between i^{th} point in $\boldsymbol{P}^{\mathcal{X}_c}$ and j^{th} one in $\boldsymbol{P}^{\mathcal{Y}_o}$. Then pose and match features of the partial observation $\mathcal{X}_o$ in object system can be interpolated by matrix multiplication of $\boldsymbol{A}_1$ and those of $\boldsymbol{P}^{\mathcal{Y}_o}$, respectively, to be aligned with features of $\mathcal{X}_c$ in camera coordinate system:

$$\begin{cases} \boldsymbol{F}_p^{\mathcal{X}_c} = \boldsymbol{F}_{p1}^{\mathcal{X}_c} \\ \widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o} = \boldsymbol{A}_1 \times \boldsymbol{F}_{p1}^{\mathcal{Y}_o} \end{cases}, \begin{cases} \boldsymbol{F}_m^{\mathcal{X}_c} = \boldsymbol{F}_{m1}^{\mathcal{X}_c} \\ \widetilde{\boldsymbol{F}}_m^{\mathcal{X}_o} = \boldsymbol{A}_1 \times \boldsymbol{F}_{m1}^{\mathcal{Y}_o} \end{cases}. \tag{5}$$

Through feature alignment, $\widetilde{\boldsymbol{P}}^{\mathcal{X}_o}$ is expected to be decoded out from $\widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o}$:

$$\widetilde{\boldsymbol{P}}^{\mathcal{X}_o} = \texttt{MLP}(\widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o}). \tag{6}$$

Supervisions on the reconstruction of $\widetilde{\boldsymbol{P}}^{\mathcal{X}_o}$ guide the learning of deep correspondence in P2P-FDA module.

P2P-FDA module (1) learns deep correspondence of the partial $\mathcal{X}$ in two coordinate systems, while C2C-FDA module (2) infers that of the complete $\mathcal{Y}$ via a same network structure, as shown in Fig. 3(b). We adopt dual FDA modules in our design to enable robust correspondence establishment, since they bring complementary functions: P2P-FDA module provides more accurate correspondence than that of C2C-FDA module, due to the difficulty in shape completion from partial observation for the latter module; however, C2C-FDA module plays a vital role under the condition of severe occlusions, which P2P-FDA module can hardly handle with.

3.3 Confidence-based Pose Estimation

After dual feature disengagement and alignment, we construct the pose and match feature pairs as follows:

$$\boldsymbol{F}_p = \begin{bmatrix} \boldsymbol{F}_p^{\mathcal{X}_c}, \widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o} \\ \widetilde{\boldsymbol{F}}_p^{\mathcal{Y}_c}, \boldsymbol{F}_p^{\mathcal{Y}_o} \end{bmatrix}, \boldsymbol{F}_m = \begin{bmatrix} \boldsymbol{F}_m^{\mathcal{X}_c}, \widetilde{\boldsymbol{F}}_m^{\mathcal{X}_o} \\ \widetilde{\boldsymbol{F}}_m^{\mathcal{Y}_c}, \boldsymbol{F}_m^{\mathcal{Y}_o} \end{bmatrix}. \tag{7}$$

As shown in Fig. 2, the paired match feature $\boldsymbol{F}_m$ is fed into an MLP for the learning of confidence scores $\boldsymbol{s} = \{s_i\}_{i=1}^{N_\mathcal{X}+N_\mathcal{Y}}$ to reflect the qualities of deep correspondence:

$$\boldsymbol{s} = \texttt{MLP}(\boldsymbol{F}_m). \tag{8}$$

The paired pose feature $\boldsymbol{F}_p$ is also fed into an MLP and weighted by $\boldsymbol{s}$ for precisely estimating the 6D pose $(\boldsymbol{R}, \boldsymbol{t})$:

$$\begin{aligned} \boldsymbol{R} = \texttt{MLP}(\boldsymbol{f}), \boldsymbol{t} = \texttt{MLP}(\boldsymbol{f}), \\ s.t.\ \boldsymbol{f} = \texttt{SUM}(\texttt{SoftMax}(\boldsymbol{s}) \cdot \texttt{MLP}(\boldsymbol{F}_p)), \end{aligned} \tag{9}$$

where SUM denotes summation along rows.

Rather than numerical calculation from two paired point sets, we directly regress the 6D object pose from deep pair-wise features with confidence scores, which effectively weakens the negative impact of correspondence of low quality on pose estimation, and thus realizes more precise results.

3.4 Training of Deep Correspondence Learning Network

For dual FDA modules, we supervise the reconstruction of $\widetilde{\boldsymbol{P}}^{\mathcal{X}_o} = \{\widetilde{\boldsymbol{p}}_i^{\mathcal{X}_o}\}_{i=1}^{N_\mathcal{X}}$ and $\widetilde{\boldsymbol{P}}^{\mathcal{Y}_c} = \{\widetilde{\boldsymbol{p}}_i^{\mathcal{Y}_c}\}_{i=1}^{N_\mathcal{Y}}$ to guide the learning of deep correspondence via the following objectives:

$$\mathcal{L}_{p2p} = \frac{1}{N_\mathcal{X}} \sum_{i=1}^{N_\mathcal{X}} ||\widetilde{\boldsymbol{p}}_i^{\mathcal{X}_o} - \boldsymbol{R}^{*T}(\boldsymbol{p}_i^{\mathcal{X}_c} - \boldsymbol{t}^*)||, \tag{10}$$

$$\mathcal{L}_{c2c} = \frac{1}{N_\mathcal{Y}} \sum_{i=1}^{N_\mathcal{Y}} ||\widetilde{\boldsymbol{p}}_i^{\mathcal{Y}_c} - (\boldsymbol{R}^* \boldsymbol{p}_i^{\mathcal{Y}_o} + \boldsymbol{t}^*)||, \tag{11}$$

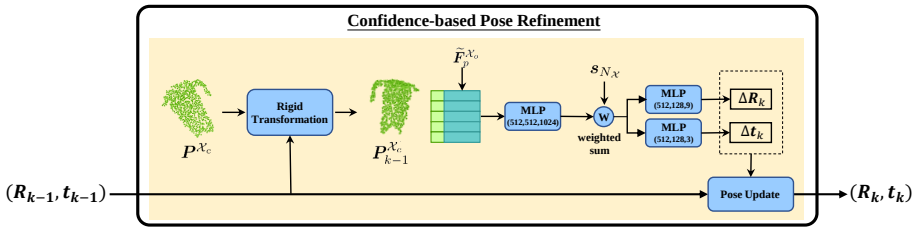

Fig. 4. An illustration of the iterative confidence-based pose estimation network.

where $\boldsymbol{P}^{\mathcal{X}_c} = \{\boldsymbol{p}_i^{\mathcal{X}_c}\}_{i=1}^{N_\mathcal{X}}$ and $\boldsymbol{P}^{\mathcal{Y}_o} = \{\boldsymbol{p}_i^{\mathcal{Y}_o}\}_{i=1}^{N_\mathcal{Y}}$ are input point sets, and $\boldsymbol{R}^*$ and $\boldsymbol{t}^*$ denote ground truth 6D pose. For the confidence-based pose estimation, we use the following objectives on top of the learning of the predicted object pose $(\boldsymbol{R}, \boldsymbol{t})$ and confidence scores $\boldsymbol{s} = \{s_i\}_{i=1}^{N_\mathcal{X}+N_\mathcal{Y}}$, respectively:

$$\mathcal{L}_{pose} = \frac{1}{N_\mathcal{Y}} \sum_{i=1}^{N_\mathcal{Y}} ||\boldsymbol{R}\boldsymbol{p}_i^{\mathcal{Y}_o} + \boldsymbol{t} - (\boldsymbol{R}^*\boldsymbol{p}_i^{\mathcal{Y}_o} + \boldsymbol{t}^*)||. \tag{12}$$

$$\begin{aligned}\mathcal{L}_{conf} = &\frac{1}{N_\mathcal{X}} \sum_{i=1}^{N_\mathcal{X}} \sigma(||\widetilde{\boldsymbol{p}}_i^{\mathcal{X}_o} - \boldsymbol{R}^T(\boldsymbol{p}_i^{\mathcal{X}_c} - \boldsymbol{t})||, s_i) \\ &+ \frac{1}{N_\mathcal{Y}} \sum_{j=1}^{N_\mathcal{Y}} \sigma(||\widetilde{\boldsymbol{p}}_j^{\mathcal{Y}_c} - (\boldsymbol{R}\boldsymbol{p}_j^{\mathcal{Y}_o} + \boldsymbol{t})||, s_{N_\mathcal{X}+j}),\end{aligned} \tag{13}$$

where $\sigma(d, s) = ds - wlog(s)$, and w is a balancing hyperparameter. We note that the objectives (10), (11) and (12) are designed for asymmetric objects, while for symmetric ones, we modify them by replacing L_2 distance with Chamfer distance, as done in [39].

The overall training objective combines (10), (11), (12), and (13), resulting in the following optimization problem:

$$\min \mathcal{L} = \lambda_1\mathcal{L}_{p2p} + \lambda_2\mathcal{L}_{c2c} + \lambda_3\mathcal{L}_{pose} + \lambda_4\mathcal{L}_{conf}, \tag{14}$$

where λ_1, λ_2, λ_3 and λ_4 are penalty parameters.

3.5 Confidence-based Pose Refinement

To take full advantages of the learned correspondence, we propose a confidence-based pose refinement network, as shown in Fig. 4, where the input point set $\boldsymbol{P}^{\mathcal{X}_c}$ is transformed with predicted pose, and paired with $\widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o}$ for residual pose estimation in an iterative manner. Specifically, assuming after $k-1$ iterations of refinement, the current object pose is updated as $(\boldsymbol{R}_{k-1}, \boldsymbol{t}_{k-1})$, and we use it for transforming $\boldsymbol{P}^{\mathcal{X}_c} = \{\boldsymbol{p}_i^{\mathcal{X}_c}\}_{i=1}^{N_\mathcal{X}}$ to $\boldsymbol{P}_{k-1}^{\mathcal{X}_c} = \{\boldsymbol{R}_{k-1}^T(\boldsymbol{p}_i^{\mathcal{X}_c} - \boldsymbol{t}_{k-1})\}_{i=1}^{N_\mathcal{X}}$; for

forming pair-wise pose features with the learned correspondence in dual FDA modules, we reuse $\widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o}$ by concatenating it with $\boldsymbol{P}_{k-1}^{\mathcal{X}_c}$. Similarly to Sec. 3.3, we feed the pose feature pairs into an MLP, and weight them by reusing the confidence scores $\boldsymbol{s}_{N_\mathcal{X}}$ (denoting the first $N_\mathcal{X}$ elements of $\boldsymbol{s}$) for estimating the residual pose $(\Delta\boldsymbol{R}_k, \Delta\boldsymbol{t}_k)$:

$$\begin{aligned} \Delta\boldsymbol{R}_k = \texttt{MLP}(\boldsymbol{f}_k), \Delta\boldsymbol{t}_k = \texttt{MLP}(\boldsymbol{f}_k), \\ s.t.\ \boldsymbol{f}_k = \texttt{SUM}(\texttt{SoftMax}(\boldsymbol{s}_{N_\mathcal{X}}) \cdot \texttt{MLP}([\boldsymbol{P}_{k-1}^{\mathcal{X}_c}, \widetilde{\boldsymbol{F}}_p^{\mathcal{X}_o}])). \end{aligned} \tag{15}$$

Finally, the pose $(\boldsymbol{R}_k, \boldsymbol{t}_k)$ of the k^{th} iteration can be obtained as follows:

$$\boldsymbol{R}_k = \Delta\boldsymbol{R}_k\boldsymbol{R}_{k-1}, \boldsymbol{t}_k = \boldsymbol{R}_{k-1}\Delta\boldsymbol{t}_k + \boldsymbol{t}_{k-1}. \tag{16}$$

4 Experiments

Datasets We conduct experiments on three benchmarking datasets, including YCB-Video [4], LineMOD [16], and Occlusion-LineMOD [3]. YCB-Video dataset consists of 92 RGB-D videos with 21 different object instances, fully annotated with object poses and masks. Following [39], we use 80 videos therein for training along with additional 80,000 synthetic images, and evaluate DCL-Net on 2,949 keyframes sampled from the rest 12 videos. LineMOD is also a fully annotated dataset for 6D pose estimation, containing 13 videos with 13 low-textured object instances; we follow the prior work [39] to split training and testing sets. Occlusion-LineMOD is an annotated subset of LineMOD with 8 different object instances, which handpicks RGB-D images of scenes with heavy object occlusions and self-occlusions from LineMOD, making the task of pose estimation more challenging; following [35], we use the DCL-Net trained on the original LineMOD to evaluate on Occlusion-LineMOD.

Implementation Details For both object observations and CAD models, we sample point sets with 1,024 points as inputs of DCL-Net; that is, $N_\mathcal{X} = N_\mathcal{Y} = 1,024$. For the training objectives, we set the penalty parameters λ_1, λ_2, λ_3, λ_4 in (14) as 5.0, 1.0, 1.0, and 1.0, respectively; w in (13) is set as 0.01. During inference, we run twice the confidence-based pose refinement for improvement of pose precision.

Evaluation Metrics We use the same evaluation metrics as those in [39]. For YCB-Video dataset, the average closest point distance (ADD-S) [45] is employed to measure the pose error; following [39], we report the Area Under the Curve (AUC) of ADD-S with the maximum threshold at 0.1m, and the percentage of ADD-S smaller than the minimum tolerance at 2cm (< 2cm). For both LineMOD and Occlusion-LineMOD datasets, ADD-S is employed only for symmetric objects, while the Average Distance (ADD) for asymmetric objects; we report the percentage of distance smaller than 10% of object diameter. Besides, we use Chamfer Distance (CD) to measure the reconstruction results.

Table 1. Ablation studies of the use of dual FDA modules on YCB-Video dataset [4]. Experiments are conducted without confidence-based weighting and pose refinement.

P2P-FDA	C2C-FDA	AUC	< 2cm	CD ($\times 10^{-3}$)	
				$\boldsymbol{P}^{\mathcal{X}_o}$	$\boldsymbol{P}^{\mathcal{Y}_c}$
×	×	94.1	97.4	–	–
✓	×	95.0	98.7	7.1	–
×	✓	94.5	98.8	–	8.2
✓	✓	**95.3**	**99.0**	**7.0**	**8.1**

Table 2. Quantitative results obtained by least-squares optimization [1] and our proposed direct regression on YCB-Video dataset [4]. Experiments are conducted without pose refinement.

		AUC	< 2cm
w/o Conf.	Least-squares Optimization [1]	94.7	98.2
	Direct Pose Regression	95.3	99.0
with Conf.	Least-squares Optimization [1]	95.4	98.3
	Direct Pose Regression	**95.8**	**99.0**

4.1 Ablation Studies and Analyses

We firstly conduct ablation studies to evaluate the efficacy of novel designs proposed in our DCL-Net. These experiments are conducted on YCB-Video dataset [4].

Effects of Dual Feature Disengagement and Alignment We conduct four experiments to evaluate the efficacy of the use of dual FDA modules: i) without any FDA modules (baseline), ii) only with P2P-FDA, iii) only with C2C-FDA, and iv) with dual modules. For simplicity, these experiments are conducted without confidence-based weighting as well as pose refinement. The quantitative results on ADD-S AUC and ADD-S< 2cm are shown in Table 1, where the reconstruction results of asymmetric objects are also reported. From the table, methods with (one or dual) FDA modules indeed outperforms the baseline, which demonstrates the importance of deep correspondence learning on pose estimation. Single P2P-FDA module achieves more accurate results than single C2C-FDA module by making better reconstructions (7.1×10^{-3} versus 8.2×10^{-3} on CD) and deep correspondence as well, and the mixed use of them boosts the performance, indicating their complementary advantages. For the last framework, we visualize the reconstruction results along with the learned correspondence of both P2P-FDA and C2C-FDA modules in Fig. 5; shape completion can be achieved for C2C-FDA module, even with severe occlusions, to build valid deep correspondence of high quality, and thus make DCL-Net more robust and reliable.

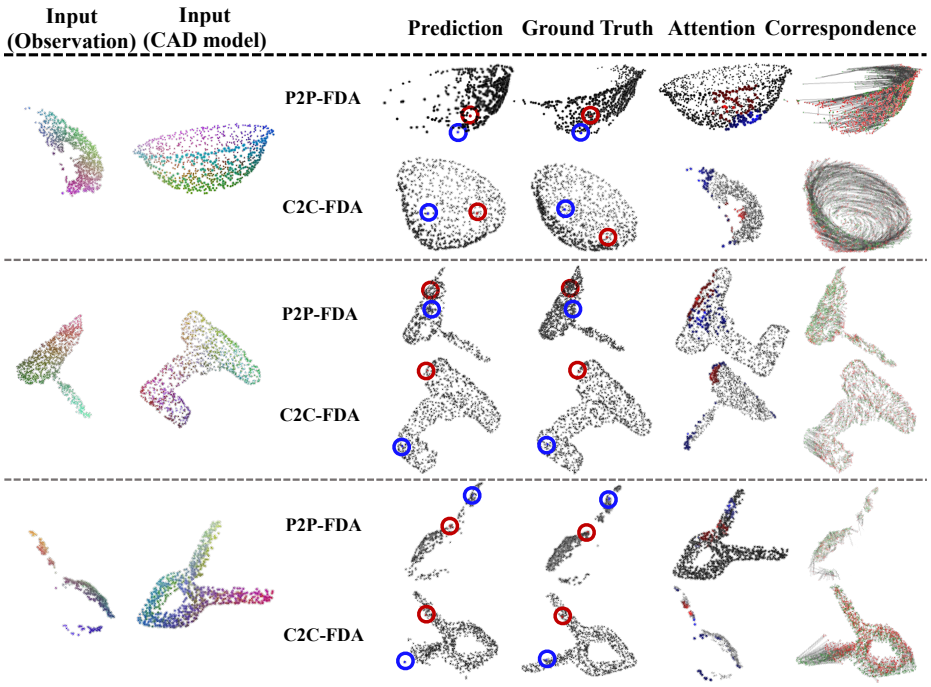

Fig. 5. Visualizations of shape predictions, attentions, and correspondence of both P2P-FDA and C2C-FDA modules on YCB-Video dataset [4]. Best view in electronic version.

We also explore the attention maps of dual FDA modules in Fig. 5. Take C2C-FDA module as an example, the predicted points are learned from the features of the input observed ones via attention maps, *i.e.*, each predicted point corresponds to the observed ones with different attention weights, and we thus colorize those corresponding points with large weights in Fig. 5; as shown in the figure, for the predicted points (red) locate at the observed parts, most of the input points with larger weights (red) could locate at the corresponding local regions, showing the qualities of attention maps, while for those at the occluded parts (blue), the corresponding points (blue) may locate scatteredly, but thanks to the correspondence learning in feature space, these points could still be completed in the C2C-FDA reconstruction results.

Effects of Confidence-based Pose Estimation Through learning deep correspondence in feature space, DCL-Net achieves direct regression of object poses, while the predictions of dual FDA modules can also establish point correspondence *w.r.t* inputs to solve poses via least-squares optimization [1]. We compare the quantitative results obtained by these two approaches (without pose refinement) in Table 2, where results of direct regression from deep feature correspon-

Table 3. Quantitative results of DCL-Net with or without pose refinement on YCB-Video dataset [4].

	AUC	< 2cm
w/o Pose Refinement	95.8	99.0
with Pose Refinement	**96.6**	**99.0**

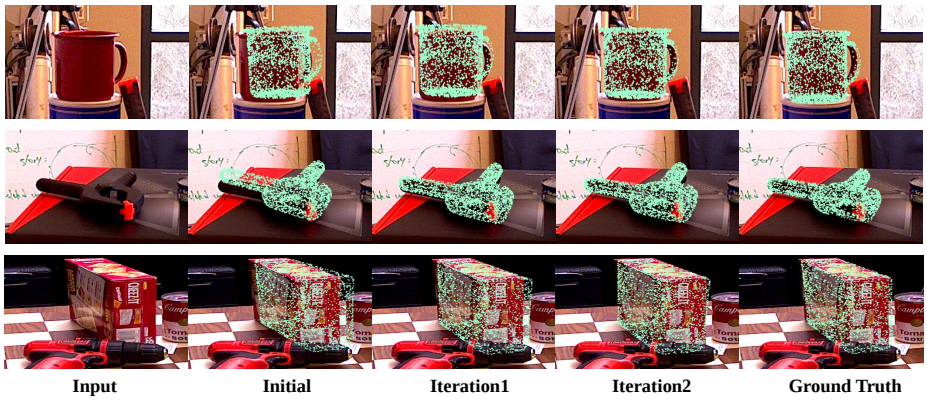

Fig. 6. Qualitative results of DCL-Net with or without pose refinement on YCB-Video dataset [4]. The sampled points of CAD models are transformed by the predicted poses and projected to 2D images.

dence outperforms those from point correspondence consistently with or without confidence scores, showing that pose estimation from feature space is less sensitive to the correspondence of low qualities, thanks to the direct objectives for the end task. Besides, we also observe that the learning of confidence scores not only measures the qualities of correspondence and decreases the influence of bad correspondence, but also helps improve the qualities themselves effectively.

Effects of Confidence-based Pose Refinement Table 4 demonstrates the efficiency of our confidence-based pose refinement for boosting the performance, *e.g.*, improvement by 0.8% on the metric of ADD-S AUC, which is also verified by the qualitative results shown in Fig. 6.

4.2 Comparisons with Existing Methods

We compare our proposed DCL-Net with the existing methods for 6D object pose estimation from RGB-D data, including those based on direct regression (*e.g.*, DenseFusion [39] and PR-GCN [47]), and those based on dense correspondence learning (*e.g.*, PVN3D [14] and FFB6D [13]).

Quantitative results on the three benchmarking datasets, including YCB-Video [4], LineMOD [16], and Occlusion-LineMOD [3], are reported in Table 4,

Table 4. Quantitative results of different methods on YCB-Video dataset [4]. The evaluation metrics are ADD-S AUC and ADD-S< 2cm. Objects with bold name are symmetric.

	PoseCNN+ICP [45]		DenseFusion [39]		G2L [5]	PVN3D [14]		PR-GCN [47]		FFB6D		DCL-Net	
	AUC	<2cm	AUC	<2cm	AUC	AUC	<2cm	AUC	<2cm	AUC	<2cm	AUC	<2cm
002_master_chef_can	95.8	**100.0**	96.4	**100.0**	94.0	96.0	**100.0**	**97.1**	**100.0**	96.3	**100.0**	96.1	**100.0**
003_cracker_box	92.7	91.6	95.5	99.5	88.7	96.1	**100.0**	**97.6**	**100.0**	96.3	**100.0**	96.4	99.4
004_sugar_box	98.2	**100.0**	97.5	**100.0**	96.0	97.4	**100.0**	**98.3**	**100.0**	97.6	**100.0**	98.1	**100.0**
005_tomato_soup_can	94.5	96.9	94.96	96.9	86.4	**96.2**	**98.1**	95.3	97.6	95.6	98.2	95.8	97.7
006_mustard_bottle	98.6	**100.0**	97.2	**100.0**	95.9	97.5	**100.0**	97.9	**100.0**	97.8	**100.0**	**98.7**	**100.0**
007_tuna_fish_can	97.1	**100.0**	96.6	**100.0**	84.1	96.0	**100.0**	97.6	**100.0**	96.8	**100.0**	97.4	**100.0**
008_pudding_box	97.9	**100.0**	96.5	**100.0**	93.5	97.1	**100.0**	**98.4**	**100.0**	97.1	**100.0**	98.2	**100.0**
009_gelatin_box	98.8	**100.0**	98.1	**100.0**	96.8	97.7	**100.0**	96.2	94.4	98.1	**100.0**	**98.9**	**100.0**
010_potted_meat_can	92.7	93.6	91.3	93.1	86.2	93.3	94.6	**96.6**	**99.1**	94.7	94.3	93.1	94.7
011_banana	97.1	99.7	96.6	**100.0**	96.3	96.6	**100.0**	**98.5**	**100.0**	97.2	**100.0**	98.1	**100.0**
019_pitcher_base	97.8	**100.0**	97.1	**100.0**	91.8	97.4	**100.0**	98.1	**100.0**	97.6	**100.0**	98.0	99.8
021_bleach_cleanser	96.9	99.4	95.8	**100.0**	92.0	96.0	**100.0**	**97.9**	**100.0**	96.8	**100.0**	97.0	**100.0**
024_bowl	81.0	54.9	88.2	98.8	86.7	90.2	80.5	90.3	96.6	96.3	**100.0**	**97.3**	**100.0**
025_mug	95.0	99.8	97.1	**100.0**	95.4	97.6	**100.0**	**98.1**	**100.0**	97.3	**100.0**	97.8	**100.0**
035_power_drill	**98.2**	99.6	96.0	98.7	95.2	96.7	**100.0**	98.1	**100.0**	97.2	**100.0**	98.0	**100.0**
036_wood_block	87.6	80.2	89.7	94.6	86.2	90.4	93.8	**96.0**	**100.0**	92.6	92.1	93.9	97.5
037_scissors	91.7	95.6	95.2	**100.0**	83.8	96.7	**100.0**	**96.7**	**100.0**	97.7	**100.0**	87.6	98.3
040_large_marker	97.2	99.7	97.5	**100.0**	96.8	96.7	99.8	97.9	**100.0**	96.6	**100.0**	97.8	99.8
051_large_clamp	75.2	74.9	72.9	79.2	94.4	93.6	93.6	87.5	93.3	96.8	**100.0**	95.7	98.6
052_extra_large_clamp	64.4	48.8	69.8	76.3	**92.3**	88.4	83.6	79.7	84.6	**96.0**	**98.6**	88.8	87.2
061_foam_brick	97.2	**100.0**	92.5	**100.0**	94.7	96.8	**100.0**	**97.8**	**100.0**	97.3	**100.0**	97.5	**100.0**
MEAN	93.0	93.2	93.1	96.8	92.4	95.5	97.6	95.8	98.5	**96.6**	**99.2**	**96.6**	99.0

Fig. 7. Qualitative results of different methods on YCB-Video dataset [4]. The sampled points of CAD models are transformed by the predicted poses and projected to 2D images.

Table 5, and Table 6, respectively, all of which show the superiority of our DCL-Net consistently in the regime of pose precision; qualitative results on YCB-

Table 5. Quantitative results of different methods on ADD(S) on LineMOD dataset [16]. Objects with bold name are symmetric.

	Implicit +ICP [36]	SSD6D +ICP [19]	PointFusion [46]	DenseFusion [39]	DenseFusion (Iterative) [39]	G2L [5]	PR-GCN [47]	DCL-Net
ape	20.6	65	70.4	79.5	92.3	96.8	**97.6**	97.4
bench	64.3	80	80.7	84.2	93.2	96.1	99.2	**99.4**
camera	63.2	78	60.8	76.5	94.4	98.2	99.4	**99.8**
can	76.1	86	61.1	86.6	93.1	98.0	98.4	**99.9**
cat	72.0	70	79.1	88.8	96.5	99.2	98.7	**100.0**
driller	41.6	73	47.3	77.7	87.0	99.8	98.8	**99.9**
duck	32.4	66	63.0	76.3	92.3	97.7	**98.9**	98.4
egg	98.6	**100**	99.9	99.9	99.8	**100.0**	99.9	**100.0**
glue	96.4	**100**	99.3	99.4	**100.0**	**100.0**	**100.0**	99.9
hole	49.9	49	71.8	79.0	92.1	99.0	99.4	**100.0**
iron	63.1	78	83.2	92.1	97.0	99.3	98.5	**100.0**
lamp	91.7	73	62.3	92.3	95.3	**99.5**	99.2	**99.5**
phone	71.0	79	78.8	88.0	92.8	98.9	98.4	**99.7**
MEAN	64.7	79	73.7	86.2	94.3	98.7	98.9	**99.5**

Table 6. Quantitative results of different methods on ADD(S) on Occlusion-LineMOD dataset [3]. Objects with bold name are symmetric.

	PoseCNN [45]	Deep-Heat [29]	SS [17]	Pix2pose [30]	PVNet [31]	Hybrid-Pose [35]	PVN3D [14]	PR-GCN [47]	FFB6D [13]	DCL-Net
ape	9.6	12.1	17.6	22.0	15.8	20.9	33.9	40.2	47.2	**56.7**
can	45.2	39.9	53.9	44.7	63.3	75.3	**88.6**	76.2	85.2	80.2
cat	0.9	8.2	3.3	22.7	16.7	24.9	39.1	**57.0**	45.7	48.1
driller	41.4	45.2	62.4	44.7	65.7	70.2	78.4	**82.3**	81.4	81.4
duck	19.6	17.2	19.2	15.0	25.2	27.9	41.9	30.0	**53.9**	44.6
egg	22.0	22.1	25.9	25.2	50.2	52.4	80.9	68.2	70.2	**83.6**
glue	38.5	35.8	39.6	32.4	49.6	53.8	68.1	67.0	60.1	**79.1**
hole	22.1	36.0	21.3	49.5	39.7	54.2	74.7	**97.2**	85.9	91.3
MEAN	24.9	27.0	27.0	32.0	40.8	47.5	63.2	65.0	66.2	**70.6**

Video dataset [4] are also provided in Fig. 7 to verify the advantages of our DCL-Net. Remarkably, on the more challenging Occlusion-LineMOD dataset, the improvements of our DCL-Net over the state-of-the-art methods of PR-GCN [47] and FFB6D [13] reach 5.6% and 4.4% on the metric of ADD(S), respectively, indicating the advantages of our DCL-Net on handling with object occlusions or self-occlusions.

Acknowledgements. This work is supported in part by Guangdong R&D key project of China (No.: 2019B010155001), and the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (No.: 2017ZT07X183). We also thank Yi Li and Xun Xu for their valuable comments.

References

1. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-d point sets. IEEE Transactions on pattern analysis and machine intelligence (5), 698–700 (1987) 10, 11
2. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. International Society for Optics and Photonics (1992) 3
3. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6d object pose estimation using 3d object coordinates. In: European conference on computer vision. pp. 536–551. Springer (2014) 1, 3, 9, 12, 14
4. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: 2015 international conference on advanced robotics (ICAR). pp. 510–517. IEEE (2015) 3, 9, 10, 11, 12, 13, 14
5. Chen, W., Jia, X., Chang, H.J., Duan, J., Leonardis, A.: G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4233–4242 (2020) 13, 14
6. Chen, W., Jia, X., Chang, H.J., Duan, J., Shen, L., Leonardis, A.: Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1581–1590 (2021) 1
7. Collet, A., Martinez, M., Srinivasa, S.S.: The moped framework: Object recognition and pose estimation for manipulation. The international journal of robotics research **30**(10), 1284–1306 (2011) 1
8. Deng, S., Liang, Z., Sun, L., Jia, K.: Vista: Boosting 3d object detection via dual cross-view spatial attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8448–8457 (2022) 1
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012) 1
10. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9224–9232 (2018) 4
11. Gu, C., Ren, X.: Discriminative mixture-of-templates for viewpoint classification. In: European Conference on Computer Vision. pp. 408–421. Springer (2010) 1, 3
12. He, C., Zeng, H., Huang, J., Hua, X.S., Zhang, L.: Structure aware single-stage 3d object detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11873–11882 (2020) 4
13. He, Y., Huang, H., Fan, H., Chen, Q., Sun, J.: Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3003–3013 (2021) 3, 12, 14
14. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11632–11641 (2020) 1, 4, 12, 13, 14
15. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient response maps for real-time detection of textureless objects. IEEE transactions on pattern analysis and machine intelligence **34**(5), 876–888 (2011) 1, 3

16. Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: 2011 international conference on computer vision. pp. 858–865. IEEE (2011) 3, 9, 12, 14
17. Hu, Y., Fua, P., Wang, W., Salzmann, M.: Single-stage 6d object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2930–2939 (2020) 14
18. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993) 1, 3
19. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In: Proceedings of the IEEE international conference on computer vision. pp. 1521–1529 (2017) 3, 14
20. Kehl, W., Milletari, F., Tombari, F., Ilic, S., Navab, N.: Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation. In: European conference on computer vision. pp. 205–220. Springer (2016) 1, 3
21. Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J.Z., Langer, D., Pink, O., Pratt, V., et al.: Towards fully autonomous driving: Systems and algorithms. In: 2011 IEEE intelligent vehicles symposium (IV). pp. 163–168. IEEE (2011) 1
22. Li, C., Bai, J., Hager, G.D.: A unified framework for multi-view multi-class object pose estimation. In: Proceedings of the european conference on computer vision (eccv). pp. 254–269 (2018) 3
23. Liebelt, J., Schmid, C., Schertler, K.: independent object class detection using 3d feature maps. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008) 1, 3
24. Lin, J., Li, H., Chen, K., Lu, J., Jia, K.: Sparse steerable convolutions: An efficient learning of se (3)-equivariant features for estimation and tracking of object poses in 3d space. Advances in Neural Information Processing Systems **34** (2021) 1
25. Lin, J., Wei, Z., Ding, C., Jia, K.: Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. arXiv preprint arXiv:2207.05444 (2022) 1
26. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3560–3569 (2021) 1
27. Marchand, E., Uchiyama, H., Spindler, F.: Pose estimation for augmented reality: a hands-on survey. IEEE transactions on visualization and computer graphics **22**(12), 2633–2651 (2015) 1
28. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016) 1, 3
29. Oberweger, M., Rad, M., Lepetit, V.: Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 119–134 (2018) 14
30. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7668–7677 (2019) 14
31. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4561–4570 (2019) 1, 3, 14

32. Rios-Cabrera, R., Tuytelaars, T.: Discriminatively trained templates for 3d object detection: A real time scalable approach. In: Proceedings of the IEEE international conference on computer vision. pp. 2048–2055 (2013) 3
33. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. International journal of computer vision **66**(3), 231–259 (2006) 1, 3
34. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011) 1, 3
35. Song, C., Song, J., Huang, Q.: Hybridpose: 6d object pose estimation under hybrid representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 431–440 (2020) 9, 14
36. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 699–715 (2018) 14
37. Tejani, A., Tang, D., Kouskouridas, R., Kim, T.K.: Latent-class hough forests for 3d object detection and pose estimation. In: European Conference on Computer Vision. pp. 462–477. Springer (2014) 3
38. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: European Conference on Computer Vision. pp. 530–546. Springer (2020) 1
39. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3343–3352 (2019) 1, 4, 8, 9, 12, 13, 14
40. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16611–16621 (2021) 2
41. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019) 1
42. Wang, Z., Jia, K.: Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1742–1749. IEEE (2019) 1
43. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3109–3118 (2015)
44. Wu, C., Chen, J., Cao, Q., Zhang, J., Tai, Y., Sun, L., Jia, K.: Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. Advances in Neural Information Processing Systems **33**, 13174–13184 (2020) 1
45. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017) 1, 3, 9, 13, 14
46. Xu, D., Anguelov, D., Jain, A.: Pointfusion: Deep sensor fusion for 3d bounding box estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 244–253 (2018) 14

47. Zhou, G., Wang, H., Chen, J., Huang, D.: Pr-gcn: A deep graph convolutional network with point refinement for 6d pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2793–2802 (2021) 1, 4, 12, 13, 14