MatchU: Matching Unseen Objects for 6D Pose Estimation from RGB-D Images

Junwen Huang[1,2] Hao Yu[1] Kuan-Ting Yu[3] Nassir Navab[1,2] Slobodan Ilic[1] Benjamin Busam[1,2,4]

[1] Technical University of Munich   [2] Munich Center for Machine Learning
[3] XYZ Robotics   [4] 3dwe.ai

Abstract

Recent learning methods for object pose estimation require resource-intensive training for each individual object instance or category, hampering their scalability in real applications when confronted with previously unseen objects. In this paper, we propose MatchU, a Fuse-Describe-Match strategy for 6D pose estimation from RGB-D images. MatchU is a generic approach that fuses 2D texture and 3D geometric cues for 6D pose prediction of unseen objects. We rely on learning geometric 3D descriptors that are rotation-invariant by design. By encoding pose-agnostic geometry, the learned descriptors naturally generalize to unseen objects and capture symmetries. To tackle ambiguous associations using 3D geometry only, we fuse additional RGB information into our descriptor. This is achieved through a novel attention-based mechanism that fuses cross-modal information, together with a matching loss that leverages the latent space learned from RGB data to guide the descriptor learning process. Extensive experiments reveal the generalizability of both the RGB-D fusion strategy as well as the descriptor efficacy. Benefiting from the novel designs, MatchU surpasses all existing methods by a significant margin in terms of both accuracy and speed, even without the requirement of expensive re-training or rendering.

1. Introduction

Object 6D pose estimation is a critical task in computer vision applications, such as robotic manipulation [39, 64], augmented reality [1, 36], and autonomous driving [22, 32]. While object 6D pose estimation with object-specific training has achieved impressive results on benchmarks [50], handling unseen objects still remains a challenge. Approaches like template matching [26], keypoint detection [8, 17, 53], surface mapping [14, 45], and reconstruction-based frameworks [29, 31, 47] have achieved high accuracy for individual objects [14, 17, 18, 47, 53]. However, these methods are not designed to handle multiple objects or generalize to objects not presented in the training data.

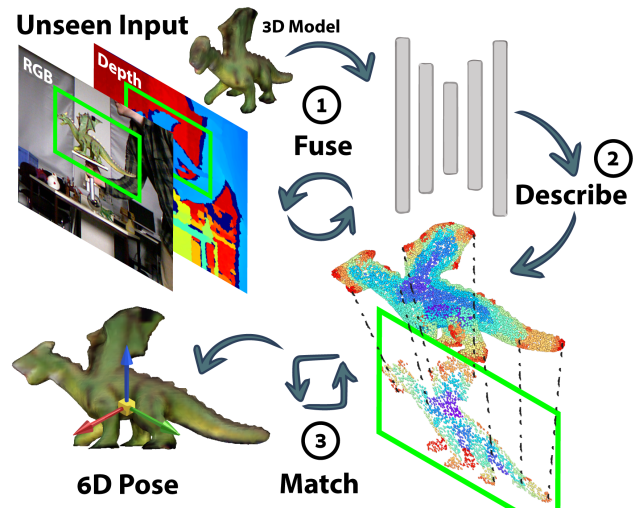Dataset-level pose estimation methods [52] can han-

Figure 1. MatchU provides a pipeline to match a previously unseen 3D CAD model of an object to an RGBD image (Top left). (Fuse) Information from RGB-D and CAD is fused. (Describe) Consumes fused information and produces generic color-aware rotation-invariant 3D descriptors. (Match) Further used for establishing correspondences as well as the 6D pose.

dle multiple objects in a dataset but struggle when faced with new instances. Similarly, category-level pose estimation methods [54] generalize to new instances within the same category but struggle with new categories. These approaches do not apply to the challenging problem of unseen object pose estimation in real-world applications where the 3D model is only available during inference time, since the models are designed to overfit the specific distribution of one object, category, or dataset. Some one-shot learning methods attempt to align object models using template matching or capture the structure from motion (SfM) of unseen objects [44, 47]. However, these approaches often require object-specific preprocessing steps. Classic approaches that target unseen object pose estimation employ handcrafted features and correspondences between CAD models and observed RGB-D images [11, 28, 42]. However, these methods introduce many pose hypotheses with high ambiguity, which require to be rated and refined iteratively, resulting in computational overhead.

Recent works have extensively delved into generic pose estimation through learning models that rate and refine pose hypotheses [3, 30], predominantly utilizing time-consuming render-and-compare strategies. This constraint limits their utility in real-world applications. Alternatively, some approaches formulate the generic 6D pose estimation problem as a point cloud registration task, benefiting from the point cloud representation backbones but neglecting crucial texture information from RGB images [6, 65], leading to less distinct point cloud-based descriptors. Therefore, they introduce ambiguities in correspondence extraction partially remedied by adding knowledge about the object symmetry during training [65].

We present MatchU, a Fuse-Describe-Match strategy for unseen object 6D pose estimation from single RGB-D images as shown in Figure 1. Our method is designed to extract rotation-invariant descriptors that can be shared across a wide range of objects, facilitating generalization to unseen objects. The extraction of rotation-invariant descriptors is crucial as it allows our method to inherently capture and model the natural symmetry of objects without relying on explicit symmetry annotations. However, rotation invariance still has some ambiguity where one point can be matched to several geometrically similar points. To address the ambiguity problem introduced by rotation invariance, we introduce a novel 2D-3D fusion module termed Latent Fusion Attention Module. This module effectively combines texture and geometric information. This results in extracting descriptors that describe both the appearance and the shape features of an object in a complementary and generic manner. Furthermore, we propose a novel Bridged Coarse-level Matching loss that leverages RGB information to enhance the learning of geometric descriptors. This loss function strengthens the association between texture and geometric features, leading to more precise and accurate matching between CAD models and RGB-D images of unseen objects. Our main contributions are:

- We propose **MatchU, a 6D pose estimation fuse-describe-match strategy** that extracts fused RGB-D input features targeted to register an unseen 3D CAD model to an object in the scene.
- We introduce a novel **Latent Fusion Attention Module** to effectively fuse texture and geometric features for generic pose estimation from RGB-D data and train MatchU with a **Bridged Coarse-level Matching Loss**.
- MatchU captures symmetries inherently by **learning a fused feature representation** without additional annotations thus reducing pose ambiguities.

2. Related Work

The majority of related work focused on 6D pose estimation of seen objects for which training data (real or synthetic) is available. However, they need to be retrained for any new object instance. There were extensions to object category pose estimation, but they cannot generalize to new unseen categories. Therefore, in recent two years several approaches, that aim at generalizing to novel unseen objects without retraining, were introduced.

2.1. Seen Object Pose Estimation.

The approaches for seen object pose estimation rely on available real or synthetic training data and train one neural network model per object or per scene. They are usually multi-stage pipelines where the core learning efforts are in establishing image-to-model (2D-3D) correspondences further used for pose estimation through PnP+RANSAC or direct regression. A larger amount of learning-based approaches consume RGB as input and only a few of them focus on RGB-D inputs facing the challenge of fusing RGB and depth information in neural networks.

RGB-D Fusion Methods are important because they profit from complementarity of two data sources and naturally improve pose accuracy as demonstrated in early works [19, 62]. In deep learning approaches, features are extracted separately from two modalities with different neural networks and their fusion is not obvious. Early approache, like PointFusion [58], extracts global RGB (CNN) and depth (PointNet) features from the patch containing the object, and fuses them with per-point depth features for 3D object bounding box detection. Later, DenseFusion [52] performs late per-point feature fusion strengthened with the global information, allowing better discrimination at the local level and resulting in better occlusion handling. Other works like PVNet3D [17] rely on DenseFusion and estimate sparse keypoints instead of dense correspondences. FFB6D [18] instead uses bidirectional fusion modules to combine modality information at earlier stages and produce stronger per-pixel fused features. Recently, DFTr [66] uses Transformers and improves the data fusion with the global semantic similarity between RGB and depth. This fusion strategy can help handle missing and noisy data caused by reflections or low-texture information.

Symmetric Objects are problematic because they look the same from different viewpoints. Correspondence-based methods have issues with visual ambiguities cause [38] as one-to-many matches define multiple equally correct poses. This has been tackled if symmetry information is known beforehand and used for data preparation [63] or in loss functions [53, 65]. Contrary to this SurfEmb [14] does not require known symmetry and learns symmetry invariant features with contrastive loss. Learned 2D-3D descriptions from SurfEmb [14] are not guaranteed to be invariant to rigid object transformations; it robustly learns quasi-invariance from a large dataset for specific objects. Additionally, predicting pose distribution [15, 24, 40] instead of a single estimate elegantly circumvents this problem.

2.2. Unseen Object Pose Estimation

Pose estimation of unseen objects considers that the neural network model is trained once and can generalize to novel unseen objects without retraining. For long, handcrafted feature matching using *point pair features*(PPF) [11] has been a competitive method in BOP challenge [50] enabling unseen object pose estimation. Its main disadvantage is efficiency due to large voting spaces and adding RGB to PPF [11] brought some benefits. Recently, Gen6D [37], OnePose [47] and OnePose++[16], utilize SfM and feature matching techniques to align a posed set of images of a given object to a target view using refined nearest neighbor image retrieval [37] or 2D-3D image matching [46].

Template-based methods like OSOP [44] and OVE6D [4] rely on representing the target object with templates. OSOP is a multi-stage pipeline, which leverages templates representing the 3D object seen from different views for segmentation, closest viewpoint selection and dense matching. OVE6D [4] is inspired by early ideas to learn specific embedding spaces for pose estimation [56], and, thus represents various 3D models together in an embedding space. ZePHyR [42] proposes hypothesis scoring, while MegaPose [30] proposes generic render-and-compare RGB pose refinement. The extracted features of OSOP [44] are not invariant to rotations and the design of the ZePHyR [42] and MegaPose [30] are computation-intensive since they need to evaluate many sampled hypotheses.

A natural way towards the generalization to unseen objects is through descriptor learning, where generic descriptors can be used to match depth pixels of the object with its 3D model. Learning strong local 3D descriptors has been exhaustively studied in the context of point cloud registration [23, 33, 43, 59]. However, for object pose estimation descriptors need not only to be unique and repeatable, but also require rotation invariance introduced in RIGA [60], YOHO [55], and RoITr [61]. Moreover, adding color information to them and maintaining generality is an additional challenge. Recent works tackling unseen object pose estimation, like Zeropose [6] and GCPose [65] rely on 3D descriptor learning. Zeropose [6] uses the foundation models ImageBind [13] and SAM [27] together with 3D-3D feature matching using descriptions from GeoTransformer [43]. GCPose [65] uses the same descriptor principle [43] but with explicit knowledge of object symmetries.

We leverage RoITr's [61] rotation-invariance by design and do not require pre-defined object symmetries for training. We aim to get the best of both worlds by designing a rotation-invariant and symmetry-aware backbone, which fuses RGB and depth information efficiently.

3. Method

3.1. Problem Formulation

The task of unseen object pose estimation aims at estimating the 6D pose between a CAD model, which is not available during training, and its partial observation from the RGB and/or depth image. In this paper, we estimate the pose of unseen objects by matching the learned descriptors between RGB-D data and its CAD model. The input of our method includes an unseen CAD model represented as a point cloud $P = \{p_i \in \mathbb{R}^3 \mid 1 \le i \le n\}$ with n points, the partial point cloud obtained from the depth channel, denoted as $Q = \{q_j \in \mathbb{R}^3 \mid 1 \le j \le m\}$ with m points, as well as its corresponding RGB image crop K of the localized object. Corresponding points $p_i \leftrightarrow q_j$ are collected in the predicted correspondence point set $\mathcal{C}$ which is used to estimate the 6D pose of the novel object, by optimizing the objective

$$\mathcal{T}^* = \arg \min_{\mathcal{T} \in \mathrm{SE}(3)} \sum_{(p_i, q_j) \in \mathcal{C}} \|(\mathcal{T}p_i - q_j)\|_2^2 \tag{1}$$

of mutual 3D correspondences. $\mathcal{T} \in \mathrm{SE}(3)$ denotes the 6D pose of the novel object in the Special Euclidean group $\mathrm{SE}(3)$ of rigid transformation in 3D space.

3.2. Method Overview

To solve for object pose, we first calculate correspondences by extracting the generic descriptors ϕ^P and ϕ^Q for the points in P and Q in latent space $\mathbb{R}^d$. By calculating the similarity between ϕ^P and ϕ^Q, we can construct the correspondence set as $\mathcal{C} = \{(p_i, q_j) \mid \phi_i^P \leftrightarrow \phi_j^Q\}$, where $\phi_i^P \leftrightarrow \phi_j^Q$ denotes matched descriptors of the point p_i and q_j, respectively. The optimization problem (Eqn. 1) is designed as a least square problem that can be robustly solved with an outlier-aware consensus algorithm such as RANSAC [12].

From the input CAD point cloud P, depth point cloud Q, and the RGB image K, our proposed method estimates a mapping function ψ that maps P and Q to generic descriptors $\phi^P = \psi(P \mid (Q, K)) \in \mathbb{R}^{n \times d}$ and $\phi^Q = \psi(Q \mid (P, K)) \in \mathbb{R}^{m \times d}$ by fusing the cross-modality information from (Q, K) and (P, K), respectively. By matching our learned generic descriptors, correspondences are established between the unseen object and its partial observation, and the object pose is finally estimated. An overview of our framework is depicted in Fig. 2.

3.3. Encoding and Fusing Descriptors

We first introduce the extraction of 3D and 2D local features, and then the cross-modality descriptor fusion.
Local 3D Feature Extraction. We employ the recent transformer-based architecture RoITr [61] as our encoder backbone to extract rotation-invariant 3D local features
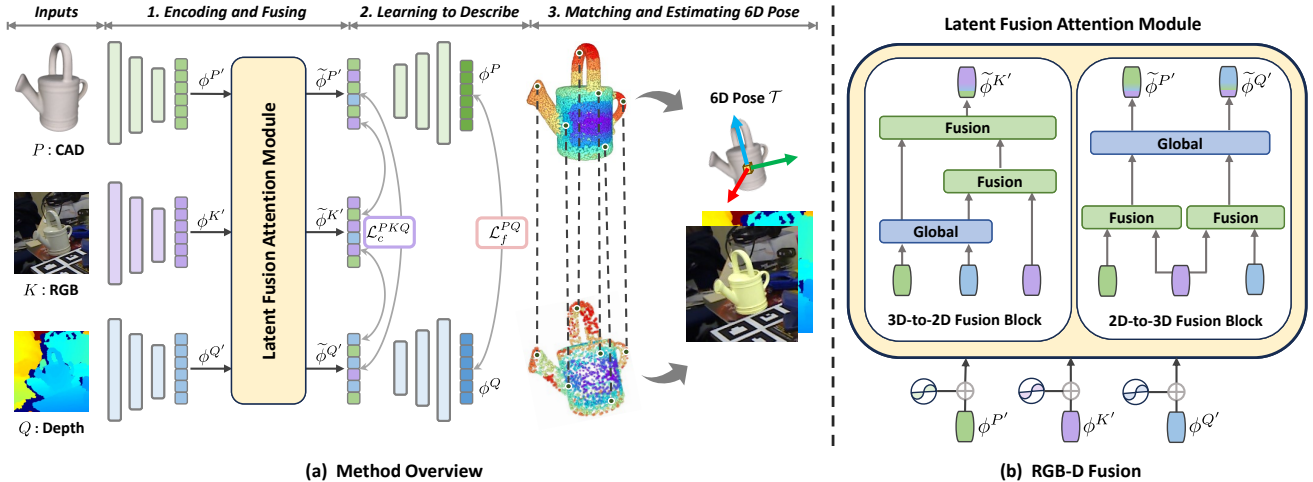
(a) Method Overview

(b) RGB-D Fusion

Figure 2. **Overview of MatchU.** Upon encountering an unseen object, we initially derive the segmented depth point cloud Q and the corresponding RGB image crop K utilizing a pre-trained generic segmentation network. Subsequently, we procure both 3D and 2D local features from the CAD point cloud P, depth point cloud Q, and the RGB image crop K. These extracted features are then amalgamated within a latent space through our innovative *Latent Fusion Attention Module*, under the guidance of a *Bridged Coarse-level Matching Loss (BCM Loss)* $\mathcal{L}c^{PKQ}$. The refined 3D descriptors $\widetilde{\phi}^{P'}$ and $\widetilde{\phi}^{Q'}$ are fed into decoders, which enhance the resolution of the descriptors to ϕ^P and ϕ^Q, this process being steered by a detailed matching loss $\mathcal{L}_f^{PQ}$. In the final stage, the 6D pose of the novel objects is deduced by aligning the descriptors within the latent space and aggregating the pose parameters $\mathcal{T}$.

from the CAD point cloud P and the partially observed point cloud Q from depth image. The inherent rotation-invariance of the descriptor provides a robust feature extraction for geometric cues and guarantees the generalizability for unseen objects. Given P and Q, our encoder down-samples the input point clouds via Farthest Point Sampling (FPS) to superpoints. They represent a well-distributed coarse representation of spatial structure from the underlying dense point cloud defined as $P' = \{p'_i \in \mathbb{R}^3 \mid 1 \leq i \leq n'\}$ and $Q' = \{q'_j \in \mathbb{R}^3 \mid 1 \leq j \leq m'\}$, where n' and m' stand for the number of superpoints in P' and Q', respectively. Following [61], for each superpoint p'_i and q'_j, we first extract the local geometric features from neighboring points within a radius of r. The local geometric cues are then projected into the latent space by a sequence of attention blocks, from which we obtain the inherently rotation-invariant local 3D geometric descriptors, denoted as $\phi_{p'_i} \in \mathbb{R}^d$, and $\phi_{q'_j} \in \mathbb{R}^d$ where d is the dimension of the latent space.

Local 2D Feature Extraction. A convolutional neural network (CNN) is used for local visual feature extraction. Following LoFTR [46], we adopt a modified encoder of FPN [35] as our CNN backbone. This 2D encoder down-samples the input image crop of size $H \times W$ to a feature map of size $\frac{H}{8} \times \frac{W}{8}$, while simultaneously projecting the local textural information into a d-dimension latent space consistent with the 3D geometry features. The image's local feature map is then flattened into $\phi^{K'} = \{k_t \in \mathbb{R}^d \mid 1 \leq t \leq \frac{H}{8} \times \frac{W}{8}\}$, where we denote the 2D superpixels as K' and the 2D superpixel features as $\phi^{K'}$.

Latent Fusion Attention Module. After extracting 3D and 2D local features, we fuse the encoded 3D and 2D context in latent space via our proposed Latent Fusion Attention Module. To keep the generalizability of our network and avoid overfitting on object-specific features, we propose to fuse the 3D and 2D features in a coarse-level latent space and leverage a *3D-to-2D Fusion Block* as well as a *2D-to-3D Fusion Block* for fusing the information in two perspectives. We leverage the *Latent Fusion Transformers* (green layers) and *Global Transformers* (blue layers) in these two fusion blocks as shown in Figure 2 (b).

Previous methods [17, 18, 52] usually interpolating features w.r.t. their spatial relationships explicitly. In MatchU, we use the positional encoding in the attention mechanism to incorporate spatial awareness and implicitly align different modalities. For 2D features, we follow DETR [5] to encode the spatial information of the 2D feature map into the feature space. As for 3D, instead of encoding the raw position of the points [66], we propose to use the pose-agnostic Point Pair Features (PPFs) [11] as the position representation following [61], which guarantees the geometric rotation-invariance and generalizability for unseen object pose estimation.

The *Latent Fusion Transformer* is designed to fuse the 2D superpixel features and 3D superpoint features in the latent space, which consists of a series of self-attention and cross-attention layers. Following [46], we adopt the linear attention [25] for all the self- and cross attention layers with the goal of lower computational complexity. We stack g self- and cross-attention layers for each Latent Fusion trans-

former in practice. The *Global Transformer* is designed to aggregate the global context of the 3D and 3D features, for which we follow the design of RoITr [61].

Details for both *3D-to-2D Fusion Block* and *2D-to-3D Fusion Block* are illustrated in Fig. 2. For the *3D-to-2D Fusion Block*, we first aggregate CAD $\phi^{P'}$ and depth $\phi^{Q'}$ superpoint features with a *Global Transformer*. Then the RGB feature $\phi^{K'}$ is fused with the global-aware depth and CAD feature sequentially to get the final cross-modal 2D feature $\widetilde{\phi}^{K'}$ for each superpixel by *Latent Fusion Transformer*. For the *2D-to-3D Fusion Block*, we first separately enhance both CAD $\phi^{P'}$ and depth $\phi^{Q'}$ superpoint features with RGB features through *Latent Fusion Transformer*. A *Global Transformer* then co-injects this information to provide the 2D-aware 3D superpoint features $\widetilde{\phi}^{P'}$ and $\widetilde{\phi}^{Q'}$ for both CAD and depth.

3.4. Learning to Describe

In order to guide the learning of the fused descriptors, we propose several loss functions. With the latent features learned from RGB images as the bridge between the latent spaces of the CAD and depth point clouds, we define *Bridged Coarse-level Matching Loss* (BCM Loss), which significantly facilitates the unification of two different 3D-based latent spaces, and helps to generate more robust and reliable correspondences between superpoints. Moreover, a fine-level matching loss is also introduced to guide the refinement of superpoint matches to point correspondences.

Bridged Coarse-level Matching Loss. To ensure the effectiveness of RGB-based 2D information in the latent space, the key is to provide the supervision signal from both 2D and 3D modality by establishing the cross-modal matches between 2D and 3D features. The alignment between the superpoints P' and Q' can be obtained via ground-truth transformation matrix and nearest neighbor search. Ground-truth 3D-2D correspondences between superpoints and superpixels are inherent in the RGB-D pair.

We adopt the Circle Loss [48], which maximizes the similarity of the positive pairs of the superpoints, as well as minimizes the similarity of the negative pairs of the superpoints. Specifically, for each superpoint $p'_i \in P'$, and $q'_j \in Q'$, we can calculate the overlap $\mathcal{V}$ between p'_i and q'_j as:

$$\mathcal{V}(p'_i, q'_j) = \frac{|\{\hat{p}'_u \in \hat{P}'_i \mid \exists \, \hat{q}'_v \in \hat{Q}'_j : \hat{p}'_u \leftrightarrow \hat{q}'_v\}|}{|\{\hat{p}'_u \in \hat{P}'_i\}|}, \quad (2)$$

where $\leftrightarrow$ denotes the correspondence relationship. $\hat{P}'_i$ is the group of points from P' assigned to p'_i by Point-to-Node grouping strategy [59], and $\hat{Q}'_j$ means the same for Q'. A pair of superpoints p'_i and q'_j are considered as a positive pair if and only if $\mathcal{V}(p'_i, q'_j) > \tau_r$, where τ_r is the threshold of the overlap. We sample a positive set of superpoints from

Q', and a negative set of superpoints for P'. The coarse-level superpoint Circle Loss loss for P' can be calculated with the weight of overlap, which we denote as $\mathcal{L}_c^{P'}$. We provide the detailed loss function in Appendix.

The same loss for Q' is defined similarly, and the overall loss between the superpoints P' and Q' is defined as

$$\mathcal{L}_c^{P'Q'} = (\mathcal{L}_c^{P'} + \mathcal{L}_c^{Q'})/2. \quad (3)$$

Similarly, we apply Circle Loss for 3D-2D coarse-level matching. We first project the 3D positive and negative samples into 2D plane, and then obtain the positive and negative pairs between the 2D superpixels and 3D superpoints.

The loss function between K' and Q' is defined as $\mathcal{L}_c^{K'Q'}$, and the loss function between K' and P' is defined as $\mathcal{L}_c^{K'P'}$. Then we calculate the overall Bridged Coarse-level Matching Loss between P', Q' and K' as

$$\mathcal{L}_c^{PKQ} = \lambda_b \mathcal{L}_c^{P'Q'} + (1 - \lambda_b)(\mathcal{L}_c^{K'Q'} + \mathcal{L}_c^{K'P'}), \quad (4)$$

where λ_b and $1 - \lambda_b$ is the weight for the 3D-3D and 3D-2D matching loss respectively.

Fine-level Matching Loss. In order to enhance the precision of the 3D-3D correspondence, we apply a fine-level matching loss to the CAD point cloud P and the observed point cloud Q. We use a series of decoder blocks introduced in [61], which generates denser points P and Q from the coarse-level superpoints P' and Q'. Given the superpoint correspondence, the group of fine-level point features is assigned to each superpoint through the point-to-node strategy, and the similarity matrix can be calculated between the corresponding groups. The fine-level matching is formulated as an optimal transport problem, which can be solved by the Sinkhorn algorithm [7]. A negative log-likelihood is applied to the similarity matrix to obtain the fine-level matching loss $\mathcal{L}_f^{PQ}$ between P and Q. The overall loss function for our training is defined as:

$$\mathcal{L} = \lambda_c \mathcal{L}_c^{PKQ} + (1 - \lambda_c) \mathcal{L}_f^{PQ}, \quad (5)$$

where $\mathcal{L}_c^{PKQ}$ is the Bridged Coarse-level Matching Loss and $\mathcal{L}_f^{PQ}$ is the fine-level matching loss. λ_c is the weight to balance the coarse and fine-level training.

3.5. Matching Descriptors and Estimating 6D Poses

After training on a large number of objects and images, we obtain a robust descriptor. During inference, we utilize extracted features to establish 3D-3D matches between CAD point cloud P and observation Q. Following [61], we measure similarities of normalized features using covariance analysis. We determine the top κ most correlated ones as putative 3D-3D matches of the correspondence point set $\mathcal{C}$ from which we create η pose hypotheses. For each hypothesis $\mathcal{T}_v$ with $1 \leq v \leq \eta$, we first randomly select $s \ll \kappa$

	Method	MH	Obj. Loc. unseen	Obj. Loc. seen	Refine.	LM-O	T-LESS	TUD-L	IC-BIN	YCB-V	Mean	Time(s)
(a) one hypo. w/o refine.	ZeroPose		✓			26.0	17.8	41.2	17.7	25.7	25.7	0.30
	OSOP		✓			39.3	-	-	-	**52.9**	46.1	0.47
	ZeroPose			✓		26.1	24.3	61.1	24.7	29.5	33.1	0.30
	MegaPose			✓		18.7	19.7	20.5	15.3	13.9	17.6	2.50
	Ours(Fast)		✓			**52.6**	**42.9**	**70.0**	**36.7**	50.5	**50.5**	0.07
(b) multi-hypo. w/o refine.	OSOP	✓	✓			46.2	-	-	-	54.2	50.2	5.30
	Ours(Accurate)	✓	✓			**56.2**	**50.6**	**75.6**	**42.2**	**60.8**	**57.8**	1.03
(c) unseen loc, w/ refine.	DrostPPF	✓	✓		✓	52.7	-	-	-	34.4	43.6	15.9
	PPF + Zephyr	✓	✓		✓	59.8	-	-	-	51.6	55.7	2.90
	OSOP	✓	✓		✓	48.2	-	-	-	57.2	52.7	5.44
	ZeroPose	✓	✓		✓	49.1	34.0	74.5	39.0	57.7	50.9	6.75
	ZeroPose (BOP)	✓	✓		✓	53.8	40.0	83.5	39.2	65.3	56.4	6.75
	Megapose (BOP)	✓	✓		✓	62.6	48.7	85.1	**46.7**	**76.4**	63.9	10.70
	Ours(Accurate)	✓	✓		✓	**64.4**	**52.7**	**89.8**	44.2	72.6	**64.7**	5.60
(d) seen loc., w/o refine.	OVE6D	✓		✓		49.6	52.3	-	-	-	57.5	-
	GCPose	✓		✓		65.2	**67.9**	92.6	-	-	**75.2**	-
	Ours(Accurate)	✓		✓		**66.8**	65.1	**93.1**	43.9	65.8	66.9	3.12
(e) seen loc., w/ refine.	OVE6D	✓		✓	✓	62.7	54.6	-	-	-	58.7	-
	HybridPPF	✓		✓	✓	63.1	65.5	92.0	-	-	**73.5**	-
	Megapose	✓		✓	✓	58.3	54.3	71.2	37.1	63.3	56.8	10.70
	ZeroPose	✓		✓	✓	56.2	53.3	87.2	41.8	58.4	59.4	6.75
	Ours(Accurate)	✓		✓	✓	**68.0**	**66.8**	**94.7**	**47.4**	**75.6**	**70.5**	5.60

Table 1. **Quantitative results [1] in terms of Average Recall(AR) on BOP-5 core benchmark datasets for unseen object pose estimation task. MH**: whether multiple hypotheses were adopted. **Obj. Loc.**: whether the object localization (detection or segmentation) is trained on the test objects. **Refine.**: whether the result is refined with either depth and/or RGB images.

correspondences from $\mathcal{C}$ and then solve Eq. 1 for the 6D object pose using RANSAC [12] optimization. This process speeds up the prediction process and provides us with the control parameter η to determine the efficiency of MatchU. All hypotheses are then ranked by an average score between 3D and RGB verification processes. For 3D, we calculate the score based on the Euclidean point-to-point distance between the transformed CAD model and the lifted depth map. For RGB, we follow the proposal of [30]. Our final prediction is the pose $\mathcal{T}$ with the highest score.

4. Experiments and Results

4.1. Implementation Details

During training, default settings are $\lambda_b = 0.3$ and $\lambda_c = 0.5$. For inference, we set $\kappa = 128$, $s = 64$, and initially $\eta = 20$, adjusting η to 64 to enhance accuracy during testing. To localize unseen objects in RGB-D images, we utilize a CAD-based segmentation method [41]. For 6D pose evaluation with instance-level localization, existing detection results from [6, 34, 49] are employed. Detailed implementation and more information about the network designs are provided in the Appendix.

4.2. Datasets and Evaluation Metrics

Datasets. Following ZeroPose [6] and MegaPose [30], we utilize the Google-Scanned-Objects (GSO) dataset [10] provided by MegaPose to train our model, where 850 GSO objects with around 800K rendered images are used for training and the rest 94 objects with around 200K images are for validation. To evaluate our method on unseen object pose estimation, we employ five core BOP datasets as our testing set, *i.e.* LM-O [2], T-LESS [20], TUD-L [21], IC-BIN [9], and YCB-V [57], among which the LM-O dataset is adopted for our ablation studies. All the CAD models and images in the test set are guaranteed to be unseen during training.

Metrics. We adopt the Average Recall (AR) in the standard benchmark BOP [21, 50] as our main evaluation metric. It calculates the average recall of three pose errors by varying the thresholds in a determined range. We also adopt the average distance metric (ADD) as our secondary metric for a fair comparison with baselines in Table 2. ADD calculates the average point distance between the point clouds of the object CAD model with ground-truth and estimated pose. We report the accuracy of distance less than 10% of the objects' diameter (ADD-0.1d) as [17, 18, 66].

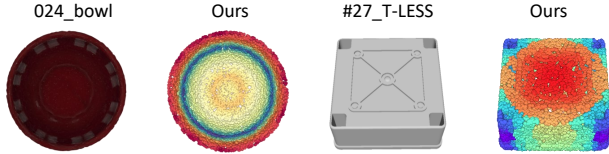[1] The numbers and timings are either from the original papers [6, 44] or BOP Challange [50].

Figure 3. **t-SNE visualization of our descriptors for symmetric objects.** We showcase the capability of capturing both continuous and discrete symmetries without external annotation.
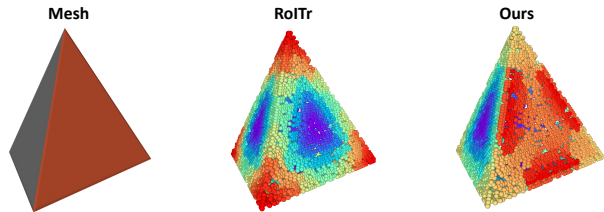

Figure 4. **t-SNE visualization of a regular tetrahedron with different colored faces** (left). Our method can extract distinct descriptors for the red face (right) while RoITr cannot (center).

4.3. Evaluation on 5 BOP Core Benchmark Datasets

The evaluation results of our method are shown in Tab. 1. Compared with all the baselines, we achieve state-of-the-art performance on the task of unseen object pose estimation. We denote our method using only 1 pose hypothesis as Ours (Fast) and as Ours (Accurate) when multiple hypotheses are leveraged. For a fair comparison, we run our method under 5 specific settings in accordance with the baselines. In *(a)*, we compare our method with Zero-Pose [6], OSOP [44], and MegaPose [30] with only one hypothesis during the inference stage and without any refinement. Under *(a)*, Ours (Fast) performs significantly better and faster on average. Note that in *(a)*, we use a generic instance detection/segmentation network for unseen objects [41]. This demonstrates the robustness of our method against noisy detection initialization, which often occurs in real applications. In *(b)*, we further improve our results by introducing more pose hypotheses (20 by default) as we described in Section 3.5. We compare our method with OSOP [44] with the same number of hypotheses and exhibit superior performance. In *(c)*, by adding the ICP refinement, our model achieves the best results among all the methods on both the overall and the per-dataset evaluation. Notably, in *(c)*, our method requires less time to perform one inference compared to other baselines [6, 30] with rendering-based refinement. In *(d)*, we compare our method with OVE6D [4] and GCPose [65] with an identical object detection/segmentation network. Our method consistently surpasses the baselines on T-LESS and TUD-L datasets. In *(e)*, we finally improve our results by using a trained detector and incorporating ICP refinement. We showcase the qualitative results in Figure 5.

Method	Object Loc.	Pose	ADD-0.1d
PVN3D [17]	seen	seen	63.2
FFB6D [18]	seen	seen	66.2
DFTr [66]	seen	seen	**77.7**
Ours(1hypo)	seen	seen	68.4
Ours(20hypo)	seen	seen	75.7
Ours(1hypo)	unseen	unseen	61.7
Ours(20hypo)	unseen	unseen	**70.8**

Table 2. **Quantitative evaluation of 6D pose (ADD-0.1d) on the LM-O dataset for seen object pose estimation task.**

4.4. Capturing Symmetry and Texture

We visualize our descriptors for symmetric objects with t-SNE [51]. As shown in Figure 3, our learned descriptors can capture continuous (024_bowl) and discrete (#27_T-LESS) symmetries, which is credited to the rotation-invariant property of our design. Compared with GC-Pose [65], which relies on the supervision of symmetry labels, our descriptors recognize the symmetry even without any external symmetry annotations. Moreover, we visualize "tetX" from SYMSOL [40] dataset, a regular tetrahedron with one red and three white faces in Figure 4. The descriptors extracted by RoITr show the same distribution on all 4 faces due to their geometric similarity, introducing ambiguity issues for matching and leading to incorrect pose estimation potentially. In contrast, our method extracts distinct descriptors on the textured face. This indicates that our method not only describes the geometric property of the objects but also captures the texture information which further eliminates ambiguities in pose estimation.

4.5. Comparison with RGB-D Fusion Pipelines

To demonstrate the effectiveness of our proposed RGB-D fusion mechanism, we compare MatchU with recent RGB-D fusion approaches [17, 18, 52, 66] on LM-O dataset. As shown in Tab. 2, although our method is specifically designed for unseen object pose estimation, it still outperforms PVN3D [17] as well as FFB6D [18], and achieves comparable results with the current state-of-the-art method DFTr [66] by increasing the number of hypotheses, this demonstrates the efficacy of our RGB-D fusion mechanism. Moreover, our method trained without the test objects even outperforms most of the baselines that have seen them. This result further confirms the generalizability of our method.

4.6. Ablation Study

Key Design Principles. First, we replace the BCM Loss $\mathcal{L}_c^{PKQ}$ with only a 3D coarse matching loss $\mathcal{L}_c^{P'Q'}$. The performance drops obviously as shown in the first row of Table 3, proving the effectiveness of BCM Loss in guiding the descriptor learning. Second, we mask out the input

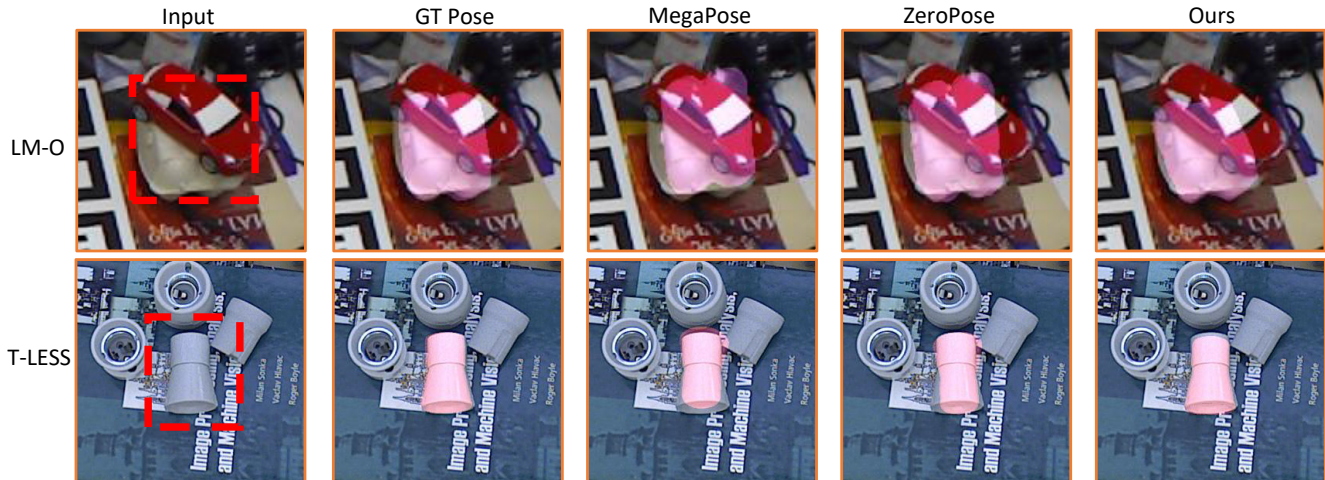|	Input	GT Pose	MegaPose	ZeroPose	Ours
LM-O					
T-LESS					

Figure 5. **Qualitative results of 6D pose estimation of our method in comparison with Megapose and ZeroPose.** The upper row shows an egg box which is heavily occluded in the LM-O dataset. Our method is robust to handle occlusion while other methods flip the poses by mistake. The lower row shows a highly ambiguous object that other methods put the pose upside down but ours predicts accurately.

Method	Mean AR (BOP-5)
Ours *w/o* BCM Loss	48.1
Ours *w/o* RGB Input	39.6
RoITr *w/* RGB Init.	42.6
Ours	**50.5**

Table 3. **Ablation study of our key designs on BOP-5 datasets**.
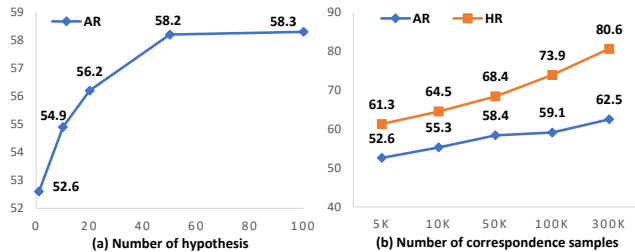

Figure 6. (a) AR with different numbers of hypotheses. (b) AR of final pose and HR of pose hypotheses with different numbers of correspondence samples.

RGB image. A sharp decrease in performance (2nd row) indicates the efficacy of our RGB-D fusion, as well as the pivotal role that RGB information plays in our pipeline. Third, we take the original RoITr model and initialize the point features with the RGB values to further demonstrate the superiority of our RGB-D fusion. RoITr *w/* RGB surpasses Ours *w/o* RGB, but is still inferior to our full pipeline by a large margin.

Influence of the Number of Hypotheses. As shown in Figure 6 (a), our method benefits from increasing the number of hypotheses. However, the performance saturates when the number passes 50, indicating 50 hypotheses could cover true poses for most cases. To balance the computation cost and performance, we use 20 as the default.

Quality of Pose Hypotheses. To investigate the quality of our pose hypotheses, we define the Hit Recall (HR) as the ratio of testing set whose ground truth pose is included in our proposed hypotheses. Specifically, the top 128 correspondences are used as the sampling pool, and one hypothesis is computed through 3 correspondences randomly selected from it. This procedure is repeated to generate multiple hypotheses. We report the HR in comparison with the Average Recall (AR) by varying the number of correspondence samples. As shown in Figure 6 (b), the AR exhibits a lower number since it only considers the top-1 scored pose hypothesis. When considering all the hypotheses in evaluation, our method achieves over 80% HR, which reflects the

potential of our descriptors in generating the correct poses.

5. Conclusion

We present MatchU, a Fuse-Describe-Match framework for unseen object pose estimation from single RGB-D images. Our method first extracts rotation-invariant descriptors from 3D point clouds of CAD model and depth map. Then, the multi-modal fusion of texture and geometry is achieved through a Latent Fusion Attention Module. A Bridged Coarse-Level Matching Loss is introduced to utilize latent features from RGB images to connect descriptions of partial observations and full object geometry. MatchU inherently models object symmetry without explicit annotations. MatchU surpasses all existing methods for unseen object pose estimation by a large margin on standard benchmarks. Certainly, it relies on external object localization and could be could be affected by their erroneous results. In the future, incorporating such modules into the pipeline to build end-to-end training might further improve our results. We believe that by closing the gap to object-specific baselines, MatchU constitutes an important step forward to truly scalable 6D pose estimation of unseen objects.

References

[1] Ronald T. Azuma. A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, 1997. 1

[2] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014. 6

[3] Benjamin Busam, Hyun Jun Jung, and Nassir Navab. I like to move it: 6d pose estimation as an action decision process. *arXiv preprint arXiv:2009.12678*, 2020. 2

[4] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In *CVPR*, 2022. 3, 7

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4

[6] Jianqiu Chen, Mingshan Sun, Tianpeng Bao, Rui Zhao, Liwei Wu, and Zhenyu He. Zeropose: Cad-model-based zero-shot pose estimation, 2023. 2, 3, 6, 7

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 5

[8] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *CVPR*, 2022. 1

[9] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *CVPR*, 2016. 6

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022. 6

[11] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *CVPR*, 2010. 1, 3, 4

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 3, 6

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3

[14] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *CVPR*, 2022. 1, 2

[15] Rasmus Laurvig Haugaard, Frederik Hagelskjær, and Thorbjørn Mosekjær Iversen. Spyropose: Se (3) pyramids for object pose distribution estimation. In *ICCV*, 2023. 2

[16] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *NeurIPS*, 2022. 3

[17] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *CVPR*, 2020. 1, 2, 4, 6, 7

[18] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *CVPR*, 2021. 1, 2, 4, 6, 7

[19] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *ICCV*, 2011. 2

[20] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *WACV*, 2017. 6

[21] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018. 6

[22] Sabera Hoque, Shuxiang Xu, Ananda Maiti, Yuchen Wei, and Md. Yasir Arafat. Deep learning for 6d pose estimation of objects — a case study for autonomous driving. *Expert Systems with Applications*, 2023. 1

[23] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *CVPR*, 2021. 3

[24] Thorbjørn Mosekjær Iversen, Rasmus Laurvig Haugaard, and Anders Glent Buch. Ki-pode: Keypoint-based implicit pose distribution estimation of rigid objects. *arXiv preprint arXiv:2209.09659*, 2022. 2

[25] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 4

[26] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017. 1

[27] 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

[28] Rebecca König and Bertram Drost. A hybrid approach for 6dof pose estimation. In *ECCV*, 2020. 1

[29] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *ECCV*, 2020. 1

[30] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *CoRL*, 2022. 2, 3, 6, 7

[31] Fu Li, Shishir Reddy Vutukur, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In *ICCV*, 2023. 1

[32] Peiliang Li, Tong Qin, and Shaojie Shen. Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving. In *ECCV*, 2018. 1

[33] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *CVPR*, 2022. 3

[34] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *ICCV*, 2019. 6

[35] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[36] Jonathan Linowes and Krystian Babilinski. *Augmented reality for developers: Build practical augmented reality applications with unity, ARCore, ARKit, and Vuforia*. Packt Publishing Ltd, 2017. 1

[37] 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. In *ECCV*, 2022. 3

[38] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *ICCV*, 2019. 2

[39] Matthew T. Mason. Toward robotic manipulation. *Annual Review of Control, Robotics, and Autonomous Systems*, 2018. 1

[40] Kieran Murphy, Carlos Esteves, Varun Jampani, Srikumar Ramalingam, and Ameesh Makadia. Implicit-pdf: Nonparametric representation of probability distributions on the rotation manifold. *arXiv preprint arXiv:2106.05965*, 2021. 2, 7

[41] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *ICCV*, 2023. 6, 7

[42] Brian Okorn, Qiao Gu, Martial Hebert, and David Held. Zephyr: Zero-shot pose hypothesis rating. In *ICRA*, 2021. 1, 3

[43] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*. 3

[44] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *CVPR*, 2022. 1, 3, 6, 7

[45] 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 *CVPR*, 2022. 1

[46] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 3, 4

[47] 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 *CVPR*, 2022. 1, 3

[48] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020. 5

[49] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *CVPR*, 2020. 6

[50] Martin Sundermeyer, Tomáš Hodaň, Yann Labbe, Gu Wang, Eric Brachmann, Bertram Drost, Carsten Rother, and Jiří Matas. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In *CVPR*, 2023. 1, 3, 6

[51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 7

[52] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *CVPR*, 2019. 1, 2, 4, 7

[53] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *CVPR*, 2021. 1, 2

[54] 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 *CVPR*, 2019. 1

[55] Haiping Wang, Yuan Liu, Zhen Dong, Wenping Wang, and Bisheng Yang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. *arXiv preprint arXiv:2109.00182*, 2021. 3

[56] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *CVPR*, 2015. 3

[57] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter . Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems*, 2018. 6

[58] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*, 2018. 2

[59] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. In *NeurIPS*, 2021. 3, 5

[60] Hao Yu, Ji Hou, Zheng Qin, Mahdi Saleh, Ivan Shugurov, Kai Wang, Benjamin Busam, and Slobodan Ilic. Riga: Rotation-invariant and globally-aware descriptors for point cloud registration, 2022. 3

[61] Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant transformer for point cloud matching. In *CVPR*, 2023. 3, 4, 5

[62] Kuan-Ting Yu, Shih-Huan Tseng, and Li-Chen Fu. Learning hierarchical representation with sparsity for rgb-d object recognition. In *IROS*, 2012. 2

[63] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *ICCV*, 2019. 2

[64] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker Jr, Alberto Rodriguez, and Jianxiong Xiao. Multiview self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In *ICRA*, 2017. 1

[65] Heng Zhao, Shenxing Wei, Dahu Shi, Wenming Tan, Zheyang Li, Ye Ren, Xing Wei, Yi Yang, and Shiliang Pu. Learning symmetry-aware geometry correspondences for 6d object pose estimation. In *ICCV*, 2023. 2, 3, 7

[66] Jun Zhou, Kai Chen, Linlin Xu, Qi Dou, and Jing Qin. Deep fusion transformer network with weighted vector-wise key-points voting for robust 6d object pose estimation. *arXiv preprint arXiv:2308.05438*, 2023. 2, 4, 6, 7