DISP6D: Disentangled Implicit Shape and Pose Learning for Scalable 6D Pose Estimation

Yilin Wen[1], Xiangyu Li[2], Hao Pan[3], Lei Yang[1,4], Zheng Wang[5], Taku Komura[1], and Wenping Wang[6]

[1] The University of Hong Kong
[2] Brown University
[3] Microsoft Research Asia
[4] Centre for Garment Production Limited, Hong Kong
[5] SUSTech
[6] Texas A&M University

Abstract. Scalable 6D pose estimation for rigid objects from RGB images aims at handling multiple objects and generalizing to novel objects. Building on a well-known auto-encoding framework to cope with object symmetry and the lack of labeled training data, we achieve scalability by disentangling the latent representation of auto-encoder into shape and pose sub-spaces. The latent shape space models the similarity of different objects through contrastive metric learning, and the latent pose code is compared with canonical rotations for rotation retrieval. Because different object symmetries induce inconsistent latent pose spaces, we re-entangle the shape representation with canonical rotations to generate shape-dependent pose codebooks for rotation retrieval. We show state-of-the-art performance on two benchmarks containing textureless CAD objects without category and daily objects with categories respectively, and further demonstrate improved scalability by extending to a more challenging setting of daily objects across categories.

Keywords: 6D pose estimation, scalability, disentanglement, symmetry ambiguity, re-entanglement, sim-to-real

1 Introduction

Estimating the 6D pose of objects from a single RGB image is fundamental in fields like robotics and scene understanding. While efficient learning-based methods have been developed [47,59,29], a common assumption with many of these works is that a specialized network is trained for each object, which makes it expensive to process multiple objects by switching and streaming to respective networks, and renders it impossible to handle novel objects without re-training.

Recent works improve the capability of a single network for processing multiple objects through different ways. For example, a series of works [56,53,7] perform category-level pose estimation, by learning to map input pixels (and point

Work partially done during internships of Y. Wen and X. Li with Microsoft Research Asia. Code and data are available at: `https://github.com/fylwen/DISP-6D`.

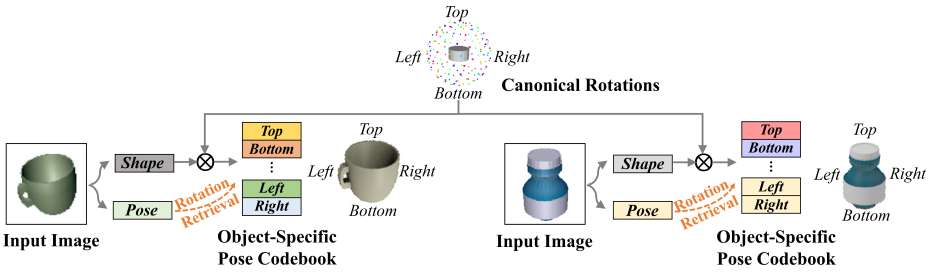


Fig. 1: **Disentanglement for pose estimation.** Images of objects are mapped to latent representations for object shape and pose, respectively. Due to different object symmetries, query pose codes must refer to object-specific pose codebooks (symmetries marked by code color) for rotation retrieval, which are generated by re-entangling canonical rotations with object shapes.

clouds) to corresponding points of a canonically aligned object, and computing pose registration based on the correspondences. However, these works assume that the space of canonically aligned objects for a given category is sufficiently regular to learn with neural networks, which does not hold for different objects across categories. Moreover, the point-wise correspondences are ambiguous under object symmetries, which may hinder the performance of these methods. On the other hand, Multipath-AAE [48] builds on the auto-encoding framework [49,50] to learn pose embeddings for different objects, by using a specific decoder for each object. Therefore Multipath-AAE is not restricted by the categorical shape alignment regularity, yet the network complexity becomes prohibitive as the number of training objects gets large. In addition, the single latent representation encoding mixed information of diverse objects under different poses may not be sufficiently accurate for pose estimation.

We present DISP6D – an approach to train a single network that processes more objects simultaneously (Fig. 2). As we build on the auto-encoding framework [50], objects do not need category labels and the symmetry ambiguity is automatically handled. Meanwhile, we extend [50] by *disentangling* object shape and pose in the latent representation; therefore we avoid per-object decoders and reduce the network training complexity significantly. The disentanglement allows the latent pose code of an arbitrary object to be compared with a pose codebook indexed by canonical rotations for retrieval of the object rotation (see Figs. 1, 2), where the learned latent poses are more accurate for RGB-based pose estimation than codes mixing shape and pose information.

Learning such a disentangled representation faces a critical challenge: the different symmetries of objects do not admit one pose codebook applicable to all objects. To understand this difficulty, consider that the cup in Fig. 1 has distinctive codes for representing the left and right views, but the rotational symmetry of the bottle demands an identical code for the two views. This exemplifies the frequent infeasibility of disentangling an input image into *independent* latent factors by a neural network [3], the factors being shape and pose in our case.

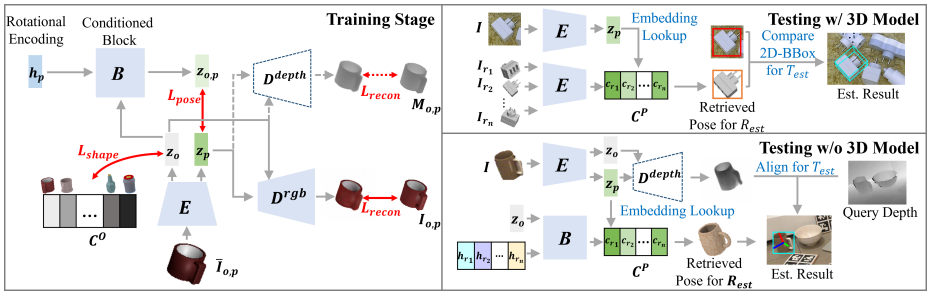


Fig. 2: **Network structures** in the training (left) and testing stage (right) for different settings. If testing objects have available 3D models (or not), we train an RGB decoder only (or plus a depth decoder) (left). During test stage, object rotation is *purely RGB-based estimation* by retrieving from the codebook $\mathcal{C}^P$, which is constructed by encoding the given object views (top right), or by shape code conditioned generation (bottom right). Translation is computed by pinhole camera (top right) or by depth comparison (bottom right).

To solve this dependent disentanglement problem, we model the shape-pose dependency by introducing a module that *re-entangles* the shape and rotation and generates an object-conditioned pose codebook respecting the object symmetry, against which the query latent pose code is compared for pose retrieval. In addition, to facilitate generalization to novel objects, we take advantage of the decoupled latent shape space and apply contrastive metric learning, which encourages objects with similar geometry to have similar shape codes. By training the system with diverse shapes, novel objects can be robustly processed by referring to similar training objects with proximate latent shape codes.

We evaluate our approach by training on synthetic data only and testing on real data. Our approach allows for evaluations of two different settings proposed by previous works, i.e., the textureless CAD objects without category labels proposed by [48] and the daily objects with specified categories by [56], on which we compare favorably than state-of-the-art methods that similarly work with RGB images for rotation estimation. In addition, we extend to a more challenging setting of daily objects without leveraging the category information by mixing the objects from [56], on which our approach preserves competitive performance. These results demonstrate the improved scalability of our method. Finally, extensive ablation studies confirm the effectiveness of disentangled shape and pose learning and other design choices.

2 Related Works

6D Pose Estimation There is a massive literature on instance-level pose estimation from RGB(D) images (see [33] for a survey). These works can be roughly classified into three streams, i.e., by direct pose regression [59,29,2], by registering 2D and 3D points [4,47,51,44,42,22,34], and by template retrieval

[21,49,50,61,57]. For instance-level pose estimation, learning-based methods train a specialized network for each testing object.

Wang *et al.* [56] propose a shared 3D shape space (NOCS) for all instances from the same category, where the objects are pre-aligned and normalized into a common coordinate system. Variations among the instances in the NOCS space are expected to be smooth and predictable, to make the NOCS mapping learnable when trained on large scale categorical datasets like ShapeNet [5]. For pose estimation, the pixels of a detected object are mapped to 3D points in the NOCS space, which are registered with the input depth image to find the 6D rigid transformation along with scaling. Grabner *et al.* [15] use a similar canonical object coordinate representation for category level 3D model retrieval.

Subsequent works improve the categorical pipeline by modeling the shape differences inside a category adaptively, with many of them fusing depth with RGB input for more accurate translation and scale estimation [11,53,6,7,9,35]. Specifically, within the RGB-input domain, Chen *et al.* [11] propose an analysis-by-synthesis approach to minimize the difference between the input image and a 2D object view synthesized by neural rendering, by gradient descent on both shape and pose variables. All these category-level approaches train different network branches for each category to learn and utilize the intra-category shape consistency.

In comparison, our scalable approach can accommodate categories of different symmetries with a common network path that learns the inter- and intra-categorical features adaptively (Fig. 1). Similarly, StarMap [63] and PoseContrast [60] work on the cross-category setting for estimating only the 3D rotation; however, they do not address object symmetries. LatentFusion [41] does not assume categorical objects either, but requires multiple view images for neural reconstruction before pose estimation.

Multipath-AAE [48] works under a different assumption: the novel test objects share little shape consistency with training objects but have 3D models available, which is practical for industrial manufacturing settings. Multipath-AAE extends the augmented auto-encoder approach [50] by sharing an encoder to learn the latent pose embedding and assigning to each object a separate decoder, which bypasses the large shape differences across objects and enables auto-encoding. The shared encoder therefore learns pose-aware features that generalize to different objects. This setting is followed by Pitteri *et al.* [45] who use learned local surface embedding for pose estimation, and Nguyen *et al.* [39] who improve robustness by modeling occlusion. Compared with [48], our disentanglement of shape and pose allows the auto-encoding without multi-path decoders for different objects, thus making the framework more scalable. However, the disentanglement into independent factors is challenging to learn and we propose re-entanglement to generate shape conditioned pose codebook for feasible learning.

Disentangled Representation Learning Disentangled representations are a key objective for interpretable and generalizable learning [1,36]. Previous works encourage disentangled representation learning by unsupervised learning [19,10].

Recently, focus has been given to the conditions under which learned representations can be disentangled [37,18,30], with the finding that quite frequently the direct mapping to disentangled independent factors is unattainable for neural networks [3]. Our discussion on scalable 6D pose estimation exemplifies the situation: the disentanglement of object shape and pose as independent factors is prevented by different object symmetries. We provide a solution to the disentanglement problem by re-entangling the independent factors so that a neural network mapping can be learned.

3 Method

As shown in Fig. 2, our overall framework is an auto-encoder that learns to encode an RGB image of the observed object to its latent shape code and object-dependent pose code separately, where the latent pose code is compared with a codebook of implicit rotation representations for fast pose estimation. Therefore, our approach obtains the object rotation purely from RGB input; depth input and reconstruction are optionally used only to remove translation/scale ambiguity when the object size is unknown (Sec. 4).

3.1 Disentangled Shape and Pose Learning

Given the input RGB image $\mathbf{I}_{o,p} \in \mathbb{R}^{3\times H\times W}$ for an object o under pose $p \in \mathrm{SE}(3)$, the encoder E maps $\mathbf{I}_{o,p}$ to a low-dimensional latent code $E(\mathbf{I}_{o,p}) = (\boldsymbol{z}_o, \boldsymbol{z}_p) \in \mathbb{R}^{2d}$ with $d \ll H\times W$, where $\boldsymbol{z}_o, \boldsymbol{z}_p \in \mathbb{R}^d$ encode the implicit shape and pose representations, respectively.

The decoder D^{rgb} tries to recover the input image from latent codes. Since we expect $\boldsymbol{z}_o$ and $\boldsymbol{z}_p$ to encode the overall object appearance and the view-specific appearance respectively, we borrow ideas from generative models [13,27,28] and use the AdaIN modulation [26] in the decoder to condition the per-view reconstruction on the object code; the detailed decoder structure can be found in the supplementary document. Moreover, we have tested by switching the roles of $\boldsymbol{z}_o$ and $\boldsymbol{z}_p$ for the decoder and found degraded performance (see supplemental).

Since we use only synthetic data for training, to narrow the domain gap between synthetic and real data, we follow [48,49] and adopt data augmentations that randomly change the color and scaling of an input image $\mathbf{I}$ to obtain the augmented image $\bar{\mathbf{I}}$, and aim to recover the canonical image $\mathbf{I}$ by auto-encoding. The loss function of the auto-encoding task therefore is

$$L_{recon} = \sum_{o,p} ||\mathbf{I}_{o,p} - D^{rgb}(E(\bar{\mathbf{I}}_{o,p}))||^2. \tag{1}$$

Note that our design accommodates different objects by sharing the same pair of encoder-decoder E and D^{rgb}, and hence is different from [48] that assigns to each object an individual decoder and previous instance-level approaches that train a specialized network for each object.

3.2 Contrastive Metric Learning for Object Shapes

The key to the generalization of pose estimation to a novel object is to exploit its similarity with the training objects, so that its generated pose codebook (Sec. 3.3) can capture its symmetry by referring to that of similar training objects. To learn such similarity relationships, we build a metric space for the shape codes of training objects by contrastive metric learning [58,40,16,8].

Denote the training object set as $\mathcal{O} = \{o_i\}_{i\in[N_O]}$, where N_O is the number of training objects. Similar to [58], to learn the contrastive metric among shape codes, we establish a shape embedding $\mathcal{C}^O \in \mathbb{R}^{N_O\times d}$ containing codes $\{\boldsymbol{c}_i \in \mathbb{R}^d\}_{i\in[N_O]}$, each corresponding to a training object. We then define the proximity of $\boldsymbol{c}_i$ to $\boldsymbol{z}_o$ in the form of probability distribution as

$$\Pr(\boldsymbol{c}_i|\boldsymbol{z}_o) = \frac{\exp(\hat{\boldsymbol{c}}_i \cdot \hat{\boldsymbol{z}}_o/\tau)}{\sum_{j=1}^{N}\exp(\hat{\boldsymbol{c}}_j \cdot \hat{\boldsymbol{z}}_o/\tau)} \tag{2}$$

where $\tau = 0.07$ is a temperature parameter controlling the sharpness of the distribution, and $\hat{\boldsymbol{a}} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}$ denotes normalized unit-length vectors.

The target distribution given o is simply a one-hot vector $\boldsymbol{w}^o \in \{0,1\}^{N_O}$, with $\boldsymbol{w}^o_i = 1$ if $o = o_i$ and the rest entries being zero. The contrastive metric loss for learning the shape space is then defined as

$$L_{shape} = -\sum_{o,p}\sum_{i=1}^{N_O} \boldsymbol{w}^o_i \log \Pr(\boldsymbol{c}_i|\boldsymbol{z}_o). \tag{3}$$

To minimize the above loss, while $\boldsymbol{z}_o$ is updated by the SGD solver during each training step, we update the shape embedding $\mathcal{C}^O$ by the exponential moving average (EMA) with decay rate d_s, thus making $\boldsymbol{c}_o$ a smoothed history of $\boldsymbol{z}_o$. Details of the EMA update can be found in the supplementary document.

3.3 Re-entanglement of Shape and Pose

The pose code $\boldsymbol{z}_p$ is compared with a codebook of sampled canonical orientations to retrieve the object rotation (Figs. 1, 2). As noted in Fig. 1, different object symmetries demand object-specific pose codebooks. To generate such a conditioned pose codebook, we propose a distributed representation of rotations and a transformation that entangles rotations with shape code in a generalizable way.

Rotational Position Encoding We need to distinguish between different rotations in a canonical pose representation. Inspired by the positional encoding in sequence models [55], we have adopted the 4D hyper spherical harmonics (HSH) rotation encoding. The HSH is a set of orthogonal basis functions on the 4D hypersphere that mimic the sine/cosine wave functions for positional encoding in sequence models: it is a distributed vector representation that can extend to high dimensions ($d = 128$ in our case), has a multi-spectrum structure that encodes both high frequency and low frequency variations of rotations, and has periodic

structures with fixed linear transformations for relative rotations [62,43]. Denoting the HSH function as $Z_{nl}^{m}(\beta,\theta,\phi)$, with $\beta \in [0,2\pi]$, $\theta \in [0,\pi]$, $\phi \in [0,2\pi]$ as the in-plane rotation, zenith and azimuth angles respectively and l,m,n as polynomial degrees, we obtain the 128-dim vector encoding $\boldsymbol{h}_p$ by ranging over $n \in [0,\cdots,6]$ with $0 \leq l \leq n$, $0 \leq m \leq l$. Details of the construction can be found in the supplemental document.

Conditioned Pose Code Generation We design a conditional block B to entangle the object code $\boldsymbol{z}_o$ with the rotational position encoding $\boldsymbol{h}_p$ of rotation p and output a pose code $\boldsymbol{z}_{o,p} = B(\boldsymbol{z}_o, \boldsymbol{h}_p)$ comparable with $\boldsymbol{z}_p$ (Fig. 2).

Entanglement is a recurring topic in machine learning, with implementation techniques like parameter generation [46,54,13] that boil down to a tensor product structure [52,38]. Therefore, we introduce a 3rd-order learnable tensor $\mathbf{W} \in \mathbb{R}^{d\times d\times d}$ and apply the following two-step transformation B to obtain the entangled pose code:

$$\boldsymbol{z}'_{o,p} = \mathbf{W}\left(FC(sg(\boldsymbol{z}_o)), FC(\boldsymbol{h}_p)\right), \quad \boldsymbol{z}_{o,p} = \texttt{FFN}(\boldsymbol{z}'_{o,p}), \tag{4}$$

where $FC(sg(\boldsymbol{z}_o)), FC(\boldsymbol{h}_p) \in \mathbb{R}^d$ are the pre-processing of $\boldsymbol{z}_o$ and $\boldsymbol{h}_p$, $sg(\cdot)$ is to stop gradient back-propagation as the shape code $\boldsymbol{z}_o$ is a pre-condition not to be updated by pose learning (see Sec. 5.5, Tab. 2 for an ablation), and $\mathbf{W}(\cdot,\cdot)$ denotes the tensor contraction along its first two orders. A feed-forward residual block `FFN` is followed to generate the final pose code $\boldsymbol{z}_{o,p}$.

To synchronize the pose representation computed via the conditional block with that learned by the encoder, we minimize the cosine distance between $\boldsymbol{z}_{o,p}$ and $\boldsymbol{z}_p$ during training:

$$L_{pose} = -\sum_{o,p} \hat{\boldsymbol{z}}_{o,p} \cdot \hat{\boldsymbol{z}}_p. \tag{5}$$

In summary, our total training loss combines the reconstruction loss (Eq. (1)), the contrastive loss for shape space (Eq. (3)) and the synchronization loss between pose representations from B and E (Eq. (5)), with weights λ_1, λ_2:

$$L = L_{recon} + \lambda_1 L_{shape} + \lambda_2 L_{pose}.$$

4 Inference under Different Settings

In the test stage, we estimate rotation purely from RGB input, which takes three steps (Fig. 2, right): Given the query image crop $\mathbf{I}$ bounding the object of interest, we first obtain its latent shape and pose codes as $(\boldsymbol{z}_o, \boldsymbol{z}_p) = E(\mathbf{I})$, then build a pose embedding $\mathcal{C}^P \in \mathbb{R}^{N_P\times d}$ with each row $\boldsymbol{c}_q \in \mathbb{R}^d$ corresponding to the rotation q from a set of N_P canonical rotations $\mathcal{R} \subset SO(3)$, and finally retrieve the estimated pose as $q^* = \arg\max_{q\in\mathcal{R}} \hat{\boldsymbol{z}}_p \cdot \hat{\boldsymbol{c}}_q$. Translation (and scale) is estimated subsequently, which may use depth data to remove scale ambiguity.

Previous works on scalable pose estimation towards novel objects have assumed two different application scenarios as discussed below, on which our framework can be flexibly adapted and achieve state-of-the-art performances. We also present an extended setting to better explore the scalability of our approach.

Setting I: Novel Objects in a Given Category A series of works [56,11,53] assume that the novel testing objects are from a specific category but have no 3D models available. Therefore, for pose retrieval we compute $\mathcal{C}^P = \{B(\boldsymbol{z}_o, \boldsymbol{h}_q)\}_{q\in\mathcal{R}}$ from the sampled canonical rotations $\mathcal{R}$ and the shape code $\boldsymbol{z}_o$.

As the testing objects have no specific sizes in this setting, to remove the 2D-3D scale ambiguity and estimate translation and scale properly, we require the input depth map and compare it with a decoded canonical depth map. The estimation of translation and scale involves a simple outlier point removal process and mean depth comparison for translation estimation and bounding box comparison for scale estimation; for details please refer to the supplemental document. As shown in Fig. 2, the depth decoder D^{depth} is simply an additional branch parallel to the RGB decoder, supervised to reconstruct a canonical depth map $\mathbf{M}_{o,p} \in \mathbb{R}^{1\times H\times W}$ for the rotated object at a fixed distance away from the camera. The reconstruction loss in Eq. (1) is updated to be:

$$L_{recon} = \sum_{o,p} ||\mathbf{I}_{o,p} - D^{rgb}(E(\bar{\mathbf{I}}_{o,p}))||^2 + ||\mathbf{M}_{o,p} - D^{depth}(E(\bar{\mathbf{I}}_{o,p}))||^2 \qquad (6)$$

Comparison in Sec. 5.2 shows our improved rotation accuracy and robustness to object symmetries.

Setting II: Novel Objects with 3D Models Multipath-AAE [48] works with a set of CAD objects with drastic geometric differences and no specific category consistency. However, the 3D models of novel testing objects are accessible, as is common in applications like industrial manufacturing [48,45].

In this setting, we follow previous auto-encoding frameworks [48,50] to construct an offline pose codebook with the CAD model. Specifically, we first render images $\mathbf{I}_q$ of the given object under the reference orientations q and then obtain $\mathcal{C}^P = \{\boldsymbol{z}_q\}_{q\in\mathcal{R}}$, with $\boldsymbol{z}_q$ the pose code part of $E(\mathbf{I}_q)$. Given the physical size and camera intrinsics, translation is obtained purely from RGB input with the pinhole camera model. The decoder D^{rgb} is not used during the test stage. As shown in Sec. 5.4, our disentangled auto-encoder learns highly discriminative pose encoding that performs even better than per-object trained auto-encoders, and generalizes well to novel objects with largely different shapes.

Setting III (Extension): Novel Objects across Categories without 3D Models We further challenge our method on an extension of setting I by combining objects of all categories in [56] into one set. Without referring to predefined category labels in training and testing, the task has never been addressed before in previous works[56,11,53]. As shown in Sec. 5.3, our disentangled auto-encoder enables a straightforward extension to this cross-category setting with marginal performance degrading compared to setting I, which demonstrates the scalability of our approach.

	NOCS [56]	Chen et al. [11]	Ours-per	Ours-all
Only synthetic training data	×	✓	✓	✓
Only RGB for rotation est.	×	✓	✓	✓
Extension to cross-category	×	×	×	✓

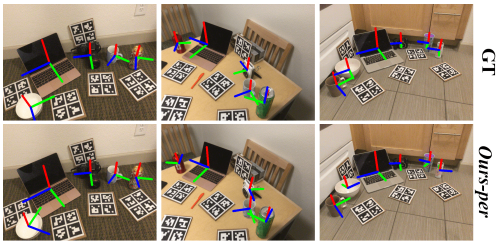


Fig. 3: Scope of compared methods on settings I and III (left), and qualitative cases of *Ours-per* (right). *All methods use query depth for translation estimation.*

5 Experiments

5.1 Setup

We resize the input images to $H \times W = 128 \times 128$, use a latent code dimension $d = 128$, and set $d_s = 0.9995$ for the EMA decay, $\lambda_1 = 0.004$, $\lambda_2 = 0.002$ for balancing the loss terms. We use the Adam optimizer [31] with default parameters and a learning rate of 0.0002, and train 50k iterations for settings I, II, and 150k iterations for setting III, with a batch size of 64 to convergence. Detailed network structure and training data preparation are in the supplementary document.

5.2 Setting I: Novel Objects in a Given Category

Dataset and Metrics The benchmark of [56] has two parts, i.e., CAMERA containing synthetic data and REAL275 containing real data, that span 6 categories of objects (*bottle, bowl, camera, can, laptop, mug*) situated in daily indoor scenes. Furthermore, the objects in a category have diverse scales, and due to the inherent 2D-3D scale ambiguity, the estimation of translation plus scaling is only possible when additional cues like depth are given.

We use the synthetic CAMERA dataset with 1085 objects for training and evaluate on the real test set of REAL275, and follow [11] to report the average precision (AP) at different thresholds of rotation and translation errors. Note that while [11] uses input depth for improved translation estimation, it assumes a fixed scale and thus does not address scale estimation. Nevertheless, for completeness we report our scale estimation result by measuring 3D IoU precision in the supplemental document.

Baselines The most relevant baseline is [11], as both methods train on synthetic data only and test on real data, and estimate rotation based on RGB input only and use depth only for translation estimation. Another baseline is the earlier [56], which however trains on both real and synthetic data and relies on input depth for rotation estimation. All three methods use the same 2D detection backbone Mask-RCNN adopted from [56]. We summarize the differences in scopes of three methods in Fig. 3(left) where our method in this setting is denoted *Ours-per*,

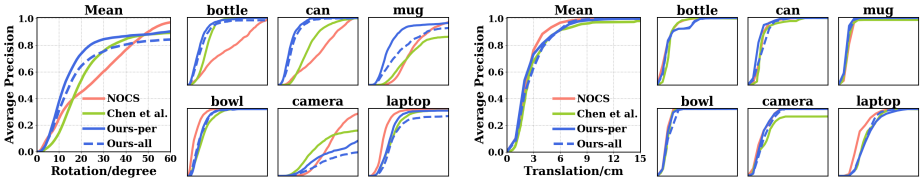


Fig. 4: **Comparison on REAL275** of average precision (ranging from 0 to 1) at different rotation error (left, ranging from 0 to 60°) or translation error (right, ranging from 0 to 15cm) thresholds. We report *Ours-per* of setting I per-category level and *Ours-all* of setting III combining all 6 categories.

and defer an empirical discussion of more category-level methods taking RGB-D input for rotation estimation [6,53,35,7] to the supplemental.

Pose Codebook 5K reference rotations are obtained by K-means clustering on the CAMERA training set rotations. Generating a pose codebook from 5K HSH codes takes 0.04s on a GTX 1080 GPU and can be batched for more objects.

Results As shown in Fig. 4, compared with Chen *et al.*[11], our rotation estimation has increased AP with a significant margin when the error threshold is below 40°; meanwhile, both methods have comparable performances on translation estimation. Compared with NOCS[56], our margin is even more significant throughout the range of 10° to 45° for rotation estimation. Qualitative results are visualized in Fig. 3. Among the different categories, we perform better in the classes of bottle, can and mug, which have strong partial symmetries and our method handles robustly. However, the camera category poses difficulty to our method; the main reason is that subtle textures are needed to distinguish vastly different poses, e.g., the front and back of a camera are quite similar for flat lens, but there are few objects out of the totally 74 objects in training set to cover such texture diversities. In comparison, both [11] and [56] use optimization to search for rotation and are more resilient to severe train/test disparities. For scale estimation, our 3D IoU accuracy is comparable to [56] (see supplemental).

5.3 Setting III (Extension): Novel Objects across Categories without 3D Models

We further challenge our method on the extended setting that combines all 6 categories of the NOCS benchmark into one set, without referring to category labels in training and testing; the trained network is denoted *Ours-all*. As we learn a metric shape space without the need for category labels (Sec. 3.2), we expect our method to extend to this cross-category setting without much difficulty.

As shown in Fig. 4, for rotation estimation, *Ours-all* achieves improved results than Chen *et al.*[11] for error thresholds $< 28°$, and NOCS [56] for error thresholds in $10° \sim 40°$, even though [11,56] train per-category network modules to exploit the intra-category consistency. Meanwhile, the lower performance compared with *Ours-per* can be attributed to the confusion of shape-conditioned

pose learning introduced by the increased cross-category shape variances, as for example under certain views a mug with an occluded handle looks quite similar to a can or bowl, but they are forced to generate pose codes with different symmetries. Qualitative cases are given in the supplemental.

Although none of the previous works [11,56] are designed to address this setting, for a better understanding of the challenge, we adapt and retrain NOCS [56] by using a single head for all categories (*i.e., NOCS-all*); without per-category correspondence consistency, we find that *NOCS-all* performs poorly especially for rotation estimation. We also retrain PoseContrast [60] under our setting, which is the state-of-the-art for cross-category rotation estimation. Results show that [60] does not handle objects with different symmetries as well as we do. Details are given in the supplemental.

5.4 Setting II: Novel Objects with 3D Models

Dataset and Metrics Following [48], we evaluate on T-LESS [23] which contains 30 textureless industrial parts with very different shapes and symmetries (see the supplementary for a visualization). Accuracy is measured by the recall rate of visible surface discrepancy metric $e_{VSD} < 0.3$ [24] at distance tolerance $20mm$, among test instances with visible portion >10%.
Baselines We compare with Multipath-AAE[48], Pitteri *et al.* [45], and Nguyen *et al.* [39]. All these methods share the same setting by training jointly on only the first 18 objects and testing on all 30 objects, using CAD models from TLESS.
Pose Codebook We follow [49,50,48] to build for each test object an offline pose codebook with 92232 reference rotations, that is formed by combining 36 in-plane rotations and 2562 equidistant spherical views sampled via [20].
Results We first report in Tab. 1(a) the accuracy for all test instances with 2D GT bounding boxes. We outperform Multipath-AAE [48] by 4% on average for the novel objects (*i.e.,* Obj 19-30) and 5% for the trained objects (*i.e.,* Obj 1-18), although Multipath-AAE [48] assigns separate decoders for the 18 training objects and optionally uses the GT mask to eliminate background noise for better performances. We also outperform the concurrent work by Nyugen *et al.* [39]. For a more complete evaluation, we further compare with [49,50] which train for each of the 30 objects a specific auto-encoder, and find our result still outperforms it by 3% on the 18 training objects of ours. These results show that our disentanglement learning improves the auto-encoder framework and generalizes to objects with different shapes and symmetries (see Sec. 5.5, Fig. 6 for detailed analysis).

We then report in Tab. 1(b) the evaluation under the full 2D detection and pose estimation pipeline, by adopting Mask-RCNN [17] from [32] as the 2D detector and following the single object single instance protocol [24]. Our result improves over that of the comparing methods by a significant margin of around 12%. Our qualitative cases are in Fig. 5 and the per-object recall rates are given in the supplementary.
Instance-Level Estimation Although we focus on scalable pose estimation for novel test objects, it is possible to apply our framework to an instance-level task

Table 1: **Comparison on T-LESS.** Reported are the average recall rates with $e_{VSD} < 0.3$. All methods were trained with only the first 18 objects, except AAE[49,50] which trains individual networks for each of the 30 objects.

(a) w/ 2D GT bboxes, † for using GT mask

Ave. on	Obj 1-18	Obj 19-30	Obj 1-30
AAE[49,50]	62.57	**66.63**	64.19
Multipath-AAE[48]	51.75	52.49	52.04
Multipath-AAE[48]†	60.75	59.89	60.41
Nguyen *et al.*[39]	59.62	57.75	58.87
Ours	**66.14**	**64.42**	**65.45**

(b) w/ MaskRCNN [17] detection.

Ave. on	Obj 1-30
Multipath-AAE[48]	23.51
Pitteri *et al.* [45]	23.27
Ours	**35.36**

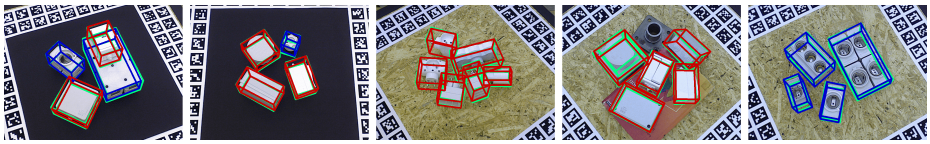


Fig. 5: **Qualitative results on T-LESS** of setting II. We denote our estimations in blue (trained objects) and red (unseen objects), and GT poses in green.

where all test objects are given for training. We provide such a limit case study in the supplementary, and compare with more instance-level pose estimation methods on the BOP leader board [49,50,42,34,32,22]. Our approach provides fast yet accurate pose estimations that can be further improved by refinement.

5.5 Ablation Study

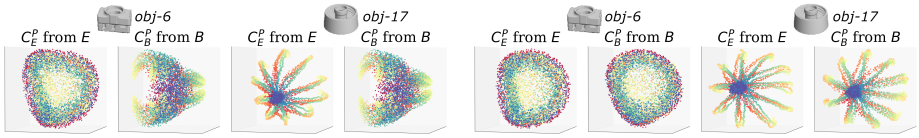
Shape Conditioned Pose Code Generation We first discuss the necessity to generate shape-dependent pose codes. To this end, we separate shape codes from pose codebook generation by replacing the 3rd-order tensor $\mathbf{W}$ in Eq. (4) with a multi-layer perceptron $\texttt{MLP}$ that takes only the HSH encoding as input, *i.e.* $\texttt{MLP}(FC(\boldsymbol{h}_p))$. The $\texttt{MLP}$ has four layers of width $[1024, 1024, 1024, 128]$ and thus more trainable weights than $\mathbf{W}$. The average precision on setting III reported in Tab. 2 (2nd, 6th rows) shows that the performance significantly drops when the shape code is separated from pose code generation, indicating the difficulty of learning independent latent representations of shape and pose.

To further visualize the effectiveness of pose code generation, given an object o, we inspect two sets of latent pose representations: $\mathcal{C}_E^P = \{\boldsymbol{z}_p\}_{p\in\mathcal{R}}$ generated by the encoder E and $\mathcal{C}_B^P = \{\boldsymbol{z}_{o,p}\}_{p\in\mathcal{R}}$ by the conditioned block B. $\mathcal{R}$ has 8020 rotations from a combination of 20 in-plane rotations and 401 quasi-equidistant views sampled via [14]. Ideally, the two sets of latent codes should coincide with each other, so that they can be compared for effective rotation estimation.

We show in Fig. 6 for two T-LESS training objects: the box-like Obj-6 and the cylinder-like Obj-17, where with our entanglement of shape and pose information, $\mathcal{C}_B^P$ well synchronize with $\mathcal{C}_E^P$ for objects with different degrees of symmetry,

Table 2: **Ablation tests** on the design of shape conditioned pose code generation and contrastive learning for object shape. Reported are mAP at different rotation error thresholds (in degrees) for mixed categories of REAL275 (setting III).

Design of B	w/ L_{shape}	AP_5	AP_{10}	AP_{15}	AP_{20}	AP_{30}	AP_{60}
$\texttt{MLP}(FC(\boldsymbol{h}_p))$	✓	4.7	15.7	28.9	36.9	47.0	72.4
$\texttt{MLP}(FC(sg(\boldsymbol{z}_o)), FC(\boldsymbol{h}_p))$	✓	7.5	27.3	47.8	61.8	74.9	84.3
$\mathbf{W}\,(FC(\boldsymbol{z}_o), FC(\boldsymbol{h}_p))$	✓	6.6	26.6	47.8	62.0	**76.6**	**87.5**
$\mathbf{W}\,(FC(sg(\boldsymbol{z}_o)), FC(\boldsymbol{h}_p))$	×	2.8	15.2	33.6	48.9	67.6	81.3
$\mathbf{W}\,(FC(sg(\boldsymbol{z}_o)), FC(\boldsymbol{h}_p))$	✓	**9.1**	**30.9**	**50.7**	**64.4**	75.3	84.3



(a) w/o shape condition (b) w/ shape condition

Fig. 6: **Top three PCA projections of pose codes $\mathcal{C}_E^P$ and $\mathcal{C}_B^P$** from encoder E and condition block B for two T-LESS objects. Point colors (blue→green→red) encode rotations as viewpoints change from north pole to south pole. The shape conditioned pose codes well capture the symmetries and synchronize with encoder outputs (b), but unconditioned pose codes fail (a).

though for Obj-6 a global rotation of the PCA projections between $\mathcal{C}_B^P$ and $\mathcal{C}_E^P$ exists due to the nearly isotropic distribution of latent codes. On the contrary, when the shape code is isolated from generating the pose codebook, it becomes difficult for $\mathcal{C}_B^P$ to follow the pattern of $\mathcal{C}_E^P$ for different objects. Such contrast demonstrates the necessity of our entanglement. We further discuss in the supplementary for objects with texture solving the rotational ambiguity, where our pose codes can well capture the textural difference.

We then move on to validate the design of combining pose and shape. An intuitive idea is to simply concatenate the shape and pose rotational encoding and process by an MLP, *i.e.* $\texttt{MLP}(FC(sg(\boldsymbol{z}_o)), FC(\boldsymbol{h}_p))$, with $\texttt{MLP}$ having four layers of width $[1024, 1024, 1024, 128]$. The comparison in Tab. 2 (3rd, 6th rows) shows that the 3rd-order tensor outperforms $\texttt{MLP}$, thus verifying our design choice.

Finally, we validate the necessity to treat $\boldsymbol{z}_o$ as a pre-condition for pose code generation, by allowing gradients to be backpropagated through the conditioned pose code generation module to $\boldsymbol{z}_o$ instead. Tab. 2, 4th and 6th rows, show that pre-conditioning by stop gradient $sg(\boldsymbol{z}_o)$ performs better for rotation error thresholds $\leq 20°$, demonstrating its recognition of subtle pose differences.

Contrastive Metric Learning for Object Shapes The mAP in Tab. 2 (5th, 6th rows) demonstrates our gain from the contrastive metric learning of the shape space, where with the shape loss L_{shape} the generalization to unseen

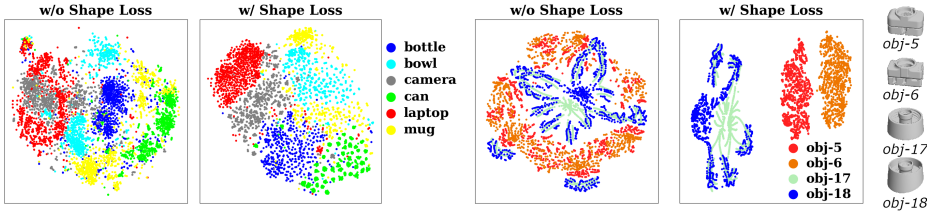


Fig. 7: **t-SNE embedding of shape codes z_o** for training images of six CAMERA categories (left) and four T-LESS objects (right). With contrastive metric learning the shape spaces show better regularity w.r.t. shape similarities.

objects is significantly improved. We also visualize the shape codes z_o with t-SNE in Fig. 7, for training samples from the CAMERA objects and 4 T-LESS objects. With shape space metric learning, we observe much better intra-category clustering and inter-category separation on CAMERA, though the network is unaware of category labels in this setting (setting III). For the T-LESS objects, the introduction of L_{shape} not only well separates the box-like objects (Obj-5,6) from the cylinder-like objects (Obj-17,18), but also recognizes the detailed geometric differences between Obj-5 and Obj-6; in comparison, the shape codes for different objects are mixed together without shape space metric learning.

6 Conclusion

We have presented a simple yet scalable approach for 6D pose estimation that generalizes to novel objects unseen during training. Building on an auto-encoding framework that handles object symmetry robustly, we achieve scalability by disentangling the latent code into shape and pose representations, where the shape representation forms a metric space by contrastive learning to accommodate novel objects, and the pose code is compared with canonical rotations for pose estimation. As disentanglement into independent shape and pose spaces is fundamentally difficult due to different object symmetries, we re-entangle shape code with pose codebook generation to avoid the issue. We obtain state-of-the-art results on two established settings when training with synthetic data only, and extend to a cross-category setting to further demonstrate scalability.

Limitation and Future Work We mainly focus on learning for rotation estimation from a single RGB image, while the translation estimation can be further improved by fully exploiting the input depth with neural networks, as discussed in [53,35]. Extending to multiview input for improved robustness under severe occlusion and inaccurate 2D detection is also a promising direction.

Acknowledgement This work was partially supported by the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative.

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8) (2013)
2. Billings, G., Johnson-Roberson, M.: Silhonet: An rgb method for 3d object pose estimation and grasp planning. arXiv preprint arXiv:1809.06893 (2018)
3. Bouchacourt, D., Ibrahim, M., Deny, S.: Addressing the topological defects of disentanglement via distributed operators (2021)
4. Brachmann, E., Michel, F., Krull, A., Ying Yang, M., Gumhold, S., et al.: Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3364–3372 (2016)
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository (2015)
6. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
7. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2773–2782 (2021)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
9. 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)
10. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems 29, pp. 2172–2180 (2016)
11. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category level object pose estimation via neural analysis-by-synthesis. In: European Conference on Computer Vision (ECCV) (2020)
12. Denninger, M., Sundermeyer, M., Winkelbauer, D., Zidan, Y., Olefir, D., Elbadrawy, M., Lodhi, A., Katam, H.: Blenderproc. arXiv preprint arXiv:1911.01911 (2019)
13. Dumoulin, V., Perez, E., Schucher, N., Strub, F., Vries, H.d., Courville, A., Bengio, Y.: Feature-wise transformations. Distill **3**(7), e11 (2018)
14. González, Á.: Measurement of areas on a sphere using fibonacci and latitude–longitude lattices. Mathematical Geosciences **42**(1), 49 (2010)
15. Grabner, A., Roth, P.M., Lepetit, V.: Location field descriptors: Single image 3d model retrieval in the wild. In: 2019 International Conference on 3D Vision (3DV). pp. 583–593. IEEE (2019)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2961–2969 (2017)
18. Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., Lerchner, A.: Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230 (2018)
19. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)
20. Hinterstoisser, S., Benhimane, S., Lepetit, V., Fua, P., Navab, N.: Simultaneous recognition and homography extraction of local patches with a simple linear classifier. In: BMVC. pp. 1–10 (2008)
21. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision. pp. 548–562. Springer (2012)
22. Hodaň, T., Baráth, D., Matas, J.: EPOS: Estimating 6D pose of objects with symmetries. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. IEEE Winter Conference on Applications of Computer Vision (WACV) (2017)
24. Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., et al.: Bop: Benchmark for 6d object pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 19–34 (2018)
25. Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: BOP challenge 2020 on 6D object localization. European Conference on Computer Vision Workshops (ECCVW) (2020)
26. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
27. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
28. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
29. 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)
30. Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: Proceedings of International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 108. PMLR (2020)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
32. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: Cosypose: Consistent multi-view multi-object 6d pose estimation. In: European Conference on Computer Vision. pp. 574–591. Springer (2020)
33. Lepetit, V.: Recent advances in 3d object and hand pose estimation. arXiv preprint arXiv:2006.05927 (2020)

34. Li, Z., Wang, G., Ji, X.: Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7678–7687 (2019)
35. 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. arXiv preprint arXiv:2103.06526 (2021)
36. Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
37. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97. PMLR (2019)
38. Martyn, J., Vidal, G., Roberts, C., Leichenauer, S.: Entanglement and tensor networks for supervised image classification (2020)
39. Nguyen, V.N., Hu, Y., Xiao, Y., Salzmann, M., Lepetit, V.: Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6771–6780 (June 2022)
40. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
41. Park, K., Mousavian, A., Xiang, Y., Fox, D.: Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
42. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. arXiv preprint arXiv:1908.07433 (2019)
43. Pasha Hosseinbor, A., Chung, M.K., Koay, C.G., Schaefer, S.M., van Reekum, C.M., Schmitz, L.P., Sutterer, M., Alexander, A.L., Davidson, R.J.: 4d hyperspherical harmonic (hyperspharm) representation of surface anatomy: A holistic treatment of multiple disconnected anatomical structures. Medical Image Analysis **22**(1), 89–101 (2015)
44. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4561–4570 (2019)
45. Pitteri, G., Bugeau, A., Ilic, S., Lepetit, V.: 3D Object Detection and Pose Estimation of Unseen Objects in Color Images with Local Surface Embeddings. In: 15th Asian Conference on Computer Vision. Kyoto (virtual conference), Japan (Nov 2020)
46. Platanios, E.A., Sachan, M., Neubig, G., Mitchell, T.: Contextual parameter generation for universal neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 425–435 (2018)
47. Rad, M., Lepetit, V.: Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3828–3836 (2017)
48. Sundermeyer, M., Durner, M., Puang, E.Y., Marton, Z.C., Vaskevicius, N., Arras, K.O., Triebel, R.: Multi-path learning for object pose estimation across domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 13916–13925 (2020)

49. 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)
50. Sundermeyer, M., Marton, Z.C., Durner, M., Triebel, R.: Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. International Journal of Computer Vision pp. 1–16 (2019)
51. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6d object pose prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 292–301 (2018)
52. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Computation **12**(6), 1247–1283 (2000). https://doi.org/10.1162/089976600300015349
53. Tian, M., Ang Jr, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (August 2020)
54. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Proc. Eur. Conf. Computer Vision (ECCV) (2020)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
56. 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 Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019)
57. Wen, Y., Pan, H., Yang, L., Wang, W.: Edge enhanced implicit orientation learning with geometric prior for 6d pose estimation. IEEE Robotics and Automation Letters (IROS) **5**(3) (2020)
58. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3733–3742 (2018)
59. 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)
60. Xiao, Y., Du, Y., Marlet, R.: Posecontrast: Class-agnostic object viewpoint estimation in the wild with pose-aware contrastive learning. In: 2021 International Conference on 3D Vision (3DV). pp. 74–84 (2021). https://doi.org/10.1109/3DV53792.2021.00018
61. Zhang, H., Cao, Q.: Detect in rgb, optimize in edge: Accurate 6d pose estimation for texture-less industrial parts. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 3486–3492. IEEE (2019)
62. Zhao, L.: Spherical and spheroidal harmonics: Examples and computations (2017)
63. Zhou, X., Karpur, A., Luo, L., Huang, Q.: Starmap for category-agnostic keypoint and viewpoint estimation. In: European Conference on Computer Vision (ECCV) (2018)