Zero-Shot Category-Level Object Pose Estimation

Walter Goodwin[1⋆], Sagar Vaze[2⋆], Ioannis Havoutis[1], and Ingmar Posner[1]

[1] Oxford Robotics Institute, University of Oxford
[2] Visual Geometry Group, University of Oxford
firstname@robots.ox.ac.uk

Abstract. Object pose estimation is an important component of most vision pipelines for embodied agents, as well as in 3D vision more generally. In this paper we tackle the problem of estimating the pose of novel object categories in a zero-shot manner. This extends much of the existing literature by removing the need for pose-labelled datasets or category-specific CAD models for training or inference. Specifically, we make the following contributions. First, we formalise the zero-shot, category-level pose estimation problem and frame it in a way that is most applicable to real-world embodied agents. Secondly, we propose a novel method based on semantic correspondences from a self-supervised vision transformer to solve the pose estimation problem. We further re-purpose the recent CO3D dataset to present a controlled and realistic test setting. Finally, we demonstrate that all baselines for our proposed task perform poorly, and show that our method provides a *six-fold improvement* in average rotation accuracy at 30 degrees. Our code is available at https://github.com/applied-ai-lab/zero-shot-pose.

Fig. 1: Zero-shot category-level pose estimation enables the alignment of different instances of the same object category, without any pose labels for that category or any other. For each category, the estimated pose of the first object, relative to the second, is visualised through transformation of the first object's point cloud.

1 Introduction

Consider a young child who is presented with two toys of an object category they have never seen before: perhaps, two toy aeroplanes. Despite having never seen examples of 'aeroplanes' before, the child has the ability to understand the spatial

⋆ These authors contributed equally

relationship between these related objects, and would be able to align them if required. This is the problem we tackle in this paper: the zero-shot prediction of pose offset between two instances from an object category, without the need for any pose annotations. We propose this as a challenging task which removes many assumptions in the current pose literature, and which more closely resembles the setting encountered by embodied agents in the real-world. To substantiate this claim, consider the information existing pose recognition algorithms have access to. Current methods make one (or more) of the following assumptions: that evaluation is performed at the *instance-level* (i.e there is no intra-category variation between objects) [41]; that we have access to *labelled pose datasets* for all object categories [3, 9, 18, 27, 38, 42, 45]; and/or that we have access to a realistic *CAD model* for each object category the model will encounter [8,17,44].

Meanwhile, humans can understand pose without access to any of this information. How is this possible? Intuitively, we suggest humans use an understanding of *object parts*, which generalise across categories, to correspond related objects. This process can be followed by using *basic geometric primitives* to understand the spatial relationship between objects. Humans typically also have a coarse *depth estimate* and can inspect the object from *multiple viewpoints*.

In this paper, we use these intuitions to build a solution to estimate the pose offset between two instances of a given category. We perform 'zero-shot' pose-estimation in the sense that our models have never seen pose-labelled examples of the test categories, and neither do they rely on category-specific CAD models. We first make use of features extracted from a vision transformer (ViT [13]), trained in a self-supervised manner on large scale data [6], to establish semantic correspondences between two object instances of the same category. Prior work has demonstrated that self-supervised ViTs have an understanding of object parts which can transfer to novel instances and categories [4, 36]. Next, using a weighting of the semantic correspondences, we obtain a coarse estimate of the pose offset by select an optimal viewpoint for one of the object instances. Having obtained semantic correspondences and selected the best view, we use depth maps to create sparse point clouds for each object at the corresponding semantic locations. Finally, we align these point clouds with a rigid-body transform using a robust least squares estimation [35] to give our final pose estimate.

We evaluate our method on the CO3D dataset [28], which provides high-resolution imagery of diverse object categories, with substantial intra-category variance between instances. We find that this allows us to reflect a realistic setting while performing quantitative evaluation in a controlled manner. We consider a range of baselines which could be applied to this task, but find that they perform poorly and often fail completely, demonstrating the highly challenging nature of the problem.

In summary, we make the following contributions:

- We formalise a new and challenging setting for pose estimation, which is an important component of most 3D vision systems. We suggest our setting closely resembles those encountered by real-world embodied agents (Sec. 3).

– We propose a novel method for zero-shot, category-level pose estimation, based on semantic correspondences from self-supervised ViTs (Sec. 4).
– Through rigorous experimentation on a devised CO3D benchmark, we demonstrate that our method facilitates zero-shot pose alignment when the baselines often fail entirely (Sec. 5).

2 Related Work

2.1 Category-level pose estimation

While estimating pose for a single object instances has a long history in robotics and computer vision [41], in recent years there has been an increased interest in the problem of *category-level* pose estimation, alongside the introduction of several category-level datasets with labelled pose [2, 38–40]. Approaches to category-level pose estimation can be broadly delineated into: those defining pose explicitly through the use of reference CAD models [17,29,31,31,44]; those which learn category-level representations against which test-time observations can be in some way matched to give relative pose estimates [7, 8, 11, 27, 33, 37, 38, 45]; and those that learn to directly predict pose estimates for a category from observations [3, 9, 18, 42].

Most methods (e.g. [8, 24, 27, 33, 38]) treat each object category distinctly, either by training a separate model per category, or by using different templates (e.g. CAD models) for each category. A few works (e.g. [42, 45]) attempt to develop category-agnostic models or representations, and several works consider the exploitation of multiple views to enhance pose estimation [20,21]. In contrast to existing works in category-level pose estimation, we do not require any pose-labelled data or CAD models in order to estimate pose for a category, and tackle pose estimation for unseen categories.

2.2 Few-shot and self-supervised pose estimation

There has been some recent work that notes the difficulty of collecting large, labelled, in-the-wild pose datasets, and thus seeks to reduce the data burden by employing few-shot approaches. For instance, Pose-from-Shape [44] exploits existing pose-labelled RGB datasets, along with CAD models, to train an object-agnostic network that can predict the pose of an object in an image, with respect to a provided CAD model. Unlike this work, we seek to tackle an in-the-wild setting in which a CAD model is not available for the objects encountered. Self-supervised, embodied approaches for improving pose estimation for given object instances have been proposed [12], but require extensive interaction and still do not generalise to the category level. Few-shot approaches that can quickly fine-tune to previously unseen categories exist [34, 42], but still require a non-trivial number of labelled examples to fine-tune to unseen categories, while in contrast we explore the setting in which no prior information is available. Furthermore, recent work has explored the potential for unsupervised methods with equivariant inductive biases to infer category-level canonical frames without labels [23],

and to thus infer 6D object pose given an observed point cloud. This method, while avoiding the need for pose labels, only works on categories for which it has been trained. Finally, closest in spirit to the present work is [14], who note that the minimal requirement to make zero-shot pose estimation a well-posed problem is to provide an implicit canonical frame through use of a reference image, and formulate pose estimation as predicting the relative viewpoint from this view. However, this work can only predict pose for single object instances, and does not extend to the category level.

2.3 Semantic descriptor learning

A key component of the presented method to zero-shot category level pose estimation is the ability to formulate semantic keypoint correspondences between pairs of images within an object category, in a zero-shot manner. There has been much interest in semantic correspondences in recent years, with several works proposing approaches for producing these without labels [1, 4, 6, 22]. Semantic correspondence is particularly well motivated in robotic settings, where problems such as extending a skill from one instance of an object to any other demand the ability to relate features across object instances. Prior work has considered learning dense descriptors from pixels [15] or meshes [32] in a self-supervised manner, learning skill-specific keypoints from supervised examples [26], or robust matching at the whole object level [16]. The descriptors in [15, 26, 32] are used to infer the relative pose of previously unseen object instances to instances seen in skill demonstrations. In contrast to these robotics approaches, in our method we leverage descriptors that are intended to be category-agnostic, allowing us to formulate a zero-shot solution to the problem of pose estimation.

3 Zero-shot Category-Level Pose Estimation

In this section, we formalise and motivate our proposed zero-shot pose estimation setting. To do this, we first outline the generic problem of object pose estimation. 6D pose estimation is the regression problem of, given an image of the object, regressing to the offset (translation and rotation) of the object with respect to some frame of reference. This frame of reference can be defined implicitly (e.g in the supervised setting, the labels are all defined with respect to some 'canonical' frame) or explicitly (e.g with a reference image). In either case, pose estimation is fundamentally a relative problem. In the zero-shot setting we consider, the frame of reference cannot be implicitly defined by labels: we do not have labelled pose for any objects. Therefore, the pose estimation problem is that of aligning (computing the pose offset between) two instances of a given category.

In our proposed setting, we assume access to N views of a target object, as well as depth information for both objects, and suggest that these constraints reflect practical settings. For objects in the open-world, we are unlikely to have realistic CAD models or labelled pose training sets. On the other hand, many embodied agents are fitted with depth cameras or can recover depth (up to a

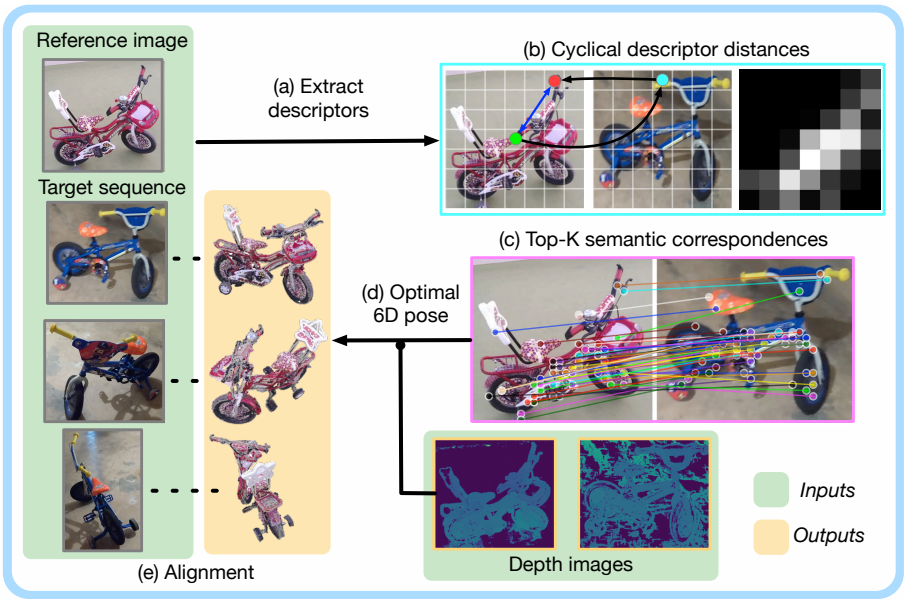

Fig. 2: Our method for zero-shot pose estimation between two instances of an object, given a reference image and a sequence of target images. In our method, we: **(a)** Extract spatial feature descriptors for all images with a self-supervised vision transformer (ViT). **(b)** Compare the reference image to all images in the target sequence by building a set of cyclical distance maps (Sec. 4.1). **(c)** Use these maps to establish K semantic correspondences between compared images and select a suitable view from the target sequence (Sec. 4.2). **(d)** Given the semantic correspondences and a suitable target view, we use depth information to compute a rigid transformation between the reference and target objects (Sec. 4.3). **(e)** Given relative pose transformations between images in the target sequence, we can align the point cloud of the *reference image* with *the entire target sequence*.

scale) from structure from motion or stereo correspondence. Furthermore, real-world agents are able to interact with the object and hence gather images from multiple views.

Formally, we consider a reference image, $I_{\mathcal{R}}$, and a set of target images $I_{\mathcal{T}_{1:N}} = \{I_{\mathcal{T}_1} ... I_{\mathcal{T}_N}\}$, where $I_i \in \mathbb{R}^{H \times W \times 3}$. We further have access to depth maps, $D_i \in \mathbb{R}^{H \times W}$ for all images. Given this information, we require a model, $\mathcal{M}$, to output a single 6D pose offset between the object in the reference image and the object in the target sequence, as:

$$T^* = \mathcal{M}(I_{\mathcal{R}}, I_{\mathcal{T}_{1:N}} | \quad D_{\mathcal{R}}, D_{\mathcal{T}_{1:N}}) \tag{1}$$

Finally, we note that, in practice, the transformations between the target views must be known for the predicted pose offset to be most useful. These

transformations are easily computed by an embodied agent and can be used to, given an alignment between $I_{\mathcal{R}}$ and *any* of the target views, align the reference instance with the entire target sequence.

4 Methods

In this section, we detail our method for zero-shot pose estimation. First, semantic correspondences are obtained between the reference and target object (Sec. 4.1). These correspondences are used to select a suitable view for pose estimation from the N images in the target sequence (Sec. 4.2). Finally, using depth information, the correspondences' spatial locations are used to estimate the pose offset between the reference and target object instances (Sec. 4.3).

4.1 Self-supervised semantic correspondence with cyclical distances

The key insight of our method is that semantic, parts-based correspondences generalise well between different object instances within a category, and tend to be spatially distributed in similar ways for each such object. Indeed, a parts-based understanding of objects can also generalise between categories; for instance, 'eyes', 'ears' and 'nose' transfer between many animal classes. Recent work has demonstrated that parts-based understanding emerges naturally from self-supervised vision transformer (ViT) features [4,6,36], and our solution leverages such a network with large scale pre-training [6]. The ViT is trained over ImageNet-1K, and we assume that it carries information about a sufficiently large set of semantic object parts to generalise to arbitrary object categories.

As described in Sec. 3, the proposed setting for pose estimation considers a relative problem, between a reference object (captured in a single image) and a target object (with potentially multiple views available). We compare two images (for now referred to as I_1, I_2) by building a 'cyclical distance' map for every pixel location in I_1 using feature similarities. Formally, consider $\Phi(I_i) \in \mathbb{R}^{H' \times W' \times D}$ as the normalised spatial feature of an image extracted by a ViT. Letting u be an index into $\Phi(I_1)$ as $u \in \{1...H'\} \times \{1...W'\}$, we find it's cyclical point u' as:

$$u' = \underset{w}{\operatorname{argmin}}\, d(\Phi(I_1)_w, \Phi(I_2)_v) \quad | \quad v = \underset{w}{\operatorname{argmin}}\, d(\Phi(I_1)_u, \Phi(I_2)_w) \tag{2}$$

Here $d(\cdot, \cdot)$ is the L2-distance, and a cyclical distance map is constructed as $C \in \mathbb{R}^{H' \times W'}$ with $C_u = -d(u, u')$. Using the top-K locations in C, we take features from $\Phi(I_1)$ and their nearest neighbours in $\Phi(I_2)$ as correspondences. This process is illustrated in Fig. 2b.

The cyclical distance map can be considered as a soft mutual nearest neighbours assignment. Mutual nearest neighbours [4] between I_1 and I_2 return a cyclical distance of zero, while points in I_1 with a small cyclical distance can be considered to 'almost' have a mutual nearest neighbour in I_2. The proposed cyclical distance metric has two key advantages over the hard constraint. Firstly,

while strict mutual nearest neighbours gives rise to an unpredictable number of correspondences, the soft measure allows us to ensure K semantic correspondences are found for every pair of images. We find having sufficient correspondences is critical for the downstream pose estimation. Secondly, the soft constraint adds a spatial prior to the correspondence discovery process: features belonging to the same object part are likely to be close together in pixel space.

Finally, following [4], after identifying an initial set of matches through our cyclical distance method, we use K-Means clustering on the selected features in the reference image to recover points which are spatially well distributed on the object. We find that well distributed points result in a more robust final pose estimate (see supplementary). In practise, we select the top-$2K$ correspondences by cyclical distance, and filter to a set of K correspondences with K-Means.

4.2 Finding a suitable view for alignment

Finding semantic correspondences between two images which view (two instances of) an object from very different orientations is challenging. For instance, it is possible that images from the front and back of an object have no semantic parts in common. To overcome this, an agent must be able to choose a suitable view from which to establish semantic correspondences. In the considered setting, this entails selecting the best view from the N target images. We do this by constructing a correspondence score between the reference image, $I_{\mathcal{R}}$, and each image in the target sequence, $I_{\mathcal{T}_{1:N}}$. Specifically, given the reference image and an image from the target sequence, the correspondence score is the sum the of the feature similarities between their K semantic correspondences. Mathematically, given a set of K correspondences between the j^{th} target image and the reference, $\{(u_k^j, v_k^j)\}_{k=1}^K$, this can be written as:

$$j^* = \underset{j \in 1:N}{\operatorname{argmax}} \quad \sum_{k=1}^{K} -d(\Phi(I_{\mathcal{R}})_{u_k^j}, \Phi(I_{\mathcal{T}_j})_{v_k^j}) \tag{3}$$

4.3 Pose estimation from semantic correspondences and depth

The process described in Sec. 4.1 gives rise to a set of corresponding points in 2D pixel coordinates, $\{(u_k, v_k)\}_{k=1}^K$. Using depth information and camera intrinsics, these are unprojected to their corresponding 3D coordinates, $\{(\mathbf{u}_k, \mathbf{v}_k)\}_{k=1}^K$, where $\mathbf{u}_k, \mathbf{v}_k \in \mathbb{R}^3$. In the pose estimation problem, we seek a single 6D pose that describes the orientation and translation of the target object, relative to the frame defined by the reference object. Given a set of corresponding 3D points, there are a number of approaches for solving for this rigid body transform. As we assume our correspondences are both noisy and likely to contain outliers, we use a fast least-squares method based on the singular value decomposition [35], and use RANSAC to handle outliers. We run RANSAC for up to 10,000 iterations, with further details in supplementary. The least squares solution recovers a 7-dimensional transform: rotation $\mathbf{R}$, translation $\mathbf{t}$, and a uniform scaling parameter λ, which we found crucial for dealing with cross-instance settings. The

least-squares approach minimises the residuals and recovers the predicted 6D pose offset, T^* as:

$$T^* = (\mathbf{R}^*, \mathbf{t}^*) = \underset{(\mathbf{R},\mathbf{t})}{\operatorname{argmin}} \sum_{k=1}^{K} \mathbf{v}_k - (\lambda \mathbf{R}\mathbf{u}_k + \mathbf{t}) \tag{4}$$

5 Experiments

5.1 Evaluation Setup

Dataset, CO3D [28]: To evaluate zero-shot, category-level pose estimation methods, a dataset is required that provides images of multiple object categories, with a large amount of intra-category instance variation, and with varied object viewpoints. The recently released Common Objects in 3D (CO3D) dataset fulfils these requirements with 1.5 million frames, capturing objects from 50 categories, across nearly 19k scenes [28]. For each object instance, CO3D provides approximately 100 frames taken from a 360^{o} viewpoint sweep with handheld cameras, with labelled camera pose offsets. The proposed method makes use of depth information, and CO3D provides estimated object point clouds, and approximate depth maps for each image, that are found by a Structure-from-Motion (SfM) approach applied over the sequences [30]. We note that, while other pose-oriented datasets exist [2,40,41], we find them to either be lacking in necessary meta-data (e.g no depth information), have little intra-category variation (e.g be instance level), contain few categories, or only provide a single image per object instance. We expand on dataset choice in the supplementary.

Labels for evaluation: While the proposed pose estimation method requires no pose-labelled images for training, we label a subset of sequences across the CO3D categories for quantitative evaluation. We do this by assigning a category-level canonical frame to each selected CO3D sequence. We exclude categories that have infinite rotational symmetry about an axis (e.g 'apple') or have an insufficient number of instances with high quality point clouds (e.g 'microwave'). For the remaining 20 categories, we select the top-10 sequences based on a point cloud quality metric. Point clouds are manually aligned within each category with a rigid body transform. As CO3D provides camera extrinsics for every frame in a sequence with respect to its point cloud, these alignments can be propagated to give labelled category-canonical pose for every frame in the chosen sequences. Further details are in the supplementary.

Evaluation setting: For each object category, we sample 100 combinations of sequence pairs, between which we will compute pose offsets. For the first sequence in each pair, we sample a single reference frame, $I_{\mathcal{R}}$, and from the second we sample N target frames, $I_{\mathcal{T}_{1:N}}$. We take $N = 5$ as our standard setting, with results for different numbers of views in Tab. 2 and the supplementary. For each pair of sequences, we compute errors in pose estimates between the ground truth

and the predictions. For the rotation component, following standard practise in the pose estimation literature, we report the median error across samples, as well as the accuracy at 15^{o} and 30^{o}, which are given by the percentage of predictions with an error less than these thresholds. Rotation error is given by the geodesic distance between the ground truth predicted rotations [19].

'Zero-shot' pose estimation: In this work, we leverage models with large-scale, self-supervised pre-training. The proposed pose estimation method is 'zero-shot' in the sense that it does not use labelled examples (either pose labels or category labels) for any of the object categories it is tested on. The self-supervised features, though, may have been trained on images containing unlabelled instances of some object categories considered. To summarise, methods in this paper **do not require** labelled pose training sets or CAD models for the categories they encounter during evaluation. They **do require** large-scale unsupervised pre-training, depth estimates, and multiple views of the target object. We assert that these are more realistic assumptions for embodied agents (see Sec. 3).

5.2 Baselines

We find very few baselines in the literature which can be applied to the highly challenging problem of pose-detection on unseen categories. Though some methods have tackled the zero-shot problem before, they are difficult to translate to our setting as they require additional information such as CAD models for the test objects. We introduce the baselines considered.

PoseContrast [42] : This work seeks to estimate 3D pose (orientation only) for previously unseen categories. The method trains on pose-labelled images and assumes unseen categories will have both sufficiently similar appearance and geometry, and similar category-canonical frames, to those seen in training. We adapt this method for our setting and train it on all 100 categories in the ObjectNet3D dataset [43]. During testing, we extract global feature vectors for the reference and target images with the model, and use feature similarities to select a suitable view. We then run the PoseContrast model on the reference and selected target image, with the model regressing to an Euler angle representation of 3D pose. PoseContrast estimates pose for each image independently, implicitly inferring the canonical frame for the test object category. We thus compute the difference between the predicted the pose predictions for the reference and chosen target image to arrive at a relative pose estimate.

Iterative Closest Point (ICP): ICP is a point cloud alignment algorithm that assumes no correspondences are known between two point clouds, and seeks an optimal registration. We use ICP to find a 7D rigid body transform (scale, translation and rotation, as in Sec. 4.3) between the reference and target objects. We use the depth estimates for each image to recover point clouds for the two instances, aggregating the N views in the target sequence for a maximally complete target point cloud. We use these point clouds with ICP. As ICP is known

to perform better with good initialisation, we also experiment with initialising it from the coarse pose-estimate given by our 'best view' method (see Sec. 4.2) which we refer to as 'ICP + BV'.

Image Matching: Finally, we experiment with other image matching techniques. In the literature, cross-instance correspondence is often tackled by learning category-level keypoints. However, this usually involves learning a different model for each category, which defeats the purpose of our task. Instead, we use category-agnostic features and obtain matches with mutual nearest neighbours between images, before combining the matches' spatial locations with depth information to compute pose offsets (similarly to Sec. 4.3). We experiment both with standard SIFT features [25] and deep features extracted with an ImageNet self-supervised ResNet-50 (we use SWaV features [5]). In both cases, we select the best view using the strength of the discovered matches between the reference and target images (similarly to Sec. 4.2).

5.3 Implementation Details

In this work we use pre-trained DINO ViT features [6] to provide semantic correspondences between object instances. Specifically, we use ViT-Small with a patch size of 8, giving feature maps at a resolution of 28×28 from square 224×224 images. Prior work has shown that DINO ViT features encode information on generalisable object parts and correspondences [4, 36]. We follow [4] for feature processing and use 'key' features from the 9th ViT layer as our feature representation, and use logarithmic spatial binning of features to aggregate local context to at each ViT patch location. Furthermore, the attention maps in the ViT provide a reasonable foreground segmentation mask. As such, when computing cyclical distances, we assign infinite distance to any point which lands off the foreground at any stage in the reference-target image cycle (Sec. 4.1), to ensure that all correspondences are on the objects of interest. We refer to the supplementary for further implementation details on our method and baselines.

5.4 Main Results

We report results averaged over the 20 considered categories in CO3D in the leftmost columns of Tab. 1. We first highlight that the the baselines show poor performance across the reported metrics. ICP and SIFT perform most poorly, which we attribute to them being designed for within-instance matching. Alignment with the SWaV features, which contain more semantic information, fares slightly better, though still only reports a 7.5% accuracy at 30^{o}. Surprisingly, we also found PoseContrast to give low accuracies in our setting. At first glance, this could simply be an artefact of different canonical poses — between those inferred by the model, and those imposed by the CO3D labels. However, we note that we *take the difference* between the reference and target poses as our pose prediction, which cancels any constant-offset artefacts in the canonical pose.

	All Categories			Per Category (Acc30 ↑)				
	Med. Err (↓)	Acc30 (↑)	Acc15 (↑)	Bike	Hydrant	M'cycle	Teddy	Toaster
ICP	111.8	3.8	0.7	3.0	6.0	1.0	1.0	7 .0
SIFT	129.4	4.0	1.5	3.0	11.0	1.0	1.0	0.0
SWaV	123.1	7.5	3.3	13.0	9.0	8.0	5.0	7.0
ICP+BV	109.3	5.4	1.2	6.0	5.0	8.0	5.0	5.0
PoseContrast	111.5	6.9	1.1	2.0	4.0	13.0	4.0	12.0
Ours (K=30)	60.2	43.5	29.0	63.0	24.0	80.0	33.0	39.0
Ours (K=50)	**53.8**	**46.3**	**31.1**	**71.0**	**26.0**	**82.0**	**39.0**	**42.0**

Table 1: We report Median Error and Accuracy at 30^{o}, 15^{o} averaged across all 20 categories. We also report Accuracy at 30^{o} broken down by class for an illustrative subset of categories. We provide full, per category breakdowns in the supplementary.

Meanwhile, our method shows substantial improvements over all implemented baselines. Our method reports less than half the Median Error aggregated over all categories, and further demonstrates a *six-fold increase* in accuracy at 30^{o}. We also note that this improvement cannot solely be attributed to the scale of DINO's ImageNet pre-training: the SWaV-based baseline also uses self-supervised features trained on ImageNet [5], and PoseContrast is initialised with MoCo-v2 [10] weights, again from self-supervision on ImageNet.

We find that performance varies substantially according to the specific geometries and appearances of individual categories. As such, in the rightmost columns of Tab. 1, we show per-category results for an illustrative subset of the selected classes in CO3D. We find that textured objects, which induce high quality correspondences, exhibit better results (e.g 'Bike' and 'Motorcycle'). Meanwhile, objects with large un-textured regions (e.g 'Toaster') proved more challenging.

The results for 'Hydrant' are illustrative of a challenging case. In principle, a hydrant has a clearly defined canonical frame, with faucets appearing on only three of its four 'faces' (see Fig. 3). However, if the model fails to identify all three faucets as salient keypoints for correspondence, the object displays a high degree of rotational symmetry. In this case, SIFT, which focuses exclusively on appearance (i.e it does not learn semantics), performs higher than its average, as the hydrant faucets are consistently among the most textured regions on the object. Meanwhile, our method, which focuses more on semantics, performs worse than its own average on this category.

5.5 Making use of multiple views

The number of target views : A critical component of our setting is the availability of multiple views of the target object. We argue that this is important for the computation of zero-shot pose offset between two object instances, as a single image of the target object may not contain anything in common with the reference image. An important factor, therefore, is the number of images avail-

able in the target sequence. In principle, if one had infinite views of the target sequence, and camera transformations between each view, the pose estimation problem collapses to that of finding the best view. However, we note that this is unrealistic. Firstly, running inference on a set of target views is expensive, with the computational burden generally scaling linearly with the number of views. Secondly, collecting and storing an arbitrarily large number of views is also expensive. Finally, the number of views required to densely and uniformly sample viewpoints of an object is very high, as it requires combinatorially sampling with respect to three rotation parameters.

In this work we experiment with the realistic setting of a 'handful' of views of the target object. In Tab. 2, we experiment with varying N in $\{1, 3, 5\}$ instances in the target sequence. In the bottom three rows, we show the performance of our full method as N is varied and find that, indeed, the performance increases with the number of available views (further results are in supplementary). However, we find that even from a *single view*, our method can outperform the baselines with access to five views.

Only finding best target view : We disambiguate the 'coarse' and 'fine' pose-estimation steps of our method (Sec. 4.2 and Sec. 4.3 respectively). Specifically, we experiment with our method's performance if we assume the reference image is perfectly aligned with the selected best target view. We show these figures as 'Ours-BV' in the top rows of Tab. 2. It can be seen that this part of our method alone can substantially outperform the strongest baselines. However, we also show that the subsequent fine alignment step using the depth information (Sec. 4.3) provides an important improvement in performance. For instance, when $N = 5$, this component of our method boosts Acc30 from 35.4% to 46.3%.

How to pick the best target view : Here, we explore the importance of our particular recipe for arriving at the optimal target view. First, we experiment with a standard baseline of choosing the target view which maximises the similarity with respect to the ViT's global feature vector (termed *'GlobalSim'*). We also try maximising the Intersection-over-Union of the foreground masks, as provided by the ViT attention maps, of the reference and target frames (*'SaliencyIoU'*). Finally, we try maximising the IoU between the foreground mask of a target object and its cyclical distance map with respect to the reference image. The intuition here is to recover a target view where a large proportion of the foreground object pixels have a unique nearest neighbour in the reference image (*'CyclicalDistIoU'*). We present the results of these findings in Tab. 3.

5.6 Qualitative Results

In Fig. 3 we provide qualitative alignment results for four object categories, including for 'Hydrant', which we include as a failure mode. The images show the reference image and the best view from the target sequence, along with the semantic correspondences discovered between them. We further show the point cloud for the reference image aligned with the target sequence using our method.

Method	Med. Err	Acc30	Acc15
Ours-BV (N=1)	92.8	12.6	3.7
Ours-BV (N=3)	69.5	26.1	8.0
Ours-BV (N=5)	61.1	35.4	10.6
Ours (N=1)	97.3	23.8	13.4
Ours (N=3)	63.9	38.9	26.2
Ours (N=5)	**53.8**	**46.3**	**31.1**

Table 2: We experiment with varying numbers of images available in the target sequence (N). Even with only one view, our method substantially outperforms existing baselines with access to multiple views. We further show the utility of pose alignment from the best view ('Ours') over simply choosing the best view with our method ('Ours-BV').

Method	Med. Err	Acc30	Acc15
CyclicalDistIOU	94.7	22.3	13.1
GlobalSim	71.9	36.8	23.7
SaliencyIOU	87.0	29.1	18.1
CorrespondSim	**53.8**	**46.3**	**31.1**

Table 3: We ablate different methods for selecting the best view from the target sequence, from which we perform our final pose computation. Compared to a other intuitive options for this task, we demonstrate the importance of our proposed best view selection pipeline for downstream performance.

Specifically, we first compute a relative pose offset between the reference image and the best target view, and then propagate this pose offset using camera extrinsics to the other views in the target sequence. Finally, we highlight the practical utility of this setting. Consider a household robot wishing to tidy away a 'Teddybear' (top row) into a canonical pose (defined by a reference image). Using this method, the agent is able to view the toy from a number of angles (in the target sequence), align the reference image to an appropriate view, and thus understand the pose of the toy *from any other angle*.

6 Conclusion

Consideration of limitations: We have have proposed a model which substantially outperforms existing applicable baselines for the task of zero-shot category-level pose detection. However, absolute accuracies remain low and we suggest fall far short of human capabilities. Firstly, our performance across the considered classes is 46.3% Acc30 with 5 views available. We imagine these accuracies to be substantially lower than the a human baseline for this task. Secondly, though single view novel category alignment is highly challenging for machines, humans are capable of generalising highly abstract concepts to new categories, and thus would likely be able to perform reasonably in a single view setting.

Final Remarks: In this paper we have proposed a highly challenging (but realistic) setting for object pose estimation, which is a critical component in most 3D vision pipelines. In our proposed setting, a model is required to align two instances of an object category without having any pose-labelled data for training. We further re-purpose the recently released CO3D dataset and devise a test setting which reasonably resembles the one encountered by a real-world embodied

Fig. 3: Example results for the categories **Teddybear**, **Toybus**, **Car**, **Hydrant**. Depicted are the correspondences found between the reference image and the best-matching frame from the target sequence found following Sec. 4.2. To the right, the estimated pose resulting from these correspondences is shown as an alignment between the reference object (shown as a rendered point cloud) and the target sequence. **Hydrant** depicts a failure mode — while the result looks visually satisfying, near rotational symmetry (about the vertical axis) leads to poor alignment.

agent. Our setting presents a complex problem which requires both semantic and geometric understanding, and we show that existing baselines perform poorly on this task. We further propose a novel method for zero-shot, category-level pose estimation based on semantic correspondences and show it can offer a six-fold increase in Acc30 on our proposed evaluation setting. We hope that this work will serve as a spring-board to foster future research in this important direction.

7 Acknowledgment

The authors gratefully acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (`http://dx.doi.org/10.5281/zenodo.22558`). Sagar Vaze is funded by a Facebook Research Scholarship. We thank Dylan Campbell for many useful discussions.

References

1. Aberman, K., Liao, J., Shi, M., Lischinski, D., Chen, B., Cohen-Or, D.: Neural best-buddies: Sparse cross-domain correspondence. ACM Transactions on Graphics (2018)
2. Ahmadyan, A., Zhang, L., Ablavatski, A., Wei, J., Grundmann, M.: Objectron: A Large Scale Dataset of Object-Centric Videos in theWild with Pose Annotations. In: CVPR (2021)
3. Akizuki, S.: ASM-Net : Category-level Pose and Shape Estimation Using Parametric Deformation. In: BMVC (2021)
4. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep ViT Features as Dense Visual Descriptors (2021)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: ICCV (2021)
7. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6D object pose and size estimation. In: CVPR (2020)
8. Chen, K., Dou, Q.: SGPA: Structure-Guided Prior Adaptation for Category-Level 6D Object Pose Estimation. In: ICCV (2021)
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: CVPR (2021)
10. Chen, X., Fan, H., Girshick, R.B., He, K.: Improved baselines with momentum contrastive learning (2020), `https://arxiv.org/abs/2003.04297`
11. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category Level Object Pose Estimation via Neural Analysis-by-Synthesis. In: ECCV (2020)
12. Deng, X., Xiang, Y., Mousavian, A., Eppner, C., Bretl, T., Fox, D.: Self-supervised 6D Object Pose Estimation for Robot Manipulation. In: ICRA (2020)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
14. El Banani, M., Corso, J.J., Fouhey, D.F.: Novel object viewpoint estimation through reconstruction alignment. In: CVPR
15. Florence, P.R., Manuelli, L., Tedrake, R.: Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation. In: CoRL (2018)
16. Goodwin, W., Vaze, S., Havoutis, I., Posner, I.: Semantically Grounded Object Matching for Robust Robotic Scene Rearrangement. In: ICRA (2021)
17. Grabner, A., Roth, P.M., Lepetit, V.: 3D Pose Estimation and 3D Model Retrieval for Objects in the Wild. In: CVPR (2018)
18. Gupta, S., Arbeláez, P., Girshick, R., Malik, J.: Inferring 3D Object Pose in RGB-D Images (2015)
19. Huynh, D.Q.: Metrics for 3d rotations: Comparison and analysis. J. Math. Imaging Vis. (2009)
20. Kanezaki, A., Matsushita, Y., Nishida, Y.: RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews from Unsupervised Viewpoints. In: CVPR (2018)
21. Kundu, J.N., Rahul, M.V., Ganeshan, A., Babu, R.V.: Object pose estimation from monocular image using multi-view keypoint correspondence. In: ECCV (2019)

22. Lee, J., Kim, D., Ponce, J., Ham, B.: SFNET: Learning object-aware semantic correspondence. In: CVPR (2019)
23. Li, X., Weng, Y., Yi, L., Guibas, L., Abbott, A.L., Song, S., Wang, H.: Leveraging SE(3) Equivariance for Self-Supervised Category-Level Object Pose Estimation. In: NeurIPS 2021 (2021)
24. Lin, Y., Tremblay, J., Tyree, S., Vela, P.A., Birchfield, S.: Single-stage Keypoint-based Category-level Object Pose Estimation from an RGB Image (2021)
25. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision (2004)
26. Manuelli, L., Gao, W., Florence, P., Tedrake, R.: kPAM: KeyPoint Affordances for Category-Level Robotic Manipulation. In: International Symposium on Robotics Research (ISRR) (2019)
27. Pavlakos, G., Zhou, X., Chan, A., Derpanis, K.G., Daniilidis, K.: 6-DoF Object Pose from Semantic Keypoints. ICRA (2017)
28. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In: ICCV (2021)
29. Sahin, C., Kim, T.K.: Category-level 6D object pose recovery in depth images. In: ECCV (2019)
30. Schonberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: CVPR (2016)
31. Shi, J., Yang, H., Carlone, L.: Optimal Pose and Shape Estimation for Category-level 3D Object Perception. In: Robotics: Science and Systems XVII (2021)
32. Simeonov, A., Du, Y., Tagliasacchi, A., Tenenbaum, J.B., Rodriguez, A., Agrawal, P., Sitzmann, V.: Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation (2021)
33. Tian, M., Ang, M.H., Lee, G.H.: Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation. In: ECCV (2020)
34. Tseng, H.Y., De Mello, S., Tremblay, J., Liu, S., Birchfield, S., Yang, M.H., Kautz, J.: Few-shot viewpoint estimation. In: BMVC (2020)
35. Umeyama, S.: Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (1991)
36. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Generalized category discovery. CVPR (2022)
37. Wang, A., Kortylewski, A., Yuille, A.: NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation. In: ICLR (2021)
38. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.: Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. CVPR (2019)
39. Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S.: Objectnet3D: A large scale database for 3D object recognition. In: ECCV (2016)
40. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond PASCAL: A benchmark for 3D object detection in the wild. In: WACV (2014)
41. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In: Robotics: Science and Systems XIV (2018)
42. Xiao, Y., Du, Y., Marlet, R.: PoseContrast: Class-Agnostic Object Viewpoint Estimation in the Wild with Pose-Aware Contrastive Learning. In: 3DV (2021)

43. Xiao, Y., Marlet, R.: Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. In: ECCV (2020)
44. Xiao, Y., Qiu, X., Langlois, P.A., Aubry, M., Marlet, R.: Pose from Shape: Deep pose estimation for arbitrary 3D objects. In: BMVC (2019)
45. Zhou, X., Karpur, A., Luo, L., Huang, Q.: StarMap for Category-Agnostic Keypoint and Viewpoint Estimation. In: ECCV (2018)