2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

ZS6D: Zero-shot 6D Object Pose Estimation using Vision Transformers

Philipp Ausserlechner[1], David Haberger[1], Stefan Thalhammer[2], Jean-Baptiste Weibel[1] and Markus Vincze[1]

***Abstract*— As robotic systems increasingly encounter complex and unconstrained real-world scenarios, there is a demand to recognize diverse objects. The state-of-the-art 6D object pose estimation methods rely on object-specific training and therefore do not generalize to unseen objects. Recent novel object pose estimation methods are solving this issue using task-specific fine-tuned CNNs for deep template matching. This adaptation for pose estimation still requires expensive data rendering and training procedures. MegaPose for example is trained on a dataset consisting of two million images showing 20,000 different objects to reach such generalization capabilities. To overcome this shortcoming we introduce ZS6D, for zero-shot novel object 6D pose estimation. Visual descriptors, extracted using pre-trained Vision Transformers (ViT), are used for matching rendered templates against query images of objects and for establishing local correspondences. These local correspondences enable deriving geometric correspondences and are used for estimating the object's 6D pose with RANSAC-based P*n*P. This approach showcases that the image descriptors extracted by pre-trained ViTs are well-suited to achieve a notable improvement over two state-of-the-art novel object 6D pose estimation methods, without the need for task-specific fine-tuning. Experiments are performed on LMO, YCBV, and TLESS. In comparison to MegaPose, we improve the Average Recall on all three datasets and compared to OSOP we improve on two datasets. The code is available at `https://github.com/PhilippAuss/ZS6D`.**

I. INTRODUCTION

Robotics, and service robotics, in particular, have the potential to profoundly transform our society. However, enabling semantic manipulation requires estimating the poses of objects, which presents substantial challenges for a constantly increasing set of objects. Contemporary pose estimation methods [1], [2], [3], [4], [5] are trained for specific objects and do not generalize to unseen ones. Their little flexibility and adaptability require re-training every time the set of objects that need to be handled by the robot changes.

Recent novel object pose estimation approaches provide a feasible solution to this problem by matching query images against rendered templates of the object models [7], [8], [9], [10]. Such deep template matching requires task-specific fine-tuning. Diverse object models are used for rendering training data, with e.g. BlenderProc [11], which is used for training multi-staged CNN pipelines. In the case of MegaPose [1], two million scene-level training images featuring $20,000$ different object instances are rendered. These images are used for training their deep template matcher. Using such strategies partially alleviates the need for object-specific training, yet assumes that the set of training objects is enough to generalize to arbitrary real-world objects. This strategy becomes untractable to obtain a system that handles all objects.

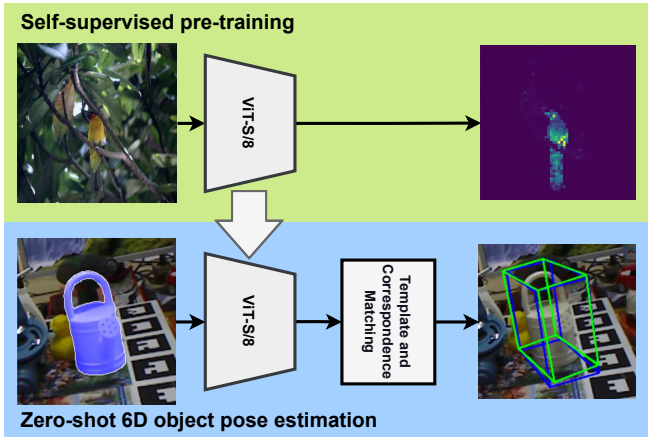

Fig. 1. **Zero-shot 6D pose estimation** Descriptors produced by a self-supervised ViT [6] are discriminative enough for novel object $6D$ pose estimation, without any task-specific fine-tuning in a zero-shot manner.

In this work we hypothesize that self-supervised pre-trained Vision Transformers [12], [6] (ViT) are bound to overcome the requirement of task-specific fine-tuning, since recent works indicate the generality of their extracted descriptors [13], [6], [14], [10]. In order to verify our hypothesis we present ZS6D, Figure 1, a method for zero-shot $6D$ object pose estimation. Our method extracts image descriptors to match a query image against rendered object templates. Subsequently, local correspondences between the query and the matched template are computed to derive geometric correspondences and estimate the pose using RANSAC-based [15] P*n*P [16]. In practice, we use colored object coordinates, i.e. object vertex locations mapped to RGB values [2]. As these are defined in object space and derived using the matched local correspondences, the retrieved pose grants a higher accuracy than the available templates provide. This allows our approach to overcome noise in the template matching and achieve accurate object poses with as few as 200 object templates. We provide experiments showing that using ViTs for zero-shot $6D$ object pose estimation alleviates the requirement for both training data and

*This work was supported by the EU-program EC Horizon 2020 for Research and Innovation under grant agreement No. 101017089, project TraceBot and under grant agreement No. 101120823 project MANiBOT funded by the European Union.

[1]All authors are with Vision for Robotics Laboratory, Automation and Control Institute, TU Wien, Austria {ausserlechner, haberger, weibel, vincze}@acin.tuwien.ac.at

[2]Stefan Thalhammer is with the Industrial Engineering Department, University of Applied Sciences Technikum Vienna, Austria stefan.thalhammer@technikum-wien.at

model fine-tuning. Besides, ZS6D also improves the Average Recall (AR) [17] on two of three tested standard datasets, in comparison to the state of the art. Our contributions to the field of $6D$ object pose estimation are the following:

- We present a pose estimation method that estimates $6D$ object poses from RGB input without depth information in a zero-shot fashion. The presented method improves over the state of the art for novel object pose estimation on two standard datasets using the Average Recall [17].
- We demonstrate that pre-trained Vision Transformers (ViT) improve over task-specific fine-tuned CNNs for novel object $6D$ pose estimation.

The paper proceeds as follows: we present relevant state-of-the-art methods in Section II, our proposed evaluation scheme in Section III, and our experimental results in Section IV before presenting our conclusions in Section V.

II. Related Work

This section presents the state of the art for $6D$ object pose estimation with the main focus on novel object pose estimation. Following that self-supervised Vision Transformers as discussed.

Contemporary methods for pose estimation [4], [3], [1], [2], [18], [19], [20], [21], [22] rely on object-specific training and a preceding object detection stage which also has to be trained separately. These approaches do not scale well, since they have to be trained for every new object. In contrast, [23], [5], [24] scale better because they are trained for an entire set of objects simultaneously, integrating object detection and pose estimation in a single stage. Nevertheless, all of these methods lack practicality in many real-world scenarios, since it is not feasible to re-train for every new set of objects encountered. Recent single reference image pose estimation methods like Pope [25] and Goodwin et al. [26] leverage the descriptors produced by self-supervised ViTs to estimate the relative rotation between the reference image and the detected object. However, these approaches are insufficient for robotic applications, since a $6D$ pose is required for object manipulation.

Novel object pose estimation: We refer to the problem of estimating the pose of unseen objects during training as novel object pose estimation. A classical approach to this problem is the Point Pair Feature (PPF) method [27]. It leverages depth information by approximating local geometries of the query image and uses it as a hash to match the object model. DeepIM [28] is one of the first approaches that leveraged CNN-based features to iteratively refine the pose of a template compared to the query image. Another noteworthy step towards novel object pose estimation comes from Sundermeyer et al. [29], which uses a common encoder that generalizes to unseen objects and extracts descriptive image features. Ngyuen et al. [9] revisits the idea of template matching by applying CNN-based features to estimate the rotation of unseen objects from query images. Thalhammer et al. [10] extends this scheme and demonstrates that ViTs outperform CNNs for template matching. With the exception of DeepIM, these approaches only estimate a rotation, which is not sufficient for robotic interaction. More recent methods like OSOP [7] deploy a task-specific fine-tuned CNN to derive dense correspondences between the query image and a large set of templates, $5K$ in the case of LMO. Another noteworthy approach is MegaPose [8] which relies on an initial template-matching followed by an iterative refinement. They use a CNN which is trained on a large-scale dataset with more than $20,000$ objects and two million images, which allows them to effectively generalize deep template matching to unseen objects. Our method differs from these approaches by relying solely on a self-supervised pre-trained ViT, with no requirement for pose estimation-specific fine-tuning and a comparably small set of templates (up to 300).

Vision Transformers: In natural language processing, transformers [32] are the dominant architecture due to their capability to be trained on large-scale datasets in a self-supervised manner. Many efforts were made to transfer this architecture to the Vision domain [33], resulting in the Vision Transformer (ViT) [12]. These models perform comparably better than CNNs, but their advantages really materialize with self-supervised training. This procedure allows them to generalize well to novel tasks and makes them robust against dataset biases [6]. Such foundational Computer Vision models show comparable results to the state of the art for supervised models in tasks like object classification, segmentation [30], and image retrieval. Latest publications [26], [25], [10] show that ViTs can be applied without fine-tuning for object pose estimation, with respect to a reference image to estimate a $3D$ pose. We show in our experiments that those foundational Computer Vision models can be applied to obtain the full $6D$ pose of unseen objects without any fine-tuning.

III. ZS6D

In this section, we propose our method for **Z**ero-**S**hot **6D** object pose estimation, named ZS6D. It solely relies on an object model and the descriptors extracted from RGB input by a pre-trained ViT, which is not trained or fine-tuned for the task of object pose estimation. ZS6D processes segmented objects by extracting dense descriptors and comparing them to pre-rendered templates to identify the closest match. It then matches local correspondences between the object and the template to derive 2D-3D correspondences, enabling the determination of the 6D object pose. This procedure requires no training or fine-tuning and is executed with a self-supervised pre-trained ViT. Figure 2 provides an abstract visualization of the pose inference. In the following subsections, we describe how the objects are segmented in the query image, how the best matching template is selected, and how the local correspondences are obtained.

A. Object Detection and Segmentation

ZS6D assumes the availability of object instance segmentation masks and a query image I_s. The segmentations are generated with the zero-shot $2D$ approach of CNOS [31]. When looking at a new scene, the Segment Anything Model (SAM) of [30] is employed to generate object segmentation

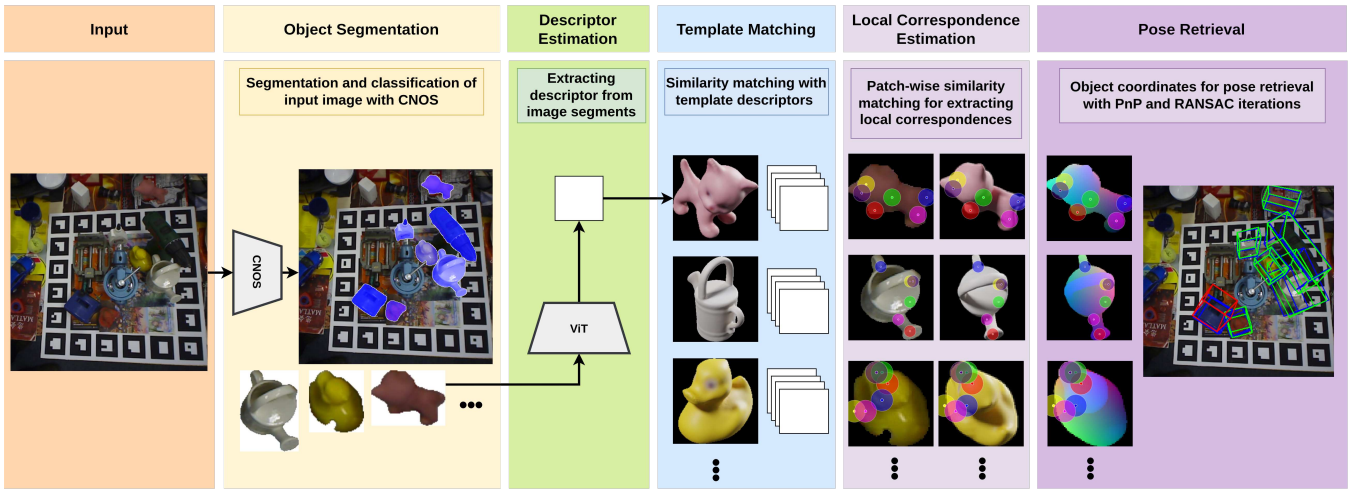

Fig. 2. **Overview of ZS6D** The diagram depicts the stages of the ZS6D pose estimation pipeline. Initially, Segment Anything [30] segmentations [31] are used to isolate the object of interest. Then, dense visual descriptors are extracted [13] from the segmented object, followed by a comparison against pre-rendered template descriptors using cosine similarity. The image further illustrates the process of matching local correspondences between the selected template and the segmented region, which enables the derivation of 2D-3D correspondences from the template's colored object coordinates. The final step is the application of a P*n*P [16] algorithm with RANSAC [15] iterations to obtain the $6D$ object pose.

proposals denoted as $\{I_p \mid I_p \subseteq I_S\}$. The descriptors D_p^{seg} of each object proposal as well as the template descriptors D_t^{seg} are generated by a single forward pass through a ViT [14]. The cosine similarities between all the template descriptors D_t^{seg} and D_p^{seg} are calculated to recognize the object within the proposal, after aggregating the similarity scores by object class. The class of the object proposal is determined by the highest aggregated score.

B. Global Descriptor Estimation

To estimate the image descriptors we apply a self-supervised ViT [6]. The core operation of a ViT [12] is the attention mechanism [33]. We define the input image $X \in \mathbb{R}^{n \times d}$ as a sequence of patches $(x_1, x_2, \ldots, x_n)$. The aim of self-attention is to estimate the interaction between all n patches. Therefore, we define three learnable weight matrices: $\mathbf{W}^Q \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}^K \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W}^V \in \mathbb{R}^{d \times d_v}$. These matrices allow us to transform the input sequence $\mathbf{X}$ into Queries $\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}$, Keys $\mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}$, and Values $\mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}}$ respectively. The self-attention is then computed as:

$$Z = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_q}})\mathbf{V} \tag{1}$$

The ViT itself consists of multiple self-attention layers, therefore creating multiple options for choosing a viable image descriptor. For example, CNOS [31] uses the class token which is a vector that gets passed together with the image patches through the network and serves as a global image embedding. In [10], the authors show that patch-wise token embeddings are more suitable for 3D pose estimation than the class token. Furthermore, the authors of [13] empirically show that the key token K embedding from layer 9 of the ViT is the most suitable for global image description. The authors argue that the shallow layers are the best to represent global geometric information. We follow their argumentation and use these as image descriptors.

C. Template Matching

The descriptor of the query image D_p is compared against a set of template descriptors $\{D_t, \forall t \in T\}$, both created by a single forward pass through the ViT. Similar to classical approaches we assume uniform coverage of the viewing space. Thus we rely on rendered object views [34], [35]. In Section IV, we present detailed ablations justifying the 300 uniformly distributed views we chose to ensure comprehensive coverage of the object model. To estimate the closest template we compute the cosine similarity between the descriptor of the object proposal and each descriptor from the set of templates, according to:

$$\max \langle D_t, D_p \rangle, \forall t \in T \tag{2}$$

The template with the highest similarity score is used for estimating local correspondences.

D. Local Correspondence Estimation

The matched template provides a coarse pose estimate with the maximum accuracy limited by the view coverage of the object. As an example, directly retrieving the rotation of the input object requires a large number of templates, e.g. $21,672$ templates for TLESS [9]. Furthermore, retrieving the object's translation using the ratio of the estimated bounding box to the rendered template depends on the rotational error between the query image and template, thus translation error increases with rotation error. In order to circumvent these issues, we estimate and match local correspondences between query and template images. ViTs [12] treat images as local patches and estimate relations between these, to obtain local descriptors. We aim to match corresponding patches between query I_p and the template images $\{I_t\}$. For

Fig. 3. **Template and corresponding object coordinates** Visualization of a template of LMO's cat and the corresponding colored object coordinates.

this purpose we adopt the key k token from layer 11 as our patch descriptor, a deeper layer found to yield superior performance when estimating local correspondences [13]. Here, $Q = \{q_i\}$ represents descriptors of the template, while $P = \{p_i\}$ denotes descriptors from the segmented region of the query image. To identify the optimal correspondences, we compute the nearest neighbor NN according to the cosine similarity for each descriptor in Q and P, retaining only those correspondences, termed LC, where the patch demonstrates the highest congruence in both directions:

$$LC(Q,P) = \{(q,p) \,|\, q \in Q,\, p \in P, \\ NN(p,Q) = q \,\wedge\, NN(q,P) = p\} \tag{3}$$

E. Pose Retrieval

From the correspondence estimation stage, only robust patch pairs are left. Given that we know the pose of the object of interest in the rendered templates, we can obtain the corresponding 3D coordinates in object space by looking up the values of the templates' colored object coordinates. Each local correspondence, therefore, yields colored object coordinates [2], that we use in the P*n*P [16] algorithm with RANSAC iterations [15] to recover the final $6D$ pose of the segmented target object from the query image I_p.

IV. Experiments

In this Section, we discuss the experimental setup. We compare our method to the state of the art for novel object $6D$ pose estimation on three of the core datasets of the Benchmark for $6D$ Object Pose Estimation challenge [17] (BOP). Additionally, we provide ablation studies evaluating the impact of the segmentation quality, as well as selecting the optimal number of views for template generation, and the optimal number of local correspondences for object coordinate estimation.

A. Datasets

We evaluate our ZS6D on three of the core BOP datasets [17], LMO [36], YCBV [37], and TLESS [38]. These three datasets reflect standard challenges for object pose estimation, occlusion in the case of LMO, strong illumination changes for YCBV, and texture-less objects for TLESS. Since our method infers poses in a zero-shot fashion, we do not require the training sets and for testing the respective test sets as they are used in BOP.

B. Implementation Details

BlenderProc [11] is used for rendering templates since it is considered the standard tool for that purpose [8], [9], [10]. For each object, we uniformly sample views on a regular icosahedron. We use 300 templates per object unless stated otherwise. Colored object coordinates are used as geometric correspondences, Figure 3, for pose retrieval with P*n*P [2]. For global descriptor estimation (Section III-B) and local correspondence estimation (Section III-D), $ViT - S/8$ [6] is used, where 8 refers to number of pixels for each side of the patches. We use the weights pre-trained on ImageNet1k [39]. Since ZS6D does not require task-specific training and solely relies on the self-supervised pre-trained $ViT - S/8$, we refer to [6] for additional architecture and training details. The input image resolution to both stages is 224×224. A descriptor size of 384 and 6528 is used for global descriptor and local correspondence estimation, respectively. Small patch sizes are crucial for estimating meaningful local correspondences in order to robustly match corresponding patches between query and template images using nearest neighbors. All experiments report the Average Recall AR metric of the BOP [17], which is the standard for $6D$ object pose estimation. Concerning inference time ZS6D takes an average of 522ms per instance, and results are obtained on LMO [36]. It is worth mentioning that ZS6D is not optimized in terms of runtime.

C. Object Segmentation

For all presented experiments we use the segmentation masks provided by CNOS [31] unless stated otherwise. The templates for classifying the SAM masks [30] are rendered from the provided object models. As proposed by the authors of CNOS, $V = 42$ viewpoints on a regular icosahedron are used for generating templates to ensure a uniformly distributed view coverage of the object. Subsequently, template descriptors D_t are computed using DINOv2 [14]. The class token is used as descriptor D_t and its dimension is $N_O \times V \times C$. We follow their hyperparameter configuration of $C = 1024$.

D. Main Results

This section presents our main results for zero-shot novel object $6D$ pose estimation. Table I reports results for pose initialization without a refinement stage, using RGB as input. A comparison is provided against novel object pose estimators MegaPose [8] and OSOP [7]. MegaPose and OSOP apply task-specific fine-tuned CNNs for object pose estimation, while our results are obtained without any fine-tuning, using a pretrained $ViT - S/8$ in a zero-shot manner. We improve the AR on all three datasets compared to Megapose, despite it using a larger number of templates (520 compared to 300 for our method) and relying on detections of Mask R-CNN [40], trained on the synthetic physically-based rendered (PBR) data of the target objects. The relative improvement

Fig. 4. **Qualitative results** Visualized are object poses as $3D$ bounding boxes, blue indicates ground truth, green true positives, and red false positives on LMO, YCBV, and TLESS.

Method	Novel object	Zero-shot	**LMO**	**YCBV**	**TLESS**
CosyPose [1]	✗	✗	0.536	0.333	0.520
GDR-Net [4]	✗	✗	0.672	0.755	0.512
MegaPose [8]	✓	✗	0.187	0.139	0.197
OSOP [7]	✓	✗	0.274	0.296	**0.403**
ZS6D (**Ours**)	✓	✓	**0.298**	**0.324**	0.210

TABLE I

EVALUATION FOR UNSEEN OBJECTS RESULTS ARE PROVIDED USING THE AVERAGE RECALL (AR) OF BOP. WE COMPARE INITIAL POSE ESTIMATING ACCURACY, WITHOUT REFINEMENT, USING RGB AS INPUT. RESULTS ARE REPORTED FOR THE NOVEL OBJECT POSE ESTIMATORS MEGAPOSE [8] AND OSOP [7] AND THE INSTANCE-SPECIFIC ONES, COSYPOSE [1] AND GDR-NET [4].

Method	Mask Origin	LMO	YCBV	TLESS
ZS6D (**Ours**)	CNOS [31]	0.298	0.324	0.210
ZS6D (**Ours**)	ground truth	0.527	0.499	0.460

TABLE II

INFLUENCE OF SEGMENTATION MASKS POSE ESTIMATION RESULTS WITH CNOS [31] AND GROUND TRUTH MASKS. USING A SINGLE POSE HYPOTHESIS FOR EVALUATION.

is 59% on LMO, 133% on YCBV, and 7% on T-LESS. Compared to instance-specific methods, like CosyPose [1] and GDR-Net [4], the AR for pose initialization with novel object pose estimators trails behind.

Evaluating against OSOP [7], we improve the AR on LMO and YCBV for a single hypothesis and multiple hypotheses. On TLESS, OSOP reports a higher AR score. Table II shows significantly improved AR of our methods on TLESS when using the ground truth masks, which suggests that the segmentation masks generated by CNOS [31] are less accurate for TLESS than for LMO and YCBV. Additionally estimating the patch-wise local correspondences on texture-less objects exhibiting symmetries leads to ambiguities. OSOP in contrast proposes a custom segmentation stage that matches the observations against object templates. We show qualitative results in Figure 4, correctly estimated poses are visualized in green, incorrect ones in red, and the ground truth in blue. The influence of the segmentation masks on the pose estimates is discussed in the next section.

E. Ablations

In this section, we present ablations to further investigate the contributing factors to the methods' performance. We conduct three central ablation studies to determine the impact of the segmentation masks, the number of views for template

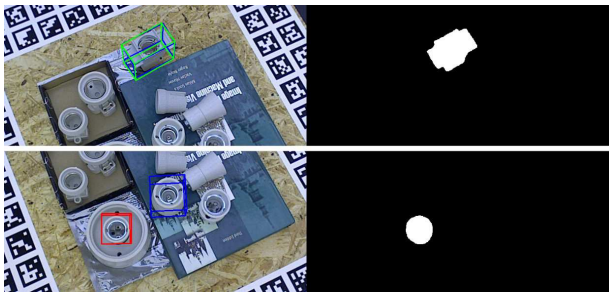

Fig. 5. Pose Estimation Accuracy Relative to Mask Quality in TLESS: The top displays successful pose estimation linked to an accurate mask. The bottom illustrates a failed pose estimation due to a wrong mask, highlighting challenges in TLESS's similar objects leading to segmentation errors and subsequent pose estimation inaccuracies by ZS6D.

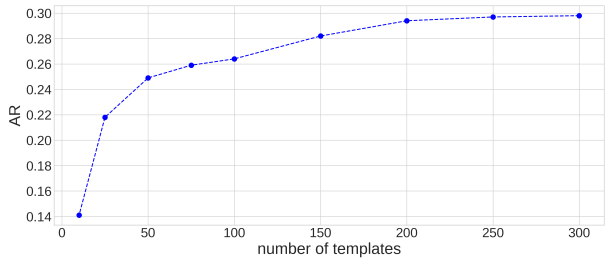

Fig. 6. **Number of templates** Impact of the number of templates per object. Reported is the AR score on LMO. Beyond approximately 200 templates, there are no significant improvements in the AR score; the performance improvement curve asymptotically flattens, indicating a plateau in efficiency gains.

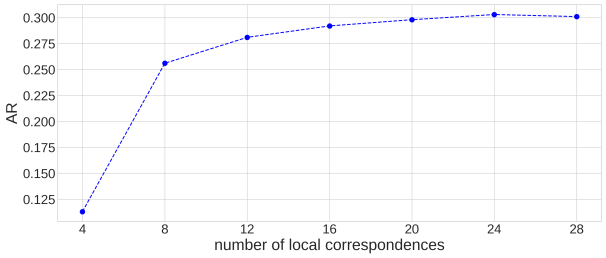

Fig. 7. **Number of correspondences** Impact of the number of extracted local correspondences. Reported is the AR score on LMO. The AR score shows a flattening trend around 20-30 extracted correspondences, beyond which it asymptotically levels off, indicating diminishing returns on further increases in correspondence count.

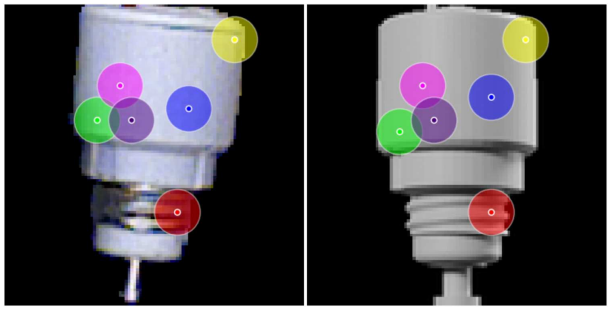

Fig. 8. **Local correspondences** Visualization of matched correspondences of the query image (left) and the matched template (right). Matching colors corresponding to the same local correspondence.

generation, and the number of local correspondences to extract for sub-sequential pose retrieval.

1) Mask Quality: We evaluate our ZS6D with the CNOS and compare it to the ground truth masks in order to disentangle the pose estimation accuracy of ZS6D from the detection stage. Results for LMO, YCBV, and TLESS are provided in Table II. The respective improvements using the ground truth masks are 77%, 54%, and 119%. These results indicate that large improvements are to be expected when obtaining more accurate segmentation masks as input to the presented method. Especially for TLESS, the AR score when using CNOS masks is far off the theoretically obtainable upper bound. Figure 5 shows a representative example of CNOS [31] providing a segmentation mask of the wrong object. Consequently, a considerable performance increase is expected when the masks are more robustly estimated.

2) Number of Templates: Figure 6 ablates the influence of the number of templates reporting the AR score on LMO. The instance segmentation masks generated using CNOS are used as location priors. A significant increase in accuracy is observable up to 200 templates. Since ZS6D derives colored object coordinates based on local correspondence matching the retrieved pose grants a higher accuracy than the available templates provide, as indicated in Figure 8. The accuracy is asymptotically approaching a maximum at 300 views.

3) Number of Correspondences: Figure 7 ablates the number of local correspondences used for deriving object coordinates. Results are provided on LMO using the AR score as a validation metric. Considering the number of extracted local correspondences we observe a very similar behavior to the number of templates. The AR score rapidly increases with the number of local correspondences and flattens out around 20 correspondences. A higher number of local correspondences is not always feasible, due to the constraints enforced by Equation 3. Additionally, using more correspondences increases the likelihood of wrong matches. This is partially compensated by the RANSAC [15] iterations.

V. Conclusions

We propose a zero-shot $6D$ object pose estimation method, which does not rely on task-specific fine-tuning and enables estimating poses of unseen objects. The presented evaluations show that foundational Computer Vision models, precisely self-supervised ViTs are well-suited for extracting general image descriptors, and as such enable pose retrieval. To be precise, we present results on LMO, YCBV, and TLESS, where we show that we can improve over results obtained by task-specific fine-tuned CNNs. The current work focuses on generating initial pose hypotheses, without applying a refinement stage. Future work will thus investigate how to refine pose hypotheses in a zero-shot fashion.

References

[1] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 574–591.

[2] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7668–7677.

[3] Y. Su, M. Saleh, T. Fetzer, J. Rambach, N. Navab, B. Busam, D. Stricker, and F. Tombari, "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6738–6748.

[4] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 611–16 621.

[5] S. Thalhammer, T. Patten, and M. Vincze, "Cope: End-to-end trainable constant runtime object pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2860–2870.

[6] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[7] I. Shugurov, F. Li, B. Busam, and S. Ilic, "Osop: A multi-stage one shot object pose estimation framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6835–6844.

[8] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," *arXiv preprint arXiv:2212.06870*, 2022.

[9] V. N. Nguyen, Y. Hu, Y. Xiao, M. Salzmann, and V. Lepetit, "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*, 2022, pp. 6771–6780.

[10] S. Thalhammer, J.-B. Weibel, M. Vincze, and J. Garcia-Rodriguez, "Self-supervised vision transformers for 3d pose estimation of novel objects," *arXiv preprint arXiv:2306.00129*, 2023.

[11] M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. Katam, and A. Lodhi, "Blenderproc: Reducing the reality gap with photorealistic rendering," in *International Conference on Robotics: Sciene and Systems, RSS 2020*, 2020.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep vit features as dense visual descriptors," *arXiv preprint arXiv:2112.05814*, vol. 2, no. 3, p. 4, 2021.

[14] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] F. Moreno-Noguer, V. Lepetit, and P. Fua, "Accurate non-iterative o(n) solution to the pnp problem," in *2007 IEEE 11th International Conference on Computer Vision*. Ieee, 2007, pp. 1–8.

[17] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, *et al.*, "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

[18] X. Liu, S. Iwase, and K. M. Kitani, "Kdfnet: Learning keypoint distance field for 6d object pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4631–4638.

[19] S. Thalhammer, M. Leitner, T. Patten, and M. Vincze, "Pyrapose: Feature pyramids for fast and accurate object pose estimation under domain shift," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 909–13 915.

[20] Y. Lin, J. Tremblay, S. Tyree, P. A. Vela, and S. Birchfield, "Single-stage keypoint-based category-level object pose estimation from an rgb image," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1547–1553.

[21] J. Richter-Klug and U. Frese, "Handling object symmetries in cnn-based pose estimation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 850–13 856.

[22] M. Jawaid, E. Elms, Y. Latif, and T.-J. Chin, "Towards bridging the space domain gap for satellite pose estimation using event sensing," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 866–11 873.

[23] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: 6d pose object detector and refiner," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1941–1950.

[24] T. Hodan, D. Barath, and J. Matas, "Epos: Estimating 6d pose of objects with symmetries," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 703–11 712.

[25] Z. Fan, P. Pan, P. Wang, Y. Jiang, D. Xu, H. Jiang, and Z. Wang, "Pope: 6-dof promptable pose estimation of any object, in any scene, with one reference," *arXiv preprint arXiv:2305.15727*, 2023.

[26] W. Goodwin, S. Vaze, I. Havoutis, and I. Posner, "Zero-shot category-level object pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 516–532.

[27] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *2010 IEEE computer society conference on computer vision and pattern recognition*. Ieee, 2010, pp. 998–1005.

[28] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 683–698.

[29] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel, "Multi-path learning for object pose estimation across domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 916–13 925.

[30] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[31] V. N. Nguyen, T. Hodan, G. Ponimatkin, T. Groueix, and V. Lepetit, "Cnos: A strong baseline for cad-based novel object segmentation," *arXiv preprint arXiv:2307.11067*, 2023.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[33] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in neural information processing systems*, vol. 32, 2019.

[34] A. Aldoma, M. Vincze, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu, and G. Bradski, "Cad-model recognition and 6dof pose estimation using 3d cues," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 585–592.

[35] T. Hodaň, X. Zabulis, M. Lourakis, Š. Obdržálek, and J. Matas, "Detection and fine 3d pose estimation of texture-less objects in rgb-d images," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 4421–4428.

[36] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*. Springer, 2014, pp. 536–551.

[37] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[38] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-less: An rgb-d dataset for 6d pose estimation of texture-less objects," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 880–888.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.